GAFT: GEOMETRY-AWARE FINE-TUNING FOR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting Vision Foundation Models (VFMs) in low-label regimes remains challenging despite their strong transferable representations. While semi-supervised learning (SSL) can mitigate label scarcity, most existing approaches treat unlabeled data primarily as noisy supervision, without explicitly accounting for the geometry of the learned feature space. In this work, we develop a theoretical framework showing that unlabeled data improves generalization by revealing the structure and density of the representation space, yielding a model-adaptive and data-adaptive bound in which labeled samples act as geometric anchors and local prediction robustness governs expected risk. We then propose GAFT (Geometry-Aware Fine-Tuning) which introduces geometry-aware regularization terms that integrate seamlessly into existing SSL pipelines. Extensive experiments on the Visual Task Adaptation Benchmark (VTAB), across multiple fine-tuning strategies, demonstrate that GAFT consistently improves both accuracy and robustness over strong baselines, highlighting the effectiveness of geometry-aware SSL for VFMs.

1 INTRODUCTION

Vision Foundation Models (VFMs) pretrained at scale have emerged as a powerful default for visual recognition, producing representations that transfer remarkably well across diverse downstream tasks (Radford et al., 2021; Oquab et al., 2024; Abnar et al., 2022). However, adapting these models becomes nontrivial in the low-label regime: with only a handful of labeled examples per class, naive fine-tuning often overfits, while freezing the representation can underfit domain shift (Mai et al., 2025; Jia et al., 2022; Zhai et al., 2019; Dong et al., 2024). In many realistic scenarios, labeled data is scarce, yet unlabeled samples from the target domain are abundant and inexpensive to collect. This imbalance naturally motivates semi-supervised fine-tuning as a practical means of improving adaptation without incurring heavy annotation costs (Mey & Loog, 2023).

Most existing semi-supervised learning (SSL) approaches leverage unlabeled data by converting it into an auxiliary source of supervision, most commonly through pseudo-labeling (PL) (Lee, 2013; Xie et al., 2020b; Pham et al., 2021; Wang et al., 2023; Chen et al., 2023; Gan & Wei, 2024; Zhang et al., 2025). In this paradigm, unlabeled examples contribute to training only once the model assigns them sufficiently confident labels. While effective in some regimes, this view treats unlabeled data primarily as *noisy labels*, which is known to be brittle: confirmation bias can reinforce early mistakes (Arazo et al., 2020; Loh et al., 2023; Chen et al., 2022a), and confidence thresholds are often miscalibrated, especially under distribution shift or limited supervision (Guo et al., 2017; Vishwakarma et al., 2025).

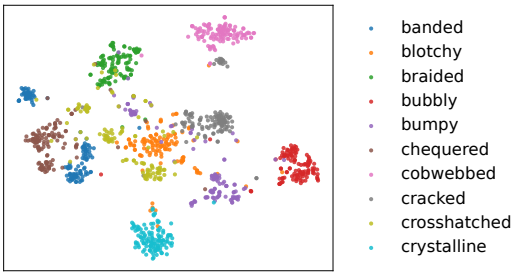


Figure 1: t-SNE visualization of DINOv2 embeddings for 10 selected DTD classes, colored by ground-truth label. Even before task-specific fine-tuning, the feature space is already organized into semantically coherent neighborhoods.

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

In contrast, due to their large-scale pretraining, VFMs often organize target-domain data into semantically meaningful neighborhoods in representation space even before adaptation (Caron et al., 2021; Oquab et al., 2024). Figure 1 provides a concrete example of this phenomenon.

Our theoretical results formalize the implication: rather than acting as surrogate labels, unlabeled samples can *reveal the geometry and density of the data distribution in representation space*, thereby identifying which inputs should behave similarly under the predictor. From our analysis, generalization error can be decomposed into a weighted loss over labeled anchors and additional alignment and robustness terms measuring local prediction sensitivity, with weights set by unlabeled coverage of each region: what matters is how well anchors align with high-density regions and how smoothly the model behaves around them.

Motivated by this insight, our contributions are threefold:

- We develop a **theoretical framework to bound the expected generalization error** by an empirical loss over labeled anchors plus two prediction-space discrepancy terms computable from unlabeled samples alone. Unlike the cluster and manifold assumptions underpinning classical graph-based SSL (Rigollet, 2007; Belkin et al., 2006; Chapelle et al., 2006), our bound imposes *no* global structural assumption on the input distribution: *any* partition of the input space that assigns one labeled anchor per region validates it, making it an explicit choice for the actual implementation.
- Based on the bound, **GAFT (Geometry-Aware Fine-Tuning)** is introduced with a reweighted supervised loss component and two regularizers which are term-for-term empirical realizations of precisely the bound's optimizable quantities, computed over a partition refreshed as the backbone adapts. GAFT is *plug-and-play* to any existing SSL objective, requiring neither the architectural changes of generative or ladder-style SSL (Kingma et al., 2014; Rasmus et al., 2015) nor specialized augmentation (Xie et al., 2020a; Berthelot et al., 2019).
- We conduct empirical evaluation across various VTAB-style few-shot tasks (Zhai et al., 2019), PEFT mechanisms, and backbone families. Results show that adding GAFT improves accuracy and robustness over strong baselines, e.g., traditional label propagation (Zhou et al., 2004; Iscen et al., 2019) and Laplacian regularization (Ando & Zhang, 2006), as well as pseudo-labeling-based approaches (Lee, 2013; Zhang et al., 2025), especially in label-scarce settings.

2 Related Work

2.1 Semi-Supervised Learning for Foundation Models

Most classical deep SSL methods were developed for training models from scratch (van Engelen & Hoos, 2020; Wang et al., 2023; Xie et al., 2020a; Berthelot et al., 2019; Tarvainen & Valpola, 2017), and how well they transfer to *adapting* pre-trained VFMs under extreme label scarcity is still under-explored. Among the few works studying this setting directly, Gan & Wei (2024) shows naive fine-tuning inherits systematic biases from the SSL process and proposes bias-reducing calibration; Zhang et al. (2025) propose PEFT-ensemble schemes (PET and V-PET) that combine multiple adapters to improve pseudo-label quality under few-shot labels; and Drozdova et al. (2024) fine-tune a VFM semi-supervisedly through a content-style decomposition of its representation. In all of them, unlabeled samples still enter the objective only as confidence-thresholded pseudo-labels (Sohn et al., 2020; Zhang et al., 2021; Chen et al., 2023), and the pretrained geometry is never used as a constraint in its own right.

2.2 Manifold-Based Regularization

A long line of SSL theory and algorithms is motivated by the *manifold* or *smoothness* principle: nearby points under the data geometry should have similar labels. Early formulations are transductive and graph-based (Joachims, 1999; Belkin & Niyogi, 2001; Zhu et al., 2003), and were later placed in a common geometric framework (Belkin et al., 2006; Ando & Zhang, 2006; Chapelle et al., 2006), where unlabeled data shapes a geometry (e.g., via a graph Laplacian) that regularizes the classifier. Modern deep SSL methods incorporate geometry implicitly through entropy minimization (Grandvalet & Bengio, 2005), label-propagation objectives (Zhou et al., 2004), teacher-graph neighbor

constraints (Luo et al., 2018), interpolation consistency training (Verma et al., 2022), or local adversarial smoothness (Miyato et al., 2019). These penalize disagreement between *pairs* of samples in a graphical setting, whereas we penalize discrepancy in *prediction* space against the *labeled anchor* of each region. To our knowledge, no prior work uses the pretrained VFM embedding space to define such a task-specific local structure during few-shot adaptation.

2.3 Efficient Foundation Models Adaptation

Pretrained VFMs, both vision-language (Radford et al., 2021) and self-supervised (Oquab et al., 2024), provide strong transferable representations for downstream recognition. In the few-shot regime, however, full fine-tuning can overfit while frozen features can be insufficient under domain shift. Many techniques counter this (Zhou et al., 2022a;b; Gao et al., 2024; Zhang et al., 2022), but *parameter-efficient fine-tuning* (PEFT) remains the popular choice (Mai et al., 2025); LoRA (Hu et al., 2022), AdaptFormer (Chen et al., 2022b), and visual prompt tuning (Jia et al., 2022) adapt large backbones with few trainable parameters, with later variants refining the update's parameterization (Dong et al., 2024). This literature is predominantly supervised, leaving the target domain's unlabeled pool unused.

3 Generalization Bound with Unlabeled Data

In this section, we prove a finite-sample bound (Theorem 3.5) that controls the expected risk of an arbitrary predictor by an empirical *anchor risk*, two predictor-dependent terms, and an h-independent residual, then lift it onto the embedding space of a pretrained backbone (Corollary 3.6). The result motivates the regularizers of Section 4.1.

3.1 Preliminaries

Notation. $\|\cdot\|$ denotes the ℓ_2-norm. $|S|$ denotes the cardinality of a finite set S, and $[K] := \{1, \dots, K\}$ for $K \geq 1$. We consider an input space $\mathcal{X}$ and an output space $\mathcal{Y} \subseteq \mathbb{R}^{C_\ell}$. The labeling function $y^* : \mathcal{X} \to \mathcal{Y}$ is unknown but deterministic. Let P_{xy} be a data distribution on $\mathcal{X} \times \mathcal{Y}$ and P_x its marginal on $\mathcal{X}$, with the relation $(x, y) \sim P_{xy} \implies y = y^*(x)$.

We write predictors as $h : \mathcal{X} \to \mathbb{R}^{C_\ell}$ and view them as members of a hypothesis class $\mathcal{H}$. The supervised loss is induced by a *cost function* $c : \mathbb{R}^{C_\ell} \times \mathcal{Y} \to \mathbb{R}_+$ via $\ell(h, x) := c(h(x), y^*(x))$ and we write $\ell(h, (x, y)) := c(h(x), y)$ when the label is named explicitly. The expected risk of h is $F(P_{xy}, h) := \mathbb{E}_{(x,y)\sim P_{xy}}[\ell(h, (x, y))]$.

Problem setting. Let $D = \{(s_i, y_i)\}_{i=1}^m$ be a labeled set of m i.i.d. samples from P_{xy} with m small (few-shot), and let $D^u = \{x_j\}_{j=1}^u$ be an unlabeled set of $u \gg m$ i.i.d. samples from P_x. We seek a predictor h minimizing $F(P_{xy}, h)$ using both D and D^u. We now define the concepts used in the analysis.

Definition 3.1. Let $h : \mathcal{X} \to \mathbb{R}^{C_\ell}$ be a predictor. For finite nonempty sets $A, B \subset \mathcal{X}$ and a measurable region $\mathcal{B} \subseteq \mathcal{X}$ with $P_x(\mathcal{B}) > 0$, we define:

$$d(A, B \mid h) := \frac{1}{|A|\,|B|} \sum_{a\in A} \sum_{b\in B} \|h(a) - h(b)\|, \tag{1}$$

$$d(A, \mathcal{B} \mid h) := \frac{1}{|A|} \sum_{a\in A} \mathbb{E}_{x\sim P_x}\left[\|h(a) - h(x)\| \,\middle|\, x \in \mathcal{B}\right]. \tag{2}$$

When $A = \{a\}$ is a singleton we write $d(a, B \mid h)$ and $d(a, \mathcal{B} \mid h)$. The same notation, with h replaced by y^*, applies to distances in label space.

Definition 3.2. Given a labeled set $D = \{(s_i, y_i)\}_{i=1}^m$, an *anchor partition* is a measurable partition $\Gamma = \{\mathcal{X}_i\}_{i=1}^m$ into m nonempty regions such that $s_i \in \mathcal{X}_i$ for each $i \in [m]$. We refer to s_i as the *anchor* of region $\mathcal{X}_i$.

3.2 Main Analysis

To support the analysis of the bound below, a number of assumptions regarding the cost function c should be mentioned. A more detailed discussion of such assumptions is left to Appendix A.1.

Assumption 3.3. The cost satisfies $c(z, y) \in [0, C]$ for all $z \in \mathbb{R}^{C_\ell}$ and $y \in \mathcal{Y}$.

Assumption 3.4. There exist constants $L_p, L_y \geq 0$ such that, for all $z, z' \in \mathbb{R}^{C_\ell}$ and all $y, y' \in \mathcal{Y}$,

$$|c(z, y) - c(z', y)| \leq L_p \, \|z - z'\|, \qquad |c(z, y) - c(z, y')| \leq L_y \, \|y - y'\|. \tag{3}$$

We write $L := \max(L_p, L_y)$.

Theorem 3.5 (Generalization bound with unlabeled data)**.** *Let $h \in \mathcal{H}$ be any predictor. Let $D = \{(s_i, y_i)\}_{i=1}^m$ be a labeled set and $\Gamma = \{\mathcal{X}_i\}_{i=1}^m$ any anchor partition. Draw an unlabeled set D^u of u i.i.d. samples from P_x, and define $D_i^u := D^u \cap \mathcal{X}_i$, $u_i := |D_i^u|$, $\hat{\pi}_i := \frac{u_i}{u}$ so that $(u_1, \ldots, u_m)$ is a* multinomial *count vector with parameters u and $(P_x(\mathcal{X}_1), \ldots, P_x(\mathcal{X}_m))$. Suppose the cost function c underlying ℓ satisfies Assumptions 3.3 and 3.4 with constants C and $L = \max(L_p, L_y)$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of D^u,*

$$F(P_{xy}, h) \leq \sum_{i=1}^m \hat{\pi}_i \, \ell(h, s_i) + L \sum_{i=1}^m \hat{\pi}_i [d(s_i, D_i^u \mid h) + d(D_i^u, \mathcal{X}_i \mid h)] + g_u, \tag{4}$$

where the residual below does not depend on h.

$$g_u := L \sum_{i=1}^m \hat{\pi}_i [d(s_i, D_i^u \mid y^*) + d(D_i^u, \mathcal{X}_i \mid y^*)] + C \sqrt{\frac{m \ln 4 - 2 \ln \delta}{u}}$$

Here, the anchor partition is a *free parameter* of the bound: any choice of Γ yields a valid inequality; the tightness of the bound depends on how well the chosen partition aligns with the data geometry.

Proof sketch. Write $p_i := P_x(\mathcal{X}_i)$, let $F_i(h) := \mathbb{E}_{x \sim P_x}[\ell(h, x) \mid x \in \mathcal{X}_i]$ be the within-region population risk, and let $\widehat{F}_i(h) := u_i^{-1} \sum_{x \in D_i^u} \ell(h, x)$ be its empirical counterpart. Adding and subtracting these two quantities decomposes the risk as $F(P_{xy}, h) = \sum_i \hat{\pi}_i \, \ell(h, s_i) + A_1 + A_2 + A_3$, where

$$A_1 = \sum_i \big(p_i - \hat{\pi}_i\big) F_i(h), \qquad A_2 = \sum_i \hat{\pi}_i \big(\widehat{F}_i(h) - \ell(h, s_i)\big), \qquad A_3 = \sum_i \hat{\pi}_i \big(F_i(h) - \widehat{F}_i(h)\big),$$

and the three terms are bounded separately. A_1 is a sampling-error term: the counts $(u_1, \ldots, u_m)$ are multinomial, so the Bretagnolle–Huber–Carol inequality controls $\sum_i |p_i - \hat{\pi}_i|$ with probability at least $1 - \delta$, and Assumption 3.3 turns this into the $C\sqrt{(m \ln 4 - 2 \ln \delta)/u}$ term of g_u. A_2 is the anchor-approximation error, the cost of replacing each unlabeled point by the anchor of its region; Assumption 3.4 bounds it by $L \sum_i \hat{\pi}_i [\, d(s_i, D_i^u \mid h) + d(s_i, D_i^u \mid y^*)\,]$. A_3 is the empirical-to-population gap within each region and is bounded the same way, by $L \sum_i \hat{\pi}_i [\, d(D_i^u, \mathcal{X}_i \mid h) + d(D_i^u, \mathcal{X}_i \mid y^*)\,]$. The Lipschitz step in A_2 and A_3 must absorb variation in the label function as well as the predictor, so each contributes one h-dependent and one y^*-dependent discrepancy; the latter cannot be influenced by training, which is why they collect into g_u alongside the concentration term. Summing the three bounds gives (4). The full argument is left in Appendix A.2, with an asymptotic decomposition in Appendix A.4.

Regarding the assumptions. Both assumptions constrain the cost function rather than the network, and both are standard in generalization analysis (Bartlett et al., 2017; Hou et al., 2023; He & Goldfeld, 2025), holding for common classification losses. Though the standard cross-entropy loss do not hold, as detailed in Section 4.1, GAFT's supervised term is implemented with the clipped cross-entropy of Remark A.1 at $\epsilon = 10^{-4}$, for which Assumption 3.3 holds with $C = \log(1/\epsilon) \approx 9.21$ and Assumption 3.4 with $L_y = \mathcal{O}(\log(1/\epsilon))$.

Comparison to previous works. Classical SSL theory licenses unlabeled data through structural assumptions on P_x, the *cluster assumption* (Rigollet, 2007) or *manifold regularization* (Belkin et al., 2006), while more recent analyses assume expansion or augmentation invariance (Wei et al., 2021; Jiang et al., 2023; Verma et al., 2022), and lower-bound results show that unlabeled data cannot universally help without some such structure (Tifrea et al., 2023). Theorem 3.5 is different in three ways.

- **(i) Predictor-induced geometry:** $d(s_i, D_i^u \mid h)$ and $d(D_i^u, \mathcal{X}_i \mid h)$ are measured in the predictor's *output* space rather than in input space, so the bound is phrased in how smoothly the *learned* representation varies within a region, making use of VFMs' powerful capacities.
- **(ii) Data-adaptive weighting:** the weights $\hat{\pi}_i = u_i/u$ are the empirical mass of each region, so regions densely populated by unlabeled samples reweight their anchor losses upward and sparse ones downward, instead of labeled samples entering symmetrically.
- **(iii) No global structure:** no cluster or manifold assumption on P_x is required, only the existence of *some* anchor partition for which g_u is small, which makes the choice of partition an explicit design knob.

In the era of large pretrained foundation models, this perspective suggests that effective semi-supervised fine-tuning should **(a)** improve h's within-region prediction smoothness using unlabeled data, **(b)** select a partition (e.g., via the pretrained embedding f) that makes g_u small, and **(c)** reweight labeled supervision by empirical unlabeled coverage, and we translate these into a concrete objective in Section 4.1.

3.3 Backbone-induced Geometry

Modern deep predictors decompose as $h = v \circ f$, where $f : \mathcal{X} \to \mathbb{R}^d$ is a feature extractor (e.g., a pretrained VFM) and $v : \mathbb{R}^d \to \mathbb{R}^{C_\ell}$ is a small head. Under mild Lipschitz regularity of v, the bound's geometric terms can be expressed in terms of f-distances rather than h-distances.

Corollary 3.6 (Backbone-induced geometry)**.** *Suppose the predictor decomposes as $h = v \circ f$, where $v : \mathbb{R}^d \to \mathbb{R}^{C_\ell}$ is L_v-Lipschitz. Then under the conditions of Theorem 3.5, with probability at least $1 - \delta$ over the draw of D^u,*

$$F(P_{xy}, h) \le \sum_{i=1}^{m} \hat{\pi}_i \, \ell(h, s_i) + L \, L_v \sum_{i=1}^{m} \hat{\pi}_i [d(s_i, D_i^u \mid f) + d(D_i^u, \mathcal{X}_i \mid f)] + g_u, \tag{5}$$

Corollary 3.6 resolves a tension: Theorem 3.5 is stated in terms of distances induced by h, while in practice we want to construct a partition using the pretrained backbone f for its rich representation capacity. The two formulations are equivalent up to a factor of L_v, which is bounded for any practical head and is absorbed into the regularization weight. This justifies the design choice in Section 4.2 of constructing Γ by clustering in pretrained f-space – the bound's geometric terms are then directly controlled by smoothness in f-space. The proof appears in Appendix A.3.

4 GAFT: Geometry-Aware Method for Fine-Tuning

4.1 Geometry-Aware Objective Function

We translate the h-dependent components of Theorem 3.5 into a tractable training objective by replacing each with a directly computable empirical surrogate. Two of the three are already empirical and can be optimized directly: the anchor risk $\sum_i \hat{\pi}_i \, \ell(h, s_i)$ and the alignment term $\sum_i \hat{\pi}_i \, d(s_i, D_i^u \mid h)$. The third, $\sum_i \hat{\pi}_i \, d(D_i^u, \mathcal{X}_i \mid h)$, contains a population expectation over $\mathcal{X}_i$ that we cannot evaluate, so we replace the population mean over $\mathcal{X}_i$ with the sample mean over D_i^u.

Definition 4.1 (Alignment regularizer)**.**

$$\mathcal{R}_{\text{align}}(h) \;:=\; \sum_{i=1}^{m} \hat{\pi}_i \, d(s_i, D_i^u \mid h) \;=\; \frac{1}{u} \sum_{i=1}^{m} \sum_{x \in D_i^u} \|h(s_i) - h(x)\|. \tag{6}$$

Definition 4.2 (Robustness regularizer)**.**

$$\mathcal{R}_{\text{robust}}(h) \;:=\; \sum_{i=1}^{m} \hat{\pi}_i \, d(D_i^u, D_i^u \mid h) \;=\; \sum_{i=1}^{m} \frac{1}{u \, u_i} \sum_{x, x' \in D_i^u} \|h(x) - h(x')\|. \tag{7}$$

The two regularizers are natural sample-based estimators of the two h-dependent geometric terms in Theorem 3.5, weighted by empirical region masses $\hat{\pi}_i$. Both terms default to 0 when a cluster has no unlabeled sample.

GAFT objective. Combining the anchor risk with the two empirical geometric surrogates yields the GAFT training objective:

$$\mathcal{L}_{\mathrm{GAFT}}(h) := \sum_{i=1}^{m} \hat{\pi}_i \, \ell(h, s_i) + \lambda_1 \, \mathcal{R}_{\mathrm{align}}(h) + \lambda_2 \, \mathcal{R}_{\mathrm{robust}}(h), \tag{8}$$

with non-negative coefficients λ_1, λ_2 to tune the amount of regularization in practice. The first term is a *reweighted* empirical loss on D, each anchor s_i weighted by $\hat{\pi}_i = u_i/u$: it reduces to standard empirical risk when regions are balanced ($u_i = u/m$), and otherwise lets anchors in densely-covered regions exert proportionally more influence, as Theorem 3.5 prescribes. The two regularizers play distinct roles: $\mathcal{R}_{\mathrm{align}}$ pulls each unlabeled point toward its anchor's prediction, while $\mathcal{R}_{\mathrm{robust}}$ encourages within-region smoothness of h on the unlabeled pool.

It is important to note that GAFT does not replace any SSL approach. Rather, the point of GAFT is that it is plug-and-play, meaning that the base SSL method is left entirely unmodified and $\mathcal{L}_{\mathrm{GAFT}}$ is simply added to its objective. Let $\mathcal{L}_{\mathrm{base}}$ be the base SSL objective, a GAFT run optimizes

$$\mathcal{L}_{\mathrm{total}}(h) = \underbrace{\mathcal{L}_{\mathrm{base}}(h)}_{\text{unmodified base SSL objective}} + \mathcal{L}_{\mathrm{GAFT}}(h). \tag{9}$$

The anchor-reweighted term in (8) is thus part of GAFT's own objective and sits alongside the base method's supervised loss rather than modifying it. Every baseline in Section 5 is accordingly run exactly as described above.

4.2 Dynamic Geometry Construction

Since vision foundation models are explicitly trained to map semantically related inputs to nearby points in embedding space (Radford et al., 2021; Caron et al., 2021; Oquab et al., 2024), and Corollary 3.6 licenses computing the bound's geometric terms there, f-space proximity provides a strong default signal for partitioning the input domain. We therefore instantiate the anchor partition Γ by nearest-anchor assignment in f-space (Section 4.2.1), refresh it periodically as f adapts during fine-tuning, and reconcile the global partition with mini-batch SGD (Section 4.2.2).

4.2.1 Geometry Construction

For each input $x \in \mathcal{X}$, let $\mathbf{z}(x) := f(x) \in \mathbb{R}^d$ denote its backbone embedding, where d is the embedding dimension of the chosen VFM. Following standard practice for VFM embeddings, we measure proximity by cosine distance, $d_{\cos}(\mathbf{z}, \mathbf{z}') := 1 - \frac{\mathbf{z}^\top \mathbf{z}'}{\|\mathbf{z}\| \, \|\mathbf{z}'\|}$. Each labeled sample s_i serves as the anchor of its own region, and each unlabeled point $x \in D^u$ is assigned to its nearest anchor under cosine distance: $\pi(x) := \arg\min_{i \in [m]} d_{\cos}\big(\mathbf{z}(x), \, \mathbf{z}(s_i)\big)$. The resulting nearest-anchor rule induces a Voronoi-style partition $\Gamma(\mathcal{X}) = \bigsqcup_{i=1}^{m} \mathcal{X}_i$ with $s_i \in \mathcal{X}_i$ by construction, together with the regional unlabeled subsets $D_i^u := \{x \in D^u : \pi(x) = i\}$ that appear in $\mathcal{R}_{\mathrm{align}}$ and $\mathcal{R}_{\mathrm{robust}}$.

This construction satisfies Definition 3.2 – the regions are measurable, nonempty, and partition $\mathcal{X}$ – so Theorem 3.5 applies with these specific $\mathcal{X}_i$ and D_i^u, and by Corollary 3.6 the bound's geometric terms are controlled in f-space up to the head's Lipschitz constant. This justifies using the pretrained backbone rather than the still-evolving predictor h for the partition.

4.2.2 Dynamic Updates and Mini-batch Geometry

Periodic refresh. The backbone f continues to adapt during fine-tuning, so a partition Γ constructed at initialization gradually drifts out of alignment with the evolving embedding geometry. To track this drift, we recompute embeddings for $D \cup D^u$ and reassign π every I epochs, following the procedure of Section 4.2.1; the interval I trades off geometric currency against per-refresh compute.

Mini-batch realization. The partition Γ is global, but to train in mini-batches, $\mathcal{R}_{\mathrm{align}}$ and $\mathcal{R}_{\mathrm{robust}}$ must be computed from the samples present in each batch. Given $B = B_\ell \cup B_u$, we restrict the global partition to the active samples only, which admits two sampling strategies (illustrated in Figure 5, Appendix C.2):

- **Independent sampling + batch filtering**: B_ℓ and B_u are drawn independently from D and D^u, then projected onto the partition via π. A restricted region may contain unlabeled samples but no anchor; $\mathcal{R}_{\text{align}}$ is then omitted while $\mathcal{R}_{\text{robust}}$ is still computed.
- **Labeled-first + cluster-conditioned unlabeled sampling**: B_ℓ is drawn first, and each unlabeled sample in B_u is then drawn from $\bigcup_{s_i \in B_\ell} D_i^u$. Every restricted region in the batch is guaranteed to have an anchor and its unlabeled neighborhood, so both terms are well-defined.

Appendices C.1, C.2, C.5, and C.6 study the resulting trade-offs in λ_1, λ_2, I, and batching strategy empirically.

5 Empirical Results

We evaluate GAFT on few-shot image classification: accuracy on six VTAB tasks across two PEFT mechanisms and two SSL paradigms (Section 5.2), robustness under common image corruptions (Section 5.3), and behaviour under skewed unlabeled class distributions and a newer backbone family (Section 5.4). Further experiments and ablations are in Appendix C.

5.1 Experimental Setup

Datasets and baselines. Following prior work (Zhang et al., 2025), we focus on a representative subset of VTAB tasks (Zhai et al., 2019) –covering texture (Cimpoi et al., 2014), scenes (Xiao et al., 2010), remote sensing (Cheng et al., 2017), and compositional reasoning (Johnson et al., 2017), as well as structured-prediction and medical-imaging tasks (Geiger et al., 2012; Cuadros & Bresnick, 2009)– each under a task-specific few-shot label budget (shots per class), known to be more challenging than standard SSL/FSL benchmarks (Vinyals et al., 2016; Snell et al., 2017) for adapting large pretrained vision models. We subsample the labeled set from the training set, treat the remaining training images as the unlabeled pool, and report accuracy on the held-out test split.

We compare GAFT against (i) labeled-only PEFT fine-tuning, (ii) traditional pseudo-labeling/consistency SSL baselines adapted to VFMs, namely FixMatch (Sohn et al., 2020), FlexMatch (Zhang et al., 2021), SoftMatch (Chen et al., 2023), FineSSL (Gan & Wei, 2024), and self-training (Xie et al., 2020b), and (iii) VFM-oriented ensemble baselines PET and V-PET, which leverage PEFT diversity to improve pseudo-label quality. Baseline results are taken from Zhang et al. (2025).

Implementation details. Unless otherwise stated, we adopt the same backbone and optimization recipe across methods for a controlled comparison, using AdamW (Loshchilov & Hutter, 2019) with similar fine-tuning defaults to the baselines, and report final-epoch top-1 accuracy, as done in standard SSL research (Sohn et al., 2020; Gan & Wei, 2024), averaged over three random seeds. All hyperparameters, including the method-specific λ_1, λ_2, I, mini-batch strategy, and u_ratio (unlabeled-to-labeled ratio within a batch), are in Appendix B.2.

5.2 Main Results

DINOv2 (ViT-B/14) is fine-tuned through LoRA (Hu et al., 2022) and AdaptFormer (Chen et al., 2022b), with + *GAFT* variants adding $\mathcal{L}_{\text{GAFT}}$ to each baseline's objective as in (9) and leaving every other component intact. Across both PEFT strategies, the + GAFT variants improve on their baselines on most datasets. For the V-PET variants, the strongest gains arise on RESISC45 and CLEVR-C, two tasks whose intra-class variation and non-natural image statistics depart most sharply from the natural-image distribution DINOv2 is pretrained on (Oquab et al., 2024). This is consistent with Theorem 3.5: when pretraining alone does not place semantically related samples in tight clusters, pseudo-labeling baselines suffer, but the unlabeled pool still carries information GAFT can exploit.

5.3 Against-Corruption Robustness

For GAFT, a natural empirical question is whether this within-region smoothness extends to test-time inputs that differ from training inputs through low-level perturbations. We probe this by evaluating models on ImageNet-C-style corruptions (Hendrycks & Dietterich, 2019), comprising $|\mathcal{C}| = 18$

Table 1: Top-1 accuracy (%) of fine-tuned DINOv2 models on six VTAB tasks. PET + GAFT and V-PET + GAFT denote the results when using our proposed geometry-aware loss for the existing methods, PET and V-PET, respectively. Avg. is the mean over the six datasets. **Bold** marks the best result per column within each PEFT block.

		DTD 3-shot	SUN397 3-shot	RESISC45 1-shot	CLEVR-C 10-shot	KITTI 5-shot	Retinopathy 40-shot	Avg
LoRA	Labeled-only	61.0	59.7	48.6	35.9	46.7	40.7	48.77
	FixMatch	57.4	56.6	36.6	35.7	14.8	38.1	39.87
	FlexMatch	52.8	55.7	33.3	37.1	26.2	36.6	40.28
	SoftMatch	53.6	60.4	46.6	33.6	27.4	32.1	42.28
	FineSSL	66.9	54.6	69.8	24.8	31.9	31.0	46.50
	ST	52.3	45.4	54.7	36.3	33.2	42.6	44.08
	PET	66.7	58.6	55.8	33.1	42.6	41.0	49.63
	PET + GAFT	**70.41**	63.77	62.86	32.05	55.73	**42.92**	54.62
	V-PET	67.8	67.1	66.0	36.8	58.2	39.2	55.85
	V-PET + GAFT	69.17	**68.82**	**78.95**	**51.84**	**61.67**	42.16	**62.10**
Adapt Former	Labeled-only	63.0	58.9	47.0	27.3	52.3	34.4	47.15
	FixMatch	59.5	57.9	26.4	32.7	19.8	38.7	39.17
	FlexMatch	57.1	56.0	51.3	33.4	34.3	33.4	44.25
	SoftMatch	58.7	60.0	53.1	32.9	24.5	33.5	43.78
	FineSSL	64.0	56.0	62.3	25.8	31.8	17.7	42.93
	ST	62.3	59.0	54.1	27.4	43.9	34.9	46.93
	PET	67.2	63.5	56.2	33.1	45.2	40.2	50.90
	PET + GAFT	**69.34**	64.69	62.85	37.40	58.68	**42.47**	55.91
	V-PET	68.2	67.7	66.3	36.8	**59.2**	39.4	56.27
	V-PET + GAFT	69.13	**68.38**	**75.61**	**48.14**	58.90	40.34	**60.08**

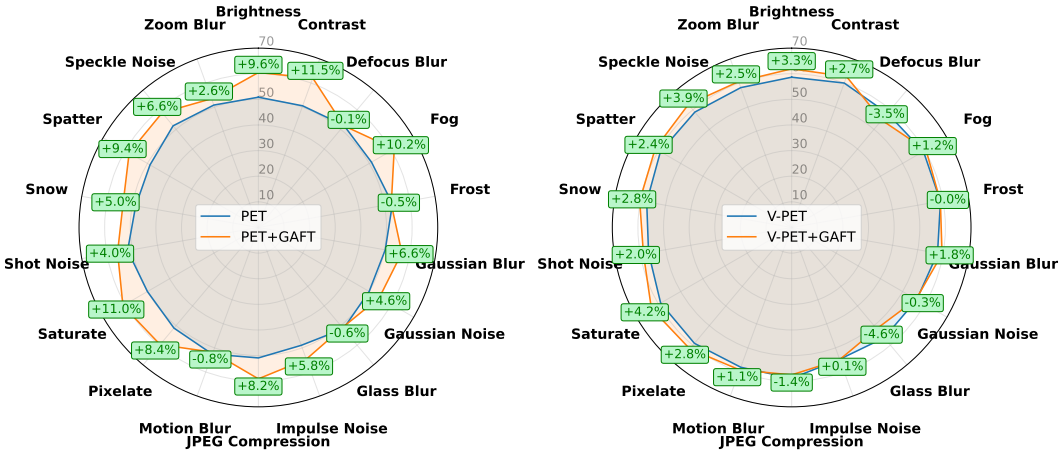


Figure 2: Robustness of DINOv2 to common image corruptions on **KITTI** when fine-tuned with **AdaptFormer**. Each spoke is a corruption type; values are per-corruption accuracies averaged over five severities ($\overline{\mathrm{Acc}}(c)$). Annotations report the absolute change in $\overline{\mathrm{Acc}}(c)$ of (+GAFT) relative to the corresponding baseline, with negative values denoting a decrease. Average gain: *(PET)* 5.64%, *(V-PET)* 1.17%.

corruption types in four families (noise, blur, weather, digital) at $|\mathcal{S}| = 5$ severity levels, applied *only at test time* to the held-out test split of DTD and KITTI with the training procedure left unchanged; Appendix C.11 details the pipeline.

Writing $\mathrm{Acc}(c, s)$ for top-1 accuracy under corruption c at severity s, we report the severity-averaged $\overline{\mathrm{Acc}}(c) := \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathrm{Acc}(c, s)$. Figure 2 compares PET and V-PET to their + GAFT counterparts

Table 2: Top-1 accuracy (%) under skewed unlabeled data. A ratio of r means one class has $r\times$ as many unlabeled samples as the other. DINOv2 backbone, AdaptFormer PEFT. **Bold** marks the higher accuracy per column within each dataset.

Dataset	Method	r=2	r=4	r=6	r=8
DTD (3-shot)	PET	**92.50**	87.50	81.25	80.00
	PET + GAFT	90.00	**90.00**	**86.25**	**82.50**
KITTI (5-shot)	PET	60.19	59.24	60.66	60.19
	PET + GAFT	**62.09**	**59.72**	**62.80**	**61.37**
CLEVR-Count (10-shot)	PET	54.09	53.31	**53.23**	51.84
	PET + GAFT	**55.82**	**53.39**	52.92	**52.34**

on KITTI using AdaptFormer; KITTI using LoRA and DTD results are in Appendix C.11. GAFT improves the corruption-averaged accuracy in every panel, with broadly distributed gains across the blur, noise, weather, and digital families.

5.4 Further Results

Robustness under class imbalance. A natural concern with anchor-weighted supervision is whether GAFT amplifies existing class-frequency bias when the unlabeled pool is skewed. To probe this, we construct two-class subsets of DTD (3-shot), KITTI (5-shot), and CLEVR-Count (10-shot) and introduce controlled imbalance of ratio r in the unlabeled data. Table 2 reports top-1 accuracy for PET vs. PET + GAFT fine-tuned with DINOv2 and AdaptFormer. GAFT maintains a lead in 7 of 8 imbalance settings on KITTI and CLEVR-Count, with the only exception at $r = 6$ on CLEVR-Count where the gap is within 0.31%. On DTD, PET outperforms PET + GAFT at the mildest imbalance (r=2) but GAFT recovers at $r \geq 4$. This supports Theorem 3.5: the weights $\hat{\pi}_i = u_i/u$ reflect unlabeled coverage around each anchor, not class frequency in the labeled set, so anchors of a minority class in dense regions retain proportional weight while anchors of an over-represented class in sparse regions are not given outsized influence.

Generalization to a newer backbone. To verify that GAFT is not tailored to the DINOv2 or CLIP embedding families, we additionally evaluate it on Qwen3-VL-Embedding-2B (Li et al., 2026; Bai et al., 2025), a more recent vision-language encoder pretrained with a different objective and architecture than either, fine-tuning it with LoRA and AdaptFormer using self-training as the base SSL method before applying GAFT. Table 3 shows ST + GAFT improves over ST under both PEFT strategies.

Table 3: Top-1 accuracy (%) on DTD (3-shot) with Qwen3-VL-Embedding-2B (Li et al., 2026; Bai et al., 2025).

Method	LoRA	AdaptFormer
ST	59.15	50.05
ST + GAFT	**60.74**	**51.76**

6 Conclusion

In this work, we present GAFT, a geometry-aware framework for semi-supervised fine-tuning of VFMs under label scarcity. Our central argument is that unlabeled data is most useful not as a source of noisy pseudo-labels, but as a probe of the geometry that pretraining has already induced: Theorem 3.5 formalizes this by bounding expected risk through anchor-weighted supervision and two predictor-induced smoothness terms, and GAFT's regularizers are direct empirical realizations of those terms. Experiments on VTAB-style benchmarks confirm consistent gains in accuracy and robustness across backbones, PEFT strategies, and SSL paradigms, supporting the core insight that geometry-aware fine-tuning is beneficial in the foundation-model era.

Future directions. In line with the broader SSL literature (van Engelen & Hoos, 2020; Gan & Wei, 2024; Zhang et al., 2025; Mai et al., 2025), our evaluation focuses on image classification to align with existing benchmarks. Extension to dense-prediction tasks and to modalities beyond static images is left to future work.

AI USE STATEMENT

We used generative AI tools selectively to assist in implementing methods and setting up datasets and training pipelines (under close human supervision), to discover and survey relevant literature (all entries manually verified by the authors), to polish the prose and grammar of the manuscript, and to assist with manuscript preparation for this submission (consistency checking of reported numbers across tables, verification of bibliographic entries, and LaTeX formatting).

We did not use generative AI tools to develop the theoretical framework, formulate the mathematical claims, or write the proofs; we did not use generative AI tools to propose or refine the research hypotheses, or to interpret experimental results. All AI-assisted output was reviewed by the authors: every number reported in this paper was traced back to its source experiment log, and all bibliographic entries were verified against the original publications. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies semi-supervised fine-tuning of vision foundation models and poses no direct ethical harms. For our work, no new data was collected or released. The VTAB benchmark is publicly available and de-identified; it is used here under its released terms with no re-identification attempted.

We acknowledge that GAFT amplifies the model's reliance on pretrained backbone geometry. This means that representational biases inherited from the web-scale pretraining of DINOv2, CLIP, or Qwen3-VL-Embedding can be reinforced under label scarcity, because the geometry used for partitioning reflects whatever biases are encoded in the backbone. Users deploying GAFT with backbone models whose pretraining data is not well-understood should audit the resulting representations for bias before deployment. Compute cost for this work was approximately 180 GPU-hours across consumer hardware (2× RTX 3090, 1× RTX PRO 6000 Blackwell), which is modest by current standards.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of the reported results.

Theory: All assumptions (Assumptions 3.3 and 3.4) are stated in Section 3; complete proofs of Theorem 3.5 and Corollary 3.6 appear in Sections A.2 and A.3; assumption justification for the clipped cross-entropy is given in Remark A.1 and Section A.1.

Setup: Dataset details, splits, and shot counts are in Section B.1 (Table 4); full hyperparameter tables are in Section B.2.

Baselines: Traditional-SSL and PET/V-PET numbers are sourced from Zhang et al. (2025) and are cited rather than independently reproduced; this is stated explicitly in Section 5.2.

Compute: Wall-clock overhead is reported in Table 13.

ACKNOWLEDGMENTS

REFERENCES

Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. In *International Conference on Learning Representations (ICLR)*, 2022.

Rie Kubota Ando and Tong Zhang. Learning on graph with Laplacian regularization. In *Advances in Neural Information Processing Systems 19 (NeurIPS 2006)*, 2006.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision (ECCV)*, 2020.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14 (NeurIPS 2001)*, 2001.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7: 2399–2434, 2006.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019. doi: 10.48550/arXiv.1905.02249.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (eds.). *Semi-Supervised Learning*. The MIT Press, 2006. ISBN 9780262033589. doi: 10.7551/MITPRESS/9780262033589.001.0001.

Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2022a.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. SoftMatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022b.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Jorge Cuadros and George Bresnick. EyePACS: An adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 2009.

Wei Dong, Xing Zhang, Bihui Chen, Dawei Yan, Zhijun Lin, Qingsen Yan, Peng Wang, and Yang Yang. Low-rank rescaled vision transformer fine-tuning: A residual design approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16101–16110. IEEE/CVF, 2024.

Mariia Drozdova, Vitaliy Kinakh, Yury Belousov, Erica Lastufka, and Slava Voloshynovskiy. Semi-supervised fine-tuning of vision foundation models with content-style decomposition, 2024.

Kai Gan and Tong Wei. Erasing the bias: Fine-tuning foundation models for semi-supervised learning. In *International Conference on Machine Learning (ICML)*, 2024.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, 2005.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330. PMLR, 2017.

H. He and Z. Goldfeld. Information-theoretic generalization bounds for deep neural networks. *IEEE Transactions on Information Theory*, 2025.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations (ICLR)*, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and Perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

S. Hou, P. Kassraie, A. Kratsios, A. Krause, and J. Rothfuss. Instance-dependent generalization bounds via optimal transport. *Journal of Machine Learning Research*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022*, volume 13693 of *Lecture Notes in Computer Science*, pp. 709–727. Springer, 2022. doi: 10.1007/978-3-031-19827-4_41.

Yangbangyan Jiang, Xiaodan Li, Yuefeng Chen, Yuan He, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. MaxMatch: Semi-supervised learning with worst-case consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5970–5987, 2023.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. DataDream: Few-shot guided dataset generation. In *European Conference on Computer Vision (ECCV)*, 2024.

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, 2014. doi: 10.48550/arXiv.1406.5298.

Dong-Hyun Lee. Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks, 2013. ICML 2013 Workshop on Challenges in Representation Learning.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL `https://arxiv.org/abs/2601.04720`.

Charlotte Loh, Rumen Dangovski, Shivchander Sudalairaj, Seungwook Han, Ligong Han, Leonid Karlinsky, Marin Soljacic, and Akash Srivastava. Mitigating confirmation bias in semi-supervised learning via efficient Bayesian model averaging. *Transactions on Machine Learning Research*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8896–8905, 2018. doi: 10.1109/CVPR.2018.00927.

Zheda Mai, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Quang-Huy Nguyen, Li Zhang, and Wei-Lun Chao. Lessons and insights from a unifying study of parameter-efficient fine-tuning (PEFT) in visual recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 14845–14857, June 2025.

Alexander Mey and Marco Loog. Improved generalization in semi-supervised learning: A survey of theoretical results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4747–4767, 2023. doi: 10.1109/TPAMI.2022.3198175.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

Lan-Cuong Nguyen, Quan Nguyen-Tri, Bang Tran Khanh, Dung D. Le, Long Tran-Thanh, and Khoat Than. Provably improving generalization of few-shot models with synthetic data. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, volume 267 of *Proceedings of Machine Learning Research*, Vancouver, Canada, 2025. arXiv:2505.24190v2.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, pp. 1–31, 2024.

Hieu Pham, Qizhe Xie, Zihang Dai, Minh-Thang Luong, and Quoc V. Le. Meta Pseudo Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11557–11568, 2021. doi: 10.1109/CVPR46437.2021.01139.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks, 2015.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:1369–1392, 2007.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arXiv.1703.05175.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

Alexandru Tifrea, Gizem Yüce, Amartya Sanyal, and Fanny Yang. Can semi-supervised learning use all the data effectively? a lower bound perspective. *Advances in Neural Information Processing Systems*, 36:21960–21982, 2023.

Victor G. Turrisi da Costa, Nicola Dall'Asen, Yiming Wang, Nicu Sebe, and Elisa Ricci. Diversified in-domain synthesis with efficient fine-tuning for few-shot classification. *arXiv preprint arXiv:2312.03046*, 2023. URL https://arxiv.org/abs/2312.03046.

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, 2020. doi: 10.1007/s10994-019-05855-6.

Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145: 90–106, 2022.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016. doi: 10.48550/arXiv.1606.04080.

Harit Vishwakarma, Yi Chen, Satya Sai Srinath Namburi GNVV, Sui Jiet Tay, Ramya Korlakai Vinayak, and Frederic Sala. Rethinking confidence scores and thresholds in pseudolabeling-based SSL. In *International Conference on Machine Learning*, 2025.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. FreeMatch: Self-adaptive thresholding for semi-supervised learning. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. doi: 10.48550/arXiv.2205.07246.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, 2020a. doi: 10.48550/arXiv.1904.12848.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020b.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2019.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Ping Zhang, Zheda Mai, Quang-Huy Nguyen, and Wei-Lun Chao. Revisiting semi-supervised learning in the era of foundation models. In *Proceedings of the Thirty-Ninth Conference on Neural Information Processing Systems (NeurIPS 2025), Poster*, 2025.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*. Springer, 2022. doi: 10.1007/978-3-031-19833-5_29.

Dengyong Zhou, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 2004.

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 (CVPR)*, pp. 16816–16825, 2022a.

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, 2022b.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

A Theoretical Proofs & Additional Remarks

We collect all proofs in this appendix. Notation follows Section 3. In particular, $D = \{(s_i, y_i)\}_{i=1}^m$ is the labeled set, $\Gamma = \{\mathcal{X}_i\}_{i=1}^m$ is an anchor partition, D^u is an i.i.d. sample of size u from P_x, $D_i^u := D^u \cap \mathcal{X}_i$, $u_i := |D_i^u|$, $\hat{\pi}_i := u_i/u$, $p_i := P_x(\mathcal{X}_i)$, and $L := \max(L_p, L_y)$.

A.1 Assumptions

The analysis of Theorem 3.5 requires two structural conditions (Assumptions 3.3 and 3.4) on the cost function c.

Remark A.1. Cost functions vary in how they relate to these two assumptions. For example, mean squared error $c(z, y) = \|z - y\|^2$ satisfies both. Vanilla cross-entropy with one-hot labels, $c(z, y) = -\log \sigma(z)_y$ (where σ is the softmax), satisfies neither: the loss diverges as $\sigma(z)_y \to 0$, and the gradient with respect to y is unbounded over the simplex.

A standard remedy is *probability clipping*,

$$c_\epsilon(z, y) \;:=\; -\sum_{k=1}^{C_\ell} y_k \,\log\max\big\{\sigma(z)_k,\, \epsilon\big\}, \tag{10}$$

which bounds the loss by $\log(1/\epsilon)$, satisfying Assumption 3.3 with $C = \log(1/\epsilon)$. Differentiating in y gives $\partial c_\epsilon/\partial y_k = -\log\max\{\sigma(z)_k, \epsilon\} \in [-\log(1/\epsilon),\, 0]$, so the gradient has ℓ_2-norm at most $\sqrt{C_\ell}\,\log(1/\epsilon)$ and Assumption 3.4 holds with $L_y = \mathcal{O}\big(\log(1/\epsilon)\big)$.

Instantiating the assumptions in practice. The supervised term of the GAFT objective is implemented with exactly this clipped cross-entropy c_ϵ, using $\epsilon = 10^{-4}$ throughout every experiment reported in this work. Assumption 3.3 therefore holds by construction with $C = \log(1/\epsilon) = \log 10^4 \approx 9.21$, and Assumption 3.4 holds with $L_y = \mathcal{O}\big(\log(1/\epsilon)\big)$. The GAFT objective is then added to the unmodified base SSL approach.

A.2 Proof of Theorem 3.5

We show Theorem 3.5 holds with probability at least $1 - \delta$ over the draw of D^u, conditional on D and Γ.

Decomposition. Define the within-region population risk and its empirical estimator:

$$F_i(h) \;:=\; \mathbb{E}_{x\sim P_x}[\ell(h, x) \mid x \in \mathcal{X}_i]\,, \qquad \widehat{F}_i(h) \;:=\; \frac{1}{u_i}\sum_{x\in D_i^u} \ell(h, x),$$

with the convention $\widehat{F}_i(h) := 0$ if $u_i = 0$. Adding and subtracting,

$$\begin{aligned} F(P_{xy}, h) = \underbrace{F(P_{xy}, h) - \sum_i \hat{\pi}_i F_i(h)}_{A_1} \;+\; \underbrace{\sum_i \hat{\pi}_i F_i(h) - \sum_i \hat{\pi}_i \widehat{F}_i(h)}_{A_3} \\ +\; \underbrace{\sum_i \hat{\pi}_i \widehat{F}_i(h) - \sum_i \hat{\pi}_i \ell(h, s_i)}_{A_2} \;+\; \sum_i \hat{\pi}_i \ell(h, s_i). \end{aligned} \tag{11}$$

We bound A_1, A_2, and A_3 separately.

Bounding A_1. Since Γ is a measurable partition, the law of total expectation gives $F(P_{xy}, h) = \sum_i p_i\, F_i(h)$, so

$$A_1 \;=\; \sum_{i=1}^m \big(p_i - \hat{\pi}_i\big)\, F_i(h).$$

Conditional on Γ, the count vector $(u_1, \dots, u_m)$ is multinomial with parameters u and $(p_1, \dots, p_m)$. The Bretagnolle–Huber–Carol inequality for multinomial distributions states that, for any $t > 0$,

$$\Pr\left(\sum_{i=1}^{m} |p_i - \hat{\pi}_i| \geq t\right) \leq 2^m \exp\bigl(-u\,t^2/2\bigr).$$

Setting the right-hand side equal to δ and solving for t yields $t = \sqrt{(2m\ln 2 - 2\ln\delta)/u} = \sqrt{(m\ln 4 - 2\ln\delta)/u}$. By Assumption 3.3, $|F_i(h)| \leq C$, so on the event $\sum_i |p_i - \hat{\pi}_i| \leq t$ we have

$$|A_1| \leq C\sum_{i=1}^{m} |p_i - \hat{\pi}_i| \leq C\sqrt{\frac{m\ln 4 - 2\ln\delta}{u}}.$$

This event holds with probability at least $1 - \delta$.

Bounding A_2. Expanding $\widehat{F}_i$:

$$A_2 = \sum_i \hat{\pi}_i \frac{1}{u_i} \sum_{x \in D_i^u} \bigl(\ell(h, x) - \ell(h, s_i)\bigr) = \sum_i \hat{\pi}_i \frac{1}{u_i} \sum_{x \in D_i^u} \bigl(c(h(x), y^*(x)) - c(h(s_i), y^*(s_i))\bigr).$$

For each summand, by adding and subtracting $c(h(s_i), y^*(x))$ and applying Assumption 3.4:

$$\begin{aligned} c(h(x), y^*(x)) - c(h(s_i), y^*(s_i)) &= \bigl[c(h(x), y^*(x)) - c(h(s_i), y^*(x))\bigr] + \bigl[c(h(s_i), y^*(x)) - c(h(s_i), y^*(s_i))\bigr] \\ &\leq L_p \,\|h(x) - h(s_i)\| + L_y \,\|y^*(x) - y^*(s_i)\| \\ &\leq L\,\bigl(\|h(x) - h(s_i)\| + \|y^*(x) - y^*(s_i)\|\bigr). \end{aligned}$$

Substituting and using Definition 3.1,

$$A_2 \leq L\sum_{i=1}^{m} \hat{\pi}_i [d(s_i, D_i^u \mid h) + d(s_i, D_i^u \mid y^*)].$$

Bounding A_3. Writing $F_i(h) - \widehat{F}_i(h)$ as a sample average over D_i^u:

$$A_3 = \sum_i \hat{\pi}_i \bigl(F_i(h) - \widehat{F}_i(h)\bigr) = \sum_i \hat{\pi}_i \frac{1}{u_i} \sum_{x' \in D_i^u} \bigl(F_i(h) - \ell(h, x')\bigr),$$

where in the last step we used that $F_i(h)$ does not depend on x'. For each fixed $x' \in D_i^u$,

$$F_i(h) - \ell(h, x') = \mathbb{E}_{x \sim P_x}[\,\ell(h, x) - \ell(h, x') \mid x \in \mathcal{X}_i\,].$$

Applying Assumption 3.4 inside the expectation, exactly as for A_2, gives

$$F_i(h) - \ell(h, x') \leq L\,\mathbb{E}_{x \sim P_x}\bigl[\,\|h(x) - h(x')\| + \|y^*(x) - y^*(x')\| \,\bigm|\, x \in \mathcal{X}_i\bigr].$$

Substituting and using Definition 3.1,

$$A_3 \leq L\sum_{i=1}^{m} \hat{\pi}_i [d(D_i^u, \mathcal{X}_i \mid h) + d(D_i^u, \mathcal{X}_i \mid y^*)].$$

Combining. Summing the three bounds and the residual $\sum_i \hat{\pi}_i\,\ell(h, s_i)$ in (11) yields, with probability at least $1 - \delta$,

$$F(P_{xy}, h) \leq \sum_{i=1}^{m} \hat{\pi}_i\,\ell(h, s_i) + L\sum_{i=1}^{m} \hat{\pi}_i [d(s_i, D_i^u \mid h) + d(D_i^u, \mathcal{X}_i \mid h)] + g_u,$$

with g_u as in the theorem statement. This completes the proof. □

A.3 Proof of Corollary 3.6

By assumption, $h = v \circ f$ with $v : \mathbb{R}^d \to \mathbb{R}^{C_\ell}$ being L_v-Lipschitz. For any $x, x' \in \mathcal{X}$,

$$\|h(x) - h(x')\| = \|v(f(x)) - v(f(x'))\| \leq L_v \,\|f(x) - f(x')\|.$$

Applying this inequality termwise to $d(s_i, D_i^u \mid h)$ and $d(D_i^u, \mathcal{X}_i \mid h)$ from Definition 3.1 yields

$$d(s_i, D_i^u \mid h) \leq L_v\, d(s_i, D_i^u \mid f), \qquad d(D_i^u, \mathcal{X}_i \mid h) \leq L_v\, d(D_i^u, \mathcal{X}_i \mid f).$$

Substituting into Theorem 3.5 and absorbing the factor L_v into the multiplicative constant gives Corollary 3.6, with the same residual g_u. □

A.4 Asymptotic Decomposition

When the unlabeled pool is plentiful, $u \to \infty$, the empirical weights $\hat{\pi}_i$ converge almost surely to the population masses $p_i := P_x(\mathcal{X}_i)$, the multinomial concentration term in g_u vanishes, and Theorem 3.5 approaches a clean deterministic decomposition. The same decomposition admits a separate, purely deterministic derivation that does not invoke unlabeled data, Lipschitz, or boundedness assumptions.

Lemma A.2 (Risk decomposition)**.** *Let $\Gamma = \{\mathcal{X}_i\}_{i=1}^m$ be any measurable partition of $\mathcal{X}$ with $s_i \in \mathcal{X}_i$, and write $p_i := P_x(\mathcal{X}_i)$. Define the local sensitivity*

$$\mathrm{Rob}(h) \;:=\; \sum_{i=1}^{m} p_i \, \mathbb{E}_{x \sim P_x}\big[\, |\ell(h,x) - \ell(h,s_i)| \,\big|\, x \in \mathcal{X}_i \big]\,.$$

Then

$$F(P_{xy}, h) \;\le\; \sum_{i=1}^{m} p_i \, \ell(h, s_i) \;+\; \mathrm{Rob}(h). \tag{12}$$

Lemma A.2 is not a finite-sample bound: it does not invoke unlabeled data explicitly, nor does it require Assumptions 3.3 and 3.4. It records the elementary fact that any risk decomposes into anchor risk plus within-region loss-sensitivity – the form GAFT's two regularizers implicitly target. Theorem 3.5 can be read as the corresponding finite-sample story: it quantifies how unlabeled samples allow $\mathrm{Rob}(h)$ to be estimated empirically and how the multinomial fluctuations of $\hat{\pi}_i$ around p_i enter the rate. This gives the cleanest reading of what the bound is selecting for.

Proof. By the partition property and the law of total expectation,

$$\begin{aligned}
F(P_{xy}, h) \;&=\; \sum_{i=1}^{m} p_i \, \mathbb{E}_{x \sim P_x}[\ell(h,x) \mid x \in \mathcal{X}_i] \\
&= \sum_{i=1}^{m} p_i \, \ell(h, s_i) + \sum_{i=1}^{m} p_i \, \mathbb{E}_{x \sim P_x}[\ell(h,x) - \ell(h,s_i) \mid x \in \mathcal{X}_i] \\
&\le \sum_{i=1}^{m} p_i \, \ell(h, s_i) + \sum_{i=1}^{m} p_i \, \mathbb{E}_{x \sim P_x}\big[\, |\ell(h,x) - \ell(h,s_i)| \,\big|\, x \in \mathcal{X}_i \big] \\
&= \sum_{i=1}^{m} p_i \, \ell(h, s_i) + \mathrm{Rob}(h),
\end{aligned}$$

where the inequality uses $\mathbb{E}[Z] \le \mathbb{E}[|Z|]$ pointwise inside each conditional expectation. □

A.5 Remarks

Remark A.3. Setting h constant within each region zeros the h-dependent geometric terms in Theorem 3.5, but does not minimize the bound, a characteristic we call *non-collapse*: the anchor-risk term $\sum_i \hat{\pi}_i \, \ell(h, s_i)$ requires $h(s_i) \approx y_i$ for every labeled anchor. Together, the two terms select for predictors that are simultaneously locally smooth on the unlabeled pool and aligned to the labeled supervision – exactly the desideratum of effective few-shot adaptation.

Remark A.4. The residual g_u in Theorem 3.5 has two components. The concentration tail $C\sqrt{(m \ln 4 - 2 \ln \delta)/u}$ vanishes at rate $O(u^{-1/2})$. The remaining label-dependent terms, $L \sum_i \hat{\pi}_i [d(s_i, D_i^u \mid y^*) + d(D_i^u, \mathcal{X}_i \mid y^*)]$, are irreducible from the perspective of the predictor: they depend on y^* and on the choice of partition, but not on h. GAFT does not optimize g_u directly; instead, the choice of partition controls it implicitly. A partition in which neighboring points share labels – i.e., y^* varies slowly within each $\mathcal{X}_i$ – makes g_u small. This is the operative assumption of any geometry-aware SSL method, and it is the assumption we instantiate in Section 4 via foundation-model embeddings.

Table 4: Summary of datasets used in this work's benchmark. Note that the number of training samples is a total of both labeled and unlabeled data points, with the number of labeled samples coming from the shot count of each dataset.

Dataset	Num. Classes	Train Size	Val Size	Test Size	Num. Shots
DTD	47	3,008	752	1,880	3
SUN397	397	49,601	17,401	21,750	3
RESISC45	45	20,160	5,040	6,300	1
CLEVR-C	8	56,000	14,000	15,000	10
KITTI	4	5,416	1,354	711	5
Retinopathy	5	36,500	9,207	42,670	40

B Experimental Details

B.1 Datasets

All experiments in this work are evaluated on the Visual Task Adaptation Benchmark (VTAB) (Zhai et al., 2019), a standardized suite for measuring how well pretrained vision representations adapt to downstream tasks under data-scarce supervision. Compared to classical SSL benchmarks (e.g., CIFAR), VTAB spans a broader set of visual domains and task types, and therefore better reflects the challenge of adapting modern VFMs when the downstream distribution deviates from the pretraining data.

Following the evaluation motivation and few-shot configuration emphasized in Zhang et al. (2025), this work selects six VTAB tasks that (i) cover diverse domain shifts and (ii) remain challenging for strong pretrained backbones without extensive supervision. Concretely, the benchmark includes:

- **DTD** (textures): fine-grained material and texture recognition (Cimpoi et al., 2014).
- **SUN397** (scenes): large-scale indoor/outdoor scene classification (Xiao et al., 2010).
- **RESISC45** (remote sensing): aerial/satellite scene classification with distinct imaging statistics (Cheng et al., 2017).
- **CLEVR-Count / CLEVR-C** (synthetic reasoning): controlled synthetic scenes for counting-based visual reasoning (Johnson et al., 2017).
- **KITTI-Dist** (autonomous driving): perception in driving scenarios, derived from KITTI (Geiger et al., 2012).
- **Retinopathy** (medical imaging): diabetic retinopathy grading from fundus photographs (Cuadros & Bresnick, 2009).

Together, these datasets span natural images, specialized sensing modalities, and synthetic distributions, making them a suitable testbed for assessing whether our proposed geometry-aware loss can help fine-tuned models improve their robustness and transfer across heterogeneous domains. Example images from each dataset are shown in Figure 3, while dataset sizes and shot counts are summarized in Table 4.

B.2 Hyperparameter Settings

To ensure a fair comparison, dataset splits, shot budgets, image preprocessing, optimizer family, and training schedule are kept consistent across methods. Specifically, for labeled-only and traditional SSL baselines (FixMatch/FlexMatch/SoftMatch), we take the results from Zhang et al. (2025), where they have followed standard defaults from prior SSL literature (Sohn et al., 2020; Wang et al., 2023; Chen et al., 2023; Gan & Wei, 2024). For PET/V-PET, we reuse the same pseudo-labeling and ensemble hyperparameters as the original paper. Detailed hyperparameter choices are as follows.

B.2.1 Common optimization, training, and preprocessing hyperparameters

Table 5 lists the shared hyperparameters used across all experiments (baselines and GAFT variants), covering optimization, training schedule, and image preprocessing.



Figure 3: Example of images from VTAB, covering a wide range of application areas. From left to right, top to bottom: DTD, SUN397, RESISC45, CLEVR-C, KITTI

AdamW (Loshchilov & Hutter, 2019) is used due to its decoupled weight decay formulation, which is known to improve stability for fine-tuning large pretrained models. The learning rate controls the update step size; momentum (β_1) smooths optimization trajectories by accumulating past gradients. Weight decay regularizes updates to reduce overfitting, and the layer decay factor is set to 1 (no depth-dependent scaling), i.e., equal learning rates across layers as commonly used in VFM fine-tuning (Oquab et al., 2024; Radford et al., 2021). For preprocessing, inputs are resized to 224 to match common pretrained VFM input resolution, while the crop ratio balances augmentation with semantic preservation.

Table 5: Common training, optimization, and image preprocessing hyperparameters used in all experiments.

Category	Hyperparameter	Value
Optimization	Optimizer	AdamW
	Learning rate	0.00005
	Momentum (β_1)	0.9
	Weight decay	0.0005
	Layer decay	1
Training	Num. epochs	30
	Batch size	8
Image Processing	Resize	224
	Crop ratio	0.875

B.2.2 PEFT CONFIGURATIONS

For LoRA, the main hyperparameter is the low-rank bottleneck size `lora_bottleneck` (Hu et al., 2022), set to 4 in all experiments. This controls the rank of the injected update matrices and thus the number of trainable parameters, encouraging compact task-specific adaptation under few-shot labels.

For AdaptFormer, we use `adapter_bottleneck` $= 16$ and `adapter_scaler` $= 0.1$, with `lora_kaiming` initialization (Chen et al., 2022b). The bottleneck determines the adapter's internal capacity, while the scaling factor modulates the adapter contribution in the residual path, preserving pretrained representations early and stabilizing optimization for large VFMs.

B.2.3 TRADITIONAL SSL-STYLE HYPERPARAMETERS

For FixMatch/FlexMatch/SoftMatch-style baselines, Zhang et al. (2025) has implemented `hard_label` to be `true`, which means that argmax pseudo-labels will be used in the corresponding training paradigms (Sohn et al., 2020; Wang et al., 2023; Chen et al., 2023). A temperature of $T = 0.5$ to sharpen logits before pseudo-labeling, a confidence cutoff of 0.95 to filter unreliable pseudo-labels, and an unlabeled loss weight of 1.0 (against the traditional supervised loss) are used in order to produce the results. Baseline SSL performances are taken from the paper *Revisiting Semi-Supervised Learning in the Era of Foundation Models*; the hyperparameter values listed here are reproduced for transparency.

B.2.4 VFM-PEFT ENSEMBLE TRAINING (PET/V-PET) HYPERPARAMETERS

PET/V-PET are configured following (Zhang et al., 2025). To reiterate, PET and V-PET are introduced as an intersection of ensembling and pseudo-labeling for SSL-style training: labeled-only fine-tuned VFMs (e.g., via LoRA or AdaptFormer) produce logits on unlabeled data, which are aggregated to form (i) a hard pseudo-label (via argmax) and (ii) a soft target distribution for distillation. This ensemble-based pseudo-labeling is then used to train a model on unlabeled samples. Because GAFT is explicitly experimented as an add-on to PET/V-PET in this work, we highlight the key hyperparameters here.

For each unlabeled image x, Zhang et al. (2025) form two perturbed views: a *weakly augmented* view x^w using light, label-preserving transforms (e.g., random crop/flip, mild color jitter), and a *strongly augmented* view x^s using heavier transformations (e.g., stronger color/geometry perturbations). The weak view is intended to preserve semantic identity, while the strong view enforces invariance under more aggressive perturbations. In PET/V-PET, losses can be applied to either or both views.

Let $\hat{y}$ be the hard pseudo-label (argmax of the ensemble) and q the soft distillation target (normalized ensemble scores). Let $p(\cdot \mid x^w)$ and $p(\cdot \mid x^s)$ be the student predictions from weak/strong views, and let T be the temperature. The PET/V-PET loss is a weighted sum:

$$\begin{aligned}\mathcal{L} = w_\alpha \,\mathrm{CE}(p(\cdot \mid x^w), \hat{y}) + s_\alpha \,\mathrm{CE}(p(\cdot \mid x^s), \hat{y}) \\ + kd_{w\alpha}\,\mathrm{KL}(\log p_T(\cdot \mid x^w) \,\|\, q) + kd_{s\alpha}\,\mathrm{KL}(\log p_T(\cdot \mid x^s) \,\|\, q)\end{aligned} \tag{13}$$

where $p_T(\cdot \mid x) = \mathrm{softmax}\big(\frac{z(x)}{T}\big)$ and $z(x)$ are logits. As a result, the relevant weightings are interpreted as follows.

- w_α: weight on the cross-entropy loss using the weak view and hard pseudo-labels.
- s_α: weight on the cross-entropy loss using the strong view and hard pseudo-labels.
- $kd_{w\alpha}$: weight on the knowledge-distillation KL term for the weak view.
- $kd_{s\alpha}$: weight on the knowledge-distillation KL term for the strong view.
- T (temperature): softens/sharpens logits for KD. Lower T sharpens (more confident), higher T softens (more diffuse) the student distribution.

To summarize, all hyperparameters mentioned in Section B.2.2, B.2.3, B.2.4 are shown in Table 6. As explained, experiments with GAFT do not require any modification to the underlying SSL paradigm; therefore, all hyperparameters are kept similar to previous works for a fairer comparison.

B.2.5 GAFT-SPECIFIC HYPERPARAMETERS

GAFT introduces a small set of additional hyperparameters that control the geometry-aware regularization independently of the underlying SSL pipeline, summarised in Table 7.

Regularization weights (λ_1, λ_2). Due to constraints in resources, the alignment and robustness coefficients were *fixed* beforehand to be λ_1=0.01, λ_2=0.1 across all datasets and backbones without further tuning. Later, we conducted an additional ablation study into the impact of varying these values; results are reported in Section C.5 and C.6. Overall, the stability of performance across a broad range of λ values supports using a single fixed pair rather than per-dataset search, and means GAFT does not introduce a hidden per-task tuning cost.

Table 7: GAFT-specific hyperparameters.

Hyperparameter	Value
λ_1 (alignment)	0.01
λ_2 (robustness)	0.1
I	5 epochs
Mini-batch strategy	Labeled-first + Cluster-conditioned
u_ratio	DTD, KITTI: 12 RESISC45: 6 SUN397, CLEVR-C: 4 Retinopathy: 4

Geometry refresh interval I. The partition Γ is recomputed every I=5 epochs throughout training. This value was fixed across all experiments; its empirical effect on accuracy and partition purity is studied in Section C.1.

Table 6: Summary of hyperparameter configurations across PEFT methods, traditional SSL baselines, and VFM-PEFT ensemble training (kept consistent with prior work).

Category	Hyperparameter	Value
PEFT (LoRA)	`lora_bottleneck`	4
	Initialization	Default
	Trainable params	Low-rank only
PEFT (AdaptFormer)	`adapter_bottleneck`	16
	`adapter_scaler`	0.1
	Initialization	`lora_kaiming`
	Injection style	Residual adapter
Traditional SSL	`hard_label`	True
	Temperature T	0.5
	Confidence cutoff	0.95
	Unlabeled loss weight	1.0
	Pseudo-label type	Argmax (hard)
VFM-PEFT Ensemble (PET)	`w_alpha`	0.1
	`s_alpha`	0.1
	`kd_w_alpha`	1.0
	`kd_s_alpha`	1.0
	Temperature	1.0

Mini-batch construction. We use labeled-first sampling with cluster-conditioned unlabeled sampling (Section 4.2) as the default strategy. The unlabeled-to-labeled ratio per batch u_ratio is set per dataset based on available compute: larger datasets require a smaller ratio to keep per-iteration time tractable. At each iteration, $\mathcal{R}_{\text{align}}$ and $\mathcal{R}_{\text{robust}}$ are computed from the filtered mini-batch view of the global partition cached from the most recent refresh.

C Additional Empirical Results

This section collects additional empirical results that complement the main paper. We begin by examining two core design choices of GAFT: the geometry refresh interval I (Section C.1) and the mini-batch construction strategy (Section C.2). We then broaden the empirical scope with results on the CLIP backbone (Section C.3), a range of shot counts (Section C.4), and a sensitivity analysis of the regularization weights (λ_1, λ_2) (Section C.5). The remaining subsections address questions of method soundness: we ablate the necessity of each regularizer individually (Section C.6), report wall-clock and memory overhead (Section C.7), examine the correlation between partition purity and downstream accuracy (Section C.8), and compare against synthetic-data augmentation pipelines (Section C.9) and geometry-oriented SSL baselines (Section C.10). Finally, Sections C.11 and C.12 report adversarial robustness under the Square Attack and additional corruption-robustness results on KITTI and DTD.

C.1 Effect of geometry refresh interval

GAFT periodically refreshes the unlabeled-to-anchor assignment to track representation drift. We vary the refresh interval I (2, 5, 10 epochs, and no refresh) while fixing the rest of the training recipe, and report accuracy trajectories on DTD (3-shot) with AdaptFormer.

Figure 4 shows that a moderate refresh rate ($I = 5$) yields the best and most stable final accuracy.

More frequent refreshes ($I = 2$) can introduce mild late-stage oscillations, while infrequent refreshes ($I = 10$) underperform due to stale assignments. Disabling refresh altogether is consistently worst, confirming that geometry updates are important for keeping GAFT's neighborhood regularization aligned with the evolving representation space. We therefore use $I = 5$ epochs by default.

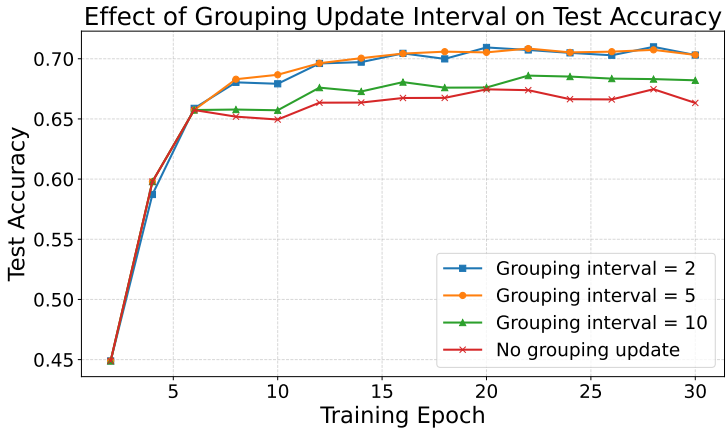
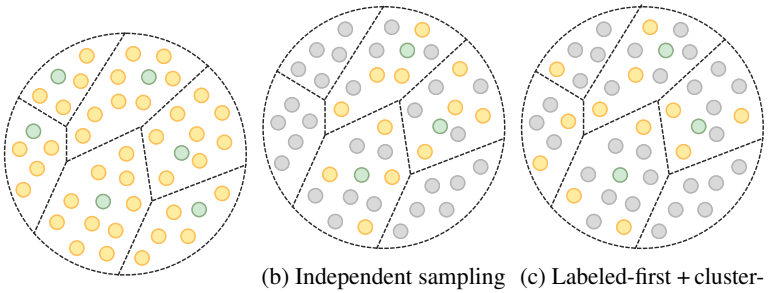


Figure 4: Effect of geometry refresh interval I on test accuracy over epochs (DTD 3-shot, AdaptFormer). The legend variable (labelled *Grouping interval* in the figure) is the geometry refresh interval I.

C.2 Effect of mini-batch composition

As discussed in Section 4.2, GAFT operates on a batch-level realization of the global partition, which makes mini-batch construction non-trivial. Figure 5 contrasts the global partition with the two strategies, which we compare under the same global anchor assignment and unlabeled-to-labeled ratio: *(i) labeled-first, cluster-conditioned sampling* (Figure 5c), where unlabeled samples are drawn only from the neighborhoods of the sampled labeled anchors; and *(ii) independent sampling, batch-filtered* (Figure 5b), where labeled and unlabeled samples are drawn independently and then grouped using the global assignment for loss computation.



(a) Global partition. (b) Independent sampling + batch filtering. (c) Labeled-first + cluster-conditioned sampling.

Figure 5: Mini-batch realizations of the anchor-induced geometry. Green: labeled anchors; yellow: unlabeled samples; gray: samples not used in the current mini-batch. Panel a is the global anchor-induced partition, built once per geometry refresh; panels b and c are the two mini-batch sampling strategies of Section 4.2.

Table 8 reports representative final-epoch top-1 accuracy results on DTD and KITTI using V-PET as the base SSL pipeline, evaluated under both AdaptFormer and LoRA, averaged across three seeds. Across three out of four configurations, the labeled-first cluster-conditioned strategy yields higher final accuracy than independent sampling with batch filtering, with the difference being more pronounced for LoRA-based fine-tuning. This suggests that explicitly anchoring each mini-batch around labeled samples and their corresponding cluster neighborhoods provides a more reliable structural signal for the proposed geometry-aware regularization.

Notably, the performance gap is smaller for AdaptFormer, which has higher intrinsic adaptation capacity than LoRA due to our choice of hyperparameters. We hypothesize that stronger adaptation modules are less sensitive to auxiliary training signals, whereas lower-capacity PEFT methods benefit more from carefully structured supervision. Based on these results, we adopt labeled-first cluster-conditioned mini-batch construction in all remaining experiments.

Table 8: Mini-batch strategy ablation using V-PET under different PEFT backbones. Final accuracy (%) on DTD and KITTI, averaged over three seeds. **Bold** marks the better of the two strategies per column.

	AdaptFormer		**LoRA**	
Mini-batch Strategy	DTD (3-shot)	KITTI (5-shot)	DTD (3-shot)	KITTI (5-shot)
Labeled-first + Cluster-conditioned	69.13	**58.90**	**69.17**	**61.67**
Independent sampling + Batch-filtered	**69.68**	58.46	68.94	60.10

C.3 Additional results on CLIP

The main paper evaluates GAFT with DINOv2 as the backbone. Since the bound in Theorem 3.5 and its backbone-induced form in Corollary 3.6 make no assumptions specific to DINOv2, GAFT should generalize to other vision foundation models. We verify this by repeating the main experiment with CLIP ViT-B/16 (Radford et al., 2021) as the backbone, fine-tuning with both LoRA and AdaptFormer on three representative VTAB tasks (DTD, RESISC45, KITTI). Hyperparameters from Section B.2 are reused without modification.

Table 9: Top-1 accuracy (%) of fine-tuned **CLIP** (ViT-B/16) on three VTAB tasks. Number in parentheses denotes shots per class. Baselines reported by Zhang et al. (2025); **bold** marks the best result per column within each PEFT block.

PEFT	**Method**	**DTD** (3)	**RESISC45** (1)	**KITTI** (5)
LoRA	ST	58.9	64.1	59.8
	PET	63.4	67.5	60.3
	PET + GAFT	66.70	78.38	69.62
	V-PET	65.6	66.6	58.0
	V-PET + GAFT	**68.72**	**79.44**	**70.18**
AdaptFormer	ST	62.0	68.8	58.2
	PET	63.8	69.5	60.5
	PET + GAFT	66.28	76.05	**68.92**
	V-PET	65.7	67.9	58.2
	V-PET + GAFT	**68.83**	**78.11**	67.65

Table 9 reports top-1 accuracy on a subset of three representative tasks. Across both PEFT mechanisms, the + GAFT variants improve their respective baselines on every dataset, with the largest gains observed on RESISC45 and KITTI – mirroring the pattern seen for DINOv2 in Table 1. The margins on these two tasks are the widest we observe anywhere: with LoRA, GAFT lifts PET by $+10.88$ points on RESISC45 (67.5 to 78.38) and by $+9.32$ points on KITTI (60.3 to 69.62). This is consistent with the reading of Corollary 3.6: the bound's geometric terms are controlled in f-space up to the head's Lipschitz constant, regardless of the specific pretraining strategy used for f. The plug-and-play property of GAFT therefore extends naturally to backbones outside the DINOv2 family.

C.4 Analyzing more few-shot settings

Traditional research in SSL often analyzes model performance across different shot counts, i.e. the number of labeled samples per class that the training process is allowed to see (Sohn et al., 2020; Chen et al., 2023; Jiang et al., 2023). In this section, we set out to conduct a similar analysis. Concretely, we fine-tune DINOv2, using AdaptFormer and LoRA, on PET and V-PET SSL baselines, while varying the shot counts, which requires us to fully re-run the pseudo-labeling process for each count. We report the final-epoch top-1 accuracy per trained model. Other hyperparameters in Section B.2 applies.

Table 10: Top-1 accuracy (%) on DTD under different shot counts. We compare PET and V-PET with and without GAFT under LoRA and AdaptFormer fine-tuning. **Bold** marks the higher accuracy within each baseline/+ GAFT pair, per column.

PEFT	Method	DTD		
		1-shot	3-shot	8-shot
AdaptFormer	PET	44.68	67.20	72.77
	PET (+GAFT)	**47.34**	**69.34**	**74.84**
	V-PET	49.79	68.20	74.63
	V-PET (+GAFT)	**57.07**	**69.13**	**77.93**
LoRA	PET	43.62	66.70	72.39
	PET (+GAFT)	**46.54**	**70.41**	**75.90**
	V-PET	49.47	67.80	74.57
	V-PET (+GAFT)	**56.44**	**69.17**	**78.19**

Table 10 studies how the proposed structure-aware regularization (GAFT) behaves as the labeled budget increases on DTD, ranging from an extreme 1-shot regime to an 8-shot setting. It can be observed that across both PEFT mechanisms and both ensemble paradigms, GAFT provides consistent improvements over the corresponding PET/V-PET baselines, with the gains being most pronounced when labeled supervision is scarce.

From the table, we consider the clearest trend to be that *GAFT yields its largest improvements in the 1-shot setting*. For AdaptFormer, PET increases from 44.68 to 47.34 (+2.66%), while V-PET improves from 49.79 to 57.07 (+7.28%). A similar pattern holds for LoRA—this behavior is consistent with the intended role of GAFT: when class supervision is extremely limited, pseudo-labeling alone is more vulnerable to confirmation bias and early miscalibration (Arazo et al., 2020; Guo et al., 2017; Rigollet, 2007), and the additional geometry-aware constraints encourage more stable within-manifold behavior, effectively acting as a stronger inductive bias on the unlabeled pool.

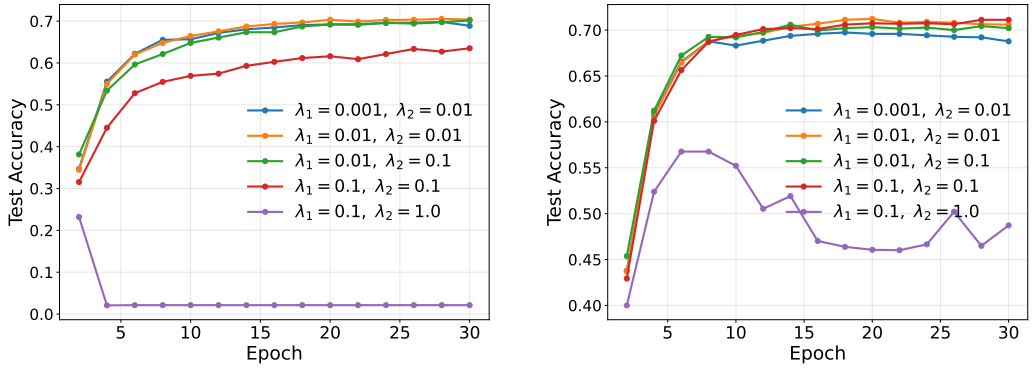
As the shot count increases beyond 1-shot, the advantage of GAFT, though not as prominent, is still significant. For PET, the improvements remain positive at higher supervision levels (e.g., AdaptFormer + PET: +2.14% at 3-shot and +2.07% at 8-shot; LoRA + PET: +3.71% at 3-shot and +3.51% at 8-shot). In our opinion, this is expected: with more labeled samples per class, the supervised signal becomes more sufficient to anchor the representation and reduce reliance on auxiliary unlabeled structure; regardless, the inclusion of GAFT still induces a reliable increase in model accuracy.

To make the trend quantitative rather than anecdotal, Table 11 collects the absolute gains of every configuration in Table 10.

Table 11: Absolute gain (percentage points) from adding GAFT, derived from Table 10. **Bold** marks the largest gain within each row.

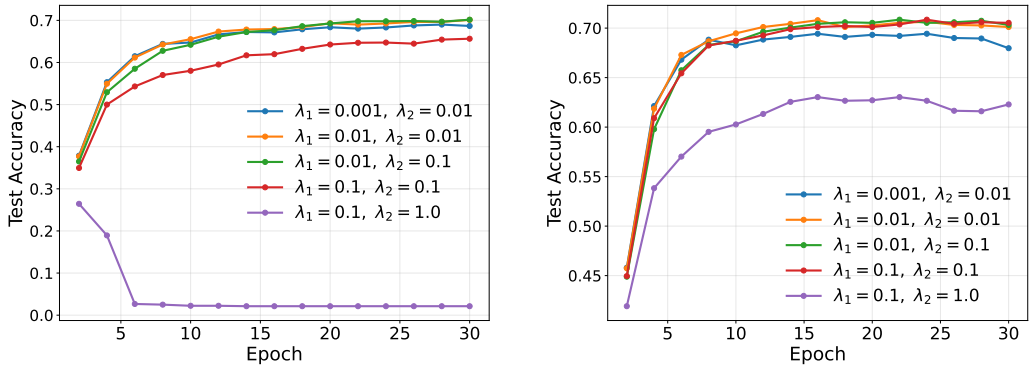
PEFT	Method	1-shot	3-shot	8-shot
AdaptFormer	PET	**+2.66**	+2.14	+2.07
	V-PET	**+7.28**	+0.93	+3.30
LoRA	PET	+2.92	**+3.71**	+3.51
	V-PET	**+6.97**	+1.37	+3.62
Mean		**+4.96**	+2.04	+3.13

The mean gain is largest at 1-shot ($+4.96$ points), and the 1-shot gain is the largest for three of the four configurations. The one exception, LoRA + PET, peaks at 3-shot ($+3.71$) rather than 1-shot ($+2.92$). Across the sweep, all twelve entries are positive.



(a) LoRA + PET

(b) LoRA + V-PET



(c) AdaptFormer + PET

(d) AdaptFormer + V-PET

Figure 6: Effect of regularization weights (λ_1, λ_2) on test accuracy across training epochs. Results are shown for two PEFT mechanisms and two ensemble paradigms on DTD (3-shot) with DINOv2.

C.5 Ablation studies on weights for geometry-aware terms

In this section, we investigate the sensitivity of the proposed geometry-aware loss to the weighting coefficients λ_1 and λ_2, which control the relative strength of the alignment term and the robustness term, respectively. While the main experiments fix these values to a single configuration, we understand that it is important to verify that the observed gains are not the result of brittle hyperparameter tuning, and that the method behaves stably under reasonable variations in the weights.

Hence, we conduct an ablation by varying (λ_1, λ_2) across a range of values and tracking the test accuracy over training epochs. All experiments are run on DTD in the 3-shot regime, fine-tuning the DINOv2 backbone, and hyperparameters in Section B.2 apply.

Figure 6 reveals several consistent trends. First, moderate values of λ_1 and λ_2 yield stable convergence and improved final accuracy, whereas overly large weights—particularly for λ_2—can substantially degrade performance. Conversely, setting both λ_1 and λ_2 to very small values (e.g., $\lambda_1 = 0.001$, $\lambda_2 = 0.01$) does not collapse training, but tends to yield suboptimal accuracy toward the final epochs, again most noticeably for LoRA under V-PET.

Second, configurations in which λ_1 and λ_2 lie on a similar scale (e.g., $\lambda_1 = 0.01$, $\lambda_2 = 0.1$) consistently achieve strong performance across both PET and V-PET and for both PEFT mechanisms. This suggests that alignment and robustness act in a complementary manner, and that the proposed method does not rely on a finely tuned asymmetry between the two regularization terms. Finally, AdaptFormer is generally less sensitive to weight selection than LoRA (6c and 6d), aligning with the intuition that higher-capacity PEFT modules (due to the choice of hyperparameters) rely less on auxiliary structural guidance.

Overall, this ablation indicates that the proposed geometry-aware regularization is robust to reasonable variations of (λ_1, λ_2), provided the weights remain within a moderate range. Based on these results, we adopt $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$ as the default configuration for all remaining experiments, offering a consistent trade-off between stability and performance across settings.

C.6 Ablation studies on necessity of both regularizers

In addition to the reweighted supervised loss, the GAFT objective in Section 4.1 introduces two regularizers, $\mathcal{R}_{\text{align}}$ and $\mathcal{R}_{\text{robust}}$, each derived from a distinct h-dependent term in Theorem 3.5. We verify that both terms contribute to the empirical gains by comparing the full objective against three reduced variants of that same objective. The reference variant disables both regularizers ($\lambda_1{=}\lambda_2{=}0$) while retaining the anchor-reweighted supervised term, so that the only quantities varied across the row are the two coefficients attached to $\mathcal{R}_{\text{align}}$ and $\mathcal{R}_{\text{robust}}$; the remaining two variants enable $\mathcal{R}_{\text{robust}}$ only ($\lambda_1{=}0$) and $\mathcal{R}_{\text{align}}$ only ($\lambda_2{=}0$). All experiments use DINOv2 on DTD (3-shot) under both PEFT mechanisms; coefficients are fixed to $\lambda_1{=}0.01$ and $\lambda_2{=}0.1$ when active.

Table 12: Top-1 accuracy (%) on DTD (3-shot, DINOv2). The first column is the GAFT objective with both regularizers disabled, which retains the anchor-reweighted supervised term; the remaining columns activate $\mathcal{R}_{\text{robust}}$, $\mathcal{R}_{\text{align}}$, and both. **Bold** marks the best result per row.

Method	$\boldsymbol{\lambda_1{=}\lambda_2{=}0}$	**+** $\mathcal{R}_{\text{robust}}$	**+** $\mathcal{R}_{\text{align}}$	**+ Both**
PET, AdaptFormer	67.47	68.86	68.29	**69.34**
V-PET, AdaptFormer	68.29	68.51	68.80	**69.13**
PET, LoRA	67.59	67.81	69.67	**70.41**
V-PET, LoRA	68.03	68.98	68.87	**69.17**

Table 12 reports the result. Each individual regularizer improves over the reference variant, confirming that both $\mathcal{R}_{\text{align}}$ and $\mathcal{R}_{\text{robust}}$ contribute useful signal. The full objective with both terms active outperforms either single-term variant in every configuration, supporting the design choice of including both so that they target complementary aspects of the geometry. Reading the reference column against the corresponding PET and V-PET rows of Table 1 additionally quantifies how much of the gain is attributable to the anchor-reweighted supervised term on its own, separately from the two geometric regularizers.

C.7 Computational and memory overhead

GAFT introduces two sources of compute beyond the baseline SSL pipeline: (i) the forward-pass and loss computations for $\mathcal{R}_{\text{align}}$ and $\mathcal{R}_{\text{robust}}$ in each iteration, and (ii) the periodic refresh of the partition Γ every I epochs (Section 4.2.2). We measure both on DTD (3,008 training samples) using a single NVIDIA RTX PRO 6000 Blackwell GPU, with $I{=}5$ over 30 epochs (yielding 6 refreshes including the initial one). Wall-clock measurements are averaged over the full training run.

Table 13: Wall-clock training time (seconds) on DTD with DINOv2. "Epoch-avg" is the average epoch time; "Total" is the full 30-epoch training time; "Avg refresh" is the average time per partition refresh.

PEFT	**Method**	**Epoch-avg** (no GAFT)	**Epoch-avg** (GAFT)	**Total** (no GAFT)	**Total** (GAFT)	**Avg refresh** (GAFT)
AdaptFormer	PET	139.83	147.17	4196.86	4416.89	13.70
	V-PET	139.80	147.04	4196.02	4412.97	13.76
LoRA	PET	142.45	150.70	4275.38	4523.03	15.25
	V-PET	141.36	147.84	4242.72	4437.12	15.04

Time overhead. Table 13 shows that GAFT adds roughly 5–6% to the per-epoch wall-clock time across all four configurations. Each partition refresh, which involves a forward pass of f over the full

training set followed by nearest-anchor assignment, takes between 13 and 16 seconds on DTD; this scales linearly in the dataset size, but remains a small fraction of the per-epoch cost owing to the use of efficient nearest-neighbour search via `faiss`. As discussed in Section C.1, the refresh interval I also does not need to be small: $I{=}5$ already yields stable performance, which further softens the trade-off.

Memory overhead. The additional memory footprint of GAFT is modest relative to the backbone's training activations. Per iteration, $\mathcal{R}_{\text{align}}$ and $\mathcal{R}_{\text{robust}}$ are computed directly from the mini-batch logits already produced by the forward pass, so they introduce only the storage required for batch-level tensors. Between refreshes, GAFT caches the labeled and unlabeled embeddings used to build Γ and reuses them across iterations; after each refresh, only the lightweight assignment map π is retained. In our experiments, this overhead is dominated by the activations of the PEFT backbone itself.

C.8 Cluster purity and downstream accuracy

Theorem 3.5 bounds expected risk in terms of two h-dependent geometric terms together with an h-independent residual g_u, the latter of which depends on how well y^* varies within each region $\mathcal{X}_i$ (Remark A.4). A partition in which neighbouring points share labels makes g_u small. While we cannot compute g_u directly without access to y^*, we can measure its empirical proxy – the *cluster purity* of Γ, defined as the fraction of unlabeled samples whose true label matches their assigned anchor's label – and ask whether higher purity tracks better downstream accuracy.

Table 14: Cluster purity (%) and test accuracy (%) across partition refreshes, for V-PET + GAFT with a CLIP backbone. The first three blocks use AdaptFormer; the final block repeats KITTI with LoRA. Test accuracy at epoch 0 is omitted as the model has not been trained. **Bold** in the final column highlights the Pearson correlation between purity and test accuracy over the trained epochs; no other cell is emphasised.

Task		Epoch 0	5	10	15	20	25	Corr.
KITTI (5)	Test acc	–	64.14	65.96	67.23	67.23	67.09	**0.9885**
	Purity	56.44	64.57	71.06	74.67	76.09	76.20	
DTD (3)	Test acc	–	61.54	63.56	65.64	67.45	68.12	**0.9922**
	Purity	46.93	62.19	66.42	71.17	73.40	74.12	
Retinopathy (40)	Test acc	–	27.84	31.63	34.19	37.41	40.93	**0.9836**
	Purity	15.74	19.32	25.64	28.01	32.16	34.09	
KITTI (5), LoRA	Test acc	–	59.92	64.76	69.20	68.35	67.23	**0.9210**
	Purity	56.66	79.40	81.95	82.89	83.63	83.96	

Table 14 reports purity and test accuracy at each refresh, along with the Pearson correlation between the two over the trained epochs. The relationship is consistent across all three benchmarks and both PEFT mechanisms: every correlation exceeds 0.92, and the three AdaptFormer settings are above 0.98. Two observations follow. First, partitions become *more* pure as training progresses, indicating that the backbone f is not merely drifting but actively producing more semantically homogeneous neighbourhoods over the unlabeled pool. Second, the correlation supports the central role of g_u in the bound: improvements in partition quality, which tighten the residual g_u, track the actual gains in downstream accuracy.

Diabetic Retinopathy is the most demanding case and therefore the most informative. Its epoch-0 purity is only 15.74%, meaning that at initialization the great majority of nearest-anchor assignments do not share their anchor's label: the pretrained CLIP geometry is close to uninformative for this task. GAFT nonetheless improves steadily, and purity more than doubles over training (15.74% to 34.09%). This matters for the concern that GAFT simply inherits an already-good pretrained geometry, since here there is very little such geometry to inherit.

Is the purity increase caused by GAFT? The trends above are consistent with GAFT tightening g_u, but on their own they do not separate that effect from the purity gains any fine-tuning would

produce. We therefore repeat the Retinopathy run with $\mathcal{L}_{\mathrm{GAFT}}$ removed, holding everything else fixed.

Table 15: Cluster purity (%) and test accuracy (%) with and without GAFT. CLIP backbone with AdaptFormer, V-PET, Diabetic Retinopathy (40-shot). **Bold** marks the higher of the two variants in each column, comparing accuracy against accuracy and purity against purity.

Variant		Epoch 5	10	15	20	25	Corr.
With GAFT	Test acc	**27.84**	**31.63**	**34.19**	**37.41**	**40.93**	0.9836
	Purity	**19.32**	**25.64**	**28.01**	**32.16**	**34.09**	
Without GAFT	Test acc	24.12	28.85	32.83	35.56	38.62	0.9901
	Purity	18.06	21.93	25.70	29.47	30.29	

Purity rises in both variants, confirming that fine-tuning alone reorganizes the embedding space. But Table 15 shows the increase is markedly larger with GAFT: by epoch 25 purity is 34.09% against 30.29%, a gain of 3.80 points, and test accuracy is 40.93% against 38.62%, a gain of 2.31 points. The purity–accuracy correlation remains above 0.98 in both variants, so the two quantities move together regardless; what GAFT changes is how far they move. This is the most direct evidence we have that the mechanism is the one the bound describes: GAFT actively improves the partition it conditions on, and the residual g_u tightens as a result.

C.9 Comparison to synthetic-data-based techniques

With the rise of generative models, recent few-shot adaptation methods have explored improving generalization by augmenting scarce labeled samples with *synthetic images* generated by text-to-image models. In this subsection, we compare GAFT-enhanced V-PET against representative synthetic-data pipelines, focusing on the DTD and SUN397 tasks, which are commonly used few-shot benchmarks in this literature. Again, hyperparameters from Section B.2 are reused here.

Table 16 reports Top-1 accuracy on DTD and SUN397, with baseline methods extracted from Table 1 of Nguyen et al. (2025). Concretely, we include CLIP (zero-shot), Real-finetune, and three prior synthetic-data-based techniques: IsSynth (He et al., 2023), DISEF (Turrisi da Costa et al., 2023), and DataDream (Kim et al., 2024). We also include the two variants (denoted "Nguyen et al.") reported in (Nguyen et al., 2025) (lightweight and full).

Although these approaches are frequently framed as "few-shot + synthetic augmentation", they effectively exploit *unlabeled* information via large pretrained models and filtering stages, and thus can be a worthy baseline for GAFT, which also utilizes such data, though in a different fashion. For example, IsSynth (He et al., 2023) generates synthetic samples by perturbing real images and filters candidates using CLIP-based scoring. DISEF (Turrisi da Costa et al., 2023) further promotes diversity through captioning and uses LoRA-based adaptation of CLIP for improved selection. DataDream

Table 16: Top-1 accuracy (%) on DTD and SUN397. Rows above the separator are from Nguyen et al. (2025). The last two rows report V-PET+GAFT under the VTAB 16-shot setting. **Bold** marks the best result per column.

Method	DTD	SUN
CLIP (zero-shot)	46.1	72.2
Real-finetune	73.9	77.1
IsSynth	75.1	77.7
DISEF	74.3	77.6
$\text{DataDream}_{\text{cls}}$	73.1	77.5
$\text{DataDream}_{\text{dset}}$	74.1	77.5
Nguyen et al. (lightweight)	75.5	77.6
Nguyen et al. (full)	74.5	77.7
V-PET+GAFT (AF)	78.62	**79.23**
V-PET+GAFT (LoRA)	**80.11**	78.87

(Kim et al., 2024) fine-tunes the generator on few-shot data to better align synthetic samples with the target dataset, and considers both class-level and dataset-level synthesis variants.

Across these pipelines, the additional supervision signal is extracted primarily in *pixel space* (via generation) and then refined by *model-based filtering* (e.g., CLIP), rather than by directly leveraging the structure of real unlabeled samples in representation space. To provide a fair comparison with GAFT, we follow Kim et al. (2024) and use 16 labeled samples per class for each dataset, before conducting GAFT-style training and extracting the final-epoch top-1 accuracy.

Table 16 shows that GAFT ubiquitously improves upon synthetic-data-based augmentation. For example, on DTD, GAFT yields strong gains: V-PET+GAFT (LoRA) reaches 80.11%, which exceeds all synthetic-data baselines reported in (Nguyen et al., 2025), including their full method (74.5%) and IsSynth (75.1%). These results suggest that leveraging GAFT through representation-space structure can be an effective method to augment few-shot training.

From here, we hypothesize that synthetic-data approaches may suffer from mismatch between generated and real images, and their performance depends heavily on the generator quality and filtering heuristics. In comparison, GAFT avoids this bias by relying only on real unlabeled samples and enforcing consistency within feature-space neighborhoods that evolve during fine-tuning.

C.10 Comparison to geometry-oriented baselines

A natural question is how GAFT relates to simpler geometry-oriented SSL methods that also exploit the structure of unlabeled data without relying on pseudo-label confidence. We consider two representatives. Label Propagation (LP) (Iscen et al., 2019) constructs a k-nearest-neighbour graph over the embeddings of labeled and unlabeled samples, propagates labels along graph edges, and uses the propagated labels as targets for further training. Laplacian Regularization (LapReg) (Ando & Zhang, 2006) instead adds a graph-Laplacian penalty that directly encourages predictions to vary smoothly along the same graph, in the spirit of manifold regularization (Belkin et al., 2006).

Because GAFT is added on top of an existing objective rather than replacing it (Section 4.1), the informative comparison is not GAFT against these methods but GAFT *applied to* them. We therefore run each baseline with and without $\mathcal{L}_{\mathrm{GAFT}}$, holding the backbone, PEFT mechanism, and optimization schedule fixed.

Table 17: Top-1 accuracy (%) of graph-based SSL baselines with and without GAFT. DINOv2 backbone with LoRA; shots in parentheses. **Bold** marks the better of each base method and its GAFT-augmented counterpart (pairwise comparison).

Method	**DTD** (3)	**KITTI** (5)
Label Propagation (LP)	57.12	35.72
LP + GAFT	**60.74**	**43.58**
Laplacian Regularization (LapReg)	56.49	42.15
LapReg + GAFT	**60.53**	**44.23**

Table 17 shows that GAFT improves both graph-based methods on both tasks: $+3.62$ and $+7.86$ points over LP on DTD and KITTI respectively, and $+4.04$ and $+2.08$ points over LapReg. This is the behavior the theory predicts. LP and LapReg both derive their signal from pairwise affinities computed in the embedding space, whereas $\mathcal{R}_{\mathrm{align}}$ and $\mathcal{R}_{\mathrm{robust}}$ constrain the predictor relative to *labeled anchors* with the bound-derived weights $\hat{\pi}_i$ of Theorem 3.5. The two mechanisms therefore act on different quantities and compose rather than compete.

C.11 Additional corruption-robustness results

The corruption-robustness analysis in Section 5.3 reports radar plots only for one PEFT-by-baseline combination on KITTI due to space constraints in the main paper. Here we provide the complete set of corruption-robustness results: the remaining KITTI radar plots (LoRA fine-tuning), and the analogous evaluation on DTD under both PEFT mechanisms. The setup – 18 corruption types in four families, 5 severity levels each, applied at test time only – is identical to that of Section 5.3.

Concretely, the training procedure is left unchanged throughout. To avoid conflating the corruption with the backbone's input normalization, we de-normalize each test input (using the ImageNet mean and standard deviation), apply the corruption operator at the chosen severity, and then re-normalize before inference.

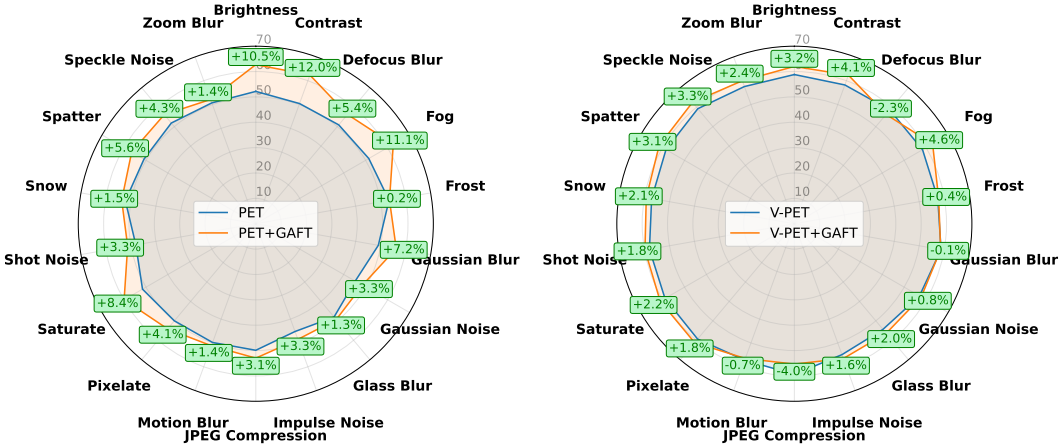


Figure 7: Robustness of DINOv2 to common image corruptions on **KITTI** when fine-tuned with **LoRA**. Each spoke is a corruption type; values are per-corruption accuracies averaged over five severities ($\overline{\mathrm{Acc}}(c)$). Annotations report the absolute change relative to the corresponding baseline, with negative values denoting a decrease. Average gain: *(PET)* 4.86%, *(V-PET)* 1.46%.

Figure 7 completes the KITTI evaluation begun in Figure 2: under LoRA fine-tuning, GAFT improves the corruption-averaged accuracy by 4.86% over PET and 1.46% over V-PET, with broad gains across the noise, blur, and weather corruption families. The pattern matches the AdaptFormer KITTI results in the main text: gains are larger for PET than V-PET, consistent with V-PET already enjoying the benefit of pseudo-label ensembling.

Figure 8 reports the same evaluation on DTD. Across all four panels GAFT improves the corruption-averaged accuracy. The DTD gains are smaller in magnitude than the KITTI gains, but their distribution across corruption families is similarly broad, with the largest gains on blur and weather corruptions – corruption types whose effect on the input is most readily smoothed out by within-region prediction consistency.

C.12 Adversarial Robustness

Besides the corruption-robustness results in Sections C.11 and 5.3, we also evaluate the more stringent setting of black-box adversarial attacks, in which the perturbation is actively optimized against the model. We use *Square Attack* (Andriushchenko et al., 2020), a query-efficient black-box adversarial attack that perturbs random square regions of an image to reduce model confidence. We use $\epsilon = 8/255$ as the perturbation budget, 1000 queries as the budget per image, and an initial square size of 0.05 of the image area, applied at test time only to the held-out test set of DTD (3-shot).

Table 18: Top-1 accuracy (%) on DTD (3-shot, DINOv2) under Square Attack (Andriushchenko et al., 2020). **Bold** marks the higher accuracy per row.

Method	Baseline	+GAFT
PET, AdaptFormer	35.96	**36.60**
V-PET, AdaptFormer	36.97	**37.29**
PET, LoRA	31.70	**33.09**
V-PET, LoRA	32.71	**32.87**

Table 18 reports top-1 test accuracy when the test set is under attack. All trained models suffer substantial accuracy loss relative to clean DTD test accuracy (Table 1), as expected for a strong black-box attack. Among the four configurations, the GAFT-augmented variants retain a higher accuracy than their respective baselines in every case. The improvement is modest, on the order of 0.2

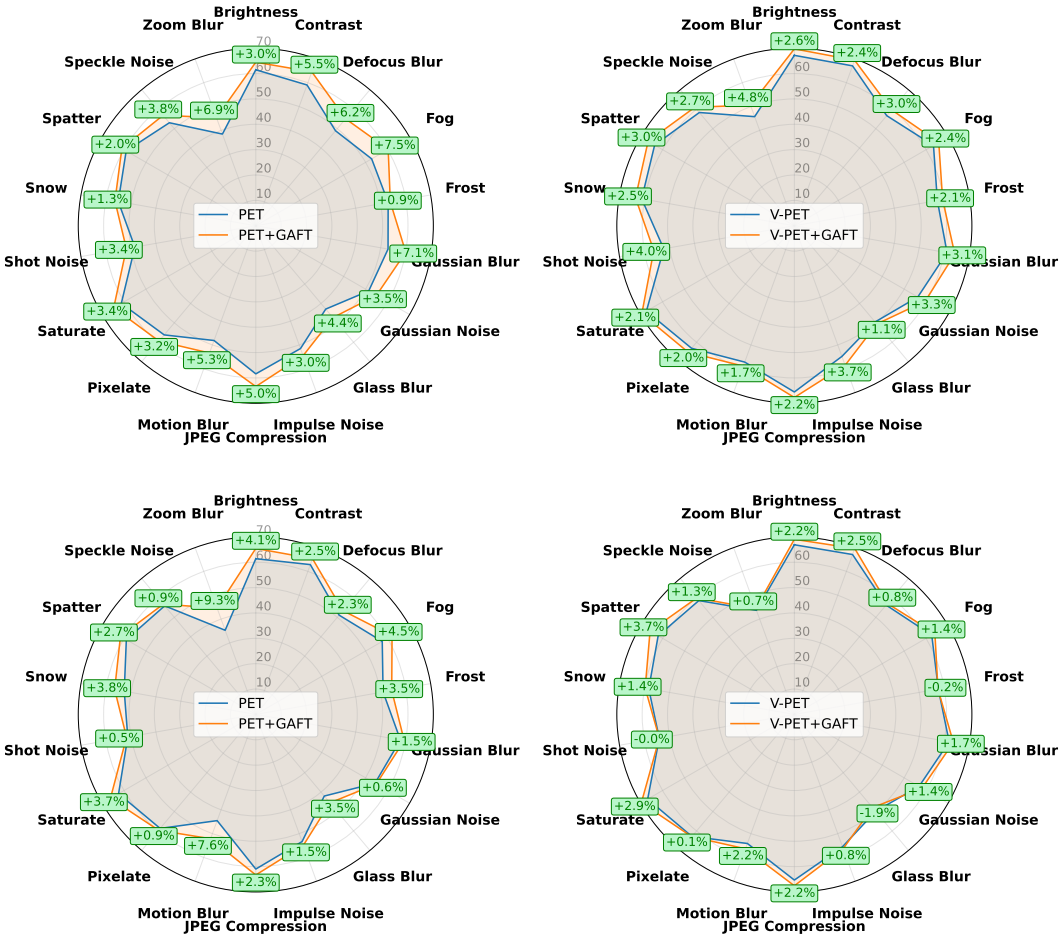


Figure 8: Robustness of DINOv2 to common image corruptions on **DTD**. Top row: AdaptFormer fine-tuning (average gain: *(PET)* 4.18%, *(V-PET)* 2.71%). Bottom row: LoRA fine-tuning (average gain: *(PET)* 3.09%, *(V-PET)* 1.29%). Annotations report the absolute change relative to the corresponding baseline, with negative values denoting a decrease.

to 1.4 points, but consistent in direction. Hence, the within-region prediction smoothness encouraged by $\mathcal{R}_{\text{robust}}$ appears to provide a small amount of resistance to black-box perturbations as a side effect, without explicit adversarial training.