TAILORING TRANSFERABILITY ESTIMATION EVALUATION TO APPLICATION PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

As the number of available pretrained models continues to grow, identifying the most suitable source models for a downstream task becomes increasingly difficult. Transferability estimation (TE) methods address this problem by ranking pretrained models without any exhaustive fine-tuning. Existing works evaluate these methods by measuring their agreement with a ground-truth ranking induced by a single score, mostly fine-tuned accuracy in classification. However, the accuracy may not be aligned with the objective of the downstream application, which may assign different importance to particular classes or errors, and thereby induce a different ranking of the same models. In this work, we argue that the evaluation of TE methods should explicitly account for downstream application preferences. To this end, we introduce a new framework in which ground-truth rankings are induced by performance scores tailored to the downstream application preferences. These scores form a family that includes the accuracy as a special case. When application preferences are uncertain, we evaluate TE methods across a distribution of application-conditioned ground-truth rankings. Through experiments, we show that conclusions drawn from conventional TE benchmarks do not necessarily hold across downstream preferences: ground-truth model rankings, their agreement with TE methods, as well as the induced ranking of TE methods vary with the downstream preferences. For binary classification, we further show how this variability can be visualized over the entire space of downstream preferences. The code is available in supplementary material.

1 INTRODUCTION

Fine-tuning deep learning models pretrained on large datasets such as ImageNet (Russakovsky et al., 2015) has become a standard paradigm in transfer learning (Kornblith et al., 2019; Chu et al., 2016). However, the downstream performance obtained after fine-tuning varies significantly with model architecture, source dataset, or pretraining strategy (Arango et al., 2024). Tailoring model selection to the downstream task is therefore critical. Yet, selecting the most suitable model by exhaustively fine-tuning all candidates is often impractical given the size of public hubs (Wightman, 2019). Transferability estimation (TE) methods address this issue by identifying promising models before fine-tuning. Given a target dataset, these methods assign a transferability score to each pretrained model. These transferability scores induce a ranking intended to predict how models will be ranked after fine-tuning.

Downstream application preferences. Existing works evaluate TE methods by computing the correlation between the predicted ranking and a fixed ground-truth ranking obtained by using a specific score on the fine-tuned models. This work focuses on classification, where it is most common to obtain the ground-truth ranking using the accuracy (Shao et al., 2022; Hu et al., 2024; Gholami et al., 2023). However, the accuracy may not reflect the downstream objectives of the application as it treats all classes and error types uniformly. In practice, application-specific trade-offs are required. For example, in skin-lesion classification, confusing two benign conditions is less clinically consequential than misclassifying a malignant lesion as benign (Yu et al., 2025). Even within the same classification task, different use cases may need different evaluations. For instance, a skin lesions classifier used for preventive screening may prioritize avoiding false negatives because missing a malignant lesion can delay diagnosis and intervention. By contrast, when its predictions are used to

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

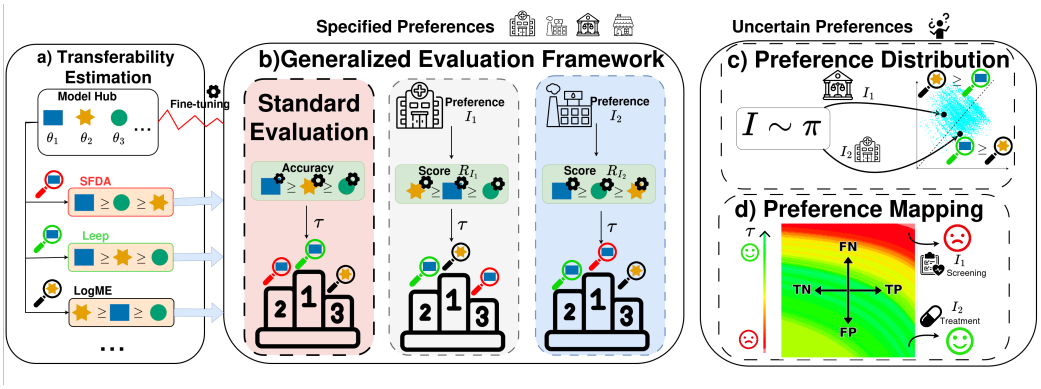


Figure 1: Overview of our application-conditioned evaluation framework. (a) TE methods rank pretrained models for a target task. (b) The ranking of models induced by each TE method is compared with a ground-truth ranking, which is induced by accuracies in standard evaluations. A ranking of TE methods is then obtained by ordering their respective agreement τ with the ground-truth ranking. However, accuracies may not reflect the priorities of a downstream application. Our framework generalizes this protocol by providing ground-truth ranking tailored to specified application preferences. (c) When the preferences are uncertain, we investigate the variability in evaluation outcomes by considering a distribution of preferences. (d) In binary classification, it is not necessary to use distributions as we can visualize this variability across a two-dimensional mapping of preferences.

inform a costly and potentially harmful treatment, avoiding false positives is more important (Gilbert et al., 2001; Aggarwal et al., 2022). Different downstream preferences can therefore induce different rankings of the same fine-tuned models. Conclusions obtained by evaluating TE methods against the accuracy-based ranking may not hold when considering other preferences.

A new application-conditioned evaluation framework. To address this limitation, we introduce an application-conditioned evaluation framework for TE, which is depicted in Fig. 1. It encompasses our contributions, which are the following: (a) We connect the evaluation of TE methods with downstream application preferences by leveraging a continuum of performance scores parametrized by preferences to obtain ground-truth rankings. To the best of our knowledge, this is the first time that TE method evaluation is performed based on an infinite family of scores instead of fixed scores. (b) We extend this to uncertain preferences by considering distributions over application preferences, enabling us to study how TE evaluation outcomes vary across possible applications. (c) For two-class classification, we show how to visualize TE method evaluations across the complete space of application preferences using Tiles. Tiles provide a powerful diagnostic tool for interpreting the variability of TE method evaluation induced by different application preferences. (d) We demonstrate the application of our evaluation framework on a popular benchmark in the literature. Our results show that different application preferences can affect the ground-truth model ranking, the measured agreement of TE methods with this ranking, and the resulting ranking of TE methods.

2 Related Work

Transferability Estimation. Numerous TE methods have been proposed to estimate how well pretrained models will perform after transfer to a target task. Earlier approaches derived transferability scores based on source-model predictions or relationships between source and target labels (Tran et al., 2019; Nguyen et al., 2020). Feature-based approaches instead rely on the structure of the feature space. They can use criteria such as class separability, probabilistic evidence, or neural collapse Bao et al. (2019); Ibrahim et al. (2023); Pandy et al. (2022); You et al. (2021); Wang et al. (2023). More recent approaches such as SFDA (Shao et al., 2022), LEAD Hu et al. (2024) or ITM Zheng et al. (2025) focus on modeling the fine-tuning process. Despite these methodological differences, existing works share the same evaluation protocol, *i.e.*, TE methods are compared to a unique ground-truth ranking, most often induced by the accuracy of the fine-tuned models (Nguyen et al., 2020; Pandy et al., 2022; You et al., 2021; Wang et al., 2023; Shao et al., 2022; Zheng et al.,

2025; Hu et al., 2024; Singh et al., 2026; Li et al., 2021; Ding et al., 2022; Huang et al., 2022; Gholami et al., 2023). However, there are some works that focus on alternative performance scores. Nguyen et al. (2020) uses the F-score F_1 for imbalanced binary classification. Ibrahim et al. (2023) and Chaves et al. (2023) respectively consider relative accuracy for target task selection and balanced accuracy in an imbalanced dataset in medical imaging.

TE Evaluation. Prior works have questioned different aspects of the evaluation protocol in TE. Agostinelli et al. (2022) shows that conclusions drawn from the evaluation of TE methods vary with experimental choices such as the target dataset, the models in the model hub, and the statistic used to compare transferability scores with downstream performance scores. Singh et al. (2026) further critiques common practices in TE evaluation, showing that the ranking of models according to the fine-tuned accuracies could almost be predicted using a static ranking of models across datasets. More recently, Claßen et al. (2026) shows that replacing accuracies with the AUC changes the ground-truth rankings in their application. Yet, the preferences of a particular application may not be adequately represented by any established performance score and may only be known qualitatively. We therefore consider a family of scores parametrized by preferences over classification outcomes and evaluate TE methods across a distribution of preferences when no single preference is specified.

Application-dependent performance-based model ranking. Model performance is commonly summarized using a few scalar scores. In classification tasks, top-k accuracy, balanced accuracy, mean-class accuracy, or macro-F_1 are popular choices (Sokolova & Lapalme, 2009). However, these scores are designed to reflect which particular aspects of performance should be emphasized and may not align with the preferences of every downstream application. Piérard et al. (2025c) recently formalized this dependence by introducing a framework for constructing families of ranking scores parametrized by application-specific importance weights. Subsequent works have primarily studied these scores in binary classification, where the space of application-dependent rankings can be represented in two dimensions (Piérard et al., 2024; 2025d; Halin et al., 2024; Piérard et al., 2025a;b), and have applied them in computer vision benchmarks Bouwmans et al. (2026); Zakharova et al. (2026). We build on this framework to tailor the reference rankings used in transferability estimation to known application preferences. We also introduce distributions over importance weights to represent uncertain preferences. We use these distributions to investigate how conclusions drawn from TE evaluation vary across plausible application preferences in multiclass classification.

3 Background on TE evaluation

Let $\mathcal{M} = \{\theta_j\}_{j=1}^{M}$ be a model hub containing M pretrained models θ_j and let $\{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}\}$ respectively denote the training and test splits of the target dataset. A TE method F assigns to each pretrained model θ_j a scalar transferability score $t_j = F(\theta_j)$ based on quantities such as features or logits obtained by applying θ_j to the train split $\mathcal{D}_{\text{train}}$ of the target dataset. It is then possible to use these scores $\{t_j\}_{j=1}^{M}$ to induce a ranking of the pretrained models according to F.

To evaluate a TE method, it is necessary to fine-tune every pretrained model on the train split of the target data $\mathcal{D}_{\text{train}}$. The fine-tuned models are then evaluated on the test split $\mathcal{D}_{\text{test}}$. Prior works then summarize the downstream performance by a scalar performance score R_I, which is typically the accuracy in classification. These performance scores also induce a ranking of fine-tuned models, which is then used as a ground truth in TE.

The quality of a TE method F is measured by the agreement between its predicted ranking and the ground-truth ranking. Common choices include Spearman rho ρ, Kendall tau τ, and weighted Kendall tau τ_w, where the latter is used to assign greater importance to disagreements involving highly ranked models. These correlation scores can then be used to rank TE methods themselves.

4 Our evaluation framework

In this section, we introduce how to connect TE methods evaluation with downstream application preferences through dedicated ground-truth rankings. Building on the work of Piérard et al. (2025c), we first construct a family of performance ranking scores parametrized by importance weights representing downstream application preferences and use them to induce application-conditioned ground-

truth rankings. We then show these scores can be used in two scenarios: when the preferences are specified or uncertain. The latter scenario requires introducing a distribution over importance weights, except in binary classification, where the entire normalized importance space, called a *Tile*, can be visualized in a single figure.

Performance ranking scores $\mathbf{R_I}$. Let $\mathcal{Y} = \{1, \dots, C\}$ denote the set of target classes. Consider the dataset $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^{N}$ on which the model is evaluated, and let $\hat{y}_j$ denote the predicted class obtained by applying the model on x_j. We represent the performance of a fine-tuned model by the empirical joint distribution $P_i(\hat{y}, y)$ of its predictions and ground-truth labels y, *i.e.*, its normalized confusion matrix. To construct performance ranking scores R_I, Piérard et al. (2025c) distinguishes two functions: the satisfaction S, which describes which outcomes $(\hat{y}, y)$ are desirable for the task, and the importance I, which represents the downstream application preferences by assigning non-negative weights to every outcome $(\hat{y}, y)$. In classification tasks, correct predictions are the desirable outcomes. We therefore fix the satisfaction as $S(\hat{y}, y) = \mathbf{1}(\hat{y} = y)$, which is the unit-diagonal matrix. The importance weights I are application-dependent and must then be specified by users to reflect the preferences of the application they are dealing with. Piérard et al. (2025c) then defines the performance ranking scores as in Eq. (1).

$$R_I(P_i) = \frac{\mathbb{E}_{P_i}[SI]}{\mathbb{E}_{P_i}[I]} = \frac{\sum_{(\hat{y},y)\in\mathcal{Y}^2} S(\hat{y}, y) I(\hat{y}, y) P_i(\hat{y}, y)}{\sum_{(\hat{y},y)\in\mathcal{Y}^2} I(\hat{y}, y) P_i(\hat{y}, y)}. \tag{1}$$

This family extends the scores used to obtain ground-truth rankings in TE methods evaluation. In particular, assigning equal importance weights I to all outcomes $(\hat{y}, y)$ recovers the accuracy. The rankings induced by diverse scores such as F_1, BA, or TNR can also be recovered (Piérard et al., 2025c). All these scores lead to interesting properties for the induced rankings[1].

Specified application preferences. When the preferences of the downstream application are known, they may be translated into importance weights $I^\star$. The corresponding score values $\{R_{I^\star}(P_i)\}_{i=1}^{M}$ then induce a ranking of the fine-tuned models that is tailored to downstream application preferences and is used as ground truth when evaluating TE methods. This formulation avoids restricting TE method evaluation to a predefined set of performance scores. Whenever the downstream application preferences can be expressed using importance weights $I^\star$, a score $R_{I^\star}$ exists, even if no standard performance score in the literature exactly expresses the desired trade-off.

As Eq. (1) shows, rescaling all importance weights I by the same positive scalar leaves the score R_I unchanged. Therefore, only the relative importance assigned to the different outcomes $(\hat{y}, y)$ needs to be specified. For example, in a medical screening task, one might assign five times more importance on classification outcomes involving sick patients compared to the others: $I^\star(\text{fn}) = I^\star(\text{tp}) = 5$ and $I^\star(\text{fp}) = I^\star(\text{tn}) = 1$.

Uncertain application preferences. When a single $I^\star$ cannot summarize downstream preferences, evaluating TE methods against one arbitrary ground-truth ranking can hide how conclusions depend on application preferences. For a given task, preferences may differ across use cases or only be known qualitatively. Hence, we represent the considered range of application preferences through a distribution π over importance weights I. Each draw $I \sim \pi$ induces its own application-conditioned ground-truth ranking against which TE methods can be evaluated. This allows us to study how conclusions drawn from TE evaluation vary across application preferences. When domain knowledge is available, π can be designed to assign greater probability to plausible preferences. Otherwise, a broad-coverage distribution can be used to assess the variability of conclusions drawn from TE method evaluation.

Binary classification and Tiles. In binary classification, the confusion matrix entries reduces to: true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn). To construct tiles, we leverage an invariance property of the performance scores R_I: independently rescaling the importance weights $\mathbf{I}^+ = (I(tn), I(tp))$ associated with correct outcomes and the importance weights

[1](A) The induced rankings remain coherent with the satisfaction associated to the task and (B) the score obtained using any convex mixture of performances lies between the lowest and largest scores of its components. Some popular performance scores, such as macro-averaged F_1, violate at least one of these properties.

$\mathbf{I}^- = (I(fp), I(fn))$ associated with incorrect outcomes yields the same ranking of models (Piérard et al., 2025c). Thus, fixing the sum of the weights in $\mathbf{I}^+$ and $\mathbf{I}^-$ to two different constants does not change the induced ranking and leaves only two free parameters $(a, b) \in [0, 1]^2$ (Piérard et al., 2024). Choosing $I(tn) + I(tp) = I(fp) + I(fn) = 1$, Eq. (2) becomes:

$$\begin{aligned} a &= \frac{I(tp)}{I(tp) + I(tn)}, \qquad b \;\; = \frac{I(fn)}{I(fp) + I(fn)}, \\ R_I(P) &= \frac{(1-a)\,P(tn) + a\,P(tp)}{(1-a)\,P(tn) + (1-b)\,P(fp) + b\,P(fn) + a\,P(tp)}. \end{aligned} \tag{2}$$

The resulting two-dimensional space is named *Tile* Piérard et al. (2024). Each point of the Tile corresponds to particular importance weights I and its associated performance score R_I. As illustrated in Fig. 2(B), several popular classification scores, such as accuracy (A), recall (TPR), specificity (TNR), precision (PPV), and F-score (F_1) are displayed at their location. Under fixed priors, the balanced accuracy (BA) is also in the Tile. It is then possible to obtain the ground-truth rankings induced by a specific R_I at each point of the Tile. This allows us to visualize TE evaluation across application preferences, removing the need for a distribution to explore the variability of TE evaluation in binary classification. As we will show in experiments, Tiles are especially useful as they directly allow to view the variability in TE evaluations.

5 Experimental Setup

We follow an experimental setup common in the literature to evaluate TE methods Shao et al. (2022); Hu et al. (2024); Gholami et al. (2023); Singh et al. (2026). More precisely, we consider eleven widely used model architectures that were pretrained on ImageNet (Russakovsky et al., 2015) and that are available in Torchvision: ResNet34 (He et al., 2016), ResNet50 (He et al., 2016), ResNet101 (He et al., 2016), ResNet152 (He et al., 2016), DenseNet121 (Huang et al., 2017), DensetNet169 (Huang et al., 2017), DenseNet201 (Huang et al., 2017), MnasNet-A1 (Tan et al., 2019), MobileNet-v2 (Sandler et al., 2018), InceptionV3 (Szegedy et al., 2016), and GoogLeNet (Szegedy et al., 2015). All these models are fine-tuned on ten popular computer vision datasets that cover a wide range of classification tasks, namely: FGVC Aircraft (Maji et al., 2013), Caltech-101 (Fei-Fei et al., 2004), Stanford Cars (Krause et al., 2013), CIFAR-10 (Krizhevsky, 2009), CIFAR-100 (Krizhevsky, 2009), DTD, Oxford 102 Flowers (Nilsback & Zisserman, 2008), Food-101 (Bossard et al., 2014), Oxford-IIIT Pets (Parkhi et al., 2012), and SUN 397 (Xiao et al., 2010). This fine-tuning is performed using SGD with momentum. As Li et al. (2020), we use a grid search on the learning rate and weight decay. The former belongs to $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ while the latter belongs to $\{0, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. Batch sizes are set to 64.

In this work, we consider seven popular TE methods, namely NCE (Tran et al., 2019), Leep (Nguyen et al., 2020), LogME (You et al., 2021), NCTI (Wang et al., 2023), SFDA (Shao et al., 2022), ETran (Gholami et al., 2023) and ITM Zheng et al. (2025). To assess the agreement between the transferability scores and the ground-truth ranking, we use the weighted Kendall tau τ_w which has been repeatedly used in the literature (You et al., 2021; Zheng et al., 2025; Hu et al., 2024).

Distribution of preferences. In this work, we are interested in the variability of the conclusions drawn from TE methods evaluation with respect to the application preferences. Therefore, instead of choosing specific importance weights, we use a broad distribution π over preferences. To construct such a distribution, we use the same invariance property of the performance scores R_I that was used to obtain tiles. Independently rescaling the importance weights $\mathbf{I}^+$ associated with correct outcomes and the importance weights $\mathbf{I}^-$ associated with incorrect outcomes yields the same ranking of models (Piérard et al., 2025c). We can therefore place the values of $\mathbf{I}^+$ and $\mathbf{I}^-$ on two simplexes and introduce two independent Dirichlet distributions:

$$\begin{aligned} &\mathbf{I}^+ \sim \text{Dirichlet}(\alpha_+), \quad I(\hat{y}, y) = C\, I^+_y, \\ &\mathbf{I}^- \sim \text{Dirichlet}(\alpha_-), \quad I(\hat{y}, y) = C(C-1)\, I^-_{\hat{y},y}, \quad \hat{y} \neq y. \end{aligned} \tag{3}$$

In this equation, the importance weights are obtained by respectively scaling the normalized importance weights $\mathbf{I}^+$ and $\mathbf{I}^-$ by C and $C(C-1)$. This scaling allows to recover popular scores. For instance, it is possible to draw $I(\hat{y}, y) = 1$ for all outcomes $(\hat{y}, y)$ and thus, recover the accuracy. In this work, we set all the concentration parameters to 1 *i.e.*, $\alpha^+ = \mathbf{1}_C$ and $\alpha^- = \mathbf{1}_{C(C-1)}$. This allows to obtain a broad distribution that samples uniformly the two simplexes.

In practice, we sample $10{,}000$ importance weights for each multiclass classification dataset. We then use Eq. (1) with these sampled importances to obtain the performance scores R_I of the different fine-tuned models. Ordering the different models according to these performance scores R_I yields the different ground-truth rankings. Let us underline that generating these ground-truth rankings requires no additional model training, as all R_I values are directly computed from the confusion matrices of the already fine-tuned models.

Tiles. We also fine-tune the different models on the ISIC2019 dataset (Codella et al., 2018). This dataset contains images of skin lesions that can is mapped into a binary classification task by grouping benign and malign classes together. More precisely, the malign class is obtained by grouping the MEL, BCC and SCC classes together. We use this dataset to illustrate the use of Tiles as diagnostic tools to visualize the variability in TE methods evaluation when dealing with binary classification. The Tiles in this work are obtained by computing the ranking scores R_I of each model for every importance I on a $1{,}000 \times 1{,}000$ grid.

6 Experiments

We apply our evaluation framework on the model hub and datasets described above. We start by expressing the variability of reference rankings when considering an entire range of application preferences. Then, we investigate how much the conclusions drawn from the evaluation of TE methods change across preferences. As we aim to isolate the effect of varying preferences, every aspect of our experiments, other than importance weights I, are fixed.

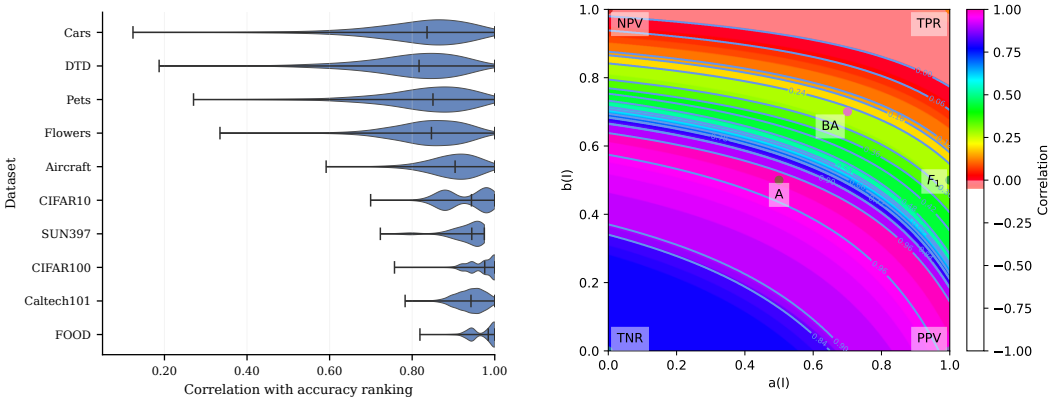
6.1 How much do reference model rankings depend on the application ?

We first investigate how much the reference rankings used in TE evaluation vary across application preferences. We compare the ranking induced by the accuracies of the fine-tuned models with the rankings induced by the performance scores R_I using weighted Kendall τ_w. The importance weights are sampled according to Eq. (3), and then yield performance scores following Eq. (1).

We can observe in Fig. 2(A) that the rankings induced by the accuracy are generally well correlated with the rankings induced by the ranking scores, as no average correlation is below 0.8. However, the extent to which ground-truth rankings vary across the sampled importance weights differs between datasets. The distributions for datasets Flowers, Pets, DTD and Cars have long left tails, reaching $\tau_w < 0.4$. It indicates that some importance weights induce ground-truth rankings that do change a lot compared to the ground-truth ranking of the accuracy.

For ISIC2019, Fig. 2(B) shows the weighted Kendall tau τ_w between the accuracy-induced ground-truth ranking and the ranking induced by R_I across the Tile. We observe that assigning greater importance on avoiding false negatives, *i.e.*, increase b, tends to produce very different ground-truth rankings. The ground-truth ranking associated to True Positive Rate (TPR) even reaches a correlation of -0.05. Balanced Accuracy (BA), a common choice for such an imbalanced dataset, also yields a different ground-truth ranking with $\tau_w = 0.27$. At the same time, large area of the Tile show high correlation $\tau_w > 0.84$.

These results indicate that while the ground-truth rankings induced by R_I are often close to the accuracy-induced ranking, the variation across application preferences is not negligible. Several multiclass datasets admit sampled preferences with much lower agreement. This is especially visible on the Tile where some preferences even lead to negative τ_w. Consequently, conclusions drawn using only the accuracy-induced ranking may not apply to an application with different preferences.



(A) Distribution of τ_w over the 10,000 sampled importances for every multiclass dataset.

(B) τ_w across the Tile for ISIC2019. Labeled points mark popular performance scores.

Figure 2: Variation of ground-truth rankings across application preferences. Weighted Kendall correlation τ_w compares each ranking induced by R_I with the ranking induced by accuracies. Low values of τ_w indicate that some application preferences induce ground-truth rankings with substantially less agreement with the accuracy-induced ranking.

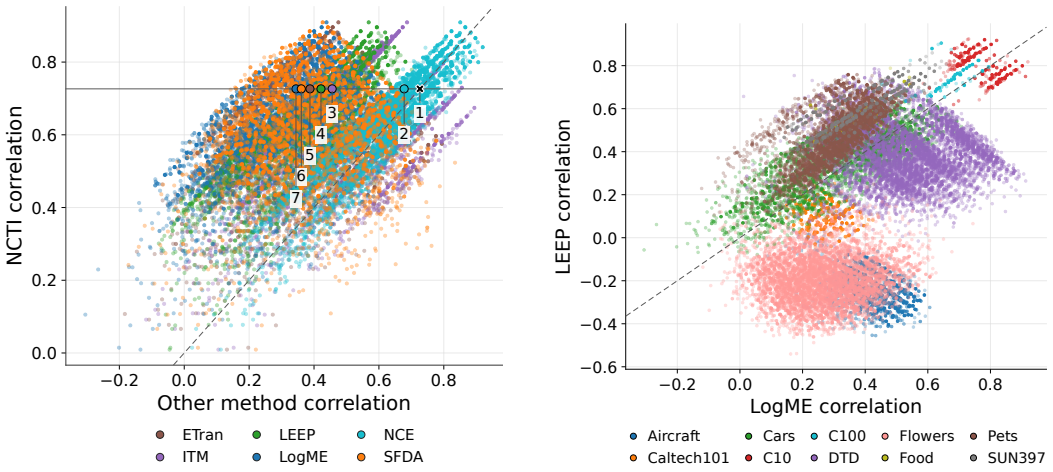
6.2 How much do conclusions about TE methods depend on the application ?

We now investigate how these changes in ground-truth rankings affect the conclusions drawn when evaluating TE methods. More specifically, we observe how the weighted Kendall tau τ_w obtained by comparing the ranking induced by the transferability scores of a TE method agrees with the ground truth sampled according to Eq. (3). Similarly to Agostinelli et al. (2022), we apply both qualitative and quantitative tools to analyze the variability in our results.

Qualitative evaluation. Figures 3(A) and 3(B) show how evaluating TE methods changes across application preferences. Each point (τ_x, τ_y) reports the weighted Kendall tau τ_w of two different TE methods with the same ground-truth ranking induced by a sampled importance weight. The plotted correlations range from below -0.2 to approximately 0.8, showing that measured agreement can vary substantially with application preferences. These point clouds are useful for pairwise comparison as a point above the diagonal line, defined by $\tau_x = \tau_y$, implies that the TE method on the vertical axis beats the other one. In Fig. 3(A), we compared the correlation of NCTI, shown on the y axis, with each other TE method, shown on the horizontal axis. NCTI was chosen because it is the best TE method according to the accuracy-induced ground-truth ranking. Thus, any points below the diagonal line represent an application preference for which the best TE method according to the accuracy-induced ground-truth ranking gets beaten by another TE method. Figure 3(B) provides the point cloud comparing LogME and Leep across the different datasets. In Flowers, LogME beats Leep across the sampled application preferences, whereas on Cars, their ordering depends on the importance weights. Note also that CIFAR10 illustrates why the variation of the reference ranking alone does not determine the variability of TE method evaluation. Despite its narrow distribution displayed in Fig. 2(A), its LogME-Leep point cloud crosses the diagonal line.

Quantitative evaluation. Table 1 extends the comparison by using the win rate of the different TE methods in every dataset. The win rate of a TE method is the fraction of sampled importance weights for which it achieves the highest τ_w among the other TE methods. We observe different results across datasets. For example, ETran wins 96% of the time on Food, whereas the wins are divided between NCTI and ITM on Pets. We can also observe that no TE method beats the other consistently across datasets, according to win rates.

Tiles. For ISIC2019, we use Tiles to visualize how the conclusions drawn from TE methods evaluation vary across application preferences. Figure 4 shows that LogME's weighted Kendall tau τ_w generally decreases when $b(I)$ increases *i.e.*, when the relative importance assigned to false negatives increases. This change also affects comparisons between TE methods. Figure 4 reports the best



(A) NCTI vs. other TE methods on Cars.

(B) LogME vs. LEEP across datasets.

Figure 3: Comparison of TE methods across application preferences. Each point correspond to one application preference and reports the weighted Kendall Tau τ_w of two TE methods with the corresponding ground-truth ranking. Points on both sides of the diagonal indicate the relative ordering of TE methods change across preferences.

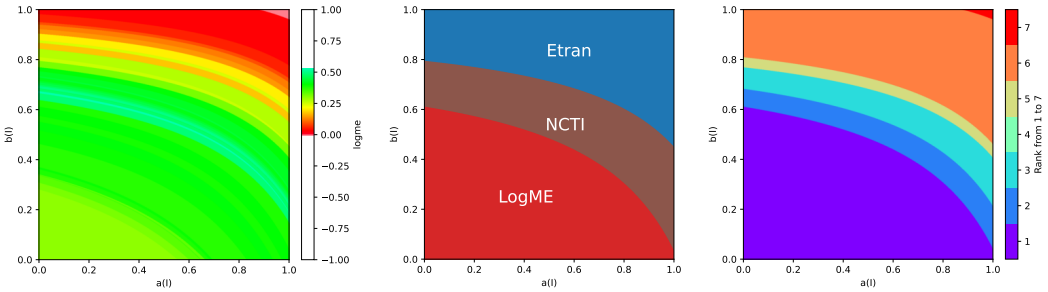
TE	Aircraft	Caltech	Cars	C10	C100	DTD	Flowers	Food	Pets	SUN397
ETran	0.00	0.00	0.03	*0.01*	**0.68**	0.00	*0.16*	**0.96**	*0.07*	*0.11*
ITM	0.00	0.00	*0.09*	0.00	0.01	**0.44**	0.12	0.00	0.38	**0.73**
LEEP	0.00	0.00	0.00	0.45	0.00	0.02	0.00	0.00	0.00	0.00
LogME	*0.01*	**0.90**	0.00	0.00	0.00	*0.07*	0.00	0.00	0.00	0.00
NCE	0.00	0.10	0.22	0.00	0.00	0.00	0.00	0.00	*0.07*	0.15
NCTI	0.13	0.00	**0.62**	0.00	0.30	0.40	0.19	0.00	**0.48**	0.00
SFDA	**0.86**	0.00	0.04	**0.53**	*0.02*	*0.07*	**0.52**	0.03	0.00	0.00

Table 1: Win rate for each transferability method across random importance experiments. For each random importance sample, a method wins when it achieves the highest correlation among the available transferability methods. Best win rates per dataset are in bold, second best are underlined, and third best are in italic.

TE method for every application preference. It can be observed that LogME is the best-performing TE method at low $b(I)$, Etran is the best at high $b(I)$, and NCTI is the best between the two. Figure 4 further shows how the rank of LogME among the TE methods deteriorates as $b(I)$ increases. We observe that LogME attains every rank from 1 to 7.

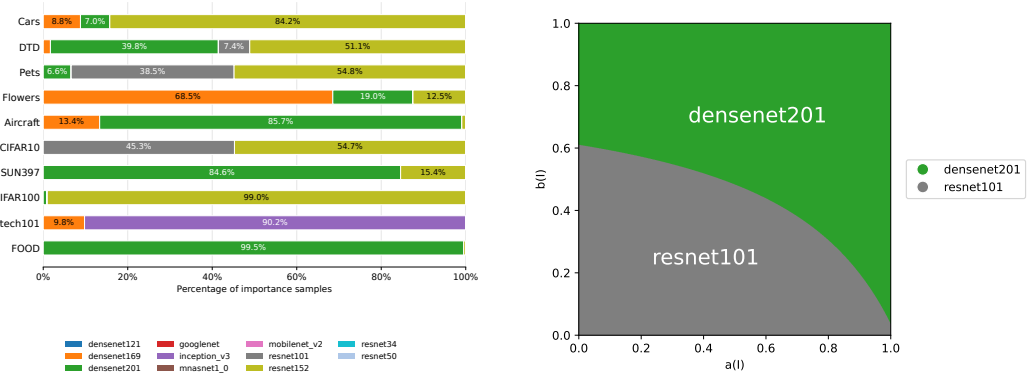
Model recommendations. A change in the highest-ranked TE method does not necessarily change the recommended model since different TE methods can recommend the same pretrained model. We thus investigate how often the recommended model changes in our evaluation framework. Figure 5(A) shows the proportion of sampled preferences for which each model is recommended. We can observe that the same model is recommended across sampled importance weights in some datasets, such as CIFAR100 or Food. On other datasets, such as Flowers or DTD, the recommended model varies across application preferences. Figure 5(B) contains a tile showing two regions for the different recommended models on ISIC2019. By comparing this figure with Fig. 4, we can see that NCTI and ETran both recommend Densenet201.

Overall, these results indicate that application preferences can change both the measured agreement between TE methods with the ground-truth rankings and the induced ranking of TE methods. The resulting model recommendation may also change, although different TE methods may recommend the same model. Using a single score cannot capture all this variability, thus highlighting the interest of the proposed evaluation framework.



(A) (B) (C)

Figure 4: Evaluation of TE methods on ISIC2019 using Tiles. Across the Tile, (A) shows the weighted Kendall tau τ_w between LogME's ranking and the ground-truth rankings induced by R_I, (B) identifies the TE method with the largest τ_w among the TE methods evaluated, and (C) shows LogME's rank among these TE methods.



(A) Frequency with which each model is recommended over the sampled importance weights.

(B) Recommended model at each location of the Tile.

Figure 5: Models recommended across application preferences. Recommendations vary less than win rates since different TE methods may recommend the same model.

7 Conclusion

In this work, we introduced a new evaluation framework that tailors the evaluation of transferability estimation (TE) methods to downstream application preferences using preference-specific ground-truth rankings. When the application preferences are specified, these rankings can be obtained from a family of performance scores. When the application preferences are uncertain, we sample them and examine how conclusions about TE methods vary. Using 10,000 sampled preferences per dataset, our experiments show that the conclusions drawn from TE methods evaluation can vary across application preferences. The correlation of TE methods, the ranking of TE methods, and the recommended model all vary. For two-class classification, we further show how to visualize this variability across application preferences over a two-dimensional space. These results highlight the interest of our evaluation framework, since using the ground-truth ranking provided by a single score may conceal conclusions relevant to other application preferences. We hope our findings motivate future works to explicitly state the application preferences underlying their evaluation of TE methods when available, or to consider a distribution of application preferences otherwise.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods and assisting with translation. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs or interpreting our results, and generating synthetic data sets, proposing or refining hypotheses, cleaning and reformatting dataset, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for modifying and creating scientific figures, drafting the initial structure of the paper, brainstorming and trying to improve the paper readability. We have reviewed all AI-assisted work. For instance, after asking AI for feedback on the experiments we were considering, we discussed with someone else for their opinion on the decisions we reached. Similarly, any piece of text translated or proposed from our ideas in our native language has been read multiple times by the different authors. Any code generated by LLMs was reviewed and compared with previous handwritten versions when available. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide our code as supplementary material. Given the provided transferability scores and confusion matrices, our results can be easily reproduced. It must however be highlighted that obtaining these transferability scores and confusion matrices can be difficult, even when using the official code of other papers. For instance, different checkpoints associated to different hash in Torchvision can lead to different transferability scores, even if they correspond to the same model architecture. Moreover, slight variations in the fine-tuning protocol can also change the ground-truth ranking and change the results.

REFERENCES

Rakesh Aggarwal, Priya Ranganathan, and C. S. Pramesh. Research studies on screening tests. *Perspectives in Clinical Research*, 13(3):168–171, July 2022. doi: 10.4103/picr.picr_111_22. URL https://doi.org/10.4103/picr.picr_111_22.

Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? In *European Conference on Computer Vision (ECCV)*, volume 13694 of *Lecture Notes in Computer Science*, pp. 303–321. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19830-4_18. URL https://doi.org/10.1007/978-3-031-19830-4_18.

Sebastian Pineda Arango, Fabio Ferreira, Arlind Kadra, Frank Hutter, and Josif Grabocka. Quicktune: Quickly learning which pretrained model to finetune and how. In *International Conference on Learning Representations (ICLR)*, pp. 13013–13032, Vienna, Austria, May 2024. URL https://openreview.net/forum?id=tqh1zdXIra.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *IEEE International Conference on Image Processing (ICIP)*, pp. 2309–2313, Tapei, Taiwan, September 2019. IEEE. doi: 10.1109/icip.2019.8803726. URL https://doi.org/10.1109/icip.2019.8803726.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://doi.org/10.1007/978-3-319-10599-4_29.

Thierry Bouwmans, Antonio Greco, Sébastien Piérard, Andrea Vincenzo Ricciardi, Carlo Sansone, Marc Van Droogenbroeck, and Bruno Vento. Illegal waste dumping detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 539–548, Tucson, Arizona, USA, March 2026.

Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila. The performance of transferability metrics does not translate to medical tasks. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, volume 14293 of *Lecture Notes in Computer Science*, pp. 105–114. Springer Nature Switzerland, October 2023. doi: 10.1007/978-3-031-45857-6_11. URL https://doi.org/10.1007/978-3-031-45857-6_11.

Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. Best practices for fine-tuning visual classifiers to new domains. In *European Conference on Computer Vision Workshops (ECCV Workshops)*, volume 9915 of *Lecture Notes in Computer Science*, pp. 435–442. Springer International Publishing, 2016. doi: 10.1007/978-3-319-49409-8_34. URL https://doi.org/10.1007/978-3-319-49409-8_34.

Niclas Claßen, Théo Sourget, Dovile Juodelyte, Rob van der Goot, and Veronika Cheplygina. Robustness of transferability estimation metrics for medical imaging. *arXiv*, abs/2608.09999, 2026. doi: 10.48550/arXiv.2608.09999. URL https://doi.org/10.48550/arXiv.2608.09999.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge hosted by the international skin imaging collaboration (ISIC). *International Symposium on Biomedical Imaging (ISBI)*, pp. 168–172, April 2018. doi: 10.1109/isbi.2018.8363547. URL https://doi.org/10.1109/isbi.2018.8363547.

Nan Ding, Xi Chen, Tomer Levinboim, Soravit Changpinyo, and Radu Soricut. PACTran: PAC-Bayesian metrics for estimating the transferability of pretrained models to classification tasks. In *European Conference on Computer Vision (ECCV)*, volume 13694 of *Lecture Notes in Computer Science*, pp. 252–268, Tel Aviv, Israël, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-19830-4_15. URL https://doi.org/10.1007/978-3-031-19830-4_15.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 178–178, Washington, District of Columbia, USA, 2004. IEEE. doi: 10.1109/cvpr.2004.383. URL https://doi.org/10.1109/CVPR.2004.383.

Mohsen Gholami, Mohammad Akbari, Xinglu Wang, Behnam Kamranian, and Yong Zhang. ETran: Energy-based transferability estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18567–18576, Paris, France, October 2023. IEEE. doi: 10.1109/iccv51070.2023.01706. URL https://doi.org/10.1109/ICCV51070.2023.01706.

Ruth Gilbert, Stuart Logan, Virginia A. Moyer, and Elizabeth J. Elliott. Assessing diagnostic and screening tests: Part 1. concepts. *Western Journal of Medicine*, 174(6):405–409, June 2001. doi: 10.1136/ewjm.174.6.405. URL https://doi.org/10.1136/ewjm.174.6.405.

Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck. A hitchhiker's guide to understanding performances of two-class classifiers. *arXiv*, abs/2412.04377, 2024. doi: 10.48550/arXiv.2412.04377. URL https://doi.org/10.48550/arXiv.2412.04377.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, Nevada, USA, June 2016. IEEE. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Zixuan Hu, Xiaotong Li, Shixiang Tang, Jun Liu, Yichun Hu, and Ling-Yu Duan. LEAD: Exploring logit space evolution for model selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28664–28673, Seattle, Washington, USA, June 2024. IEEE. doi: 10.1109/cvpr52733.2024.02708. URL https://doi.org/10.1109/CVPR52733.2024.02708.

G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger. Densely connected convolutional networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Honolulu, Hawaii, USA, July 2017. doi: 10.1109/CVPR.2017.243. URL https://doi.org/10.1109/CVPR.2017.243.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9201–9225, July 2022. URL https://proceedings.mlr.press/v162/huang22d.html.

Shibal Ibrahim, Natalia Ponomareva, and Rahul Mazumder. Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. *Machine Learning and Knowledge Discovery in Databases*, pp. 693–709, 2023. doi: 10.1007/978-3-031-26387-3_42. URL https://doi.org/10.1007/978-3-031-26387-3_42.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, Long Beach, California, USA, June 2019. IEEE. doi: 10.1109/cvpr.2019.00277. URL https://doi.org/10.1109/cvpr.2019.00277.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 554–561, Sydney, New South Wales, Australia, December 2013. IEEE. doi: 10.1109/iccvw.2013.77. URL https://doi.org/10.1109/iccvw.2013.77.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, April 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning. In *International Conference on Learning Representations (ICLR)*, pp. 1–20, Addis Ababa, Ethiopia, April 2020. URL https://openreview.net/forum?id=OyoxnmRzyzB.

Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. Ranking neural checkpoints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2662–2672, Nashville, Tennessee, USA, June 2021. IEEE. doi: 10.1109/cvpr46437.2021.00269. URL https://doi.org/10.1109/cvpr46437.2021.00269.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv*, abs/1306.5151, 2013. doi: 10.48550/arXiv.1306.5151. URL https://doi.org/10.48550/arXiv.1306.5151.

Cuong V. Nguyen, Tal Hassner, Matthias W. Seeger, and Cédric Archambeau. LEEP: a new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305, July 2020. URL http://proceedings.mlr.press/v119/nguyen20b.html.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, Bhubaneswar, India, December 2008. IEEE. doi: 10.1109/icvgip.2008.47. URL https://doi.org/10.1109/icvgip.2008.47.

Michal Pandy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using Bhattacharyya class separability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9162–9172, New Orleans, Louisiana, USA, June 2022. IEEE. doi: 10.1109/cvpr52688.2022.00896. URL https://doi.org/10.1109/cvpr52688.2022.00896.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505, Providence, Rhode Island, USA, June 2012. IEEE. doi: 10.1109/cvpr.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. The Tile: A 2D map of ranking scores for two-class classification. *arXiv*, abs/2412.04309, 2024. doi: 10.48550/arXiv.2412.04309. URL https://doi.org/10.48550/arXiv.2412.04309.

Sébastien Piérard, Adrien Deliège, Anaïs Halin, and Marc Van Droogenbroeck. A methodology to evaluate strategies predicting rankings on unseen domains. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Workshop on Big Surveillance Data Analysis and Processing (BIG-Surv)*, pp. 1–6, Nantes, France, June 2025a. IEEE. doi: 10.1109/icmew68306.2025.11152102. URL https://doi.org/10.1109/ICMEW68306.2025.11152102.

Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. Multi-domain performance analysis with scores tailored to user preferences. *arXiv*, abs/2512.08715, 2025b. doi: 10.48550/arXiv.2512.08715. URL https://doi.org/10.48550/arXiv.2512.08715.

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. Foundations of the theory of performance-based ranking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14293–14302, Nashville, Tennessee, USA, June 2025c. IEEE. doi: 10.1109/cvpr52734.2025.01333. URL https://doi.org/10.1109/CVPR52734.2025.01333.

Sébastien Piérard, Anaïs Halin, François Marelli, Simon Pernas, and Jérôme Pierre. Sorbetto: a Python library for producing classification tiles with different flavors to visualize, analyze, and compare performances in two-class classification problems. *Zenodo*, September 2025d. doi: 10.5281/zenodo.17591788. URL https://doi.org/10.5281/zenodo.17591788.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, April 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, Salt Lake City, Utah, USA, June 2018. IEEE. doi: 10.1109/cvpr.2018.00474. URL https://doi.org/10.1109/cvpr.2018.00474.

Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. Not all models are equal: Predicting model transferability in a self-challenging fisher space. In *European Conference on Computer Vision (ECCV)*, volume 13694 of *Lecture Notes in Computer Science*, pp. 286–302. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19830-4_17. URL https://doi.org/10.1007/978-3-031-19830-4_17.

Prabhant Singh, Sibylle Hess, and Joaquin Vanschoren. How NOT to benchmark your SITE metric: Beyond static leaderboards and towards realistic evaluation. In *International Conference on Learning Representations (ICLR)*, pp. 1–22, Rio de Janeiro, Brazil, April 2026. URL https://openreview.net/forum?id=ZHKVPkJMSI.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009. doi: 10.1016/j.ipm.2009.03.002. URL https://doi.org/10.1016/j.ipm.2009.03.002.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015. doi: 10.1109/CVPR.2015.7298594. URL https://doi.org/10.1109/CVPR.2015.7298594.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, Nevada, USA, June 2016. IEEE. doi: 10.1109/cvpr.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. MnasNet: Platform-aware neural architecture search for mobile. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2815–2823, Long Beach, California, USA, June 2019. IEEE. doi: 10.1109/cvpr.2019.00293. URL https://doi.org/10.1109/cvpr.2019.00293.

Anh Tran, Cuong Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1395–1405, Seoul, South Korea, October 2019. IEEE. doi: 10.1109/iccv.2019.00148. URL https://doi.org/10.1109/ICCV.2019.00148.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5526–5535, Paris, France, October 2023. IEEE. doi: 10.1109/iccv51070.2023.00511. URL https://doi.org/10.1109/ICCV51070.2023.00511.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, San Francisco, California, USA, June 2010. IEEE. doi: 10.1109/cvpr.2010.5539970. URL https://doi.org/10.1109/cvpr.2010.5539970.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143, July 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Zhen Yu, Toan D. Nguyen, Lie Ju, Yaniv Gal, Maithili Sashindranath, Paul Bonnington, Lei Zhang, Victoria Mar, and Zongyuan Ge. Hierarchical skin lesion image classification with prototypical decision tree. *npj Digital Medicine*, 8(1), January 2025. doi: 10.1038/s41746-024-01395-z. URL https://doi.org/10.1038/s41746-024-01395-z.

Anastasia Zakharova, Thierry Bouwmans, Anthony Cioppa, Adrien Deliège, Antonio Greco, Anaïs Halin, Kamil Jeziorek, Meghna Kapoor, Tomasz Kryjak, Islam Osman, Sébastien Piérard, Carlo Sansone, Mohamed S. Shehata, Renaud Vandeghen, Marc Van Droogenbroeck, and Bruno Vento. Event detection in videos: A framework for the development of new methods. *arXiv*, abs/2607.04372, 2026. doi: 10.48550/arXiv.2607.04372. URL https://doi.org/10.48550/arXiv.2607.04372.

Yaoyan Zheng, Huiqun Wang, Nan Zhou, and Di Huang. Implicit modeling for transferability estimation of vision foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 39, pp. 1–28, San Diego, California, USA, December 2025. Curran Associates, Inc. URL https://openreview.net/forum?id=QnUNr9Xuoo.