Variational Patch Gating for Training-Free Few-Shot Classification

Ahmed Radwan[1], Ahmad Abdel-Qader[1], Islam Osman[1], and Mohamed Shehata[1]

The University of British Columbia, British Columbia, Canada
{ahmedm04,aabdelqa}@student.ubc.ca,
{islam.osman,mohamed.sami.shehata}@ubc.ca

Abstract. Few-shot classification from a frozen backbone collapses when the target domain lies far from the pre-training distribution, because global embeddings map unrelated classes to nearly identical CLS tokens. Local patch descriptors are more resilient: an edge or texture gradient activates similar features whether it appears in a photograph or an X-ray. Yet part-level matching remains largely unexplored for few-shot recognition. The few methods that attempt it formulate matching as optimal transport or bipartite assignment, requiring iterative solvers whose cost grows rapidly with shots. We take a different approach. Instead of solving a matching optimization, we formulate patch-based classification as probabilistic evidence accumulation under a discriminative variational model. Each query patch carries a Bernoulli latent variable; minimizing the resulting assignment free energy yields a closed-form sigmoid gate and a softplus evidence score whose per-patch contribution saturates at $-\log(1-\pi)$, bounding clutter influence without iterative solvers or learned parameters. We evaluate extensively across twelve datasets spanning satellite, medical, fine-grained, and natural-image domains. Without any gradient updates, VPG outperforms the best fine-tuned baseline on five of ten dataset-settings and exceeds the previous training-free state of the art by $+5.35$ points on CDFSL. With lightweight episodic fine-tuning, VPG reaches 71.06% on CDFSL and 81.8% on four unseen Meta-Dataset domains.

Keywords: Few-shot classification · Training-free inference · Cross-domain generalization · Variational inference

1 Introduction

Few-shot image classification relies on global image embeddings from a pre-trained backbone. When training and test distributions overlap, prototype-based classifiers perform well. Under large domain shift, however, global representations collapse. A satellite view of a highway and an aerial shot of a river can produce nearly identical CLS tokens from a frozen ViT, making class separation impossible at the global level.

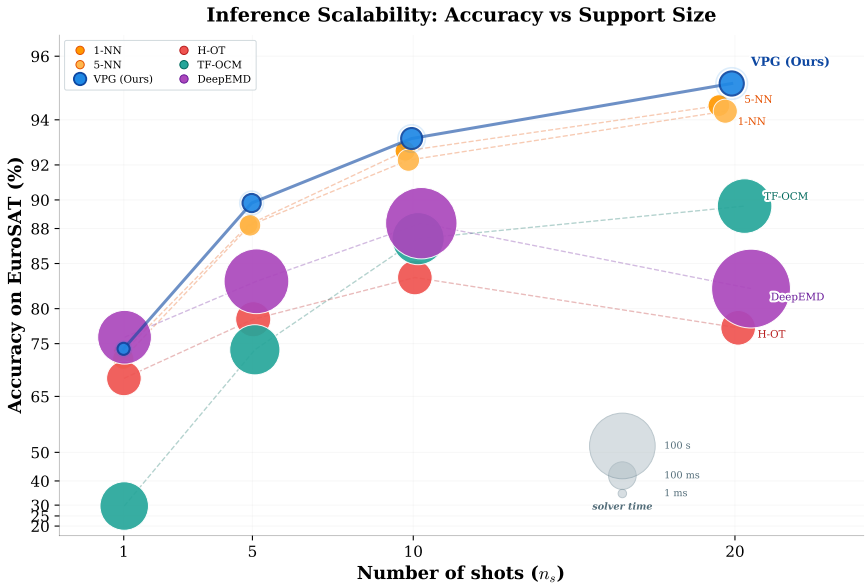

Fig. 1: Inference scalability on EuroSAT (5-way, frozen DINO ViT-S/16). Bubble area is proportional to log classification time per episode. All methods use the same frozen backbone, so times reflect the classification stage only, deconfounding backbone cost. VPG (blue) reaches the highest accuracy among part-based classifiers while remaining orders of magnitude faster. As shots grow, VPG scales linearly while LP-based solvers scale quadratically.

Local patch descriptors offer a more resilient signal. Patches of similar low-level structure (edges, textures, repeated motifs) map to similar features regardless of the source domain. If class evidence can be accumulated from discriminative patches rather than from a single global vector, classification may remain robust even when global embeddings fail.

Part-based reasoning has seen very limited adoption in few-shot learning. The few methods that attempt patch-level classification formulate it as optimal transport (DeepEMD [27], H-OT [8]) or minimum-cost bipartite assignment (TF-OCM [19]). These formulations build a cost matrix over all query-support token pairs, so their complexity grows quadratically or worse in the number of support patches. At 20-shot, the support set reaches $\sim$4 000 tokens per class; LP-based solvers become impractical and Sinkhorn iterations slow by orders of magnitude (Fig. 1).

We take a different perspective. Instead of solving a matching optimization, we formulate patch-based classification as probabilistic evidence accumulation under a foreground-clutter discriminative variational model. Each query patch carries a Bernoulli latent variable indicating foreground or clutter. Free-energy minimization produces, in closed form

- A per-patch **sigmoid gate** $r_p^\star = \sigma(\text{logit}(\pi) - s_\varepsilon)$, the variational posterior probability that the patch belongs to the class. Conditioned on the chosen prior π, its functional form emerges analytically from the variational objective.
- A per-patch **softplus evidence** $e = \text{softplus}(\text{logit}(\pi) - s_\varepsilon)$, a score that saturates at $-\log(1-\pi)$, bounding the total influence of any single patch regardless of mismatch magnitude.

Because inference is closed-form, no iterative Sinkhorn or LP solver is needed. Patch evidence reduces to matrix multiplications followed by elementwise sigmoid and softplus operations, making part-based reasoning scalable for training-free few-shot classification. We use one globally fixed configuration, $(\pi, \varepsilon, \alpha) = (0.7, 0.01, 0.6)$, on all datasets and episodes. Section 4.4 confirms stability across broad ranges of π and ε, with a broad optimum for α. To our knowledge, this is the first work to formulate patch-based few-shot recognition as probabilistic inference under a foreground-clutter discriminative variational model, yielding a bounded-evidence classifier with closed-form inference.

Contributions.

- A discriminative variational model for patch observations in few-shot classification whose assignment free-energy minimum yields a closed-form sigmoid gate and a bounded softplus evidence score.
- A unified class score that adds bounded local patch evidence and global prototype similarity under one posterior, fused with a single fixed mixing weight.
- A training-free classifier, VPG, requiring no iterative solvers and no learned parameters at inference, evaluated with one fixed configuration $(\pi, \varepsilon, \alpha)$ reused across all datasets. No other method in Table 1 operates with zero learned parameters, zero base-class training, and a single global configuration.
- Extensive evaluation across twelve datasets spanning four benchmark suites (CDFSL, a medical benchmark [23], Meta-Dataset, and in-domain miniImageNet/CUB). Without any gradient updates, VPG outperforms the best fine-tuned baseline on five of ten dataset-settings and exceeds the previous training-free state of the art by $+5.35$ points on CDFSL. We also provide a lightweight episodic fine-tuning variant for fair comparison with methods that require adaptation.

2 Related Work

Prototype-based few-shot classification. Prototypical Networks [21] classify by distance to class means and remain competitive when the feature extractor is strong. TF-OCM [19] extends this family to a training-free setting with a frozen backbone and no meta-training. VPG belongs to this family but replaces global cosine similarity with a per-patch evidence score that bounds each patch's influence on the class decision.

Transport-based part matching. A small number of methods attempt patch-level classification. DeepEMD [27] solves an Earth Mover's Distance between query and support patch sets using a linear programming solver whose runtime scales cubically in the number of support tokens. H-OT [8] uses Sinkhorn iterations with a hierarchical structure; each iteration touches the full $P_q \times P_s$ cost matrix. TF-OCM [19] segments patch embeddings via modularity optimization, then solves minimum-cost bipartite matching. All three methods share a common bottleneck: the cost matrix grows with the product of query and support patches, so classification time degrades sharply at higher shots (Fig. 1). VPG avoids this bottleneck entirely. Each query patch evaluates a closed-form soft-min against the support dictionary, so runtime scales linearly in n_s.

Cross-domain few-shot adaptation. Recent CDFSL methods improve transferability by modifying source-domain training. FLoR [28] flattens the loss landscape during pre-training to reduce domain sensitivity. CD-CLS [30] decouples the CLS token objective from patch tokens to improve cross-domain transfer. AttnTemp [29] re-calibrates attention temperatures at test time. REAP [26] inserts random register tokens into the ViT encoder to improve CLS robustness. PMF [11] meta-trains on labeled base classes before episodic fine-tuning. All of these methods operate on the global CLS embedding and require supervised base-class training or architecture modifications during pre-training. To the best of our knowledge, VPG is the first method to formulate patch-level few-shot classification as closed-form probabilistic inference, achieving competitive or superior accuracy to all CLS-based approaches while being orders of magnitude faster than iterative matching solvers (Fig. 1).

3 Method

Motivation. Most few-shot classifiers compare global embeddings and degrade when domains shift. Reasoning at the patch level is a natural alternative, but linear aggregation lets every patch vote with equal weight. Background texture, imaging artifacts, and other clutter then drive the class score by sheer count rather than discriminative quality. We derive a probabilistic model that bounds each patch's contribution, yielding a class score driven by discriminative evidence rather than clutter count.

Our classifier operates on frozen ViT patch tokens in three steps (Fig. 2). (i) We represent each class by a weighted dictionary of support tokens. (ii) We compute a soft-min evidence cost for each query patch against that dictionary. (iii) We gate each patch's contribution via a binary latent variable, solved in closed form. The sigmoid gate and softplus evidence score emerge analytically from the variational objective, conditioned on the chosen prior π.

3.1 Patch tokens and class dictionaries

Let F be a frozen Vision Transformer. Given an image X, we extract P ℓ_2-normalized patch tokens $\{x_p\}_{p=1}^{P} \subset \mathbb{R}^d$ (excluding the class token) and assign uniform query weights $a_p = 1/P$.

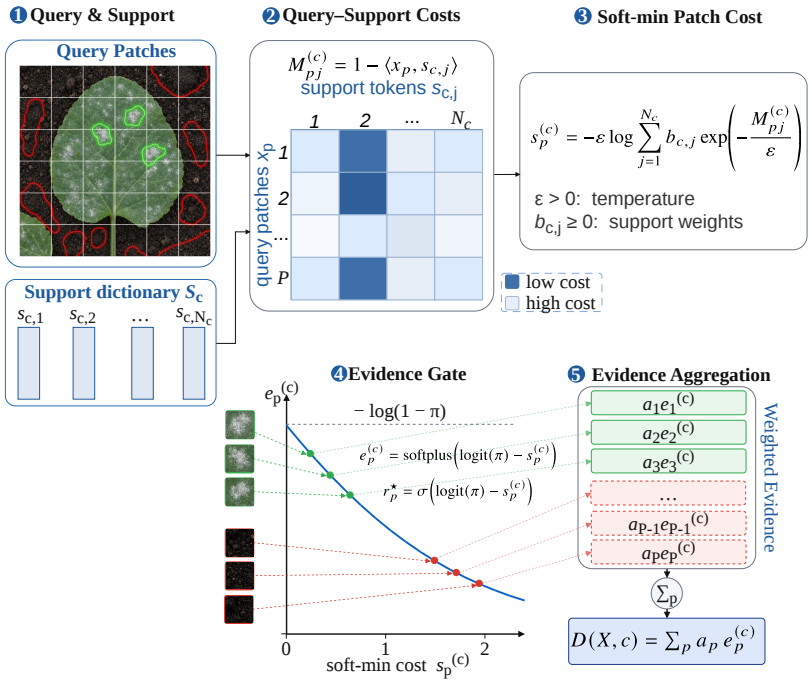

Fig. 2: VPG pipeline. Given query patches x_p and a class support-token dictionary $S_c = \{s_{c,j}\}_{j=1}^{N_c}$, we first compute query-support costs $M_{pj}^{(c)}$ and reduce them to a soft-min patch cost $s_p^{(c)}$. The cost is then mapped to a sigmoid gate $r_p^\star$ (Proposition 1) and bounded softplus evidence $e_p^{(c)}$ (Proposition 2). Low-cost informative patches (green) receive high evidence, while high-cost clutter patches (red) are suppressed. The final local evidence is the weighted aggregation $D(X,c) = \sum_p a_p e_p^{(c)}$.

For an episode with C classes, the support set of class c provides a token dictionary $S_c = \{s_{c,j}\}_{j=1}^{N_c}$ formed by concatenating patch tokens across its support images, with uniform mixture weights $b_{c,j} = 1/N_c$. We also compute a global prototype $g_c \in \mathbb{R}^d$ as the ℓ_2-normalized mean of the class-level CLS embeddings.

Ground cost. All tokens are ℓ_2-normalized. We use the cosine-distance cost

$$M_{pj}^{(c)} = 1 - \langle x_p, s_{c,j} \rangle \in [0, 2]. \tag{1}$$

3.2 Soft-min matching cost

To measure how well a query patch x_p matches class c, we define the **soft-min matching cost**

$$s_\varepsilon(x_p, S_c, b_c) \;=\; -\varepsilon \log \sum_{j=1}^{N_c} b_{c,j} \exp\Big(-\frac{M_{pj}^{(c)}}{\varepsilon}\Big), \tag{2}$$

where $\varepsilon > 0$ controls the sharpness of prototype selection: as $\varepsilon \to 0$, s_ε converges to the hard nearest-prototype cost $\min_j M_{pj}^{(c)}$; larger ε softens the minimum across prototypes. Because $M \in [0, 2]$ (Eq. 1), s_ε inherits the same ground-cost scale.[1]

We use s_ε directly as an unnormalized energy surrogate for assigning a patch to a class. Low cost means the patch matches the class dictionary well; high cost means it does not. Modeling a full generative likelihood $p(x_p \mid c) \propto \exp(-s_\varepsilon)$ would require estimating an intractable, class-dependent partition function. We bypass this by casting patch selection as a discriminative variational assignment.

3.3 Discriminative variational patch assignment

Rather than estimating class-conditional likelihoods over the feature space, we formulate patch selection as an energy-based assignment problem. Let $z_p \in \{0, 1\}$ be a latent indicator for whether query patch x_p is a discriminative foreground part for class c (z_p=1) or uninformative clutter (z_p=0), with base foreground rate $P(z_p{=}1) = \pi$.

We introduce a variational posterior $q(z_p) = \mathrm{Bern}(r_p)$, where $r_p \in [0, 1]$ is a soft assignment probability, and define the patch assignment free energy as

$$\begin{aligned} \mathcal{F}_p(r_p; c) &= \mathbb{E}_{q(z_p)}\big[z_p \cdot s_\varepsilon(x_p, S_c, b_c)\big] + \mathrm{KL}_{\mathrm{Bern}}(q(z_p) \| \pi) \\ &= r_p \cdot s_\varepsilon(x_p, S_c, b_c) + r_p \log \tfrac{r_p}{\pi} + (1{-}r_p) \log \tfrac{1-r_p}{1-\pi}. \end{aligned} \tag{3}$$

The first term linearly penalizes assigning a patch to a class when its soft-min cost s_ε is high. The KL term regularizes r_p toward the prior π, limiting overconfident assignments. This formulation operates entirely on the assignment variable r_p and avoids the need for normalized class-conditional densities over the feature space.

The per-patch assignment cost to minimize is

$$\min_{r_p \in [0,1]} \Big[r_p \cdot s_\varepsilon(x_p, S_c, b_c) \;+\; \mathrm{KL}_{\mathrm{Bern}}(r_p \| \pi) \Big]. \tag{4}$$

[1] With uniform support weights $b_j = 1/N_c$, an additive offset $\varepsilon \log N_c$ appears. This offset is class-independent for balanced support sets and vanishes as $\varepsilon \to 0$. At our operating point $\varepsilon = 0.01$, s_ε is dominated by the nearest-prototype cost and lies in $[0, 2]$ to numerical precision.

Remark 1 (Structured backgrounds). The derivation above sets the background energy to zero ($\beta \equiv 0$), treating uninformative patches as equally likely across classes. If the background is *structured* (non-uniform over patches), a natural extension is to include a patch-dependent background energy

$$\beta(x_p) \;=\; -\log p_{\mathrm{bg}}(x_p),$$

which yields the *relative* variational gate

$$r_p^\star \;=\; \sigma\Big(\mathrm{logit}(\pi) - \big(s_\varepsilon(x_p; S_c, b_c) - \beta(x_p)\big)\Big) \;=\; \sigma\big(\mathrm{logit}(\pi) + \beta(x_p) - s_\varepsilon(x_p; S_c, b_c)\big).$$

Importantly, although $\beta(x_p)$ is class-independent, it enters *inside* the gate nonlinearity and therefore changes the patch selection nonlinearly; hence it can affect the aggregated class scores and is *not* guaranteed to cancel in $\arg\max_c$ under our nonlinear evidence aggregation. In this work we set $\beta \equiv 0$ (constant-background model); estimating β from unlabeled target data (or a small validation set) is a compatible extension for handling structured clutter.

Proposition 1 (Optimal Gate). *The unique minimizer of* (4) *is*

$$r_p^\star = \sigma\big(\mathrm{logit}(\pi) - s_\varepsilon(x_p, S_c, b_c)\big)\,, \tag{5}$$

where σ is the logistic sigmoid and $\mathrm{logit}(\pi) = \log \frac{\pi}{1-\pi}$.

Proof sketch. The objective is strictly convex on $(0, 1)$. Setting $\partial F / \partial r_p = 0$ gives $s_\varepsilon + \log \frac{r_p}{\pi} - \log \frac{1-r_p}{1-\pi} = 0$, so $\log \frac{r_p}{1-r_p} = \mathrm{logit}(\pi) - s_\varepsilon$, which inverts to the sigmoid. □

Because s_ε lives on the same scale as the ground cost $M \in [0, 2]$, the prior π sets the gating threshold directly in cost units. The parameter ε controls prototype aggregation inside s_ε, not the gate steepness, so π and ε fulfil complementary roles: π determines *how much cost is tolerable*, while ε determines *how costs are pooled over prototypes*. To our knowledge, this is the first closed-form sigmoid gate for patch-based few-shot classification derived from a variational assignment objective.

Remark 2 (Cost-scale energy). Working at unit temperature on $[0, 2]$ decouples π from ε: the gating threshold is interpretable in cost units regardless of aggregation sharpness (see Appendix A for a comparison with the rescaled alternative).

Interpretation. The gate is a sigmoid applied to the difference between a prior term $\mathrm{logit}(\pi)$ (how likely any patch is foreground) and s_ε (how well the patch matches the class dictionary). Good matches (s_ε small) push $r_p^\star$ toward 1; poor matches (s_ε large) push it toward 0. A patch matched to its nearest prototype ($s_\varepsilon \approx 0$) receives $r_p^\star \approx \pi$; a patch with cost much larger than $\mathrm{logit}(\pi)$ is gated out.

3.4 The bounded-influence evidence score

Substituting $r_p^\star$ back into (4) and simplifying gives the following result (full derivation in Appendix A):

Proposition 2 (Bounded Evidence). *Define* $e(x_p; c) = \mathrm{softplus}\big(\mathrm{logit}(\pi) - s_\varepsilon(x_p, S_c, b_c)\big)$, *where* $\mathrm{softplus}(t) = \log(1 + e^t)$. *Then*

$$0 \;\le\; e(x_p; c) \;\le\; -\log(1-\pi). \tag{6}$$

Proof sketch. The optimized free energy equals $\log \frac{1}{1-\pi} - \mathrm{softplus}(\mathrm{logit}(\pi) - s_\varepsilon)$ (Appendix A.2). Defining evidence as the complement gives $e = \mathrm{softplus}(\cdot)$. Since softplus is monotone increasing with $\mathrm{softplus}(-\infty) = 0$ and the argument is bounded above by $\mathrm{logit}(\pi)$, we obtain $e \le \mathrm{softplus}(\mathrm{logit}(\pi)) = -\log(1-\pi)$. □

For $\pi = 0.7$ this cap is $-\log(0.3) \approx 1.20$: the *unweighted* per-patch evidence satisfies $e(x_p; c) \le 1.20$ regardless of match quality. Under the aggregated class evidence $D(X, c) = \sum_{p=1}^{P} a_p\, e(x_p; c)$ with $\sum_{p=1}^{P} a_p = 1$ (Eq. (9)), each patch contributes at most $a_p \cdot 1.20$ to $D(X, c)$ (in particular, $1.20/P$ for uniform $a_p = 1/P$), and the total satisfies $D(X, c) \le 1.20$. Poorly matched clutter patches contribute near zero. This bound provides a formal robustness certificate for the local branch $D(X, c)$: no single background patch can unboundedly dominate the aggregated patch evidence, regardless of domain shift. The total class evidence is

$$D(X, c) = \sum_{p=1}^{P} a_p \cdot e(x_p; c). \tag{7}$$

Limitation. Because $e \ge 0$, the score provides positive evidence for a class but never negative evidence against it. When a wrong class shares a few spurious patch matches with the query, its evidence may be elevated. This is a deliberate design choice: positive-only evidence guarantees monotonicity (adding a support image never decreases class evidence) and keeps the method training-free. Modeling explicit negative evidence, for example by subtracting a background or competing-class statistic, remains future work.

3.5 Joint local-global class score

We score each class with both local patch evidence and a global descriptor term. Let g_X be the query CLS feature and g_c the class prototype from support CLS features. We model the global branch with a cosine/vMF-style compatibility

$$p_{\mathrm{glob}}(c \mid X) \;\propto\; \exp\big(\kappa \langle g_X, g_c \rangle\big), \qquad \kappa > 0,$$

and the local branch with a bounded evidence expert

$$p_{\mathrm{loc}}(c \mid X) \;\propto\; \exp\big(D(X, c)\big),$$

where $D(X, c)$ is the variational local evidence from Secs. 3.3–3.4.

Proposition 3 (Log-linear / product-of-experts fusion). Assuming a uniform class prior $P(c) = 1/C$, a standard product-of-experts posterior,

$$p(c \mid X) \;\propto\; p_{\mathrm{loc}}(c \mid X)\, p_{\mathrm{glob}}(c \mid X)\, P(c),$$

implies class ranking by maximizing the additive score

$$S_{\mathrm{raw}}(X, c) \;=\; D(X, c) \;+\; \kappa \langle g_X, g_c \rangle. \tag{8}$$

Proof sketch. Taking log of the product gives additivity. Because $P(c)$ is uniform, $\log P(c)$ is constant across classes and drops from $\arg\max_c$. □
Remark. This fusion does *not* require conditional independence between g_X and patch observations; g_X may be any deterministic or stochastic summary of the same image X. The local branch is a discriminative variational score (Sec. 3.3); the global branch is a cosine compatibility kernel. The two are combined as complementary energy-based experts.

3.6 The foreground prior π

The prior π controls the clutter threshold. Larger π retains more patches, smaller π gates them more aggressively. The temperature ε controls the sharpness of the soft-min cost and is complementary to π: π determines *how much cost is tolerable*, ε determines *how costs are pooled*. We fix $(\pi, \varepsilon, \alpha) = (0.7, 0.01, 0.6)$ across all datasets and episodes with no per-dataset retuning; Section 4.4 confirms stability.

3.7 Final classifier

Equation (8) establishes an additive local-global score. Both branches are bounded and of comparable magnitude ($D(X, c) \in [0, -\log(1-\pi)] \approx [0, 1.2]$ for $\pi = 0.7$ and $\langle g_X, g_c \rangle \in [-1, 1]$). Because the classification decision ($\arg\max_c$) is invariant to multiplication by a strictly positive scalar, we can rescale S_{raw} by $\frac{1}{1+\kappa} > 0$ without changing predictions. Setting $\alpha = \kappa/(1+\kappa) \in (0, 1)$ yields the equivalent convex combination

$$\mathrm{logit}_c(X) = (1 - \alpha)\, D(X, c) + \alpha\, \langle g_X, g_c \rangle, \qquad \alpha \in (0, 1). \tag{9}$$

This bijection maps $\kappa \in (0, \infty)$ to a bounded, normalized weight $\alpha \in (0, 1)$, making it a rank-preserving reparameterization of the log-linear score rather than an ad-hoc heuristic. Section 4.4 confirms stability across $\alpha \in [0.4, 0.6]$. We classify as $\hat{y} = \arg\max_c \ \mathrm{logit}_c(X)$.

Algorithm. Algorithm 1 summarizes inference. The only computation beyond a single matrix multiply per class is an elementwise sigmoid and softplus, both $\mathcal{O}(P)$.

Require: Support $\{(X_i^S, y_i)\}$, query X^Q, frozen encoder F, parameters $(\varepsilon, \pi, \alpha)$
$\{x_p\} \leftarrow \text{PatchTokens}(F, X^Q)$, $g_X \leftarrow \text{Global}(F, X^Q)$
$a_p \leftarrow 1/P \quad \forall\, p$
for $c = 1, \ldots, C$ **do**
  $S_c \leftarrow \text{Concat}(\text{PatchTokens}(F, X_i^S) : y_i{=}c)$
  $b_{c,j} \leftarrow 1/N_c$; $g_c \leftarrow \text{Normalize}(\text{mean}_i\, \text{Global}(F, X_i^S))$
  $s_p^{(c)} \leftarrow -\varepsilon \log \sum_j b_{c,j} \exp(-M_{pj}^{(c)}/\varepsilon)$ ▷ soft-min cost
  $e_p \leftarrow \text{softplus}(\text{logit}(\pi) - s_p^{(c)}) \quad \forall\, p$ ▷ gated evidence
  $D_c \leftarrow \sum_p a_p\, e_p$
  $\text{logit}_c \leftarrow (1{-}\alpha)\, D_c + \alpha\, \langle g_X, g_c \rangle$ ▷ fixed additive mix
end for
return $\arg\max_c\ \text{logit}_c$

Algorithm 1: Variational Patch Gating (VPG)

Computational cost. The dominant operation is the matrix multiply $X\, S_c^\top \in \mathbb{R}^{P \times N_c}$, costing $\mathcal{O}(PN_c d)$ per class. The sigmoid and softplus are elementwise and $\mathcal{O}(P)$. Total cost across C classes is $\mathcal{O}(CP\bar{N}d)$, identical to other patch-matching methods. No iterative Sinkhorn or inner optimization loop is needed.

4 Experiments

4.1 Setup

Datasets. We evaluate on the Cross-Domain Few-Shot Learning (CDFSL) benchmark [9], comprising *EuroSAT* [10] (satellite), *CropDiseases* [16] (plant pathology), *ISIC2018* [4] (dermatoscopy), and *ChestX* [24] (radiology). We also report on HEp [18] and BCCD_WBC [20]. For Meta-Dataset out-of-domain evaluation, we follow the MDL protocol [22] on unseen domains: MS-COCO, MNIST, CIFAR-10, and CIFAR-100.

Protocol. All experiments use 5-way 5-shot classification; we report mean accuracy ± 95% confidence interval over 1 000 episodes. The backbone is a frozen DINO ViT-S/16 [2] throughout, shared across all methods in Table 1. All methods use identical image resolution (224×224) and the same episode sampler for fair comparison. VPG uses one globally fixed configuration, $(\pi, \varepsilon, \alpha) = (0.7, 0.01, 0.6)$, reused unchanged across datasets and episodes. These values follow from the derivation: $\pi = 0.7$ is a conservative foreground rate, $\varepsilon = 0.01$ gives near-hard-min matching on the $[0, 2]$ cost scale, and $\alpha = 0.6$ balances the two score branches. We selected these values once on the miniImageNet validation split and froze them for all subsequent benchmarks; no target-domain validation data was used to select or adjust them. Section 4.4 confirms that accuracy varies by less than 1% across $\pi \in [0.1, 0.9]$ and $\varepsilon \in [0.01, 0.1]$, with a broad optimum for α at 0.4–0.6.

Fine-tuning variant. For comparisons with methods that adapt per episode, we report VPG+FT, a lightweight episodic adaptation that unfreezes only the

Table 1: CDFSL 5-way 5-shot results (frozen DINO ViT-S/16 backbone, 1 000 episodes). TF: truly training-free (no base-class training). FT: episodic fine-tuning. All baselines share the same DINO ViT-S/16 backbone and evaluation protocol from [26]. Confidence intervals are shown where provided by the original authors. Best results are shown in bold red, second-best in blue.

Method	TF	FT	Venue	ChestX	ISIC2018	EuroSAT	CropDiseases	Avg
MEM-FS [23]	×	×	TIP'23	26.67	47.38	86.49	93.74	63.57
StyleAdv [7]	×	×	CVPR'23	26.97	47.73	88.57	94.85	64.53
FLoR [28]	×	×	CVPR'24	26.71	49.52	90.41	95.28	65.48
DAMIM [15]	×	×	AAAI'25	27.28	50.76	89.50	95.52	65.77
CD-CLS [30]	×	×	NeurIPS'24	27.23	50.46	91.04	95.68	66.10
AttnTemp [29]	×	×	NeurIPS'24	27.72	53.09	90.13	95.53	66.62
REAP [26]	×	×	ICML'25	27.98	52.80	90.53	95.68	66.75
TF-OCM [19]	✓	×	ECCVW'24	25.00	44.52	86.02	93.20	62.20
PMF [11]	×	✓	CVPR'22	27.27	50.12	85.98	92.96	64.08
FLoR [28]	×	✓	CVPR'24	27.02	53.06	90.75	96.47	66.83
DAMIM [15]	×	✓	AAAI'25	27.82	54.86	91.18	96.34	67.55
CD-CLS [30]	×	✓	NeurIPS'24	27.66	54.69	91.53	96.27	67.54
AttnTemp [29]	×	✓	NeurIPS'24	28.03	54.91	90.82	96.66	67.61
REAP [26]	×	✓	ICML'25	**28.80**	56.07	91.92	96.74	68.38
VPG (Ours)	✓	×	Ours	27.80±0.41	51.50±0.53	94.02±0.31	96.89±0.32	67.55
VPG (Ours)	✓	✓	Ours	28.30±0.41	**59.97±0.61**	**97.39±0.22**	**98.60±0.20**	**71.06**

last transformer block (∼2.2M of 21M total parameters) and runs 5 gradient steps with prototypical cross-entropy (lr $= 10^{-5}$, AdamW), following the protocol of PMF [11]. This takes <1 s per episode and requires no base-class meta-training, unlike all other fine-tuned methods in Table 1.

4.2 Main Results

Table 1 compares VPG against recent CDFSL methods. All baseline results are taken from [26], evaluated under identical DINO ViT-S/16 backbone and protocol.

Training-free. Without any gradient updates or base-class training, VPG achieves 67.55% average accuracy, competitive with the best non-FT competitor REAP (66.75%). Table 1 stratifies methods by training regime (TF column). Among training-free methods, VPG outperforms TF-OCM by +5.35 points. As a part-matching algorithm, VPG also exceeds every other matching-based method (DeepEMD, H-OT, TF-OCM) despite requiring no training, while running orders of magnitude faster (Fig. 1). Training-free VPG outperforms *all fine-tuned methods* in Table 1 on two of four domains. On EuroSAT, VPG scores 94.02% vs. best-FT REAP's 91.92% (+2.10); on CropDiseases, 96.89% vs. REAP+FT's 96.74%. This shows that part-level gating captures local discriminative structure that episodic adaptation cannot recover for globally-oriented classifiers. Unlike REAP and other baselines, VPG requires no supervised pre-training stage beyond the frozen self-supervised backbone.

With fine-tuning. VPG+FT reaches 71.06% average, surpassing REAP (68.38%) by 2.68 points. EuroSAT shows the largest gain (+5.47 over REAP), followed by ISIC2018 (+3.90). Episodic adaptation sharpens the patch representations that

Table 2: Five-shot results on HEp and BCCD_WBC. Baseline results are from the cited work. Best results are shown in bold red, second-best in blue. VPG+FT achieves the best performance on both datasets; training-free VPG surpasses all non-FT baselines on both datasets without any gradient updates.

Method	FT	HEp	BCCD_WBC	Avg
SimpleCNAPS [1]	×	53.15	47.06	50.10
Reptile [1]	×	51.76	50.91	51.33
TF-OCM [19]	×	62.78	66.26	64.52
ProtoNet [21]	✓	50.70	46.89	48.79
PMF [11]	✓	69.32	64.91	67.11
MEM-FS+RDA [23]	✓	71.21	66.52	68.86
VPG (Ours)	×	69.26±0.59	68.86±0.40	69.06
VPG (Ours)	✓	**76.35±0.61**	**80.30±0.72**	**78.33**

Table 3: Comparison on miniImageNet and CUB (protocol from ADC/H-OT [8], Table 1). Results marked ∘ are from original papers with supervised backbones; † from our reproduction with a DINO ViT-S/16 backbone. Confidence intervals are shown where available in the source. Best results are shown in bold red, second-best in blue. In-domain performance with a self-supervised ViT backbone far exceeds that of supervised ResNets, confirming the strength of the feature extractor.

			miniImageNet		CUB	
Method	FT	Backbone	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MAML∘ [6]	✓	ResNet18	49.61±0.92	65.72±0.77	69.96±1.01	82.70±0.65
Baseline++∘ [3]	✓	ResNet18	51.87±0.77	75.68±0.63	67.02±0.90	83.58±0.54
Neg-Cosine∘ [13]	✓	ResNet12	63.85±0.81	81.57±0.56	72.66±0.85	89.40±0.34
ProtoNet∘ [21]	✓	ResNet12	60.37±0.83	78.02±0.57	66.09±0.92	82.50±0.58
DeepEMD∘ [27]	✓	ResNet12	65.91±0.82	82.41±0.56	75.65±0.83	88.69±0.50
Free-Lunch∘ [25]	✓	ResNet12	64.73±0.44	81.15±0.42	74.13±0.86	87.91±0.45
H-OT∘ [8]	✓	ResNet12	65.63±0.32	82.87±0.43	76.28±0.40	89.87±0.36
Free-Lunch∘ [25]	✓	WRN28	68.57±0.55	82.88±0.42	79.56±0.87	90.67±0.35
H-OT∘ [8]	✓	WRN28	69.04±0.29	84.36±0.41	81.23±0.35	91.45±0.38
H-OT† [8]	✓	DINO ViT-S	83.15±0.36	93.91±0.16	87.64±0.37	95.94±0.16
VPG (Ours)	×	DINO ViT-S	85.76±0.32	96.06±0.46	86.02±0.89	95.57±0.55
VPG (Ours)	✓	DINO ViT-S	**87.15±0.76**	**99.25±0.41**	**88.05±0.24**	**99.46±0.60**

the gate then filters. Table 2 extends the comparison to two medical datasets (HEp and BCCD_WBC), and Table 3 reports in-domain results on miniImageNet and CUB under the H-OT protocol.

4.3 Unseen-Domain Meta-Dataset Results

Table 4 reports results on four unseen Meta-Dataset domains. VPG+FT reaches 81.8% average, above DIPA's 78.5%. The largest gains appear on CIFAR-10 (+11.3) and CIFAR-100 (+2.1), while MS-COCO is nearly tied. These results support the same pattern as CDFSL: bounded local evidence and global similarity remain effective under domain shift without per-domain retuning.

Table 4: Meta-Dataset unseen-domain results, varying-way 5-shot, MDL setting (baselines from DIPA [17] supplementary Table 11). SS PT: self-supervised pre-training. RN18: ResNet-18, ViT-s: ViT-small. Baseline results are reported as published; confidence intervals are shown where provided by the original authors. Best results are shown in bold red, second-best in blue. VPG+FT averages 81.8% across four unseen domains, exceeding DIPA (78.5%) despite using no supervised training.

Method	Backbone	SS PT	MS-COCO	MNIST	CIFAR-10	CIFAR-100	Average
Simple CNAPS [1]	RN18		49.2±0.8	88.9±0.4	66.1±0.7	53.8±0.9	64.5
SUR [5]	RN18		48.1±0.9	90.1±0.4	50.3±1.0	46.4±0.9	58.7
URT [14]	RN18		52.3±0.9	86.5±0.5	61.4±0.7	52.5±0.9	63.2
TSA [12]	RN18		56.0±0.8	**92.5**±0.4	72.0±0.7	64.1±0.8	71.2
DIPA [17]	ViT-s	✓	**64.64**±0.68	92.07±0.33	80.37±0.53	76.79±0.64	78.5
VPG+FT	ViT-s	✓	64.35±0.73	92.37±0.46	**91.64**±0.43	**78.91**±0.68	**81.8**

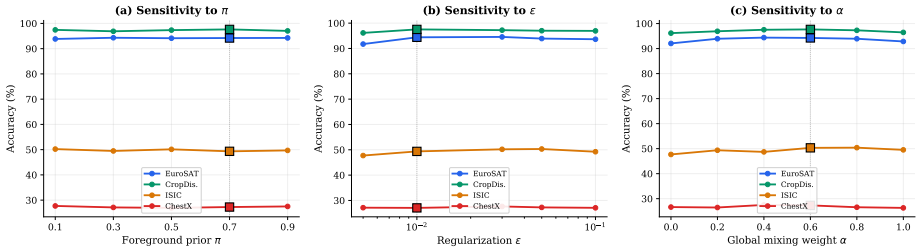

Fig. 3: Sensitivity of VPG to its three hyperparameters on CDFSL (5-way 5-shot, 600 episodes). **(a)** Accuracy varies by less than 1% across $\pi \in [0.1, 0.9]$ on every dataset. **(b)** Performance is stable for $\varepsilon \in [0.01, 0.1]$; only the extreme ε=0.005 degrades EuroSAT. **(c)** The mixing weight α has a broad optimum at 0.4–0.6; parts-only (α=0) and global-only (α=1) both lose accuracy. Squares mark the chosen configuration (π=0.7, ε=0.01, α=0.6). VPG is insensitive to π and ε across a wide range; α requires only coarse tuning.

4.4 Ablation Study

VPG is robust to all three hyperparameters (Fig. 3).

Foreground prior π and evidence temperature ε. Sweeping π from 0.1 to 0.9 at fixed ε=0.01, α=0.6 changes accuracy by less than 1% on all four datasets (Fig. 3a), because shifting the softplus cap rescales evidence but preserves class ordering. Varying ε from 0.01 to 0.1 causes less than 1% variation (Fig. 3b); only the extreme ε=0.005 degrades EuroSAT.

Global mixing weight α. Performance is more sensitive to α (Fig. 3c). Parts-only (α=0) loses ∼2% on EuroSAT and CropDiseases; global-only (α=1) is uniformly worse. The broad optimum at $\alpha \in [0.4, 0.6]$ confirms complementarity of the two branches.

Component ablation. Table 5 isolates the contribution of each component by removing one at a time while keeping all other parameters fixed at (π=0.7, ε=0.01, α=0.6).

Removing the global CLS branch (α=0) drops the average by 2.39 points; removing the local branch (α=1) costs 1.34 points, confirming complementarity.

Table 5: Component ablation on CDFSL (5-way 5-shot, 600 episodes). Each row removes one component from VPG while keeping the rest fixed. Both score branches are needed.

Configuration	EuroSAT	CropDis	ISIC	ChestX	Avg
VPG (full, α=0.6)	**94.02±0.31**	**96.89±0.32**	**51.50±0.53**	**27.80±0.41**	**67.55**
− global branch (α=0)	91.21±0.42	95.56±0.40	47.47±0.67	26.38±0.51	65.16
− local branch (α=1)	92.73±0.42	96.49±0.39	48.75±0.67	26.87±0.53	66.21
− gate & softplus (raw $-s_\varepsilon$)	93.27±0.37	96.87±0.37	49.61±0.66	27.20±0.52	66.74

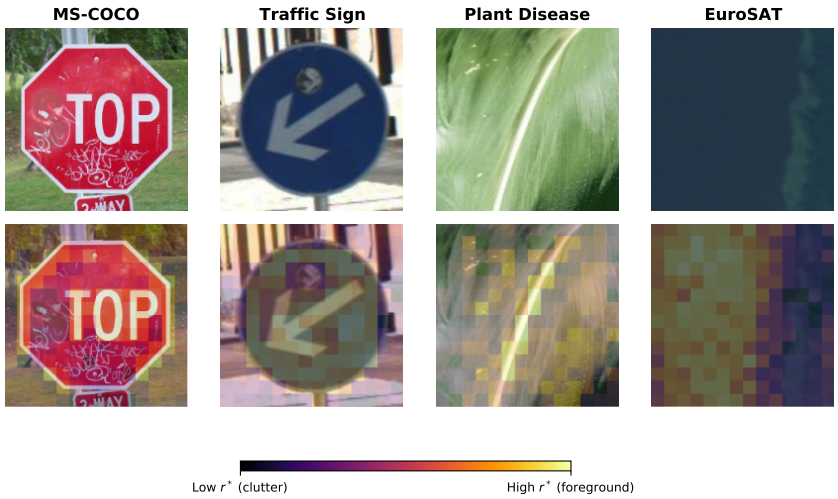

Fig. 4: Predicted-class gate $r_p^\star$ overlaid on query images from four domains. Bright patches (high $r^\star$) are treated as foreground evidence; dark patches are suppressed as clutter. The gate concentrates on semantically relevant regions without any saliency supervision.

Replacing the derived softplus with raw negative soft-min cost costs 0.81 points. The gate and softplus are not independent design choices but arise jointly from the variational objective (Propositions 1–2); removing the softplus discards the bounded-influence guarantee that the derivation provides.

4.5 Computational Efficiency

Because the soft-min partition function and sigmoid gate evaluate in *closed form*, VPG avoids iterative solvers. At 5-shot, VPG classifies an episode in 12 ms, which is 37× faster than H-OT (Sinkhorn, 453 ms) and 7 210× faster than DeepEMD (LP solver, 86.5 s). Classification time scales linearly with shots (Fig. 1). All times are measured on a single NVIDIA A100 GPU. All methods use the same frozen backbone, so we report classification time after the shared feature extraction pass to deconfound backbone cost from the classification algorithm.

4.6 Qualitative Analysis

Figure 4 visualizes the predicted-class gate $r_p^\star$ per query patch. On MS-COCO, the gate isolates the foreground object; on CropDiseases, it selects diseased tissue over dark surround. These patterns emerge from the variational posterior alone with no saliency supervision.

5 Conclusion

VPG formulates patch-based few-shot recognition as probabilistic evidence accumulation under a discriminative variational model, deriving a closed-form sigmoid gate and bounded softplus evidence that eliminate the need for iterative transport solvers. With one fixed configuration, it reaches 71.06% on CDFSL (67.55% training-free) and 81.8% across four unseen Meta-Dataset domains, while running 37× faster than Sinkhorn and 7,210× faster than LP solvers. The bounded-evidence principle makes part-level reasoning practical for training-free few-shot classification.

References

1. Bateni, P., Goyal, R., Masrani, V., Wood, F., Sigal, L.: Improved few-shot visual classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14493–14502 (2020)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
3. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
4. Codella, N.C.F., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
5. Dvornik, N., Schmid, C., Mairal, J.: Selecting relevant features from a multi-domain representation for few-shot classification. In: Computer Vision – ECCV 2020. pp. 769–786. Springer (2020). https://doi.org/10.1007/978-3-030-58607-2_45
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135. PMLR (2017)
7. Fu, Y., Xie, Y., Fu, Y., Jiang, Y.G.: StyleAdv: Meta style adversarial training for cross-domain few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24575–24584 (2023). https://doi.org/10.1109/CVPR52729.2023.02354
8. Guo, D., Tian, L., Zhao, H., Zhou, M., Zha, H.: Adaptive distribution calibration for few-shot learning with hierarchical optimal transport. In: Advances in Neural Information Processing Systems. vol. 35, pp. 6996–7010 (2022)

9. Guo, Y., Codella, N.C.F., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: Computer Vision – ECCV 2020. pp. 124–141. Springer (2020). https://doi.org/10.1007/978-3-030-58583-9_8
10. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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). https://doi.org/10.1109/JSTARS.2019.2918242
11. Hu, S.X., Li, D., Stühmer, J., Kim, M., Hospedales, T.M.: Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9068–9077 (2022)
12. Li, W.H., Liu, X., Bilen, H.: Cross-domain few-shot learning with task-specific adapters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7161–7170 (2022)
13. Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., Hu, H.: Negative margin matters: Understanding margin in few-shot classification. In: Computer Vision – ECCV 2020. pp. 438–455. Springer (2020). https://doi.org/10.1007/978-3-030-58548-8_26
14. Liu, L., Hamilton, W., Long, G., Jiang, J., Larochelle, H.: A universal representation transformer layer for few-shot image classification. In: International Conference on Learning Representations (2021)
15. Ma, R., Zou, Y., Li, Y., Li, R.: Reconstruction target matters in masked image modeling for cross-domain few-shot learning. Proceedings of the AAAI Conference on Artificial Intelligence **39**(18), 19305–19313 (2025). https://doi.org/10.1609/aaai.v39i18.34125
16. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Frontiers in Plant Science **7**, 1419 (2016). https://doi.org/10.3389/fpls.2016.01419
17. Perera, R., Halgamuge, S.: Discriminative sample-guided and parameter-efficient feature space adaptation for cross-domain few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23794–23804 (2024)
18. Qi, X., Zhao, G., Chen, J., Pietikäinen, M.: Exploring illumination robust descriptors for human epithelial type 2 cell classification. Pattern Recognition **60**, 420–429 (2016). https://doi.org/10.1016/j.patcog.2016.05.032
19. Radwan, A., Shehata, M.: TF-OCM: Training-free optimal community matching for domain generalized few-shot learning. In: Computer Vision – ECCV 2024 Workshops. Lecture Notes in Computer Science, vol. 15639, pp. 100–115. Springer (2025). https://doi.org/10.1007/978-3-031-91585-7_7
20. Shenggan: BCCD dataset: Blood cell count and detection dataset. GitHub repository (2018), https://github.com/Shenggan/BCCD_Dataset, accessed 26 June 2026
21. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)
22. Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.A., Larochelle, H.: Meta-dataset: A dataset of datasets for learning to learn from few examples. In: International Conference on Learning Representations (2020)
23. Walsh, R., Osman, I.I., Shehata, M.S.: Masked embedding modeling with rapid domain adjustment for few-shot image classification. IEEE Transactions on Image Processing **32**, 4907–4920 (2023). https://doi.org/10.1109/TIP.2023.3306916

24. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017)
25. Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: Distribution calibration. In: International Conference on Learning Representations (2021)
26. Yi, S., Zou, Y., Li, Y., Li, R.: Random registers for cross-domain few-shot learning. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 72317–72333. PMLR (2025)
27. Zhang, C., Cai, Y., Lin, G., Shen, C.: DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12203–12213 (2020)
28. Zou, Y., Liu, Y., Hu, Y., Li, Y., Li, R.: Flatten long-range loss landscapes for cross-domain few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23575–23584 (2024)
29. Zou, Y., Ma, R., Li, Y., Li, R.: Attention temperature matters in ViT-based cross-domain few-shot learning. In: Advances in Neural Information Processing Systems. vol. 37, pp. 116332–116354 (2024). https://doi.org/10.52202/079017-3694
30. Zou, Y., Yi, S., Li, Y., Li, R.: A closer look at the CLS token for cross-domain few-shot learning. In: Advances in Neural Information Processing Systems. vol. 37, pp. 85523–85545 (2024). https://doi.org/10.52202/079017-2716