arXiv:2609.10851v1 [cs.CV] 9 Sep 2026

Are We Really Doing Few-Shot Learning? A Critical Examination of Pre-Training Assumptions

Alejandro Galan-Cuenca Marcelo Saval-Calvo
Antonio Javier Gallego

University Institute for Computer Research, University of Alicante, Alicante, Spain
a.galan@ua.es, m.saval@ua.es, jgallego@dlsi.ua.es

Abstract

Few-shot learning is commonly evaluated under protocols that pre-train a model on a large auxiliary set whose classes are disjoint from the target episodes yet drawn from the same visual domain. This paper examines whether such protocols truly reflect low-data learning. We systematically compare no pre-training, class-disjoint in-domain pre-training, supervised out-of-domain pre-training, and label-free out-of-domain pre-training across eight datasets, three few-shot architectures, and multiple way-shot settings. Our results show that class disjointness alone is insufficient to remove the influence of target-domain data. In-domain pre-training improves over no pre-training by 33.41 percentage points on average, whereas supervised out-of-domain pre-training yields 23.75 percentage points, revealing a 9.66-point optimistic bias associated with domain overlap. Although out-of-domain pre-training is more realistic in applications where target-domain data are scarce, its effectiveness depends strongly on the compatibility between source and target domains. We further show that labeled source data are not strictly required, with an augmentation-based label-free strategy reaching an average gain of 27.71 percentage points and closely matching supervised out-of-domain pre-training at 27.97 percentage points. Finally, we introduce a descriptor-based source-selection strategy that estimates source-domain suitability before pre-training, reaching a median gap of only 1.37 percentage points to oracle selection. These findings highlight the need to move beyond in-domain pre-training as the default few-shot evaluation protocol, since it can overestimate performance in realistic scenarios where target-domain data are scarce.

Keywords: Few-shot learning, pre-training, out-of-domain transfer, domain shift, benchmark evaluation, unsupervised learning, source-domain selection.

1 Introduction

Few-Shot Learning (FSL) is a subfield of machine learning focused on developing models that can learn and generalize from extremely limited training data [1–3]. In classification, it is commonly formalized through the n-way k-shot framework, where a model must discriminate among n classes using only k training examples per class. Standard settings include 5-way 1-shot and 20-way 5-shot, requiring classification among 5 or 20 classes using 1 or 5 samples per class, respectively. In this framework, data is organized into a *support set*, containing the only labeled examples available as prior knowledge (i.e., the $n \times k$ training samples), and a *query set*, containing unseen examples used to evaluate generalization. This FSL setting naturally arises when collecting large volumes of labeled data is costly, limited, or impractical, including rare-species classification [4–6], medical imaging and diagnosis with scarce annotated or pathological cases [7–10], fraud detection with few positive cases [11], satellite-based Earth observation [12], wideband signal detection and recognition [13], and facial recognition with few images per individual [14–16], among many other applications.

In the FSL literature, it is common to assume, beyond the support and query sets, the availability of a labeled dataset for model pre-training. Standard benchmarks typically define class-disjoint splits: a test split containing the target classes, from which the support and query sets—commonly referred to as an *episode* or *meta-test set*—are constructed, and a pre-training split, commonly called the *meta-train set*, containing the remaining classes and samples[1] [17,18]. Although this pre-training[2] set contains classes different from those in the episode, its samples come from the same domain as the target classes, resulting in an In-Domain (ID) class-disjoint pre-training scheme. For example, if the target classes are specific bird species, pre-training may use other bird species from the same dataset.

(a) Omniglot examples.

(b) miniImageNet examples.

(c) Omniglot 5/20-way 1/5-shot results.

Figure 1: ID similarity in standard FSL benchmarks and its effect on pre-training. (a) Omniglot: 63 characters spanning 50 alphabets, grouped by similarity; each displayed character corresponds to a distinct class. (b) miniImageNet: examples of visually similar, semantically related classes. (c) Comparison of pre-training (PT) strategies on Omniglot under 5-way 1-shot and 20-way 5-shot configurations: No-PT, ID-PT, and OOD-PT. For OOD-PT, bars report the mean accuracy over seven source domains, and arrows mark the best and worst source (max/min) accuracy.

Although this scenario may appear artificial and idealized, it is the standard evaluation protocol in FSL benchmarks, where datasets such as miniImageNet [19], Omniglot [20], and CIFAR-FS [21] are partitioned following this methodology. Despite class-disjoint splits, the sets retain highly similar visual characteristics. Omniglot, for instance, contains 1,623 handwritten characters from 50 alphabets, yet all samples share similar black-and-white handwritten features (see Fig. 1a). Similarly, miniImageNet includes color photographs from 100 classes with substantial visual overlap in shapes, textures, predominant colors, and backgrounds. Its classes also belong to semantically related groups, including vehicles, animals, plants, and objects, with closely related pairs such as house finch vs. robin, meerkat vs. ferret, and catamaran vs. yawl (see Fig. 1b).

This ID pre-training setup relaxes the core premise of FSL, which is to learn when only a few labeled samples are available in the target domain. Although pre-training and target classes are disjoint, the model still benefits from the high visual similarity across classes, thereby reducing the task complexity. This raises two critical issues:

- It introduces an optimistic bias by exploiting shared distributional properties between pre-training and target classes, thus resembling conventional supervised learning, where training and test data come from the same i.i.d. distribution.

[1] Some benchmarks also define an additional validation split derived from the meta-training set, with disjoint classes, for model selection and hyperparameter tuning.

[2] We generally refer to the meta-training split as the pre-training set to emphasize its role in learning the feature extractor on a source dataset before episodic evaluation on novel classes.

- It is unrealistic to assume that in real-world FSL applications, there will be abundant ID labeled data from other (disjoint) classes, already annotated and sufficient to perform such a pre-training step.

In practice, the pre-training step will involve non-i.i.d. conditions—i.e., Out-Of-Domain (OOD) data—since obtaining large amounts of labeled ID data is difficult or even infeasible in some domains. Consequently, the pre-training distribution may differ substantially from that of the target episodes.

To assess this more realistic scenario, we compare three pre-training strategies: (i) ID pre-training, (ii) OOD pre-training, and (iii) no pre-training, where the model relies solely on the limited labeled samples in the episode. Our goal is to show that current FSL benchmark protocols introduce a systematic bias, yielding overly optimistic and therefore unfair evaluations, and to establish a more robust and realistic framework for FSL training and evaluation.

Fig. 1c illustrates the performance of the three pre-training (PT) strategies on Omniglot under 5-way 1-shot and 20-way 5-shot settings. The results show both the substantial benefit of pre-training (No-PT vs. ID/OOD-PT) and the optimistic advantage of ID-PT over OOD-PT. Note that OOD-PT results are averaged across seven heterogeneous source domains, ranging from similar handwritten character datasets to more distant natural and medical images (see Sect. 4). In the 1-shot setting, ID-PT achieves 62.3% accuracy, compared with a 43.4% OOD-PT average and drops of up to 27 points depending on the source domain. In the 20-way 5-shot setting, ID-PT reaches 61.7%, versus a 53.8% OOD-PT average, with substantial variability across pre-training domains (45.5%–72.6%).

Beyond analyzing the optimistic bias of standard ID pre-training, this work explores several strategies to cope with this issue: (i) a simple ID adaptation (IDA) step—episodic fine-tuning on the scarce $n \times k$ labeled ID support samples—to improve performance after OOD pre-training; (ii) label-free pre-training based on unsupervised and Self-Supervised Learning (SSL) techniques, eliminating the need for labeled pre-training data; and (iii) domain-similarity measures to rank candidate OOD sources for a target and select, before pre-training, the one most likely to improve downstream few-shot performance.

Experiments were conducted using three widely adopted FSL architectures—Matching Networks [19], Prototypical Networks [22], and Relation Networks [23]—which represent the field's main methodological approaches and have served as the foundation for many subsequent works. We evaluate all source–target combinations across eight datasets: three standard benchmarks—miniImageNet, Omniglot, and CIFAR-FS [19–21]—and five additional datasets spanning Greek, Egyptian, and Korean text, musical scores, and medical images.

The remainder of this paper is organized as follows. Section 2 reviews related work on FSL and pre-training; Section 3 presents the methodology and proposed approaches; Section 4 describes the datasets, evaluation protocols, and implementation details; and Section 5 reports and discusses the results. Finally, Section 6 summarizes the main findings and future research directions.

2 Related Work

Within the broad landscape of machine learning, FSL has attracted considerable attention as a data-efficient learning paradigm for generalizing from few labeled examples. Comprehensive reviews of FSL and related meta-learning approaches are provided in [1, 18, 24, 25].

Several families of methods have been proposed for few-shot generalization, including (i) data augmentation [26], (ii) generative approaches that synthesize additional training examples [27, 28], (iii) transfer learning and Domain Adaptation (DA) methods that adapt models to new target domains using labeled data from related sources or tasks [29], and (iv) meta-learning, including optimization-based methods such as Model-Agnostic Meta-Learning (MAML) [30] and

metric-based approaches [19,22,23]. Among them, meta-learning offers a more flexible and powerful framework by learning inductive biases and similarity structures that enable discrimination from few labeled examples, rather than relying mainly on the appearance or characteristics of previously available data, as is typically the case in DA, augmentation, and generative schemes. This section focuses on meta-learning-based FSL methods, which are the primary target of this work.

In visual recognition, meta-learning has become a central framework for FSL [24]. Rather than learning a single classifier over a fixed label set, it exploits experience across many tasks—typically subsets of classes—to build models that can be quickly adapted or are inherently effective on *novel* tasks with few labeled examples. In metric-based methods, query labels are inferred by comparing query embeddings with the support set through attention mechanisms (Matching Networks [19]), class prototypes (Prototypical Networks [22]), or learned similarity functions (Relation Networks [23]). Numerous refinements and alternative baselines have since been proposed, including non-episodic and nearest-neighbor pipelines [31–33], embedding adaptation methods [34], and analyses of training and evaluation choices [35,36].

A key component of most meta-learning methods is *episodic training*. Introduced by Vinyals et al. [19], it mimics few-shot evaluation by sampling episodes from a larger training pool, each comprising an n-way k-shot support set and a query set[3]. Episodic training is now widely used in metric-based, optimization-based [30,37], and memory-augmented models [38,39].

Although episodic training has become closely associated with FSL, recent work has questioned some of its assumptions. Laenen et al. [40] show that episode construction and permutations affect training dynamics and that episodic sampling is not universally beneficial. Dhillon et al. [33] and Alshamrani et al. [41] further show that standard supervised training without episodic sampling can match or outperform episodic meta-learning in some settings, particularly as the number of training classes increases.

Beyond episode sampling, growing attention has focused on the *data regime* used to develop and benchmark meta-learning methods. Standard benchmarks such as miniImageNet, Omniglot, and CIFAR-FS [19–21] assume a large labeled dataset divided into class-disjoint meta-training, validation, and meta-testing splits. Although their classes do not overlap, all splits come from the same dataset and therefore share similar visual characteristics and semantics, yielding an ID class-disjoint pre-training setup. Consequently, models are pre-trained on abundant labeled ID data before few-shot evaluation, reducing task difficulty and introducing optimistic bias in reported performance (Fig. 1c). Thus, although FSL targets low-data learning, much of the literature still assumes access to a relatively large meta-training set at least weakly aligned with the target tasks [19,22,23,31–37,42–48], among many others.

To address this issue, several works move beyond the ID regime by studying distribution shifts between meta-training and meta-testing data. This setting, commonly termed Cross-Domain FSL (CD-FSL), assumes source and target tasks from different domains and, in the canonical formulation, disjoint label spaces. In practice, CD-FSL often trains on one dataset and evaluates on tasks from another, a protocol also known as Cross-Dataset FSL [49–54]. Most works focus on improving transfer through domain-invariant representations [55], autoencoder-based disentanglement of domain-specific and invariant factors [56], cross-domain distance functions [51], multi-source training [57], and pixel-level affinity mechanisms [58].

However, using different datasets does not guarantee strict class disjointness or a genuinely OOD setting. Several popular few-shot benchmarks share parent collections—for example, miniImageNet and tieredImageNet are both derived from ImageNet—yet are sometimes treated as distinct sources in cross-dataset protocols [49,50]. Even with disjoint label sets, datasets may contain semantically overlapping categories; moreover, when both benchmarks derive from the same data source, the resulting shift may remain largely ID in terms of visual statistics

[3]During meta-training, the query set contains additional *labeled* examples from the same n classes, whose labels are used to compute the episode loss; at meta-test time, query labels are used only for evaluation.

and semantics. To prevent label leakage and near-ID shifts, we use completely disjoint datasets with explicitly non-overlapping class sets between pre-training and target tasks. This protocol also allows us to systematically analyze how the nature and degree of source–target domain similarity affect few-shot performance.

While CD-FSL methods typically assume a fixed or small set of source–target benchmarks and propose architectural or training modifications to improve cross-domain adaptation, we take a different perspective: when suitable ID data are unavailable, cross-domain pre-training is often the default. The key questions are then *which* OOD source to use for a target and how to *eliminate source-label requirements* so that any dataset can serve as pre-training data. Rather than proposing a new cross-domain meta-learner, we study the *pre-training regime* itself by comparing ID, truly OOD, and no pre-training, and analyzing the effects of (i) source-dataset choice, (ii) source-label availability, and (iii) IDA.

Recent works have also explored reducing reliance on labeled source data through SSL pre-training before few-shot adaptation. Alfaro-Contreras et al. [59] study this setting for few-shot symbol classification, while Hu et al. [60] show that external data and simple fine-tuning can substantially influence few-shot performance. Our study complements this line by separating *label supervision* from *domain alignment*. We rank candidate OOD sources using domain-similarity criteria before pre-training and compare two label-free alternatives that preserve the downstream FSL architectures, thereby isolating the effects of the data regime.

Another important distinction is whether FSL methods use frozen representations or adapt at test time (e.g., through IDA). Some approaches learn generic features so that simple classifiers, such as nearest-neighbor or linear models, perform well on frozen embeddings [31, 33], whereas gradient-based methods such as MAML [30] incorporate support-set adaptation into training. Other studies show that even simple fine-tuning on k-shot support sets can yield substantial gains [37, 60]. We isolate test-time adaptation from pre-training by evaluating each regime with and without episodic fine-tuning and analyzing how gains vary across OOD sources. This separates improvements due to algorithmic changes from those driven by more favorable data regimes and benchmark choices.

3 Method

This section formalizes the few-shot setting considered in this work and the pre-training regimes under analysis. We first define the target and source domains, the episodic evaluation protocol, and the meta-risk used to compare training regimes. We then discuss how domain shift explains the optimistic bias introduced by standard ID pre-training. Finally, we describe the proposed label-free pre-training strategies and the source-domain selection procedure.

3.1 Problem formulation

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and label spaces, respectively. A domain is defined by a joint distribution $P(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$. We distinguish the target domain $P_{\text{tgt}}(X, Y)$, from which few-shot episodes are sampled, from the source domain $P_{\text{src}}(X, Y)$ used for pre-training.

In few-shot classification, each episode $T = (S_T, Q_T)$ is defined by a support set S_T and a query set Q_T, both constructed from a subset of target classes $\mathcal{C}_T \subset \mathcal{Y}_{\text{tgt}}$, with $|\mathcal{C}_T| = n$ (i.e., n-way). The support set contains k labeled samples per class (i.e., k-shot),

$$S_T = \{(x_i, y_i)\}_{i=1}^{nk}, \quad (x_i, y_i) \sim P_{\text{tgt}}(X, Y \mid Y \in \mathcal{C}_T), \tag{1}$$

and, *in principle*, should constitute the only labeled target-domain information available to the model within the episode. The query set Q_T contains additional disjoint samples from the same classes in C_T and is used to evaluate the episode performance.

We write a few-shot predictor for episode T as

$$h_{\theta,T}(x) = g_T\big(f_\theta(x); S_T\big), \tag{2}$$

where f_θ is the learned representation and θ contains all trainable parameters shared across episodes. The function g_T denotes the episode-conditioned classifier induced by the support set S_T, whose samples are also embedded through f_θ when performing comparisons. This general formulation is intended to cover metric-based methods such as Matching Networks, Prototypical Networks, and Relation Networks, which differ in how g_T compares query and support embeddings. In methods with trainable comparison modules, such as Relation Networks, those shared parameters are treated as part of θ.

3.2 Pre-training regimes

We compare three pre-training regimes, which differ in how the shared parameters θ are initialized before target-domain episodes are evaluated:

- **No-PT:** the parameters θ are not initialized through source-domain pre-training. For each episode, θ is trained only on the scarce labeled examples in S_T.
- **ID-PT:** the model is pre-trained on labeled classes from the same visual domain as the target data, but with disjoint label sets, $\mathcal{Y}_{\text{src}} \cap \mathcal{Y}_{\text{tgt}} = \varnothing$. This corresponds to the standard class-disjoint benchmark protocol used in many FSL datasets.
- **OOD-PT:** the model is pre-trained on a source domain that differs from the target domain both in its label set and in its input distribution, i.e., $\mathcal{Y}_{\text{src}} \cap \mathcal{Y}_{\text{tgt}} = \varnothing$ and $P_{\text{src}}(X) \neq P_{\text{tgt}}(X)$.

Although ID-PT satisfies class disjointness, source and target samples are drawn from the same dataset or visual domain. Therefore, their input distributions are often highly similar, i.e., $P_{\text{src}}(X) \approx P_{\text{tgt}}(X)$, which makes the setting closer to an i.i.d. train/test split than to a strict low-data target-domain scenario. In contrast, OOD-PT introduces a domain shift between pre-training and evaluation, resulting in a more realistic—and more challenging—FSL setting.

For the ID-PT and OOD-PT regimes, we also evaluate an ID adaptation (IDA) step after pre-training. This step fine-tunes the model episodically using only the target support set S_T, which constitutes the only labeled target-domain data available at evaluation time. At each IDA episode, class-balanced support and query subsets are sampled with replacement from the examples available in S_T for each target class.

3.3 Meta-risk and optimistic bias

This section formalizes how different pre-training regimes affect the target meta-risk[4] under the same few-shot evaluation protocol, and how this can induce an optimistic bias when ID pre-training is performed. All quantities used in the formulation below are defined for a fixed n-way k-shot setting, whose dependence is omitted for clarity. For a loss function ℓ, the target meta-risk of parameters θ is defined as

$$\mathcal{R}_{\text{tgt}}(\theta) = \mathbb{E}_{T\sim p_{\text{tgt}}}\mathbb{E}_{(x,y)\sim Q_T}\left[\ell\big(h_{\theta,T}(x), y\big)\right], \tag{3}$$

where p_{tgt} denotes the episodic distribution induced by $P_{\text{tgt}}(X, Y)$ under the fixed n-way k-shot protocol, T denotes a target episode with support set S_T and query set Q_T, $(x, y) \sim Q_T$ denotes uniform averaging over query samples, and $h_{\theta,T}$ is the support-conditioned predictor

[4]The target meta-risk measures the expected query-set loss on a previously unseen target task, represented by an episode, after conditioning the predictor on that episode's support set under the fixed n-way k-shot protocol.

induced by the model parameters θ. Analogously, $\mathcal{R}_{\text{src}}(\theta)$ denotes the expected meta-risk over source-domain episodes.

The effect of the pre-training domain can be interpreted through the lens of domain adaptation theory [61]. Although classical domain-adaptation bounds are not directly formulated for class-disjoint episodic few-shot learning, they provide a useful decomposition of the factors that affect transfer from a source domain to a target domain. In particular, we consider the following adaptation-inspired relation:

$$\begin{aligned}\mathcal{R}_{\text{tgt}}(\theta) \lesssim\ & \mathcal{R}_{\text{src}}(\theta) + \tfrac{1}{2}\text{disc}_{\mathcal{H}}\big(f_\theta \# P_{\text{src}}(X), f_\theta \# P_{\text{tgt}}(X)\big) \\ & + \lambda^{\theta}_{\text{src,tgt}}.\end{aligned} \tag{4}$$

where $f_\theta \# P(X)$ denotes the embedding distribution induced by the encoder f_θ, $\text{disc}_{\mathcal{H}}$ measures the source–target discrepancy in representation space under a hypothesis class $\mathcal{H}$, and $\lambda^{\theta}_{\text{src,tgt}}$ captures the residual joint error of the best support-conditioned predictor for both domains. In this relation, these terms play different roles in our setting:

- The source risk $\mathcal{R}_{\text{src}}(\theta)$ reflects how well the model solves source-domain episodes. By itself, it does not encode whether the source is visually aligned with the target.
- The discrepancy term $\text{disc}_{\mathcal{H}}$ captures the relation between the pre-training and target domains. In ID-PT, this term is expected to be small. In OOD-PT, the source and target input distributions may differ substantially, making this term larger and source-dependent.
- The residual term $\lambda^{\theta}_{\text{src,tgt}}$ captures the remaining source–target incompatibility under the representation and support-conditioned predictor. The IDA step can reduce this mismatch by adapting the model to S_T.

This decomposition highlights the difference between class disjointness and domain disjointness. Standard ID-PT satisfies $\mathcal{Y}_{\text{src}} \cap \mathcal{Y}_{\text{tgt}} = \varnothing$, but it still exposes the representation to samples that are visually close to those later appearing in the support and query sets. Consequently, ID-PT can reduce the target meta-risk by aligning the representation with the target visual distribution before few-shot evaluation. This induces an optimistic evaluation regime.

Let θ_m denote the model parameters obtained after pre-training on source domain m, and let $\theta_\varnothing$ denote the No-PT model. We define the risk-based transfer gain of source m on target domain tgt as

$$\Delta_{\mathcal{R}}(m, \text{tgt}) = \mathcal{R}_{\text{tgt}}(\theta_\varnothing) - \mathcal{R}_{\text{tgt}}(\theta_m). \tag{5}$$

Positive values of $\Delta_{\mathcal{R}}(m, \text{tgt})$ indicate that pre-training on source m reduces the target meta-risk with respect to No-PT, whereas negative values indicate detrimental transfer.

For ID-PT, let m_{ID} denote an ID source. For OOD-PT, let m_{OOD} denote an OOD source. We define the optimistic bias induced by ID-PT as the excess transfer gain obtained from the ID source:

$$B_{\mathcal{R}}(m_{\text{ID}}, m_{\text{OOD}}, \text{tgt}) = \Delta_{\mathcal{R}}(m_{\text{ID}}, \text{tgt}) - \Delta_{\mathcal{R}}(m_{\text{OOD}}, \text{tgt}). \tag{6}$$

A positive $B_{\mathcal{R}}$ indicates that ID-PT provides an additional advantage over OOD-PT. According to Eq. (4), this advantage is consistent with the smaller source–target discrepancy expected in representation space when the representation is pre-trained on a distribution that is already visually close to the target evaluation episodes.

For ID-PT and OOD-PT, we also quantify the gain provided by IDA. Let θ_m^{IDA} denote the episode-adapted parameters obtained by applying IDA to θ_m using S_T before evaluating the corresponding query set. The risk-based contribution of IDA is

$$\Delta^{\text{IDA}}_{\mathcal{R}}(m, \text{tgt}) = \mathcal{R}_{\text{tgt}}(\theta_m) - \mathcal{R}_{\text{tgt}}(\theta_m^{\text{IDA}}). \tag{7}$$

If $\Delta_{\mathcal{R}}^{\mathrm{IDA}}(m_{\mathrm{OOD}}, \mathrm{tgt})$ is larger than $\Delta_{\mathcal{R}}^{\mathrm{IDA}}(m_{\mathrm{ID}}, \mathrm{tgt})$, this indicates that the support set contributes more to target adaptation when the representation has not already been aligned with the target domain during pre-training. Conversely, a small or negative IDA gain under ID-PT suggests that most of the improvement has already been obtained through ID pre-training, leaving little additional benefit for adaptation based only on the available $n \times k$ labeled target samples.

3.4 Label-free OOD pre-training

As discussed before, one way to mitigate the optimistic bias introduced by ID pre-training is to rely on OOD data instead. However, even when a suitable OOD source is available for a given target dataset, standard pre-training assumes that this external source is fully labeled. This requirement can be restrictive in practice and may significantly reduce the set of candidate source datasets. To remove this dependency on source annotations, we propose two label-free pre-training strategies that construct episodic supervision directly from unlabeled samples $x \sim P_{\mathrm{src}}(X)$.

3.4.1 Clustering-based pseudo-labels

Our first label-free OOD pre-training strategy, which we denote as Unsupervised Constrained K-Means (UCKM), constructs pseudo-labels by clustering the unlabeled source dataset and treating each cluster as a class. Concretely, we partition the source dataset into K_c clusters using visual features extracted from its images. This partition defines a pseudo-label space $\tilde{\mathcal{Y}} = \{1, \ldots, K_c\}$, where each sample is assigned the index of its cluster.

Since clustering is performed directly from image features, the induced pseudo-classes do not necessarily correspond to human-interpretable categories. Instead, samples may be grouped according to regularities in the source domain such as texture, color, or other latent visual attributes (see Figure 2 for an example). Nevertheless, these pseudo-classes provide a useful supervisory signal for episodic pre-training by encouraging the encoder to separate distinct visual groups.

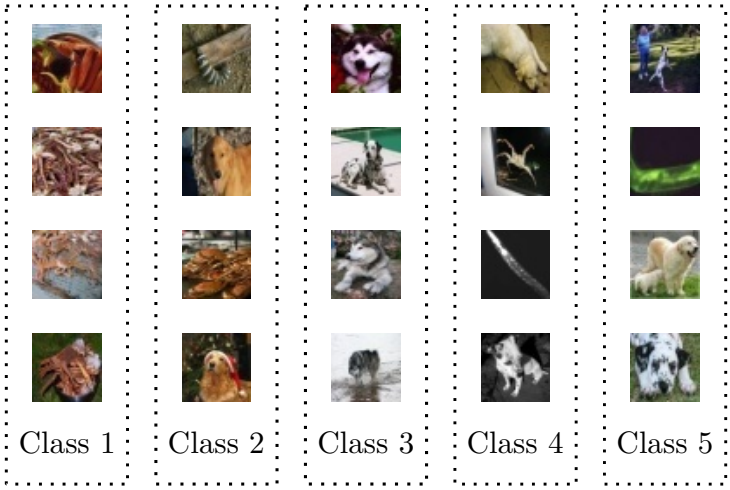


Figure 2: Examples of the pseudo-labeling process using UCKM on the miniImageNet dataset.

For episodic pre-training to be feasible, each cluster must contain enough samples to form both the support and query sets. In an n-way k-shot setting with q query samples per class, this requires at least $k + q$ samples per cluster. To enforce this constraint, we use the constrained K-means algorithm of Bradley et al. [62], which allows setting lower bounds on cluster size and thus prevents the creation of too small clusters.

After clustering, each source sample is assigned its cluster index as a pseudo-label, and episodic pre-training proceeds in the same way as in the supervised setting.

The number of clusters K_c controls the granularity of the partition: smaller values produce coarser pseudo-classes, whereas larger values yield finer groupings that may better capture the diversity of the source data. We therefore treat K_c as a hyperparameter and study its effect empirically.

3.4.2 Instance-based pseudo-labeling via augmentations

Our second label-free strategy, referred to as Unsupervised Instance-based Augmentation (UIAug), is inspired by the instance discrimination paradigm introduced in self-supervised representation learning [63–65]. In these approaches, different augmented views of the same image are treated as positive pairs, while views from different images are treated as negatives, typically under contrastive objectives. Instead of relying on pairwise comparisons between augmented samples, we adapt this principle to the episodic meta-learning setting by turning instance discrimination into a synthetic n-class classification task.

Specifically, each unlabeled image x_i from the source dataset is treated as a distinct pseudo-class with identifier $\tilde{y}_i$. For each episode, we sample n images from the unlabeled source dataset, which define the n pseudo-classes of the episode. For each selected image x_i, we generate k stochastic augmentations $\{\tilde{x}_i^{(1)}, \ldots, \tilde{x}_i^{(k)}\}$ using standard image transformations (e.g., rotations, translations, scaling, blurring, or color perturbations). These augmented samples populate the support set $\tilde{S}_T = \{(\tilde{x}_i^{(j)}, \tilde{y}_i) \mid i = 1, \ldots, n,\ j = 1, \ldots, k\}$, while the original, non-augmented images are used to form the query set $\tilde{Q}_T = \{(x_i, \tilde{y}_i) \mid i = 1, \ldots, n\}$.

This construction preserves the standard n-way k-shot episodic structure while deriving supervision directly from unlabeled data. Figure 3 illustrates the episode construction for a 5-way 5-shot example.

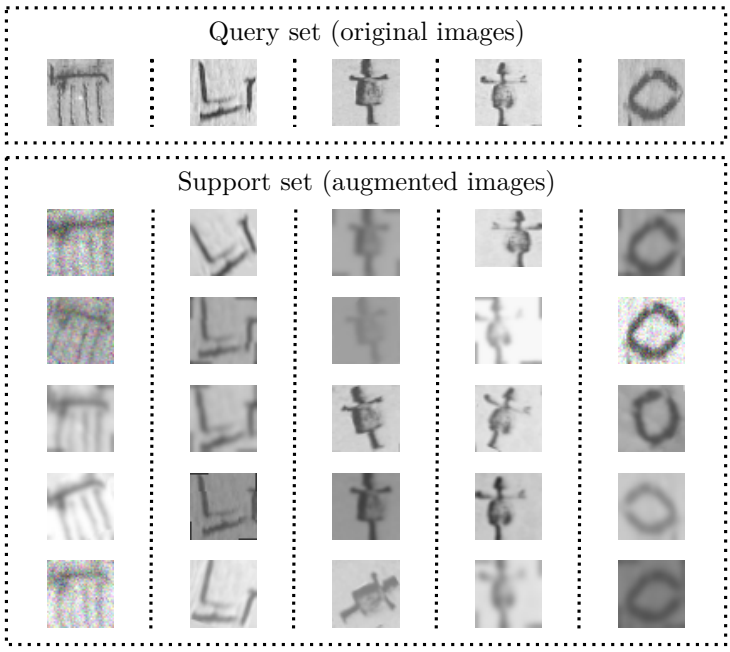


Figure 3: UIAug construction for a 5-way 5-shot episode. Each column corresponds to a pseudo-class defined by a single source instance, whose augmentations form the support set.

Unlike the clustering-based strategy, which groups different images into pseudo-classes according to their visual regularities, UIAug generates multiple transformed views of the same image and treats each original instance as its own pseudo-class. In addition, UIAug does not

require an explicit clustering stage or the specification of the number of clusters K_c, making it a simple and scalable approach.

3.5 Source-domain selection prior to pre-training

When multiple candidate OOD datasets are available, selecting the most suitable pre-training domain can substantially affect downstream performance (as shown in Fig. 1). In practice, however, identifying the best option is non-trivial, as it would require exhaustively pre-training and evaluating a separate model for each candidate source, which is computationally prohibitive. To address this limitation, we propose a *source-selection* procedure that ranks candidates for a given target using *domain-similarity* signals that can be computed *before* pre-training.

Let $\mathcal{S} = \{P_{\text{src}}^{(1)}, \ldots, P_{\text{src}}^{(M)}\}$ be a pool of M candidate source domains and P_{tgt} a fixed target domain. Ideally, source selection would choose the domain m that maximizes the transfer gain $\Delta_{\mathcal{R}}(m, \text{tgt})$, as defined in Eq. (5). This gain, however, is only revealed after learning source-specific parameters θ and evaluating them on the target domain, making it unavailable before pre-training. We therefore propose to approximate this oracle choice using a scoring function computed before pre-training:

$$\hat{m} = \underset{m \in \{1,\ldots,M\}}{\text{argmax}} \ \text{Score}\big(P_{\text{src}}^{(m)}, P_{\text{tgt}}\big). \tag{8}$$

To instantiate $\text{Score}(\cdot, \cdot)$, we consider two strategies: individual similarity descriptors and a composite score based on factor analysis, as detailed below.

3.5.1 Individual-descriptor scoring

We consider a set of J individual descriptors that quantify source–target similarity from different perspectives, including dataset-level metadata (e.g., number of classes, number of samples, or samples per class), low-level image statistics (e.g., brightness, contrast, sharpness, or colorfulness), representation-space overlap, prediction-based distance and compactness patterns, and reconstruction-based similarity. A detailed description of these descriptors is provided in Appendix A.

For each source–target pair $(P_{\text{src}}^{(m)}, P_{\text{tgt}})$, these J descriptors are arranged into a similarity vector $\mathbf{s}_{m,\text{tgt}} = [s_{m,\text{tgt}}^{(1)}, \ldots, s_{m,\text{tgt}}^{(J)}]$. All descriptors are represented numerically and standardized. In the individual-descriptor setting, each descriptor j defines its own scoring function: $\text{Score}_j\big(P_{\text{src}}^{(m)}, P_{\text{tgt}}\big) = s_{m,\text{tgt}}^{(j)}$.

To assess how informative each descriptor is for source selection, we perform an offline calibration stage in which the transfer gain $\Delta_{\mathcal{R}}(m, \text{tgt})$ is estimated for a collection of source–target pairs by running the full pre-training and target-evaluation pipeline. Using these observed gains, we quantify the association between each descriptor and downstream transfer through the Pearson correlation, computed over all considered source–target pairs (m, tgt):

$$r^{(j)} = \text{Corr}\Big(s_{m,\text{tgt}}^{(j)}, \Delta_{\mathcal{R}}(m, \text{tgt})\Big). \tag{9}$$

The magnitude of $r^{(j)}$ reflects how strongly descriptor j is associated with downstream few-shot gains. Once this calibration is fixed, source selection for a new target only requires computing the corresponding descriptor values, without exhaustively pre-training on all candidate sources.

3.5.2 Factor-analysis composite scoring

Individual descriptors may be noisy, redundant, or only partially informative when considered in isolation. We therefore construct a composite score by applying factor analysis to the standardized similarity vectors defined above. Concretely, the vectors $\mathbf{s}_{m,\text{tgt}}$ computed for all calibration

source–target pairs are stacked into a descriptor matrix whose rows correspond to source–target pairs and whose columns correspond to the J standardized descriptors. Before factor extraction, descriptors with low individual Kaiser–Meyer–Olkin (KMO) adequacy are removed [66]. The number of factors is selected using the Kaiser rule, retaining factors with eigenvalues greater than or equal to one.

Factor analysis is then applied to the retained descriptors. We evaluate different rotation configurations, including standard rotations and, during calibration, a target-aligned Procrustes rotation. Let $\mathcal{F}$ denote the set of resulting factor scores, and let z_f be the score associated with factor f. We select the factor most strongly associated with the observed transfer gains $\Delta_{\mathcal{R}}$:

$$f^\star = \arg\max_{f \in \mathcal{F}} \left| \mathrm{Corr}\big(z_f, \Delta_{\mathcal{R}}\big) \right| . \tag{10}$$

The final factor-analysis score $z_{f^\star}$ is obtained from the selected factor $f^\star$ after sign alignment according to its correlation with $\Delta_{\mathcal{R}}$, so that larger values indicate higher expected transfer gain. This score is then used as $\mathrm{Score}(\cdot,\cdot)$ in Eq. (8) to rank candidate sources.

4 Experimental setup

We evaluate the effect of pre-training under controlled few-shot conditions by systematically varying the dataset, source–target domain pair, model architecture, episodic configuration, and adaptation regime. This section details the experimental setup used for the empirical evaluation. The source code and configuration files required to reproduce all experiments are publicly available at https://github.com/Alejandro-Galan/fsl_critique.

4.1 Datasets

We use eight image classification datasets spanning heterogeneous visual domains, formats, and class distributions, summarized in Table 1. They include music notation (Capitan [59]); handwritten characters and symbols (Greek [67], Egyptian [68], TKH [69], and Omniglot [20]); medical imaging (OrganAMNIST from MedMNIST [70]); and natural images (MiniImageNet [19] and CIFAR-FS [21]). This selection provides a broad set of source–target domain relationships, ranging from closely related symbolic domains to markedly different ones, enabling analysis of how pre-training is affected by differences in semantic content, visual appearance, class count, and imbalance.

Table 1: Datasets used in the experimental evaluation. The table reports, for each dataset, its acronym, semantic domain, color mode, original image resolution (px.), number of classes, total number of samples, and samples-per-class statistics (minimum, maximum, and mean number of samples per class).

Dataset	Acr.	Semantic domain	Color mode	Original resolution	Classes	Total samples	Samples per class (min, max, mean)
Capitan [59]	CAP.	Music notation	RGB	64×64	28	17 112	$(57, 2\,737, 596)$
Egyptian [68]	Egypt.	Historical symbols	Gray	64×64	18	4 210	$(49, 447, 163)$
Greek [67]	Greek	Handwritten characters	RGB	64×64	112	168 719	$(50, 12\,725, 1\,494)$
Omniglot [20]	Omni.	Handwritten characters	Gray	105×105	1 623	32 460	$(20, 20, 20)$
TKH [69]	TKH	Handwritten characters	RGB	64×64	347	323 498	$(49, 16\,812, 908)$
OrganAMNIST [70]	oMNIST	Medical imaging	Gray	64×64	11	64 932	$(2\,407, 10\,482, 5\,348)$
CIFAR-FS [21]	CIF.	Natural images	RGB	32×32	100	60 000	$(600, 600, 600)$
MiniImageNet [19]	mINet	Natural images	RGB	84×84	100	60 000	$(600, 600, 600)$

For MiniImageNet, CIFAR-FS, and Omniglot, we use the public class-disjoint splits introduced in the original publications. For the remaining datasets, whose original partitions

were not designed for class-disjoint meta-learning evaluation, we generate a new random class-disjoint split in each bootstrap run, assigning approximately 80% of the classes to meta-training and 20% to meta-testing. Thus, all target-domain results are evaluated on classes unseen during pre-training.

Since our experimental design compares all datasets within a common episodic framework and treats each domain as both a potential pre-training source and a target, all images are rescaled to a common spatial resolution of 40×40 pixels and normalized to $[0, 1]$ by dividing pixel intensities by 255. This resolution limits interpolation artifacts in low-resolution datasets such as CIFAR-FS while preserving the visual structure of higher-resolution images after down-sampling. The standardized input enables the same architecture to be applied across all source–target pairs and prevents transfer performance differences from being attributed to resolution or intensity scaling. Accordingly, our results are intended for the proposed cross-domain comparison and may not be directly comparable to benchmarks using the original resolutions or preprocessing pipelines.

4.2 Implementation details

We evaluate three representative metric-based meta-learning architectures: Matching Networks [19], Prototypical Networks [22], and Relation Networks [23]. These models cover the main families of episodic similarity-based classification methods and provide a common basis for comparing the proposed pre-training regimes.

We follow standard few-shot evaluation settings and consider 5- and 20-way classification with 1, 5, and 10 shots, evaluating all feasible combinations across architectures, datasets, and pre-training regimes.[5] Each experiment comprises 4 000 test episodes, while pre-training, when applicable, also runs for 4 000 episodes. Results are averaged over five bootstrap runs with different random seeds and class-disjoint partitions. All compared methods are evaluated on identical experiment sets, and we report both overall averages and results disaggregated by way, shot, architecture, and target domain.

Following [22], we use an initial learning rate of 10^{-3}, a weight decay of 10^{-6}, and a learning-rate step size of 2 000 episodes. For IDA, adaptation uses only the labeled support set of each target episode and consists of three 1 000-episode stages with a learning rate of 10^{-4}.

For UIAug, we use random resized crops with scales in $[0.85, 1.0]$, rotations within $\pm 5°$, color jittering with probability 0.8, gray-scale conversion with probability 0.2, Gaussian blur, and additive Gaussian noise. Flips are excluded because orientation may be semantically relevant in medical images, alphabets, and music notation.

4.3 Evaluation metrics

For experimental evaluation, we report mean query accuracy, the standard metric in few-shot classification, together with its standard deviation across bootstrap runs. Although the methodological formulation is expressed in terms of meta-risk, accuracy provides an empirical counterpart under the 0–1 classification setting. Since our goal is to quantify the effect of each pre-training regime, we report accuracy gains with respect to the No-PT baseline. Let $\mathrm{Acc}(m, \mathrm{tgt})$ be the mean query accuracy on target domain tgt after pre-training on source domain m, and let $\mathrm{Acc}(\varnothing, \mathrm{tgt})$ denote the corresponding No-PT accuracy. The accuracy gain of pre-training on source m over No-PT is

$$\Delta_{\mathrm{Acc}}(m, \mathrm{tgt}) = \mathrm{Acc}(m, \mathrm{tgt}) - \mathrm{Acc}(\varnothing, \mathrm{tgt}), \tag{11}$$

where positive values indicate beneficial transfer, whereas negative values indicate detrimental transfer.

[5] A dataset is eligible as a pre-training source or target in 20-way experiments only if its meta-training or meta-test split, respectively, contains at least 20 classes.

The effect of IDA is measured as

$$\Delta_{\text{Acc}}^{\text{IDA}}(m, \text{tgt}) = \text{Acc}^{\text{IDA}}(m, \text{tgt}) - \text{Acc}(m, \text{tgt}), \tag{12}$$

where $\text{Acc}^{\text{IDA}}(m, \text{tgt})$ is the mean query accuracy after episodic adaptation on the target support set. This difference isolates the gain obtained from the only labeled target samples available at evaluation time.

5 Results

This section presents the experimental results from two complementary perspectives. First, we evaluate how different pre-training regimes affect few-shot performance, comparing ID-PT, supervised OOD-PT, and the proposed label-free OOD methods. Second, we examine source-domain selection in OOD-PT by analyzing how source–target relationships influence transfer and whether domain-level metrics can identify suitable pre-training sources.

5.1 Effect of Pre-Training on Few-Shot Performance

We begin by assessing the role of pre-training through Δ_{Acc} (Eq. 11), using No-PT as a common reference. This highlights the gap between ID-PT and OOD-PT, the extent to which IDA mitigates source–target mismatch, and the effectiveness of the proposed label-free OOD-PT variants.

5.1.1 ID-PT vs. OOD-PT

Figure 4 reports Δ_{Acc} for ID-PT and supervised OOD-PT before IDA. In ID-PT, source and target domains coincide while preserving class-disjoint pre-training and evaluation splits. In supervised OOD-PT, results are averaged over all source domains different from the target. Error bars reflect variability across experimental configurations, including, in the OOD-PT case, the variability induced by the choice of source domain.

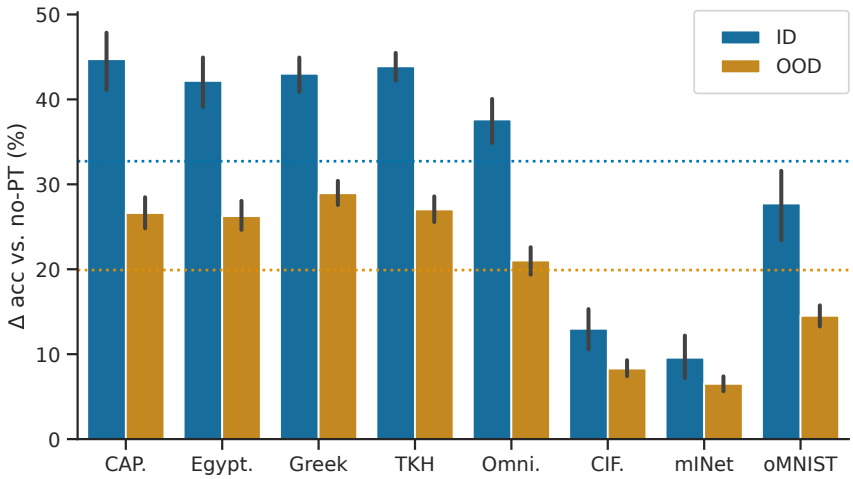


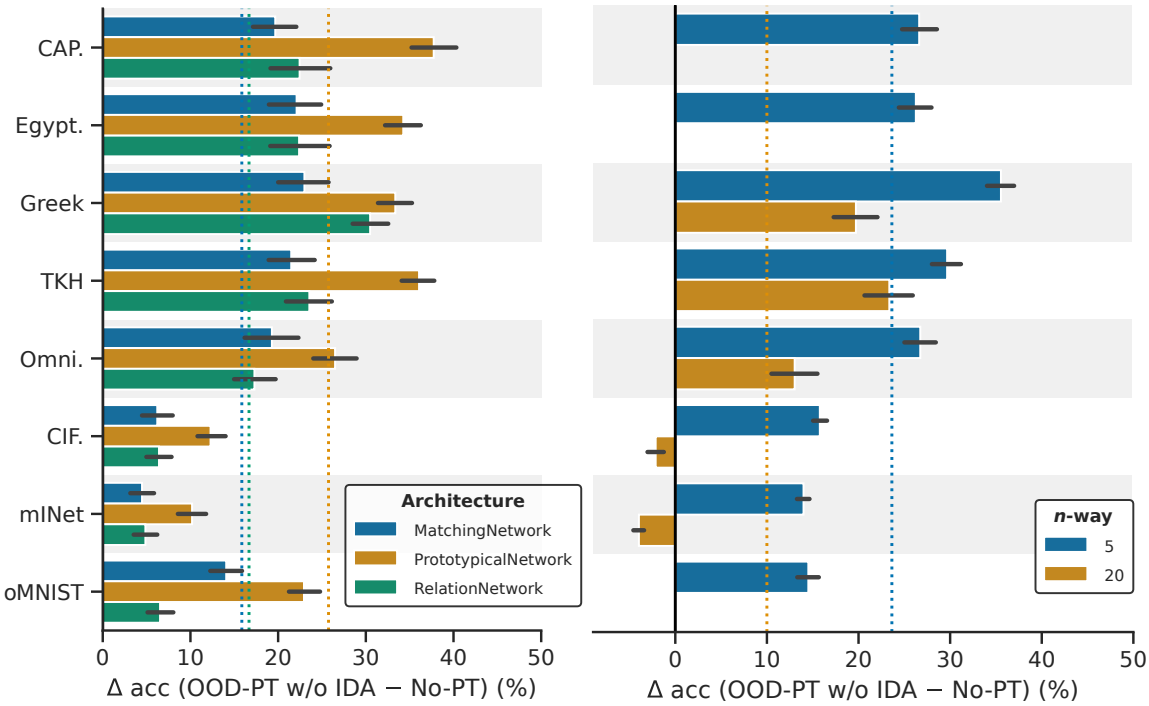
Figure 4: Accuracy gain over No-PT for ID-PT and supervised OOD-PT. Bars report Δ_{Acc} for each target dataset, averaged over architectures, bootstrap runs, n-way configurations, and k-shot settings. Error bars indicate the variability across the averaged experimental configurations, and dotted lines denote the overall mean gain for each pre-training strategy.

Both pre-training regimes outperform No-PT, but by different margins. ID-PT yields the largest gains, averaging 33.41 percentage points (pp.) across target datasets, compared with 23.75 pp. for supervised OOD-PT. This 9.66 pp. gap shows that class-disjoint pre-training

within the target domain provides a substantial advantage beyond the general benefit of pre-training.

The gap is most pronounced in domain-specific datasets. For Capitan, Egyptian, Greek, TKH, and Omniglot, ID-PT improves accuracy by approximately 37–46 pp., consistently exceeding supervised OOD-PT. The advantage is smaller for the natural-image datasets, where ID-PT yields gains of 13.43 pp. on CIFAR-FS and 9.97 pp. on MiniImageNet, with gaps over supervised OOD-PT of 2.98 and 2.02 pp., respectively. These results support our central claim that ID-PT provides an optimistic evaluation setting for FSL, whereas OOD-PT offers a more conservative and realistic performance estimate.

Figure 5 further analyzes supervised OOD-PT by architecture and n-way configuration. Prototypical Networks consistently obtain the largest gains, exceeding 30 pp. on several domain-specific datasets, whereas Matching Networks and Relation Networks show lower and more similar improvements. One possible explanation is that prototype-based classification transfers more robustly across domains because it relies on class-level representations rather than learning a more flexible similarity function, which may require more target-specific evidence and be more sensitive to noise or outliers.



(a) Grouped by architecture.

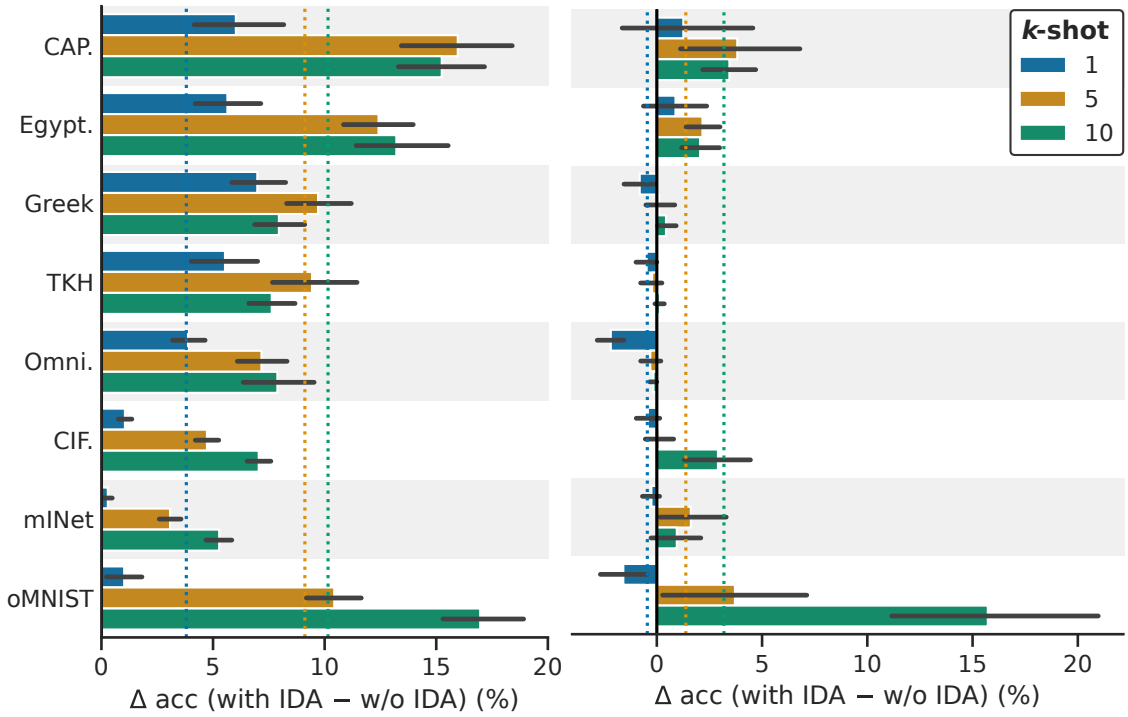
(b) Grouped by N-way.

Figure 5: Mean accuracy gain of supervised OOD-PT over No-PT for each target dataset, averaged across all OOD source domains. Results are grouped by architecture in **(a)** and by n-way configuration in **(b)**. Bars indicate mean Δ_{Acc}, error bars show variability across source domains, and dotted lines mark the overall mean across targets.

The n-way analysis shows that the benefit of supervised OOD-PT decreases as the task becomes more demanding. In the 5-way setting, it improves performance for all target datasets, with an average gain of 23.64 pp. and larger gains on domain-specific datasets. In the 20-way setting, the average gain remains positive at 10 pp., but the effect becomes more target-dependent. OOD-PT still improves performance on symbolic datasets, while yielding negative gains on the natural-image datasets. This suggests that increasing the number of target classes makes OOD transfer less reliable when source and target domains are less aligned.

5.1.2 Effect of IDA

We next examine the effect of IDA after pre-training. Figure 6 reports the adaptation gain, $\Delta_{\text{Acc}}^{\text{IDA}}$ (Eq. 12), for supervised OOD-PT and ID-PT across k-shot settings.



(a) Supervised OOD-PT. (b) ID-PT.

Figure 6: Effect of IDA under **(a)** supervised OOD-PT and **(b)** ID-PT, measured by $\Delta_{\text{Acc}}^{\text{IDA}}$ (Eq. 12) for each target dataset and k-shot setting. Bars show mean gains, error bars indicate variability across experimental configurations, and dotted lines mark the overall mean for each k-shot setting.

The main observation is that the effect of IDA is noticeably stronger under OOD-PT. It improves performance for every target dataset and k-shot setting, with an average gain of 7.69 pp. The improvement increases with the number of support samples, from 3.80 pp. in 1-shot to 9.12 pp. in 5-shot and 10.15 pp. in 10-shot. The largest gains again occur mainly in domain-specific datasets, with OrganAMNIST reaching 16.97 pp. in 10-shot and Capitan exceeding 15 pp. in both 5-shot and 10-shot. These results indicate that IDA effectively uses the available target support set to reduce source–target mismatch.

Under ID-PT, the benefit of IDA is substantially smaller, with an average gain of 1.37 pp. The 1-shot setting yields a negative average effect of -0.45 pp., with several datasets showing negligible or adverse gains. This suggests that, when pre-training is performed on the target domain, adaptation provides limited additional information and may even overfit to the small support set. Although the effect increases with k, reaching 3.19 pp. in 10-shot, it remains substantially below the gains observed under OOD-PT.

5.1.3 Label-Free OOD-PT

This section evaluates OOD pre-training when source-domain labels are not available using the two strategies introduced in Section 3.4: UCKM and UIAug.

UCKM Figure 7 analyzes the effect of the number of clusters in UCKM, with $K_c \in \{5, 20, 250, 1000\}$, using Δ_{acc} in settings with and without IDA. Overall, increasing the number of clusters pro-

vides more informative supervision during pre-training, with $K_c = 1000$ yielding the largest gains across all k-shot settings and both IDA configurations.

The effect of K_c is most pronounced without IDA, where the gap between the worst and best configurations reaches 10.16, 11.73, and 18.34 pp. for 1-, 5-, and 10-shot, respectively. With IDA, these gaps decrease to 4.80, 2.21, and 1.84 pp., indicating that adaptation reduces sensitivity to the number of clusters and partially compensates for less informative pseudo-label structures. Since $K_c = 1000$ still performs best in all settings, we use it in the remaining UCKM experiments.

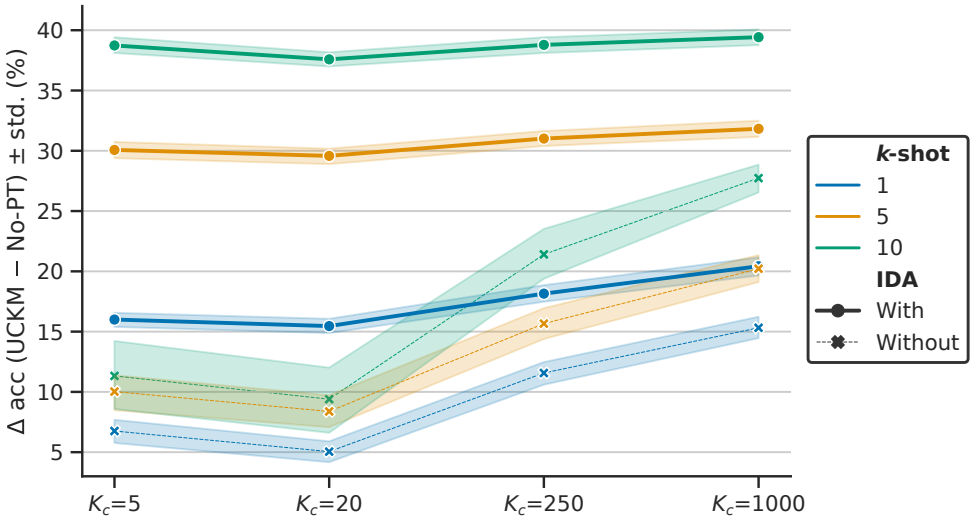


Figure 7: Effect of the number of clusters in UCKM, reported as Δ_{acc} for $K_c \in \{5, 20, 250, 1000\}$ across k-shot settings with and without IDA. Lines show the mean gain over the No-PT baseline across experimental configurations.

UIAug Figure 8 compares UIAug, supervised OOD-PT, and UCKM with $K_c = 1000$ when IDA is applied. Results are reported as Δ_{acc} over No-PT, with dotted vertical lines indicating mean gains and individual points showing results for each OOD source domain.

Overall, UIAug closely matches supervised OOD-PT, with mean gains of 27.71 and 27.97 pp., respectively, across all source–target pairs and experimental configurations. This indicates that instance-based unsupervised pre-training recovers nearly all the benefit of using source labels. By comparison, UCKM achieves a lower mean gain of 24.98 pp., suggesting that clustering-based pseudo-labels provide less consistent supervision in this setting.

The same conclusion holds when selecting the best OOD source for each target. UIAug achieves an average gain of 33.59 pp., slightly above supervised OOD-PT at 33.40 pp. and above UCKM at 31.12 pp. This further supports UIAug as an effective label-free OOD pre-training strategy. However, the gap between the average and best-source performance confirms the strong sensitivity of OOD-PT to source–target compatibility and motivates the need for source-domain selection strategies.

The results also follow the trends observed above. Domain-specific targets show the largest gains, confirming that they benefit most from external pre-training. UIAug closely matches supervised OOD-PT across targets, reproducing both its gain patterns and the ordering of source domains from lowest to highest gain. This correlation is weaker for UCKM, whose source ranking varies more and whose best source–target pairs remain below those of UIAug and supervised OOD-PT. These findings further highlight the role of source–target compatibility and confirm UIAug as a competitive label-free alternative. Results without IDA, reported in App. B, show the same overall trends.

5.2 Source-domain selection for OOD-PT

The previous analysis in Fig. 8 shows that OOD-PT effectiveness strongly depends on the pre-training source domain. Across the three OOD-PT strategies, selecting the best source yields an average improvement of 11.2 pp. over selecting the worst source, with this gap reaching 20.7 pp. in the most sensitive setting. This motivates this section, which examines whether the suitability of a candidate P_{src} for a target P_{tgt} can be estimated before pre-training. To address this question, we evaluate similarity descriptors grouped into four families: dataset-level properties, representation-space overlap, distance-based compactness, and reconstruction-based mismatch. Their definitions are provided in Sec. 3.5 and App. A.

To isolate the effect of P_{src}, results are averaged over the three OOD-PT strategies and computed without IDA, preventing target-domain adaptation from compensating for or masking the effect of source-domain choice.

5.2.1 Individual-descriptor correlations

We first assess each descriptor independently to determine whether it is sufficiently informative for source-domain selection without requiring pre-training. Figure 9 reports the Pearson correlation between each descriptor family and the downstream accuracy gain after OOD-PT.

Overall, individual descriptors show only moderate predictive power. Dataset-level properties perform best, with the average number of samples per class and the conceptual-domain indicator reaching 36.5% and 35.9%, respectively. Distance-based compactness shows similar results, with near-far and minimum-distance correlations of 35.5% and 32.1%. Among representation overlap descriptors, DINO-based GIQA diversity performs best at 31.4%, whereas FSL-encoder descriptors remain below 28%. Autoencoder reconstruction loss reaches only 16.9%.

These results indicate that several individual descriptors appear to reflect aspects of source-domain suitability, but none fully characterizes the source–target compatibility underlying transfer performance. We therefore examine whether factor analysis can combine these complementary descriptors into a more informative score.

5.2.2 Factor-analysis composite score

This section evaluates the factor-analysis score described in Sec. 3.5.2, which combines the descriptors into latent factors and selects the one most associated with transfer gains. Figure 10 reports the correlation between the seven extracted factors and the observed accuracy gains. The first factor is the most predictive, with a Pearson correlation of 74.8%, compared with 43.9% and 38.3% for the next two. The remaining factors show much weaker associations. We therefore use the first factor as the *Best factor* in the subsequent source-selection analysis.

To better understand what information is captured by the *Best factor*, Fig. 11 reports the contribution of each descriptor family. Representation-space overlap descriptors dominate, especially those based on the FSL encoder, which contribute nearly 40% of the total. Dataset-level properties also contribute substantially, showing that source-domain suitability depends on both feature-space similarity and coarse dataset characteristics. Distance-based compactness and reconstruction-based mismatch contribute only marginally.

5.2.3 Final analysis of source selection

Finally, we evaluate source selection using the proposed descriptors. For each target, each descriptor family and the *Best factor* rank the candidate sources and select the top-ranked one. We then compare the downstream accuracy gain Δ_{Acc} obtained with this source against that of the oracle source, defined as the candidate yielding the highest gain after pre-training and target evaluation. Figure 12 summarizes the resulting oracle gaps across all targets using one box plot per descriptor family and for the *Best factor*.

The results show that the *Best factor* provides a more reliable selection criterion than any individual descriptor family, with a median oracle gap of only 1.37 pp. All individual families yield median gaps above 8 pp. A descriptor-level breakdown is provided in App. C; although some individual descriptors achieve competitive results, the analysis confirms the same overall trend. Thus, no single descriptor captures source–target suitability sufficiently well, whereas combining complementary signals is key to reducing the error of automatic source selection.

To contextualize this selection behavior in terms of final few-shot performance, Fig. 13 reports the aggregate accuracy across pre-training regimes. For the three OOD-PT strategies, the bars show the performance obtained using the source selected by the *Best factor*, the dashed lines show the average over all candidate sources, and the upper and lower arrows indicate the performance obtained with the best and worst sources, respectively.

The automatic selection based on the *Best factor* consistently outperforms the average OOD source choice. Without IDA, it improves average accuracy by 6.8–8.7 pp. across OOD-PT strategies and exceeds the worst source choice by 29.5–39.8 pp. With IDA, these gains decrease to 3.0–3.5 pp. over the average and 23.0–33.6 pp. over the worst source, as target-domain adaptation partly compensates for source–target mismatch. Nevertheless, the selected source remains above the average in all settings, showing that descriptor-based selection is beneficial and avoids the substantial degradation caused by poor source choices.

The figure also confirms the main experimental trends. ID-PT achieves the highest overall accuracy, reflecting the optimistic advantage of target-domain pre-training. UCKM performs below supervised OOD-PT, particularly without IDA, but exhibits higher lower bounds and narrower source-dependent ranges. This suggests reduced sensitivity to the choice of source dataset, albeit at the cost of a lower performance ceiling. UIAug provides the strongest label-free results, closely matching supervised OOD-PT both with and without IDA. Overall, the large gap between the upper and lower arrows confirms the importance of source choice for OOD-PT.

6 Conclusions and Future Work

This paper revisits the common assumption in standard FSL evaluation that class-disjoint pre-training is sufficient to define a low-data learning problem. Our results show that this assumption is incomplete. Although ID pre-training avoids label overlap with the target classes, it still exposes the model to the same visual domain as the test episodes, producing a substantial optimistic bias. In our experiments, ID-PT improves over No-PT by 33.41 pp. on average, compared with 23.75 pp. for supervised OOD-PT, revealing an average 9.66 pp. advantage associated with ID-PT. This gap shows that pre-training is not a secondary implementation detail, but a major determinant of few-shot accuracy.

In realistic FSL scenarios, access to a large class-disjoint auxiliary dataset matching the target tasks in domain and distribution cannot be assumed, since data scarcity often characterizes the target domain itself. OOD pre-training is therefore more practical given the availability of numerous large external datasets. Our results show that OOD-PT generally outperforms No-PT, although its effectiveness depends on source–target compatibility. IDA partially mitigates this mismatch, particularly with more support samples, but does not eliminate domain shift, indicating that realistic FSL evaluation should consider both class separation and source–target domain compatibility.

We also show that labeled auxiliary data are not strictly necessary for effective OOD pre-training. UIAug nearly matches supervised OOD-PT, achieving an average gain of 27.71 pp. compared with 27.97 pp., and slightly surpassing it in some cases. This makes label-free pre-training a practical alternative when source annotations are unavailable, while also expanding the pool of datasets that can be used for pre-training.

Finally, our descriptor-based source-selection strategy provides a reliable way to estimate

the suitability of a source domain before pre-training, reducing the risk of negative transfer. The *Best factor* remains above the average source in all settings and achieves a median gap of only 1.37 pp. to the oracle selection, while outperforming the worst source by up to 39.8 pp. without IDA and 33.6 pp. with IDA. Thus, effective source selection can be performed without training on every candidate dataset, while avoiding the substantial degradation caused by poor source choices.

Future work should extend this analysis to larger and more heterogeneous benchmarks, more robust unsupervised pre-training methods, and evaluation protocols that explicitly control domain overlap between pre-training and target episodes.

Acknowledgments

This work was supported by the Generalitat Valenciana (GV), Conselleria d'Educació, Cultura, Universitats i Ocupació, through the SmallOMR project (CIAICO/2023/255). Alejandro Galan-Cuenca acknowledges the support from Programa I+D+i de la GV through grant CIACIF/2023/090.

References

[1] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, Jun. 2020.

[2] F.-F. Li, R. Fergus, P. Perona *et al.*, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 28, no. 4, pp. 594–611, 2006.

[3] M. Fink, "Object classification from a single example utilizing class relevance metrics," *Advances in neural information processing systems*, vol. 17, 2004.

[4] D. Dasgupta, A. Mondal, and P. P. Chakrabarti, "Can synthetic plant images from generative models facilitate rare species identification and classification?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7530–7540.

[5] Q. Zhang, X. Yi, J. Guo, Y. Tang, T. Feng, and R. Liu, "A few-shot rare wildlife image classification method based on style migration data augmentation," *Ecological Informatics*, vol. 77, p. 102237, 2023.

[6] S. Villon, C. Iovan, M. Mangeas, T. Claverie, D. Mouillot, S. Villéger, and L. Vigliola, "Automatic underwater fish species classification with limited data using few-shot learning," *Ecological Informatics*, vol. 63, p. 101320, 2021.

[7] H. Tang, X. Liu, S. Sun, X. Yan, and X. Xie, "Recurrent mask refinement for few-shot medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3918–3928.

[8] R. Feng, X. Zheng, T. Gao, J. Chen, W. Wang, D. Z. Chen, and J. Wu, "Interactive few-shot learning: Limited supervision, better medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2575–2588, 2021.

[9] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Proceedings of the 3rd Machine Learning for Health Symposium*, ser. Proceedings of Machine Learning Research, S. Hegselmann, A. Parziale, D. Shanmugam, S. Tang, M. N. Asiedu, S. Chang, T. Hartvigsen, and H. Singh, Eds., vol. 225. PMLR, 10 Dec 2023, pp. 353–367.

[10] A. Galan-Cuenca, A. J. Gallego, M. Saval-Calvo, and A. Pertusa, "Few-shot learning for covid-19 chest x-ray classification with imbalanced data: an inter vs. intra domain study," *Pattern Analysis and Applications*, vol. 27, no. 3, p. 69, 2024.

[11] S. Lai, J. Wu, C. Ye, Z. Wu, Z. Ma, and H. Zhou, "Iftrans: A few-shot classification method for consumer fraud detection in social sensing," *Knowledge-Based Systems*, p. 113608, 2025.

[12] S. Mohammadi, M. Belgiu, and A. Stein, "Few-shot learning for crop mapping from satellite image time series," *Remote Sensing*, vol. 16, no. 6, 2024.

[13] X. Hao, S. Yang, R. Liu, Z. Feng, T. Peng, and B. Huang, "Vslm: Virtual signal large model for few-shot wideband signal detection and recognition," *IEEE Transactions on Wireless Communications*, 2024.

[14] D. Shome and T. Kar, "Fedaffect: Few-shot federated learning for facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 4168–4175.

[15] X. Chen, X. Zheng, K. Sun, W. Liu, and Y. Zhang, "Self-supervised vision transformer-based few-shot learning for facial expression recognition," *Information Sciences*, vol. 634, pp. 206–226, 2023.

[16] Q. Zhu, Q. Mao, H. Jia, O. E. N. Noi, and J. Tu, "Convolutional relation network for facial expression recognition in the wild with few-shot learning," *Expert Systems with Applications*, vol. 189, p. 116046, 2022.

[17] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," *arXiv preprint arXiv:2203.04291*, 2022.

[18] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, no. 13s, Jul. 2023.

[19] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[20] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[21] L. Bertinetto, J. F. Henriques, P. H. S. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," 2019.

[22] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[23] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[24] A. Vettoruzzo, M.-R. Bouguelia, J. Vanschoren, T. Rögnvaldsson, and K. Santosh, "Advances and challenges in meta-learning: A technical review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 4763–4779, 2024.

[25] H. Gharoun, F. Momenifar, F. Chen, and A. H. Gandomi, "Meta-learning approaches for few-shot learning: A survey of recent advances," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–41, 2024.

[26] Y. Zhang, M. Gong, J. Li, K. Feng, and M. Zhang, "Few-shot learning with enhancements to data augmentation and feature extraction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6655–6668, 2025.

[27] Q. Luo, L. Wang, J. Lv, S. Xiang, and C. Pan, "Few-shot learning via feature hallucination with variational inference," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3962–3971.

[28] K. Li, Y. Zhang, K. Li, and Y. Fu, "Adversarial feature hallucination networks for few-shot learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 467–13 476.

[29] A. Zhao, M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, and J.-R. Wen, "Domain-adaptive few-shot learning," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1389–1398.

[30] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1126–1135.

[31] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 266–282.

[32] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," 2019.

[33] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," *arXiv preprint arXiv:1909.02729*, 2020.

[34] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[35] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," 2020.

[36] X. Luo, H. Wu, J. Zhang, L. Gao, J. Xu, and J. Song, "A closer look at few-shot classification again," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 23 103–23 123.

[37] J. Lin, H. Shao, X. Zhou, B. Cai, and B. Liu, "Generalized maml for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals," *Expert Systems with Applications*, vol. 230, p. 120696, 2023.

[38] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[39] E. Mitchell, C. Lin, A. Bosselut, C. D. Manning, and C. Finn, "Memory-based model editing at scale," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 817–15 831.

[40] S. Laenen and L. Bertinetto, "On episodes, prototypical networks, and few-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 581–24 592, 2021.

[41] S. Alshamrani, H. Alhichri, Y. Bazi, and N. Alajlan, "Prototypical networks for remote sensing scene classification : Episodic versus classical training," in *2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2023, pp. 100–105.

[42] J. Liu, L. Song, and Y. Qin, "Prototype rectification for few-shot learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 741–756.

[43] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, "Laplacian regularized few-shot learning," in *International conference on machine learning*. PMLR, 2020, pp. 11 660–11 670.

[44] M. Boudiaf, I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. Ben Ayed, "Information maximization for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2445–2457, 2020.

[45] Y. Hu, V. Gripon, and S. Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," in *International Conference on Artificial Neural Networks*. Springer, 2021, pp. 487–499.

[46] E. Lee, C.-H. Huang, and C.-Y. Lee, "Few-shot and continual learning with attentive independent mechanisms," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9455–9464.

[47] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[48] J. Luo, H. Shao, J. Lin, and B. Liu, "Meta-learning with elastic prototypical network for fault transfer diagnosis of bearings under unstable speeds," *Reliability Engineering & System Safety*, vol. 245, p. 110001, 2024.

[49] H. Xu, S. Zhi, S. Sun, V. Patel, and L. Liu, "Deep learning for cross-domain few-shot visual recognition: A survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–37, 2025.

[50] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, "A broader study of cross-domain few-shot learning," in *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*. Springer, 2020, pp. 124–141.

[51] P. Li, S. Gong, C. Wang, and Y. Fu, "Ranking distance calibration for cross-domain few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9099–9108.

[52] W.-H. Li, X. Liu, and H. Bilen, "Improving task adaptation for cross-domain few-shot learning," *arXiv preprint arXiv:2107.00358*, vol. 6, 2021.

[53] H. Wang and Z.-H. Deng, "Cross-domain few-shot classification via adversarial task augmentation," *arXiv preprint arXiv:2104.14385*, 2021.

[54] H. Cheng, S. Yang, J. T. Zhou, L. Guo, and B. Wen, "Frequency guidance matters in few-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 814–11 824.

[55] Y. Fu, Y. Fu, and Y.-G. Jiang, "Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 5326–5334.

[56] H. Liang, Q. Zhang, P. Dai, and J. Lu, "Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9424–9434.

[57] W.-H. Li, X. Liu, and H. Bilen, "Universal representations: A unified look at multiple task and domain learning," *International Journal of Computer Vision*, vol. 132, no. 5, pp. 1521–1545, 2024.

[58] H. Chen, Y. Dong, Z. Lu, Y. Yu, and J. Han, "Pixel matching network for cross-domain few-shot segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 978–987.

[59] M. Alfaro-Contreras, A. Ríos-Vila, J. J. Valero-Mas, and J. Calvo-Zaragoza, "Few-shot symbol classification via self-supervised learning and nearest neighbor," *Pattern Recognition Letters*, vol. 167, pp. 1–8, 2023.

[60] S. X. Hu, D. Li, J. Stühmer, M. Kim, and T. M. Hospedales, "Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9058–9067.

[61] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1–2, pp. 151–175, 2010.

[62] P. S. Bradley, K. P. Bennett, and A. Demiriz, "Constrained k-means clustering," Microsoft Research, Tech. Rep. MSR-TR-2000-65, May 2000. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2000-65.pdf

[63] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[64] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html

[65] D. Jiang, W. Li, M. Cao, W. Zou, and X. Li, "Speech SimCLR: Combining Contrastive and Reconstruction Objective for Self-Supervised Speech Representation Learning," in *Proc. Interspeech 2021*, 2021, pp. 1544–1548.

[66] N. Shrestha, "Factor analysis as a tool for survey analysis," *American journal of Applied Mathematics and statistics*, vol. 9, no. 1, pp. 4–11, 2021.

[67] B. Gatos, N. Stamatopoulos, G. Louloudis, G. Sfikas, G. Retsinas, V. Papavassiliou, F. Sunistira, and V. Katsouros, "Grpoly-db: An old greek polytonic document image database," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 646–650.

[68] M. Franken and J. C. van Gemert, "Automatic egyptian hieroglyph recognition by retrieving images as texts," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 765–768.

[69] H. Yang, L. Jin, W. Huang, Z. Yang, S. Lai, and J. Sun, "Dense and tight detection of chinese characters in historical documents: Datasets and a recognition guided detector," *IEEE Access*, vol. 6, pp. 30 174–30 183, 2018.

[70] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[71] I. S. Façoco, J. Rebelo, P. Matias, N. Bento, A. C. Morgado, A. Sampaio, L. Rosado, and M. Barandas, "pymdma: Multimodal data metrics for auditing real and synthetic datasets," *SoftwareX*, vol. 31, p. 102256, 2025.

[72] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[73] J. Yang, R. Xu, Z. Qi, and Y. Shi, "Visual anomaly detection for images: A systematic survey," *Procedia Computer Science*, vol. 199, pp. 471–478, 2022, the 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19.

[74] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019) - Volume 5: VISAPP*, INSTICC. SciTePress, 2019, pp. 372–380.

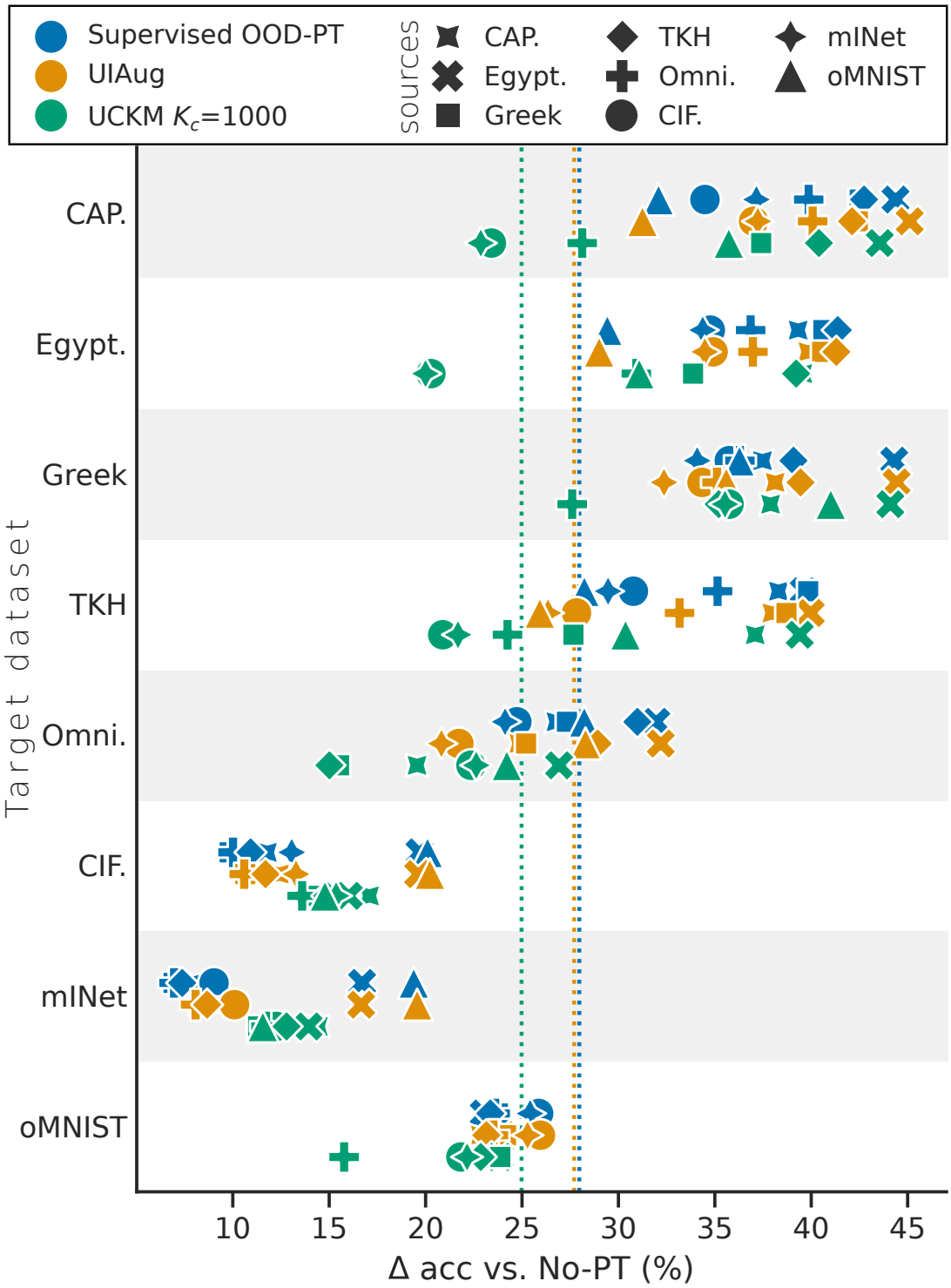


Figure 8: Accuracy gain over No-PT, Δ_{acc}, for each target dataset under supervised OOD-PT, UIAug, and UCKM with $K_c = 1000$ when IDA is applied. Marker shapes denote the OOD source domain, colors indicate the pre-training strategy, and dotted vertical lines show the mean gain for each strategy.

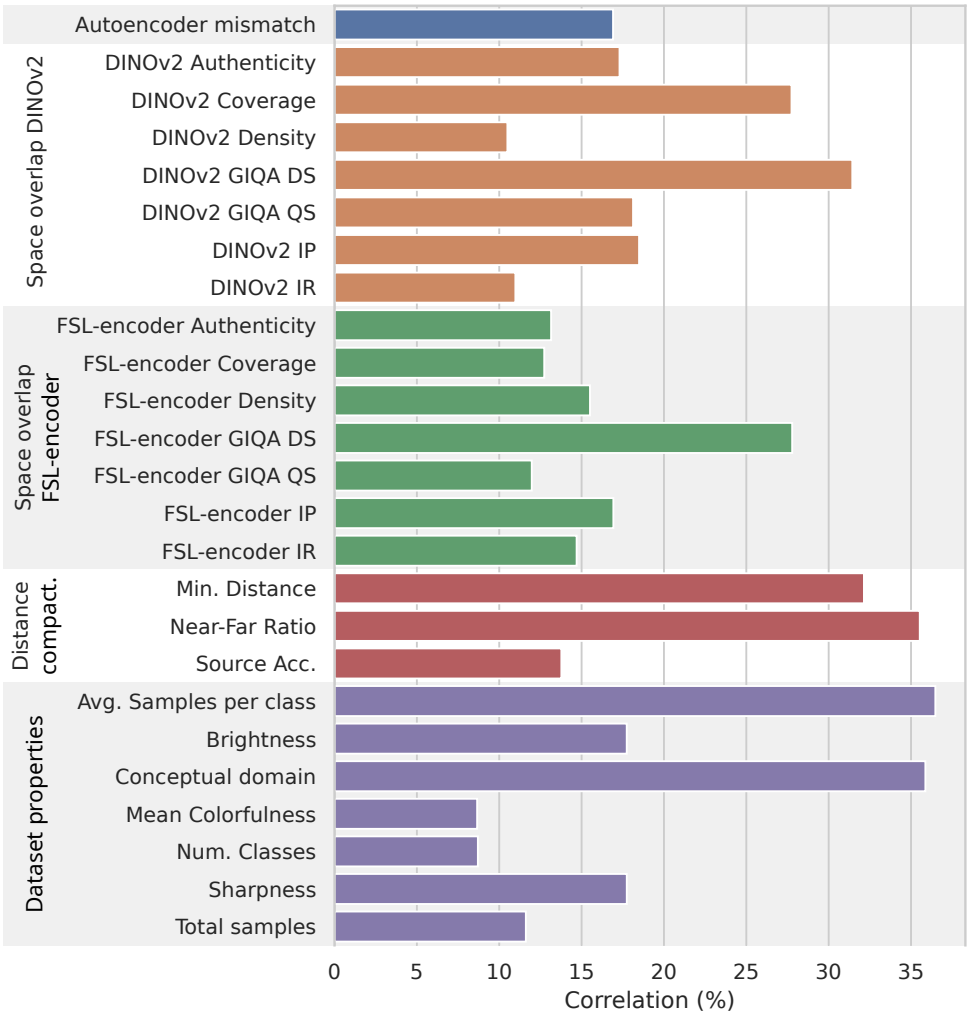


Figure 9: Pearson correlation (%) between each descriptor family and the downstream accuracy gain obtained after OOD-PT. Correlations are computed over source–target pairs using the proposed descriptors and the observed gain with respect to the corresponding No-PT baseline.

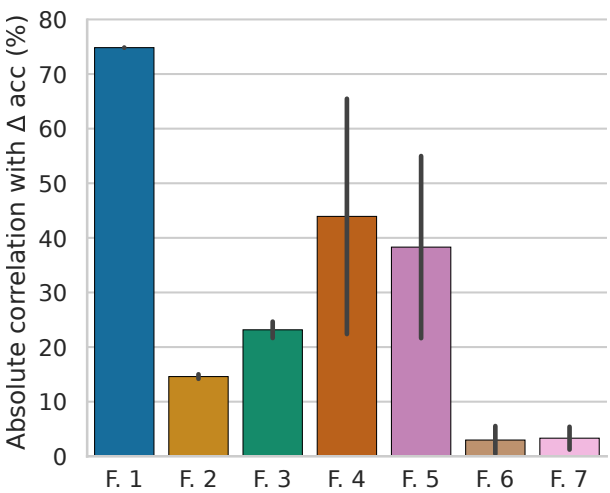


Figure 10: Pearson correlation (%) of each extracted factor with the downstream accuracy gain after OOD-PT.

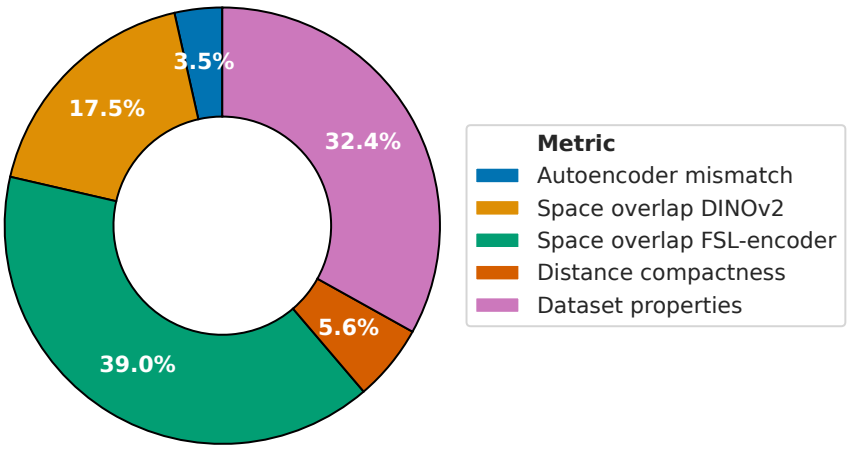


Figure 11: Relative contribution of each descriptor family to the selected *Best factor*. Contributions are obtained by aggregating the factor loadings of the descriptors belonging to each family and normalizing them across groups.

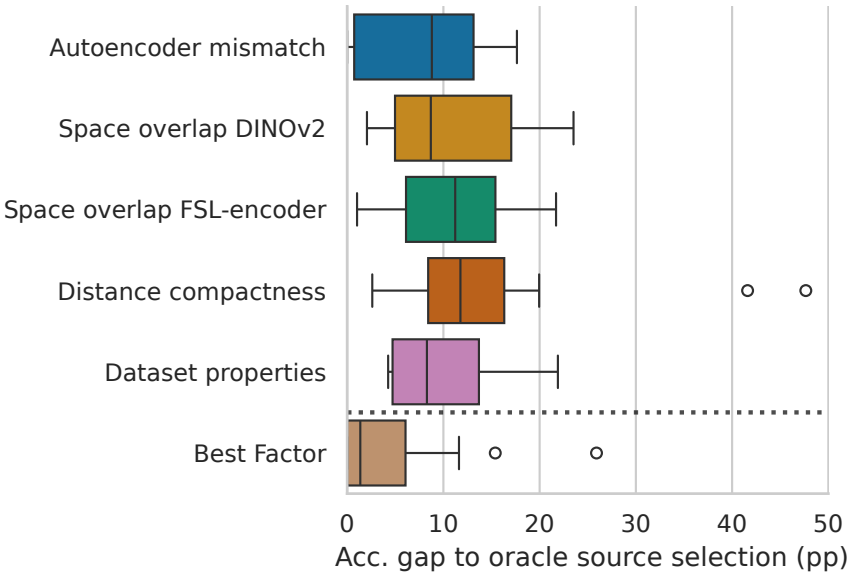


Figure 12: Evaluation of source-domain selection. Candidate sources are ranked using each descriptor family or factor-based score, and the top-ranked source is selected for OOD-PT. Box plots show the downstream accuracy-gain gap in pp. between the selected and oracle sources, defined as the best-performing source after evaluating all candidates. Lower values indicate better selection.

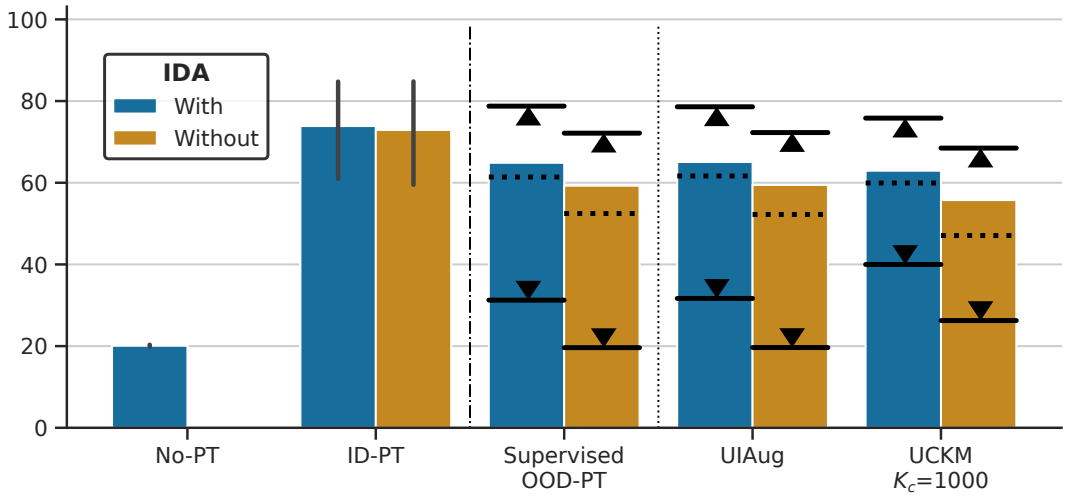


Figure 13: Comparison of pre-training regimes. Bars report mean few-shot accuracy for No-PT, ID-PT, and the three OOD-PT strategies. For OOD-PT, bars show the performance obtained with the source selected by the *Best factor*, dashed lines indicate the average over all candidate sources, while upper and lower arrows mark the best and worst choices, respectively.

A Similarity metrics

This appendix details the domain-similarity descriptors used in Sec. 3.5 to rank candidate OOD source domains before pre-training. Given a target domain P_{tgt} and a candidate source domain $P_{\text{src}}^{(m)}$, each proposed descriptor captures a different source–target property that can be computed without training a source-specific model. For each source–target pair, the descriptors defined below are collected into the vector $\mathbf{s}_{m,\text{tgt}} = [s_{m,\text{tgt}}^{(1)}, \ldots, s_{m,\text{tgt}}^{(J)}]$, where J denotes the total number of proposed descriptors. This vector is used by both the individual-descriptor scores and the factor-analysis score described in Sec. 3.5.2. Unless otherwise stated, all descriptors are standardized; descriptors that naturally measure dissimilarity are sign-aligned so that larger values consistently indicate higher expected source–target compatibility.

We organize the descriptors into four families: dataset-level statistics, representation-space overlap, distance-based compactness, and reconstruction-based mismatch.

A.1 Dataset-level statistics

The first family captures global properties of each dataset. Each domain is characterized by three low-level image statistics—brightness, sharpness, and colorfulness—and four dataset-level metadata descriptors: the number of classes, the total number of samples, the average number of samples per class, and the high-level semantic domain category defined in Table 1: music notation, historical symbols, handwritten characters, medical imaging, or natural images.

Let $\eta^{(j)}(P)$ denote the value of the j-th property of domain P. For numerical properties, the corresponding source–target descriptor is defined as the absolute difference

$$s_{m,\text{tgt}}^{(j)} = \left| \eta^{(j)}(P_{\text{src}}^{(m)}) - \eta^{(j)}(P_{\text{tgt}}) \right| . \tag{13}$$

For the semantic domain category, we instead use a binary mismatch indicator, which is equal to 0 when the source and target belong to the same category and 1 otherwise.

These descriptors are simple to compute and capture global visual and dataset-structure differences. However, they do not measure the overlap or separability of source and target samples in a feature space, which is addressed by the representation-based descriptors below.

A.2 Representation-space overlap

The second family compares source and target domains in a feature space induced by a fixed encoder $\phi(\cdot)$. Given embedded samples from $P_{\text{src}}^{(m)}$ and P_{tgt}, we compute a set of feature-based descriptors originally introduced for auditing real and synthetic datasets [71]. In this work, we use them as source–target compatibility measures by treating the target domain as the reference distribution and the candidate source domain as the evaluated distribution.

Specifically, we consider Authenticity, Coverage, Density, Generated Image Quality Assessment Diversity Score (GIQA-DS), GIQA quality score (GIQA-QS), Improved Precision (IP), and Improved Recall (IR). These descriptors address related notions of fidelity and coverage but differ in their geometric estimators and aggregation rules. IP measures the fraction of source samples lying within the estimated target manifold, whereas Density quantifies how many target neighborhoods contain each source sample. GIQA-QS provides a complementary parametric fidelity estimate by evaluating source samples under a Gaussian mixture model fitted to the target embeddings. Conversely, IR and Coverage quantify how well the source spans the target manifold using neighborhood-based criteria, while GIQA-DS estimates this reciprocal relation by evaluating target samples under a density model fitted to the source embeddings. Finally, Authenticity measures whether source samples are close to individual target samples relative to the local spacing within the target set, thereby capturing local duplication rather than general source–target proximity. Together, these descriptors characterize complementary aspects of representation-space compatibility.

In our experiments, these descriptors are computed in two types of feature spaces: (i) a generic self-supervised representation obtained with DINOv2 [72], following the original publication, and (ii) task-adapted representations extracted from the few-shot architectures considered in this work (see Sec. 4). In the latter case, a single model is trained using only the n-way k-shot support set of the target domain; this model is then kept fixed and used as a common encoder to extract representations for both the target and candidate source domains. The DINOv2 features provide a general-purpose estimate of representation-space overlap, whereas the few-shot encoders provide a task-adapted estimate conditioned on the target support set.

A.3 Distance-based compactness

The third family characterizes how compact and separable each candidate source domain appears under a target-adapted few-shot representation. These descriptors do not capture an intrinsic property of the source, but rather how its samples are organized in the representation induced by the target distribution.

For a given target domain P_{tgt}, we train a few-shot model with encoder $f_{\theta_{\text{tgt}}}$ on its n-way k-shot support set and keep it fixed to extract embeddings from all candidate source domains. Given a source episode $T = (S_T, Q_T)$ sampled from a candidate source domain $P_{\text{src}}^{(m)}$, let $\mathcal{C}_T$ denote its set of classes or pseudo-classes. For supervised OOD-PT, $\mathcal{C}_T$ is defined by the ground-truth source labels, whereas for the label-free variants it is defined by the pseudo-labels induced by the corresponding strategy (see Sec. 3.4). Following the general predictor in Eq. (2), each source query is evaluated through the target-adapted model as

$$h_{\theta_{\text{tgt}},T}(x) = g_T\big(f_{\theta_{\text{tgt}}}(x); S_T\big), \qquad x \in Q_T, \tag{14}$$

where g_T is the architecture-specific episodic classifier induced by the source support set S_T. During this inference process, for each source query image $x \in Q_T$, we collect the architecture-specific distances or dissimilarity scores computed between its embedding and the support-derived embeddings or class representatives used by g_T, denoted by $\mathcal{D}_{\theta_{\text{tgt}}}(x, S_T)$. From these scores, we derive three complementary descriptors.

The first descriptor measures the contrast between near and far candidate classes. Let q_{50} be the empirical median of $\mathcal{D}_{\theta_{\text{tgt}}}(x, S_T)$. We split the scores into a lower half, containing values up to the median, and an upper half, containing values above the median, and define the query-level *near-to-far distance ratio* as

$$\rho_{\text{Q}}(x, S_T) = \frac{\sum_{d_i \in \mathcal{D}_{\theta_{\text{tgt}}}(x,S_T):\, d_i \le q_{50}} d_i}{\sum_{d_i \in \mathcal{D}_{\theta_{\text{tgt}}}(x,S_T):\, d_i > q_{50}} d_i + \epsilon}. \tag{15}$$

where $\epsilon > 0$ is a small constant for numerical stability. Lower values indicate a more contrasted distance profile, in which the query is much closer to a subset of candidate classes than to the remaining ones, whereas higher values correspond to a flatter profile and therefore suggest weaker separability under the target-adapted representation. The final descriptor is obtained by averaging $\rho_{\text{Q}}(x, S_T)$ over query samples and sampled source episodes.

The second descriptor keeps only the *minimum distance* for each query, i.e., the distance to its nearest source representative, and aggregates this value over queries and episodes. This provides a direct nearest-prototype compactness measure, where lower values indicate that source samples tend to lie close to at least one source class representative under the target-adapted embedding.

Finally, we compute the *source accuracy*, defined as the episodic classification accuracy obtained when the target-trained few-shot model classifies source-domain query samples. Accuracy is measured against the available source labels in the supervised case or against pseudo-labels in the label-free case. This descriptor complements the geometric distance summaries by assessing whether target-adapted compactness translates into correct source-domain classification.

A.4 Reconstruction-based mismatch

The fourth family estimates source–target visual mismatch from reconstruction error, treating the comparison as an anomaly-detection problem. Following reconstruction-based anomaly-detection methods [73, 74], we assume that an autoencoder fitted to the target support set should reconstruct samples from visually similar source domains with lower error. Thus, for a target-domain setting, we train a single convolutional autoencoder a_ψ^{tgt} to reconstruct the images in the n-way k-shot target support set, without using their labels. This autoencoder is then kept fixed and used to evaluate all candidate source domains. For each source domain $P_{\text{src}}^{(m)}$, we compute the reconstruction-based mismatch descriptor as

$$D_{\text{AE}}\Big(P_{\text{src}}^{(m)}, P_{\text{tgt}}\Big) = \frac{1}{|\mathcal{X}_{\text{src}}^{(m)}|} \sum_{x_i \in \mathcal{X}_{\text{src}}^{(m)}} \frac{1}{HWC} \left\| a_\psi^{\text{tgt}}(x_i) - x_i \right\|_2^2, \tag{16}$$

where $\mathcal{X}_{\text{src}}^{(m)}$ denotes the images from the m-th candidate source domain, and H, W, and C are the image height, width, and number of channels. Low values indicate that source samples are compatible with the reconstruction patterns learned from the target support set, suggesting a low visual mismatch.

In our implementation, the autoencoder operates on grayscale input images. The encoder consists of four 3×3 convolutional layers with channel sizes 16, 32, 64, and 128, each followed by a LeakyReLU activation and 2×2 max-pooling. The resulting feature map is flattened and projected to a 512-dimensional latent representation. The decoder maps this latent code back to a feature map and reconstructs the image through four transposed-convolution layers with channel sizes 64, 32, 16, and 1, using LeakyReLU activations except for the final sigmoid output. The model is trained with a mean squared reconstruction loss using Adam.

B Robustness of UIAug without IDA

Figure B.14 reports the same comparison as in Figure 8, but without applying IDA. As expected, removing IDA reduces the absolute gains with respect to the main analysis. However, the relative behavior of the three pre-training strategies is largely preserved. Averaged over all source–target combinations and experimental configurations, supervised OOD-PT obtains a mean gain of 20.35 percentage points, while UIAug reaches 19.78 percentage points. This confirms that UIAug remains close to its supervised counterpart even without adaptation. In contrast, UCKM obtains a lower average gain of 16.51 percentage points.

The same conclusion holds when considering the best OOD source for each target. In this case, UIAug reaches an average gain of 45.78 percentage points, slightly above supervised OOD-PT, which obtains 45.02 percentage points, and clearly above UCKM, which obtains 40.04 percentage points. These results support the conclusions drawn in the main paper: UIAug provides a competitive label-free alternative to supervised OOD pre-training, while UCKM is less consistent. Moreover, the large gap between average and best-source performance confirms the importance of source–target compatibility, independently of the use of IDA.

C Descriptor-level analysis of source selection

Figure C.15 provides the descriptor-level counterpart to Fig. 12. Rather than aggregating the descriptors into conceptual families, it reports the source-selection error obtained with each individual descriptor and each extracted factor. Factor 1, denoted as the *Best factor* in the main analysis, achieves the lowest overall error, with a median oracle gap of 1.37 pp.

Among the individual descriptors, *Conceptual domain* and *DINOv2 Coverage* obtain the lowest median gaps, both at 2.41 pp., followed by *FSL-encoder GIQA-DS* at 4.02 pp. The

conceptual-domain indicator is also comparatively robust: by prioritizing sources from the same high-level semantic category as the target whenever available, it avoids many large errors associated with clearly mismatched source–target pairs. The competitive performance of the two representation-based descriptors further suggests that the extent to which a source covers the target representation space is informative of its transfer suitability. Nevertheless, their higher variability across targets supports the main conclusion: individual descriptors can provide effective selection cues, but combining complementary signals through factor analysis yields more reliable source selection overall.

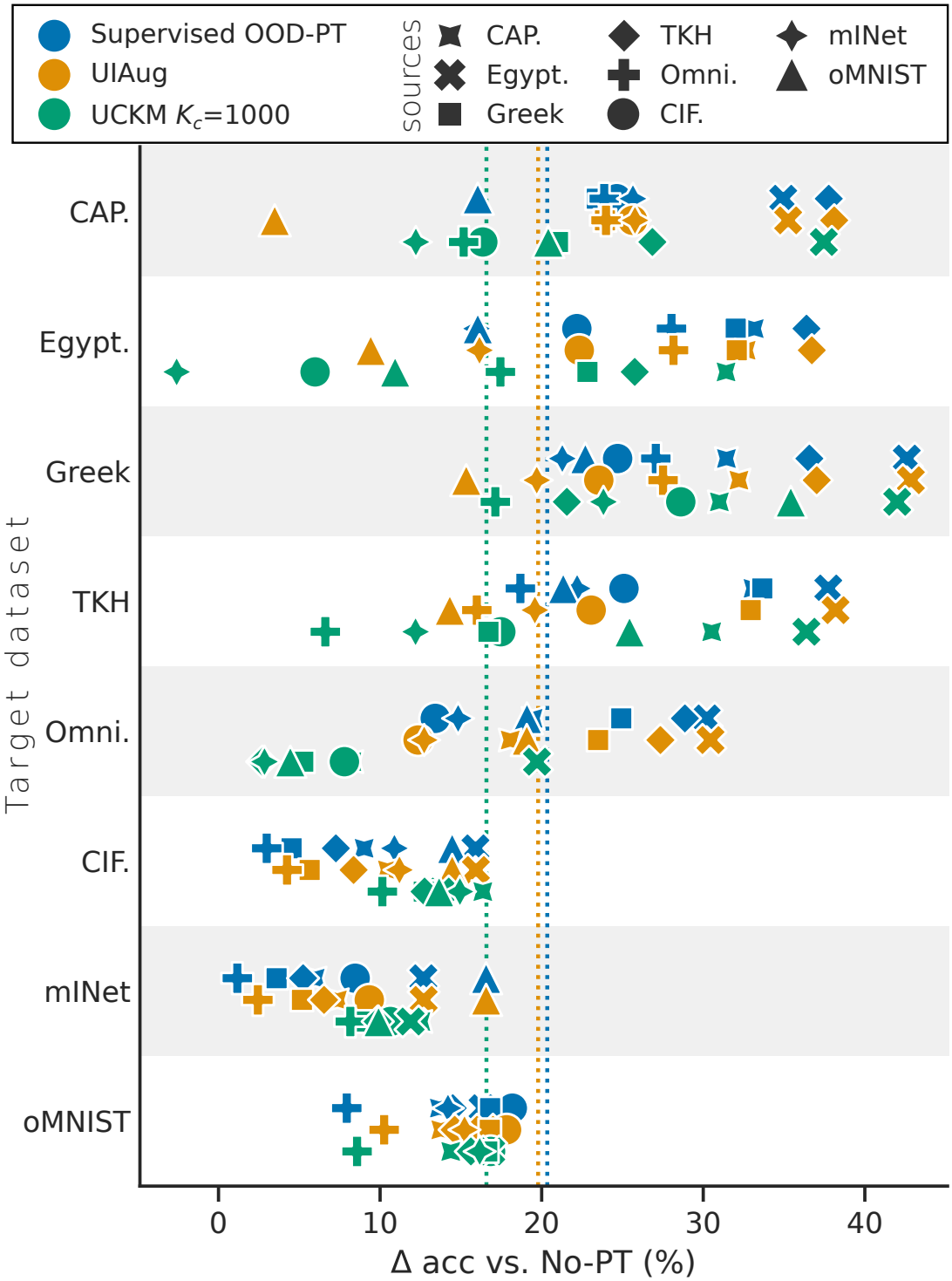


Figure B.14: Accuracy gain over the corresponding No-PT baseline, Δ_{acc}, for supervised OOD-PT, UIAug, and UCKM with $K_c = 1000$ without applying IDA. Results are shown for each target dataset, where marker shapes indicate the OOD source domain and colors indicate the pre-training strategy. Dotted vertical lines denote the average gain for each strategy.

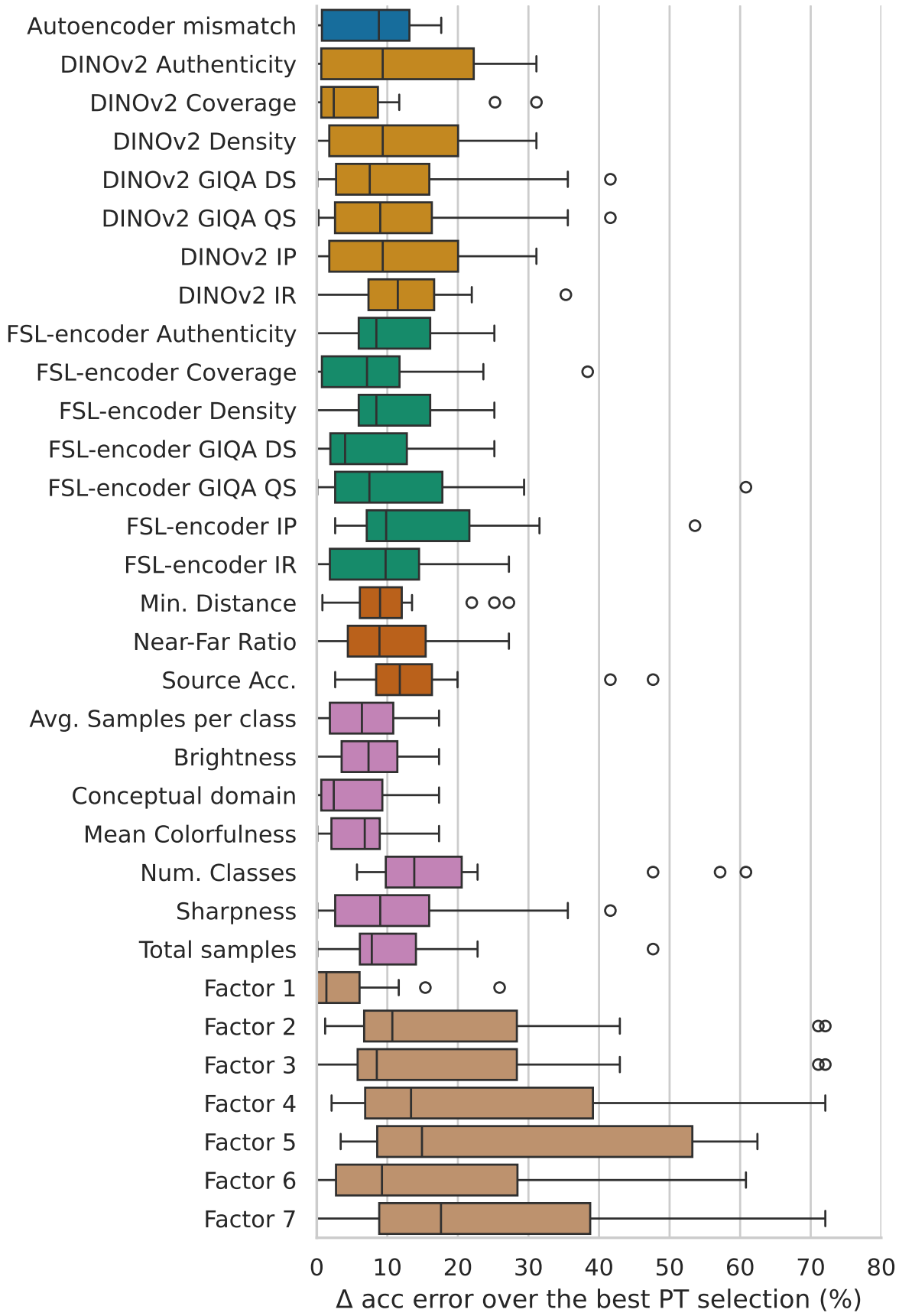


Figure C.15: Descriptor-level evaluation of source-domain selection. Candidate sources are ranked using each individual descriptor or extracted factor, and the top-ranked source is selected for OOD-PT. Box plots report the downstream accuracy-gain gap in pp. between the selected source and the oracle source, defined as the best-performing candidate after evaluating all sources. Factor 1 corresponds to the *Best factor* used in the main analysis. Results are averaged over the three OOD-PT strategies without IDA. Lower values indicate better selection.