Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Fusion of a Pre-trained Domain-Shift Branch and a Linear Probe for Cross-Domain Few-Shot Classification with Frozen Foundation Models under Unknown Target Domains

Zhenchao Wang
China Mobile (China)
Zhiguo Zhang
zhangzhiguo@cmiot.chinamobile.com
China Mobile (China)

Research Article

Keywords: cross-domain few-shot learning, frozen foundation model, domain-shift branch, linear probe, dual-branch fusion, unknown target domain, DINOv2

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11795552/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Fusion of a Pre-trained Domain-Shift Branch and a Linear Probe for Cross-Domain Few-Shot Classification with Frozen Foundation Models under Unknown Target Domains

Zhenchao Wang[1], Zhiguo Zhang[1*]

Abstract

In cross-domain few-shot classification (CD-FSL) with frozen foundation models, the target domain identity is unknown at deployment, and single-branch adaptation can incur marked degradation on some target domains. Existing fusion methods either adopt a globally fixed weight, which causes large cross-domain fluctuations, or introduce learned gates, which suffer from limited interpretability and uncontrolled selection risk in the few-shot regime. To address this, we propose ADF (Adaptive Dual-branch Fusion), which fuses a pre-trained domain-shift branch and a linear probe at the score level using a pre-defined shrunken grid {0, 0.3, 0.5} for the fusion weight alpha; when alpha = 0, the method falls back to the pure linear probe. ADF trains no gating parameters, and its selection rule is deterministic, interpretable, and reproducible. Under strictly paired evaluation on DINOv2 ViT-B/14 across five target domains, 1-shot/5-shot settings, and eight baselines, ADF attains a mean rank of 1.40 at 5-shot and ranks in the top two on all five domains; paired tests against each domain's strongest single branch detect no statistically significant disadvantage in 10/10 settings, with significant advantage in two. Ablations show that enlarging the weight grid brings no oracle accuracy gain at 5-shot but amplifies the degradation risk of large fusion weights; label-free gating yields no resolvable gain over the constrained scalar selection; and the fusion gain comes mainly from the linear probe, with the domain-shift branch providing a domain-prior fallback. Hence the main value of ADF lies not in per-domain peak accuracy but in reducing the tail risk of single-branch adaptation under unknown target domains via a small candidate set and an explicit fallback path. We conclude that ADF provides a safe, robust, and interpretable fusion solution for CD-FSL with unknown target domains.

Keywords: **cross-domain few-shot learning; frozen foundation model; domain-shift branch; linear probe; dual-branch fusion; unknown target domain; DINOv2**

1 Introduction

Cross-domain few-shot classification (CD-FSL) requires a model to classify with only a handful of labeled samples in each target domain, while a substantial distribution shift exists between the target and source domains. Large-scale self-supervised foundation models such as DINOv2 have

✉ Zhenchao Wang
284100142@qq.com
Zhiguo Zhang
zhangzhiguo@cmiot.chinamobile.com

[1] Intelligent IoT Products Department, China Mobile Internet of Things Co., Ltd., Sichuan,China

established the current best practice for this task via a frozen backbone plus lightweight adaptation: a fixed strong representation combined with light adaptation often outperforms complex meta-learning architectures.

However, an often-overlooked fact is that no single adaptation paradigm is simultaneously optimal for all target domains. Prior studies show that, across target domains, the optimal strategy may switch among linear probing only, partial fine-tuning, and no adaptation; on domains with large distribution shifts, the advantage of complex meta-learning vanishes and can even fall below a linear classifier on randomly initialized features. Consequently, when the target domain identity cannot be anticipated at deployment, committing to one fixed paradigm is itself risky: individual domains may exhibit marked degradation that average accuracy fails to capture.

In this paper we explicitly formulate this risk as a constrained weight-selection problem: given two complementary adaptation branches (one carrying a source-domain shift prior and the other relying solely on the support set), how can we select the fusion weight without using target-domain test information so that the observed selection gap on any domain remains small? To this end, we propose ADF (Adaptive Dual-branch Fusion), which fuses a pre-trained domain-shift branch and a per-episode linear probe at the score level under a pre-defined shrunken grid {0, 0.3, 0.5}; when alpha = 0, the method falls back to the pure linear probe. Its key design has three parts: (1) the fusion weight is restricted to a finite candidate grid rather than a continuous space; (2) the weight is selected solely on a held-out in-domain validation segment, with the test segment excluded from selection; and (3) $\alpha = 0$ serves as an explicit fallback, automatically degenerating to the pure probe when the domain-shift branch is systematically weaker.

For branch fusion under unknown target domains, our contributions are threefold. First, we formalize the problem as finite-candidate weight selection: a pre-trained domain-shift branch and a per-episode linear probe are fused at the score level, α is selected from a pre-defined grid only on an in-domain validation segment, and alpha = 0 provides a fallback path, avoiding the uncontrollability of continuous search and gating training. Second, we propose ADF, which realizes interpretable and reproducible dual-branch fusion without training any gating parameters. Third, under strictly paired evaluation on five target domains, 10 settings, and eight baselines, ADF ranks in the top two on all five domains at 5-shot with a mean rank of 1.40 and detects no significant disadvantage in 10/10 settings; we further show that its gain is bounded by the complementarity of the two branches.

2 Background and Related Work

2.1 Cross-Domain Few-Shot Learning

The foundational work BSCD-FSL [1] established the evaluation protocol of using ImageNet as the source domain and selecting target domains by decreasing similarity to natural images, and reported that the then state-of-the-art meta-learning methods lagged behind simple fine-tuning by 12.8 percentage points on average. Subsequent evaluations further showed that complex meta-learning tends to lose its advantage under large domain shifts; the consensus has shifted toward emphasizing representation over task structure [2,4]: rather than pursuing more intricate meta-learning mechanisms, light adaptation on strong representations is preferred. This also implies that the optimal adaptation may differ across target domains, and how to adapt without knowing the target domain identity remains an underexplored question in CD-FSL.

2.2 Efficient Adaptation on Frozen Backbones

There are three mainstream approaches to adapting frozen foundation models: (1) training a linear classification head (linear probing); (2) learning parameter-efficient adaptation modules (adapters / LoRA) [6,7]; and (3) building caches or prototype-based methods (e.g., Tip-Adapter) [5]. Self-supervised pre-trained models such as DINOv2 [3] provide strong supervision-free features, making frozen backbone plus lightweight adaptation a strong CD-FSL baseline. Linear probing is simple and robust, especially under large domain shifts; adapters learn task-relevant representations on the support set but introduce extra parameters and

overfitting risk; cache methods are training-free but depend on the quality of feature-label retrieval. These three approaches are typically used independently, and their relative merits vary with the degree of domain shift; robustly selecting among multiple adaptation paradigms remains an open problem.

2.3 Ensembling, Routing, and Model Selection

Diversity measures in ensemble learning address how much a group of classifiers can improve together [10,11]; mixture-of-experts (MoE) and dynamic routing [8] address how to assign experts per input; and validation-based model selection chooses the best model on a held-out set in the few-shot regime. Diversity measures characterize member complementarity but give no combination rule; dynamic routing assigns expert weights per sample, and its gates work well with abundant data but are noise-sensitive when labels are scarce; validation-based selection avoids gate training at the cost of extra held-out data, and the selected objects are usually discrete models. When the selection objects expand from discrete models to continuously varying fusion weights and the validation data available for selection are very limited, how to simultaneously constrain the selection loss during selection has not been systematically addressed in the literature.

In recent CD-FSL research, some works have begun to focus on robust decisions under unknown target domains. Most existing works either improve meta-learning or enhance single-model representations [9]; research on multi-branch candidate sets with constrained weight selection using a small amount of in-domain validation data and explicit characterization of the selection loss remains scarce. The key difference between our method and ordinary dynamic routing is that we do not train a gating network in the few-shot regime; instead we use a discrete grid plus validation-based selection, obtaining observable and reproducible robustness within the evaluated domains.

3 The ADF Method

3.1 Overall Framework of ADF

We consider CD-FSL on a frozen backbone: a target domain D_t contains C classes and evaluation is organized as $C_{\text{-way k-shot}}$（k in {1,5}）, each consisting of a support set $S=\{(x_i, y_i)\}$ and a query set Q with S ∩ Q = ∅. The backbone f_θ (DINOv2 ViT-B/14) is frozen throughout and extracts d-dimensional features cached per episode. ADF fuses two complementary branches: the pre-trained domain-shift branch (DMP) is pre-trained on the source domain D_s and carries the source transfer prior, while the per-episode linear probe (LP) trains a linear head on frozen features using only the current episode's support set. The two branches are linearly fused at the class-score level with weight α:

$$s_\alpha = (1-\alpha)\cdot s_{LP} + \alpha \cdot s_{DMP} \quad (1)$$

The fusion weight α is neither learned nor globally fixed; it is selected by an in-domain validation segment on a shrunken grid (α in W = {0, 0.3, 0.5}), and the selection process never touches the test segment. $\alpha = 0$ corresponds to dropping the domain-shift branch and degenerating to the pure probe, an explicit fallback path. This design makes ADF free of learnable gating parameters: weight selection is a deterministic rule that is interpretable and reproducible. Notation: let D_s and D_t denote the source and target domains; α^* (oracle) denotes the optimal weight with test-segment labels known, and $\hat{\alpha}_d$ denotes the weight selected by target domain d on the validation segment according to Eq. (3).

To formalize the risk of choosing a wrong weight, we define the observed selection gap.Let $\alpha_d^* \in W$ denote the weight achievable on target domain d when test-segment labels are known (oracle), and let $\hat{\alpha}_d \in W$ be the weight actually selected on the validation segment. The selection loss of that domain is then $acc_{test}(\alpha_d^*) - acc_{test}(\hat{\alpha}_d)$, and the cross-domain observed selection gap is

$$L_{obs} = \max_{d\in D}[acc_{test}(\alpha_d^*) - acc_{test}(\hat{\alpha}_d)] \quad (2)$$

The L_{obs} characterizes, for any target domain, the maximum accuracy drop of the validation-selected weight relative to the oracle weight; it is only a post-hoc diagnostic,

not a deployment risk upper bound. Since W contains only three candidate weights, $\hat{\alpha}_d$ differs from α_d^* by at most the grid span; the grid constrains only the number of candidates, not the magnitude of the gap itself.

The overall architecture of ADF is shown in Fig. 1. The frozen DINOv2 ViT-B/14 extracts features, the two branches output class scores in parallel, the validation segment selects the weight for score fusion, and the test set never participates in weight selection.

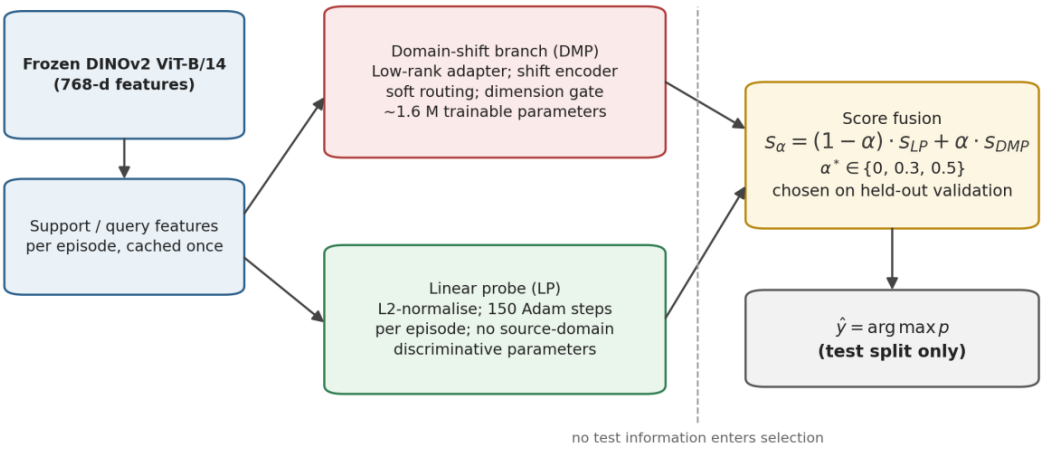


Fig. 1. Architecture of ADF

3.2 Domain-Shift-Aware Branch

When the target domain is globally shifted relative to the source (e.g., satellite versus natural images), a probe trained only on the support set may overfit in-domain noise; the domain-shift branch DMP compensates with a global prior learned from the source domain. DMP consists of a low-rank adapter, a shift encoder, and a shift-classification auxiliary head plus a domain-discriminative auxiliary head that are active only during training. The auxiliary heads serve solely as regularizers in source-domain training and are completely removed at inference, adding no inference cost; the branch outputs a class posterior s_DMP in R^C. Before evaluation, 60 episodes are consumed for test-time augmentation (TTA) statistic warm-up to update the normalization statistics of DMP for the target distribution; the DINOv2 backbone is a Vision Transformer using LayerNorm and remains frozen, so it is not updated.

3.3 Per-Episode Linear Probe

The linear probe introduces no extra parameters, is robust under large domain shifts, and depends only on the current episode's support set, so at the branch level it naturally satisfies the constraint of not using test-segment information (although ADF as a whole still requires a separate labeled in-domain validation segment for weight selection; see Section 3.4). On frozen features, each episode trains a linear head with 150 full-batch Adam steps on the support set and outputs s_{LP}; support and query sets are sampled to be strictly non-overlapping.

3.4 Validation-Based Weight Selection

The selection rule for the fusion weight α is as follows. The validation and test segments come from the same target-domain data source and are split 3:7 from 300 episodes, disjoint at the episode level. The validation segment (90 episodes) searches for the optimal weight over the shrunken grid W = {0, 0.3, 0.5} by validation accuracy (see Section 3.1 for notation and the definition of the selection gap):

$$\hat{\alpha} = argmax_{\alpha \in W} acc_{val}(\alpha) \quad (3)$$

where $acc_{val}(\alpha)$ denotes the accuracy obtained after fusing the two branches with weight α on the validation segment; the search is performed only on the validation segment and never touches the test segment.

The grid includes $\alpha = 0$ as an explicit fallback: when validation search indicates that the domain-shift branch is systematically weaker than the probe, the selected result converges to $\alpha = 0$ and the method degenerates to the pure probe, thereby avoiding the selection gap that the domain-

shift branch might introduce. Validation protocol: the validation segment consists of labeled in-domain samples, strictly non-overlapping with the support set and episode-disjoint from the test segment; its labels are used only for weight selection and represent the target-domain labeling cost required by the method, rather than a zero-label setting relying only on the support set. Each setting uses 90 validation episodes: for 5-way domains each episode has 5 classes x 15 queries, and for ChestXray 2 classes x 15 queries; label slots can reused across episodes. The validation and test segments are sampled consecutively with the same sampler and eval_seed: the first 90 episodes form validation and the last 210 form test, sharing no episode indices. Owing to sampling with replacement across episodes, samples may overlap at the instance level; within an episode, strict_dedup guarantees support-query disjointness (measured overlap rate 0.0000). CropImages has only 46 test images, so instance-level overlap is unavoidable; this domain is used only as an in-domain stress test and is excluded from the main conclusions.

Because the grid contains only three candidate weights and $\alpha = 0$ is the fallback, using a discrete grid restricts weight selection to a finite candidate set and prevents divergence from continuous search; however, the grid constrains only the number of candidate weights, not the magnitude of the accuracy gap between grid points, which is the object to be measured rather than an upper bound. The maximum single-domain selection gap observed across five target domains and 10 domain-setting pairs is 0.67 pp; this value describes only the post-hoc gap between validation-based selection and the test-segment oracle, an empirical diagnostic within the evaluation rather than a deployment risk upper bound.

3.5 Computational Complexity

The extra cost of ADF comes from two parts. The domain-shift branch has about 1.6 million trainable parameters, of which the auxiliary heads (shift classifier and domain discriminator) are active only during training. The linear probe performs 150 full-batch Adam steps per episode. Selecting α requires only a three-point search over the 90 validation episodes (the first 90 of the 300 episodes per domain-setting), which is negligible. Overall inference cost is on the same order as using either branch alone.

4 Experimental Setup

4.1 Datasets and Splits

The source domain is FC100 (60/20/20 train/validation/test classes, mutually disjoint). The five target domains span a spectrum from close to natural images to strongly shifted (Table 1): NEU-DET (industrial surface defects, 5-way 5-shot), EuroSAT (satellite remote sensing, 5-way 5-shot), ISIC2018 (dermoscopy, 5-way 5-shot), CropImages (plant leaves, 5-way 5-shot), and ChestXray (chest X-ray, 2-way 5-shot). ChestXray has only two classes, is reported separately, is excluded from the 5-way average, but is included in the cross-domain competition ranking; CropImages has only 46 test images and serves as a low-sample stress-test domain. Owing to its limited size, this dataset is sampled with replacement across episodes to meet the 300-episode evaluation requirement; within an episode, support and query sets remain strictly non-overlapping. Its results are therefore used mainly for relative in-domain comparison, and its absolute values generalize more weakly than the other four data-sufficient domains. ISIC2018 is split 80/20 into train/test with 7-class stratification (seed = 42).

4.2 Evaluation Protocol

All methods are evaluated on exactly the same episode sequence: the global random state is reset with seed = 42 and 300 strictly paired episodes are sampled consecutively, with the first 90 as validation (the first 60 of which are also used for TTA statistic warm-up) and the last 210 as test. Frozen features are extracted once and cached, and all feature-based methods share the same cache, ensuring strict episode-level pairing. Within each episode the support and query sets are strictly non-overlapping.For data-sufficient domains, each episode is 5-way (2-way for ChestXray) with k support images per class and 15 query images per class; for CropImages, because only 46 test images are available, each episode follows the code's default configuration: 5-

way, 5 support images per class (1 for 1-shot), and 15 queries per class; since one episode requires 5 x 15 = 75 query images, exceeding the available images per class, the query set is refilled with replacement from the remaining samples (support-query strictly disjoint, measured overlap rate 0.0000). The metric is top-1 accuracy, reported as the mean over the last 210 test episodes.

Table 1. Dataset statistics and evaluation settings

Target domain	Classes	Train / Test	Setting	Domain characteristic	Similarity to natural images
NEU-DET	6	1,620 / 180	5-way 5-shot / 1-shot	Industrial surface defects	Medium
ISIC2018	7	10,015 / 2,003	5-way 5-shot / 1-shot	Dermoscopy (7 classes, 80/20 split)	Low
EuroSAT	10	21,600 / 5,400	5-way 5-shot / 1-shot	Satellite remote sensing	Medium
CropImages	7	183 / 46	5-way 5-shot / 1-shot	Plant leaves (low-shot stress test)	High
ChestXray	2	5,232 / 624	2-way 5-shot / 1-shot	Chest X-ray (binary; reported separately)	Lowest

TTA statistic warm-up is applied only to the normalization layers of the DMP branch, using transductive updates [12](only the validation-segment support images are used to estimate normalization statistics, without touching the test segment); the 60 warm-up episodes take the support sets of the first 60 validation episodes and estimate the target-domain statistics (μ, σ) of the DMP normalization layers from these support images only, and warm-up and formal evaluation share the same episode sequence, introducing no extra data. Except for the DMP single-branch control, all baselines (SimpleShot, LP, LP-FT, Tip-Adapter, etc.) do not perform this statistic update and are paired with ADF episode-by-episode under the same protocol; the DMP single branch, as a component control, uses the same warm-up procedure as ADF.

Significance tests use paired bootstrap combined with a sign-flipping permutation test [14,15] at significance level 0.05. Failure to reject the null is not interpreted as equivalence. The permutation test uses 20,000 resamples, giving an analytic resolution of $p = 5 \times 10^{-5}$ [16]; hence a reported $p < 0.0001$ indicates that no more extreme permutation was observed at the sampling resolution.

4.3 Compared Methods and Implementation Details

Eight comparison methods are used: the prototype-based SimpleShot (L2-normalized nearest centroid); the probe-based LinearProbing and LP-FT; the cache-based Tip-Adapter and Tip-Adapter-F; and the adapter-based Episode-Adapter (a single bottleneck adapter trained on the support set), Adapter-Ensemble (uniform ensemble of five checkpoint adapters), and LP-after-Adapter (adapter transformation followed by probing); the domain-shift branch alone is also evaluated as a control. The backbone is DINOv2 ViT-B/14 with 224x224 inputs, position embeddings interpolated from 1370 to 257, frozen throughout; source training uses 8 epochs, batch size 2, initial learning rate 2.5×10^{-4} with cosine annealing, and the best model appears at epoch 3. All results come from the same checkpoint; the domain-shift branch and fusion involve no target-domain retraining.

5 Results and Analysis

5.1 Main Results and Cross-Domain Ranking

Tables 2 and 3 report accuracies at 5-shot and 1-shot. At 5-shot, ADF averages 80.33% over the four 5-way domains (SimpleShot 77.29%, +3.04 pp) and 80.24% over

Table 2. Top-1 accuracy at 5-shot (%, mean over the last 210 test episodes; the first 90 are used only for validation-based alpha selection and TTA warm-up)

Method	NEU-DET	EuroSAT	ISIC	CropImages	ChestXray†	Mean(4×5way)	Mean(5)
SimpleShot (CL2N)	97.49	79.56	40.79	91.33	77.87	77.29	77.41
LinearProbing	98.13	84.33	44.51	93.96	79.90	80.23	80.17
LP-FT	98.11	84.36	44.49	93.98	79.95	80.24	80.18
Tip-Adapter	96.84	75.87	39.23	91.33	71.13	75.82	74.88
Tip-Adapter-F	97.37	80.06	40.92	91.37	75.95	77.43	77.13
Episode-Adapter	97.99	83.92	43.83	93.05	79.06	79.70	79.57
Adapter-Ensemble (5)	97.45	79.52	40.61	91.45	77.75	77.26	77.36
LP-after-Adapter	98.14	84.32	44.54	93.96	79.83	80.24	80.16
ADF (ours)	**98.34**	**84.39**	**44.63**	**93.96**	**79.90**	**80.33**	**80.24**
Domain-shift-branch (alone)	97.51	79.38	40.86	90.17	78.46	76.98	77.28

Table 3. Top-1 accuracy at 1-shot (%, mean over the last 210 test episodes; the first 90 are used only for validation-based alpha selection and TTA warm-up)

Method	NEU-DET	EuroSAT	ISIC	CropImages	ChestXray†	Mean(4×5way)	Mean(5)
SimpleShot (CL2N)	90.17	61.78	29.95	73.59	63.70	63.87	63.84
LinearProbing	91.28	62.88	30.24	74.55	63.68	64.74	64.53
LP-FT	91.28	62.93	30.24	74.49	63.70	64.74	64.53
Tip-Adapter	90.17	61.78	29.95	73.59	63.70	63.87	63.84
Tip-Adapter-F	90.25	61.60	29.83	73.33	63.63	63.75	63.73
Episode-Adapter	90.76	63.25	30.42	74.15	63.46	64.65	64.41
Adapter-Ensemble (5)	90.06	61.65	30.01	73.67	63.84	63.85	63.85
LP-after-Adapter	91.30	62.77	30.16	74.70	63.86	64.73	64.56
ADF (ours)	**92.06**	**63.92**	**30.45**	**74.55**	**63.68**	**65.25**	**64.93**
Domain-shift-branch (alone)	87.21	62.12	29.83	70.03	64.49	62.30	62.74

Note: (dagger) ChestXray uses a 2-way setting, is reported separately, and is excluded from the 5-way average (Mean(4x5way) contains only the four 5-way domains) but is included in the competition ranking of Table 4; Mean(5) is the five-domain average including ChestXray

all five domains including the 2-way domain; relative to SimpleShot, the improvements on the five domains are 0.85, 4.83, 3.84, 2.63, and 2.03 pp. At 1-shot, the four 5-way domains average 65.25% (+1.38 pp over SimpleShot) and the five-domain average is 64.93% (+1.09 pp over SimpleShot's five-domain mean).

The advantage of ADF lies not in absolute per-domain accuracy but in cross-domain consistency. With the paired-test criterion of Table 5, at 5-shot one domain is significantly better, four show no significant difference, and none is significantly worse; the same pattern holds at 1-shot (10/10 with no significant disadvantage). Relative to each domain's strongest single branch, ADF's largest positive gain is +0.76 pp and its largest negative difference is -0.81 pp (not significant), with most differences on the order of 0.1 pp. As a descriptive statistic only, Table 4 shows that ADF has a mean rank of 1.40 at 5-shot with per-domain ranks 1,1,1,2,2 and no domain outside the top two. At 1-shot its mean rank is 2.40 over all five domains including ChestXray, mainly due to ChestXray: this 2-way domain ranks seventh by top-1 accuracy, its strongest single branch is the unadapted domain-shift branch (64.49%), while ADF, selecting $\alpha = 0$ on validation, obtains 63.68% (-0.81 pp mean difference, not significant, p = 0.1785; see Section 5.2).

Table 4. Competition ranking of ADF and major comparison methods (ranks based on per-domain top-1 accuracy)

Method	5-shot ranks (1 = best)	5-shot mean	1-shot ranks (1 = best)	1-shot mean
ADF (ours)	1, 1, 1, 2, 2	1.40	1, 1, 1, 2, 7	2.40
LP-FT	4, 2, 4, 1, 1	2.40	3, 3, 3, 4, 4	3.40
LinearProbing	3, 3, 3, 2, 2	2.60	3, 4, 3, 2, 7	3.80
LP-after-Adapter	2, 4, 2, 2, 4	2.80	2, 5, 5, 1, 2	3.00
Episode-Adapter	5, 5, 5, 5, 5	5.00	5, 2, 2, 5, 10	4.80
Tip-Adapter-F	9, 6, 6, 7, 9	7.40	6, 10, 9, 9, 9	8.60
SimpleShot (CL2N)	7, 7, 8, 8, 7	7.40	7, 7, 7, 7, 4	6.40
Domain-shift-branch (alone)	6, 9, 7, 10, 6	7.60	10, 6, 9, 10, 1	7.20
Adapter-Ensemble (5)	8, 8, 9, 6, 8	7.80	9, 9, 6, 6, 3	6.60
Tip-Adapter	10, 10, 10, 8, 10	9.60	7, 7, 7, 7, 4	6.40

Note: rankings are in descending order of per-domain top-1 accuracy; ties receive the same rank and subsequent ranks are skipped.

5.2 Significance Tests Against the Strongest Single Branch

To determine whether ADF's failure to rank first on every domain reflects a real disadvantage or sampling noise, Table 5 reports per-domain paired tests against the strongest single branch of each domain, judged by Holm-corrected p < 0.05 [13]. At 5-shot, only NEU-DET is significantly better (+0.20 pp, $p_{_raw} = 0.0025$, $p_{_Holm} = 0.0225$), the other four show no significant difference,and none is significantly and none is significantly worse. At 1-shot, the NEU-DET difference grows to +0.76 pp ($p_{_raw} < 0.0001$, $p_{_Holm}$ =

0.0005, significantly better); EuroSAT's nominal +0.67 pp has p_raw = 0.0445 < 0.05 but p_{Holm} = 0.3560 after correction, hence no significant difference. Paired tests were performed for all 10 settings, none shows significant disadvantage (two are significantly better; consistent with our stance, failure to reject the null is not interpreted as equivalence).

Table 5. Per-domain paired tests of ADF against the strongest single branch of each domain (last 210 test episodes)

Domain	k	ADF	Strongest-single branch	Delta (pp)	95%-CI-of Delta	p (raw)	p (Holm)	Verdict
NEU-DET	5	98.34	LP-after-Adapter 98.14	+0.20	[+0.07, +0.33]	0.0025	0.0225	ADF-significantly better
EuroSAT	5	84.39	LP-FT 84.36	+0.03	[−0.31, +0.37]	0.8640	1.0000	No significant difference
ISIC	5	44.63	LP-after-Adapter 44.54	+0.09	[−0.25, +0.43]	0.6075	1.0000	No significant difference
CropImages	5	93.96	LP-FT 93.98	−0.02	[−0.11, +0.07]	0.6780	1.0000	No significant difference
ChestXray	5	79.90	LP-FT 79.95	−0.05	[−0.16, +0.06]	0.3855	1.0000	No significant difference
NEU-DET	1	92.06	LP-after-Adapter 91.30	+0.76	[+0.39, +1.13]	<0.0001	0.0005	ADF-significantly better
EuroSAT	1	63.92	Episode-Adapter 63.25	+0.67	[+0.02, +1.32]	0.0445	0.3560	No significant difference
ISIC	1	30.45	Episode-Adapter 30.42	+0.03	[−0.54, +0.60]	0.9180	1.0000	No significant difference
CropImages	1	74.55	LP-after-Adapter 74.70	−0.15	[−0.42, +0.12]	0.2755	1.0000	No significant difference
ChestXray	1	63.68	DMP branch alone 64.49	−0.81	[−2.02, +0.35]	0.1785	1.0000	No significant difference

Except for the separately discussed 1-shot ChestXray (2-way heterogeneous setting, see below), all settings use the last 210 strictly paired test episodes for reporting and testing.

The 1-shot ChestXray setting is a heterogeneous 2-way setting (excluded from the 5-way average but included in the competition ranking); here the domain-shift branch alone has the highest numerical value. To apply a unified criterion to all 10 domain-setting pairs, we conduct a paired test under the same protocol: data are taken from the last 210 strictly paired test episodes of the same evaluation batch, and per-episode accuracies are given by the ADF and domain-shift-branch curves, with means of 63.68% and 64.49% respectively, consistent with Table 3. The sign-flipping permutation test (20,000 resamples) gives p(raw) = 0.1785, 10,000 paired bootstraps give a 95% CI for Δof [-2.02, +0.35], and the measured mean difference Δ= -0.8095 pp agrees with the -0.81 pp in Table 5; after Holm correction p = 1.0000, judged as no significant difference. At this point all 10 settings are included in one multiple-comparison family and ADF shows no significant disadvantage in 10/10 settings; failure to reject the null is not interpreted as equivalence. The last 210 episodes are used because of the 300 paired episodes per (domain, setting, seed), the first 90 form the validation segment (α selection and TTA warm-up occur before formal evaluation) and only the last 210 form the test segment consistent with the main results of Tables 2 and 3, so the paired test and main results share the same batch and test segment, preventing validation information from entering the significance decision.

Analysis of the experimental results shows that the strongest single branch differs across the five domains: LP-after-Adapter (NEU-DET 5/1-shot, ISIC 5-shot), LP-FT (EuroSAT 5-shot, CropImages 5-shot, ChestXray 5-shot), Episode-Adapter (EuroSAT/ISIC 1-shot), and the domain-shift branch (ChestXray 1-shot). First selecting a method and then applying it to all domains requires knowledge of domain identity; at 5-shot, without knowing the domain identity, ADF achieves one significantly better domain, four with no significant difference, and none significantly worse; at 1-shot the pattern is the same, with the NEU-DET difference widening significantly to +0.76 pp, EuroSAT's nominal +0.67 pp becoming non-significant after Holm correction, and ChestXray's nominal -0.81 pp also not significant, again with none significantly worse. For deployment with unknown target domains, this stability of no significant disadvantage is more meaningful than small gains in mean accuracy; it should be noted that, except for NEU-DET 1-shot (+0.76 pp), ADF's differences from each domain's strongest single branch are within 1 pp, so its value lies in the fallback mechanism and cross-domain stability rather than peak accuracy.

5.3 Ablation Studies

(1) Choice of grid size. Tables 6 and Fig. 2 show the complete five-point grid scan. Expanding the three-point grid to five points and taking the oracle on the test segment yields no improvement in the best accuracy on any of the five domains at 5-shot (the best values are identical); at 1-shot the average improves by 0.19 pp, concentrated on NEU-DET (+0.44) and ChestXray (+0.50), exactly the settings where the domain-shift branch is relatively stronger at 1-shot. Given that expanding the grid markedly amplifies selection risk (at α = 1.0 the 5-shot loss is -0.70 to -4.65 pp), the three-point grid is optimal at zero cost for 5-shot, and for 1-shot it is a reasonable trade-off that loses 0.19 pp in exchange for a marked convergence of selection risk.

(2) Fusion versus gating. Replacing a scalar weight with label-free gating is a natural idea. Table 7 and Fig. 3 report four rules (R1 confidence margin, R2 uncertainty, R3a agreement, R3b mean margin): only 5 of the 10 settings show positive gains (EuroSAT 5-shot +0.02, ISIC 5-shot +0.01, ChestXray 5-shot +0.02, ISIC 1-shot +0.07, ChestXray 1-shot +1.42 pp). Except for R1 at 1-shot ChestXray (65.10%, +1.42 pp over ADF), all positive gains are at most +0.07 pp, on the order of the random fluctuation of 210 test episodes and hence not resolvable gains; this is also the only setting where DMP alone outperforms LP. This supports our design choice: in few-shot cross-domain

scenarios, handing selection to a gate is not safer than a constrained scalar weight.

(3) Component contribution. To decompose the fusion gain, we compare ADF with its two constituent branches separately (Tables 2 and 3): at 5-shot the four-domain averages are 80.33% for ADF versus 80.23% for the pure linear probe (LP) and 76.98% for the pure domain-shift branch (DMP); at 1-shot they are 65.25% versus 64.74% and 62.30%.

Table 6. Accuracy change relative to alpha = 0 (5-shot, pp; complete five-point grid scan)

Domain	α=0	α=0.3	α=0.5	α=0.7	α=1.0	α^* (oracle)
NEU-DET	0.00	+0.16	+0.24	+0.14	−0.70	0.50
EuroSAT	0.00	+0.09	−0.28	−1.14	−4.65	0.30
ISIC	0.00	0.00	+0.16	−0.43	−3.67	0.50
CropImages	0.00	−1.03	−1.70	−2.47	−4.14	0.00
ChestXray	0.00	+0.02	−0.01	−0.12	−1.75	0.30

Note: α* (oracle) is the weight maximizing the change shown in the table, not the oracle accuracy value.

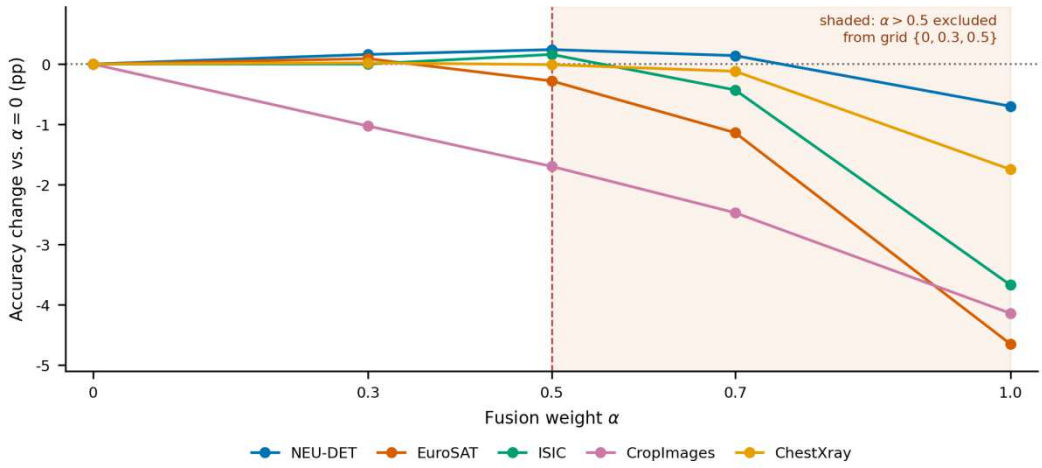


Fig. 2. Weight-grid scan (visualization of Table 6).

Table 7. Gains of four label-free gates over the fixed weight (pp)

Domain	k	R1 margin	R2 uncertainty	R3a agreement	R3b mean margin	Best Delta
NEU-DET	5	−0.03	0.00	0.00	−0.04	0.00
EuroSAT	5	−0.11	−0.01	+0.02	−0.03	+0.02
ISIC	5	−0.22	0.00	−0.03	+0.01	+0.01
CropImages	5	−0.98	0.00	0.00	0.00	0.00
ChestXray	5	+0.02	0.00	0.00	0.00	+0.02

NEU-DET	1	−0.25	−0.01	−0.03	−0.07	−0.01
EuroSAT	1	−0.05	0.00	−0.07	−0.08	−0.05
ISIC	1	−0.02	−0.12	+0.07	+0.05	+0.07
CropImages	1	−1.23	0.00	0.00	0.00	0.00
ChestXray	1	+1.42	0.00	0.00	0.00	+1.42

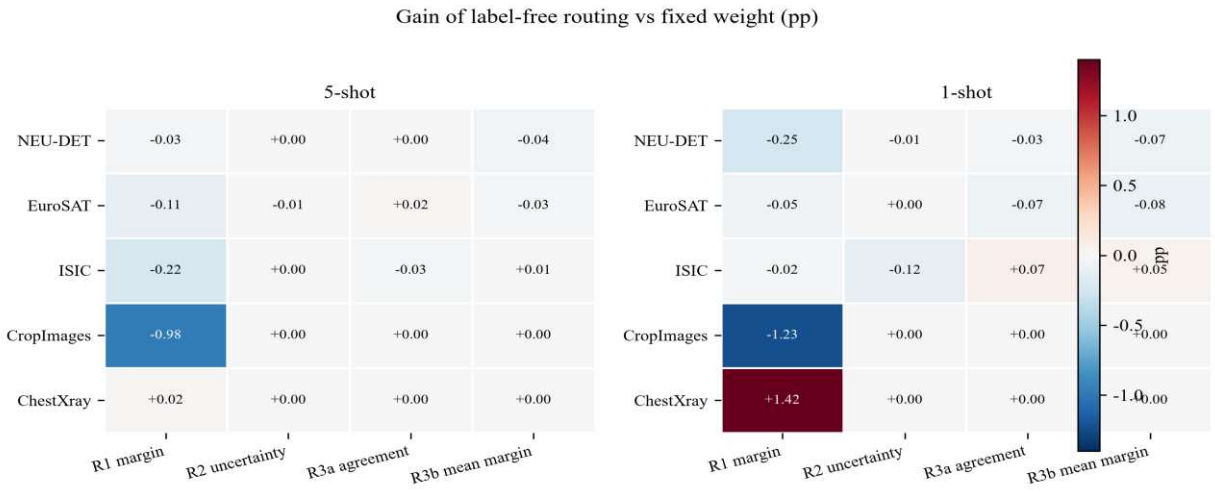


Fig. 3. Heatmap of gains of four label-free gates over the fixed weight

The linear probe contributes most of the fusion gain, while the domain-shift branch is numerically worse at every domain's α = 1.0 endpoint at 5-shot (relative to α = 0, -0.70 to -4.65 pp in Table 6), with the smallest gap on NEU-DET (-0.70 pp), suggesting that the source prior still has value in this industrial domain; 1-shot ChestXray is the only setting where DMP numerically beats LP (64.49% vs 63.68%, p = 0.1785; see Section 5.2). This explains the mechanism behind probes dominating on disagreement samples: the empirical gain of fusion concentrates on samples where LP already dominates, while DMP provides a domain prior and prevents LP from becoming unstable when the support set is tiny (e.g., 1-shot), rather than directly raising mean accuracy. All these values are means over the last 210 test episodes, from a single-backbone, single-seed (42) evaluation protocol.

The three ablations jointly support the design choices of ADF: the three-point grid is cost-free at 5-shot and markedly reduces the observed selection gap; all four label-free gates yield no resolvable gain; and the fusion gain mainly comes from the per-episode linear probe, with the domain-shift branch acting as a domain-prior fallback, while validation-based adaptive selection falls mostly on alpha in {0.3, 0.5}, avoiding the degraded pure-DMP endpoint (α = 1). ADF is therefore positioned as robust fusion that trades minimal encoding for a bounded observed selection gap within the evaluation, rather than as a method for raising peak accuracy.

5.4 Mechanism Analysis: Why Is the Gain Less Than One Percentage Point?

Tables 8, 9 and Fig. 4 explain this at the sample level. Define the disagreement set $D_{div} = \{i:\hat{y}_{LP}(i) \neq \hat{y}_{DMP}(i)\}$(agreement rate 0.667-0.978), $\pi = P(\hat{y}_{LP} = y \mid i \text{ in } D_{div})$, the accuracy of the linear probe on the disagreement set; the complementarity potential $V = (1 - DF) - \max\{acc_{LP}, acc_{DMP}\}$, i.e., the total extra gain a router could add over the best single branch; and the realization ratio $\kappa = \Delta/V$, the fraction of that gain actually obtained. The disagreement-set derivation never uses test labels; to make the two branches comparable, margins are class-wise softmax-normalized on each episode's support set before subtraction.

Tables 8, 9 and Fig. 4 reveal three points. First, potential exists but realization is minimal: V ranges from 0.67 to 8.24 pp with 95% intervals excluding zero, while the average realization ratio by ΣΔ/ΣV is only 1.3% at 5-shot and 1.8% at 1-shot. Second, the linear probe systematically dominates on disagreement samples: 9 of the 10 settings have $\pi > 0.5$, with ratios of 1.1x to 6.5x, and the two branches agree heavily on erroneous samples (Yule's Q = 0.880-0.988). Third, even with a 101-point grid on [0,1] and the oracle taken with test labels, the attainable upper bound $\kappa_{_sup}$ per setting is at most 33.3%, and the weighted average attainable upper bound($\Sigma\kappa_{_sup}.V / \Sigma V$) is only 3.1% (weighted by routable-set sample size over 10 settings;

Table 8. Mechanism analysis: complementarity potential V and realization ratio kappa (disagreement-set criterion)

Domain	k	pi	V (pp)	95% CI for V	Delta (pp)	kappa (%)	kappa_sup (%)
NEU-DET	5	0.658	0.67	[0.53, 0.80]	+0.20	29.9	33.3
EuroSAT	5	0.701	3.68	[3.37, 3.99]	+0.03	0.8	3.1
ISIC	5	0.591	8.24	[7.76, 8.72]	+0.09	1.1	1.7
CropImages	5	0.867	0.69	[0.37, 1.07]	−0.02	−2.9	0.0
ChestXray	5	0.561	5.43	[4.79, 6.08]	−0.05	−0.9	6.4
NEU-DET	1	0.698	3.16	[2.80, 3.55]	+0.76	24.1	24.5
EuroSAT	1	0.523	7.99	[7.40, 8.54]	+0.67	8.4	13.1
ISIC	1	0.512	7.92	[7.44, 8.36]	+0.03	0.4	5.0
CropImages	1	0.794	1.59	[1.25, 1.94]	−0.15	−9.4	0.0
ChestXray	1	0.475	7.57	[6.67, 8.43]	−0.81	−10.7	4.8

Table 9. Discriminative power of the routing statistic d (AUC/KS and attainable realization upper bound)

Domain	k	AUC(d)	KS(d)	kappa_thr
NEU-DET	5	0.837	0.555	32.4%
EuroSAT	5	0.696	0.298	3.1%
ISIC	5	0.567	0.111	0.5%
CropImages	5	0.516	0.235	0.0%
ChestXray	5	0.593	0.164	3.8%
NEU-DET	1	0.774	0.400	17.7%
EuroSAT	1	0.628	0.196	13.1%
ISIC	1	0.543	0.076	5.0%
CropImages	1	0.590	0.181	0.0%
ChestXray	1	0.562	0.120	4.8%

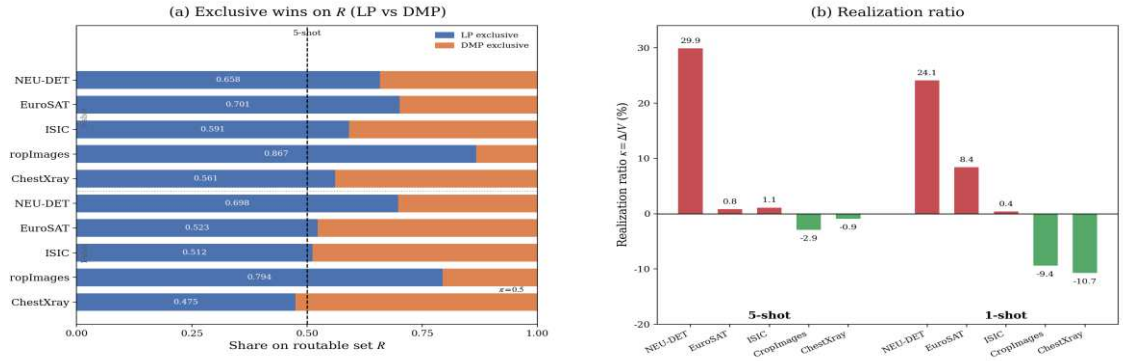


Fig. 4. Mechanism analysis: potential-realization decomposition and discriminative power of the routing statistic (visualization of Tables 8 and 9)

the single-setting peak of 33.3% occurs on NEU-DET). Those indicate that usable fusion information exists only on disagreement samples, where the probe dominates and the two branches err together, so any global weight can realize only a small fraction. ADF attains 29.9% on NEU-DET, close to that setting's attainable upper bound. The negative kappa values in Table 8 for CropImages (5-shot -2.9%, 1-shot -9.4%) and ChestXray 5-shot (-0.9%) correspond to a negative overall gain $\Delta < 0$, meaning validation selection falls back to $\alpha = 0$ or close to the pure probe, so fusion gives no gain over the pure probe, consistent with the weaker domain-shift branch in these settings (see the fallback mechanism in Section 3.4).

The experiments also delineate the applicability of ADF. When the routing statistic d = margin(p_{DMP}) - margin(p_{LP}) is highly discriminative (NEU-DET, AUC = 0.774-0.837), the realization ratio reaches 24%-30% and fusion is worthwhile; when AUC is near random (CropImages, AUC = 0.516), realization tends to zero, the optimal strategy is $\alpha = 0$ (pure probe), and ADF's validation selection automatically lands exactly there. Across the 10 settings, the Spearman rank correlation between AUC and κ_{sup} is +0.69 (n = 10 domain-settings, routable sample set |R| = 86-3177 episodes, permutation test p = 0.033), providing a quantitative basis for this boundary. From a deployment perspective, ADF suits scenarios where hardware can run two lightweight branches simultaneously, the target domain of an input is unknown before deployment, severe performance collapse on any class is unacceptable, and a small labeled in-domain validation segment can be collected for weight selection (Eq. (3) requires validation labels to compute acc_val). When no in-domain labeled data can be obtained at all, our weight-selection rule no longer applies and an unsupervised validation criterion would be needed, which we leave for future work.

6 Discussion

6.1 Design Motivation

Our experiments show that no single paradigm is optimal on all five target domains and that meta-learning methods collectively degrade under large domain shifts. In this situation, further increasing the expressiveness of one branch is offset by the uncertainty of whether that branch suits the given domain. ADF's choice is to model this uncertainty explicitly: a scalar weight estimated only in-domain, plus an observable estimate of the selection gap. It does not pursue higher mean accuracy but suppresses the tail risk of marked degradation on individual domains, and shows no significant disadvantage in per-domain paired tests.

The source of robustness must be distinguished from simple averaging. ADF's robustness is not the statistical dilution effect of averaging two branches: if it were, the fixed $\alpha = 0.5$ should be equally robust, but Table 6 shows that a fixed α suffers marked drops of 1-4 pp on several domains; moreover ADF is significantly better than the strongest single branch on NEU-DET, so it is not dragged

down by a weak branch. Robustness comes from the targeted selection of α on the validation segment. Its boundary should also be clear: at 1-shot ChestXray the domain-shift branch numerically exceeds ADF (64.49% vs 63.68%) and the only resolvable gate gain (R1, +1.42 pp) also occurs there; when the domain-shift branch is systematically stronger and validation discrimination is insufficient, validation selection may choose wrongly and the robustness no longer holds (see Section 6.3).

6.2 Relation to Existing Work

Diversity measures in ensemble learning answer how much better a combination can be [10,11]; our V and κ further answer how much of that diversity a constrained combiner can actually capture. Consistent with BSCD-FSL's simple methods winning out [1], we locate the cause in the probe dominating on disagreement samples and the two branches erring together, rather than mere representation deficiency. Compared with mixture-of-experts routing [8], our experiments show that gating is not better than a scalar weight in few-shot cross-domain settings, in line with the observation that routing degenerates toward uniform allocation when data are limited.

6.3 Applicability Conditions

ADF's benefit relies on a testable premise: the relative strength of the two branches on disagreement samples should not be overly lopsided. If the domain-shift branch is clearly weaker on every domain (e.g., CropImages), validation selects $\alpha = 0$ and the method degenerates to the pure probe, equivalent to Linear Probing; conversely, if the domain-shift branch is systematically stronger (e.g., 1-shot ChestXray), validation may still choose wrongly because of insufficient discrimination. We recommend estimating π and the routing-statistic AUC on a few episodes before use: fusion is most valuable when π clearly deviates from 1/2 and AUC is clearly above 0.5.

In summary, ADF is not a universal accuracy-improvement algorithm but a robust decision tool under risk constraints: its risk-protection capability degrades when the strength gap between the two branches on disagreement samples is large and the validation segment is too small to discriminate. In practice, we recommend estimating the routing-statistic AUC on a few episodes in advance as a precondition for deciding whether to enable this method.

7 Conclusion

We propose ADF, a dual-branch fusion method based on a pre-trained domain-shift branch and a linear probe. ADF constrains the fusion ratio with a validation-selected weight α and fuses the two branches at the score level; α is selected from the pre-defined grid $\{0, 0.3, 0.5\}$, and $\alpha = 0$ falls back to the pure linear probe. Strictly paired evaluation on DINOv2 ViT-B/14 shows that ADF ranks in the top two on all five domains at 5-shot with a mean rank of 1.40, detects no significant disadvantage in 10/10 settings, and is significantly better in two; its gain is bounded by branch complementarity, and its value lies in reducing the tail risk of single-branch adaptation under unknown target domains via a finite candidate set and a fallback path, rather than in per-domain accuracy gains. We conclude that ADF provides a safe, robust, and interpretable fusion solution for CD-FSL with unknown target domains.

Our work has boundaries: ADF requires a small labeled target-domain validation segment; the TTA of the DMP branch uses the unlabeled target-domain test stream, making the setting transductive; and experiments are based on a single backbone, a single source domain, and a single seed in the main tables. Future work includes adapting the grid upper bound to support-set size, constructing routing statistics more closely tied to classification correctness, validating the decision rule across multiple pre-trained backbones and random seeds, and designing unsupervised weight-selection criteria that require no in-domain labels.

Author contributions Zhenchao Wang: Designed the study, conducted the experiments, analyzed the data, and wrote the first draft of the manuscript. Zhiguo Zhang: provided supervisory guidance, contributed to the interpretation of results, and critically revised the manuscript for important intellectual content. All authors approved the final version.

Funding This research was not supported by any funding.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Data availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

[1] Guo Y, Codella N C F, Karlinsky L, et al. A Broader Study of Cross-Domain Few-Shot Learning[C]//ECCV 2020: 124-141.

[2] Paeedeh N, Pratama M, Ma'sum M A, et al.[Cross-domain few-shot learning via adaptive transformer networks][J]. Knowledge-Based Systems, 2024, 288: 111458.

[3] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision[J]. Transactions on Machine Learning Research, 2024. (arXiv:2304.07193).

[4] Hu S X, Li D, Stuhmer J, et al. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference[C]//CVPR 2022: 9058-9067.

[5] Zhang R, Fang R, Zhang W, et al. Tip-Adapter: Training-Free CLIP-Adapter for Better Vision-Language Modeling[C]//ECCV 2022: 493-510.

[6] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models[C]//ICLR 2022 (arXiv:2106.09685).

[7] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-Efficient Transfer Learning for NLP[C]//ICML 2019, PMLR 97: 2790-2799.

[8] Li Y, Ye H, Zhou H, et al.[TACO: Task-aware dynamic routing network for cross-domain few-shot learning][J]. Neurocomputing, 2026, 660: 131752

[9] Baik S, Choi M, Choi J, Kim H, Lee K M. Meta-Learning with Adaptive Hyperparameters[C]//NeurIPS 2020: 20755-20765.

[10] Kuncheva L I, Whitaker C J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy[J]. Machine Learning, 2003, 51(2): 181-207.

[11] Zhou Z-H. Ensemble Methods: Foundations and Algorithms[M]. Chapman & Hall/CRC, 2012.

[12] Efron B, Tibshirani R J. An Introduction to the Bootstrap[M]. New York: Chapman & Hall, 1993.

[13] Holm S. A Simple Sequentially Rejective Multiple Test Procedure[J]. Scandinavian Journal of Statistics, 1979, 6(2): 65-70.

[14] Fisher R A. The Design of Experiments[M]. Edinburgh: Oliver & Boyd, 1935.

[15] Pitman E J G. Significance Tests Which May Be Applied to Samples from Any Populations[J]. Supplement to the Journal of the Royal Statistical Society, 1937, 4(1): 119-130.

[16] Phipson B, Smyth G K. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn[J]. Statistical Applications in Genetics and Molecular Biology, 2010, 9(1): Article 39.