Transferability Estimation using Bhattacharyya Class Separability

Michal Pándy* Andrea Agostinelli Jasper Uijlings Vittorio Ferrari Thomas Mensink

Google Research

{michalpandy, agostinelli, jrru, vittoferrari, mensink}@google.com

Abstract

Transfer learning has become a popular method for leveraging pre-trained models in computer vision. However, without performing computationally expensive fine-tuning, it is difficult to quantify which pre-trained source models are suitable for a specific target task, or, conversely, to which tasks a pre-trained source model can be easily adapted to. In this work, we propose Gaussian Bhattacharyya Coefficient (GBC), a novel method for quantifying transferability between a source model and a target dataset. In a first step we embed all target images in the feature space defined by the source model, and represent them with per-class Gaussians. Then, we estimate their pairwise class separability using the Bhattacharyya coefficient, yielding a simple and effective measure of how well the source model transfers to the target task. We evaluate GBC on image classification tasks in the context of dataset and architecture selection. Further, we also perform experiments on the more complex semantic segmentation transferability estimation task. We demonstrate that GBC outperforms state-of-the-art transferability metrics on most evaluation criteria in the semantic segmentation settings, matches the performance of top methods for dataset transferability in image classification, and performs best on architecture selection problems for image classification.

1. Introduction

The goal of transfer learning is to reuse knowledge learned on a source task to help train a model for a target task. Currently, the most common form of transfer learning in computer vision is to pre-train a source model on the ILSVRC'12 dataset [55] and then fine-tune it on the target dataset [3, 14, 23, 24, 30, 35, 57, 75]. However, each target task may benefit from a different source model architecture [12, 25, 45, 53] or different source dataset [42, 46, 71]. The challenge then becomes to determine which (pre-trained) source model is most suitable for a particular target task, or to which target task a specific model can be easily adapted. Determining this by fine-tuning all combinations of source models and target datasets is computationally prohibitive.

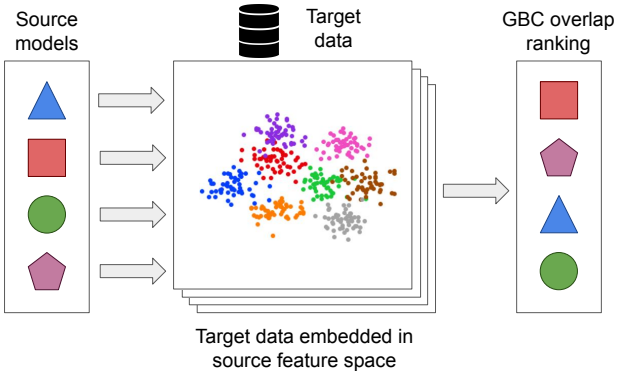

Figure 1. This figure illustrates the high-level overview of our approach. On the left, we use the pre-trained source models to embed data into the source models' feature space. On the right, we use GBC to rank these methods based on how much classes overlap in corresponding embedding spaces.

To address this problem, several recent works introduced transferability metrics [4, 38, 47, 60, 61, 72], which aim at predicting how well a source model transfers to a given target dataset. A good transferability metric is computationally efficient, and its predictions correlate well with the final performance of a model after fine-tuning on the target dataset. Typically a transferability metric is estimated by applying the source model to the target dataset to extract embeddings or predictions, which are then combined with the target ground-truth labels to measure transferability.

This paper proposes a novel transferability metric: the Gaussian Bhattacharyya Coefficient (GBC). The main idea is to measure the amount of overlap between target classes in the feature space of the source model (Fig. 1). If this overlap is small, the target classes are easily separated which means the knowledge in the source model is useful for the

*Currently at Waymo.

target task and the source model should transfer well. Conversely, if the overlap is large, the target classes are difficult to separate and the source model transfers badly to this target task. In order to estimate the amount of overlap, we apply the feature extractor of the source model to the target dataset and model each target class as a Gaussian distribution in this space. Importantly, we carefully apply regularization techniques to ensure that the Gaussian model can accurately represent each class. Then, we measure the sum of the overlaps between each pair of target classes using the Bhattacharyya coefficient. The Bhattacharyya coefficient has a closed-form solution when applied on Gaussian distributions. We use this overlap as our transferability metric.

We perform extensive experiments on two tasks. First we consider image classification, the primary focus of previous works on transferability metrics [4, 38, 47, 60, 61, 72]. Additionally, we consider a realistic transfer learning scenario for the task of semantic segmentation by considering transfer across a large variety of datasets from different image domains. Our experiments demonstrate that GBC outperforms several state-of-the-art transferability metrics: LEEP [47], LogME [72], H-score [4]. Furthermore, we demonstrate that our method is computationally efficient.

In summary, our paper makes the following contributions: 1) We introduce GBC, a new transferability metric which measures the amount of overlap between target classes in the source feature space. Since we model the target samples with per-class Gaussians, the GBC can be estimated in closed form; 2) We lift transferability experiments to a realistic transfer learning scenario for semantic segmentation; 3) We experimentally demonstrate that our GBC method outperforms other transferability metrics, including LEEP [47], LogME [72], and H-score [4].

2. Related work

While our work falls into the broad domain of transfer learning [50, 67], and relates to model selection [19, 52], and domain adaption [6, 18, 49], in this section we discuss the most relevant related work in estimating transferability metrics. We structure these along four general paradigms.

Task relatedness. Pioneering works in task relatedness [7, 33, 41] introduce symmetric measures between source and target tasks or domains. The intuition is that related tasks could be learned together more efficiently [7]. While intuitively task relatedness should correlate with transferability, these particular measures are generally hard to estimate. Moreover, the relatedness measures are symmetric, while transfer is asymmetrical: ImageNet is probably a very good source dataset for CIFAR-10, while the reverse is likely not true [42, 47].

Label comparison-based methods. LEEP [47] and NCE [61] use the labels of the source domain and the target domain to construct transferability metrics. In NCE by Tran *et al*. [61], they assume that the images of a source and target task are identical, but their labels differ. Then, they use the negative conditional entropy between the target labels and the source labels as a tranferability metric. Nguyen *et al*. propose LEEP [47], where the source model is applied to the target dataset. The resulting label predictions are utilised for computing a log-likelihood between the target labels and the source model predictions. An assumption in label comparison-based models is their dependence on the source output label space. Specifically, two source models with an identical feature extractor yet different classification heads will produce different transferability scores. In contrast source embedding-based methods rely purely on the underlying feature extractor.

Source embedding-based methods. Source embedding-based methods utilize the embeddings of target samples obtained via a pre-trained source model. Target embeddings are used together with their labels to compute various distance metrics. Cui *et al*. [16] propose to compute the Earth Mover's Distances between the class conditioned means of the embeddings. In H-score by Bao *et al*. [4], high transferability estimates are assigned to sources, where the embeddings display low feature redundancy and high inter-class variance. Li *et al*. [38] introduce $\mathcal{N}$LEEP, an extension to LEEP where the authors fit a Gaussian Mixture Model of the target data in the embedding space and use this in place of the source model's classification head to compute the LEEP score. Finally, You *et al*. [72] propose the state-of-the-art LogME score which treats each target label as a linear model with Gaussian noise, and then optimise the parameters of the prior distribution to find the average maximum (log) evidence of labels given the target sample embeddings. Our work also falls into the source embedding-based methods, but we directly consider the separability of class conditioned target embeddings.

Optimal transport. There have been several works proposing transferability estimation based on optimal transport (OT), including [2, 60]. The underlying assumption is that when in the source model's embedding space the source and target datasets have similar geometrical structures, and hence have a low OT-distance, the given source model is a suitable for the given target dataset. With [2], we share the idea to model classes as Gaussian distributions in the embeddeding space. However, OT based approaches have some serious drawbacks: (1) The method in [60] relies on parameter tuning based on ground-truth transferability scores; (2) These methods require access to the source training set; and (3) Computing the (regularized) OT distance scales quadratically in the number of data samples, which makes it practically infeasible to compute transferability scores for large datasets (including ImageNet).

3. Method

3.1. Formal description

Before we describe our method, we first provide a more formal description of the problem at hand. The goal is to estimate the transferability score $\mathcal{S}_{s\to t}$ of a *source model* m_s for a particular *target task* t. The target task t is described by a training set $\mathcal{D}_t$ containing images and ground truth label pairs (x_t, y_t).

A good transferability metric $\mathcal{S}_{s\to t}$ correlates with the accuracy $\mathcal{A}_{s\to t}$ of the target model $m_{s\to t}$. The accuracy $\mathcal{A}_{s\to t}$ is measured by evaluating $m_{s\to t}$ on the (unseen) test set of the target task $\mathcal{D}_t^{\text{test}}$. To create the target model $m_{s\to t}$, it is initialized using the weights of the source model m_s, after which it is fully fine-tuned on the target task t using the target training set $\mathcal{D}_t$. However, since fully fine-tuning $m_{s\to t}$ is computationally expensive, instead we want to predict how it will transfer using a computational efficient transferability metric $\mathcal{S}_{s\to t}$.

The source model m_s is defined by (a) the network architecture, such as ResNet50 [25] or VGG16 [58]; and (b) the training dataset used to train the source network, such as supervised classification on ImageNet [55].

For our method, we assume that we have access to the image embedding function of the source model $f_s(x)$, which returns a feature vector representation of image x. Our method only relies on the feature extractor $f_s(x)$, similar to H-score [4] and LogME [72]. In contrast, optimal transport based methods require access to the source (training) dataset $\mathcal{D}_s$ [2, 60], and LEEP requires the per target example predictions in the source label space [47].

Evaluating transferability. We evaluate the performance of the transferability metrics by evaluating the correlation between $\mathcal{S}_{s\to t}$ and $\mathcal{A}_{s\to t}$, as measured by the weighted Kendall tau rank correlation τ_w, as proposed by [72].

In contrast to the Pearson r correlation coefficient which measures a linear relation between $\mathcal{S}$ and $\mathcal{A}$, the Kendall rank correlation allows for highly non-linear relations, since it correlates rankings. The weighted Kendall correlation τ_w places higher weights on the models with the highest accuracies. This incorporates the rationale that it is more important to have the top few models correctly ranked, than the models with lower accuracies. For a more elaborate discussion on the appropriateness of the weighted Kendall tau, we refer the reader to You *et al.* [72].

We evaluate τ_w for different kinds of transferability scenarios, either fixing the target task to find the most suitable source model, or by correlating a fixed source model with different target tasks.

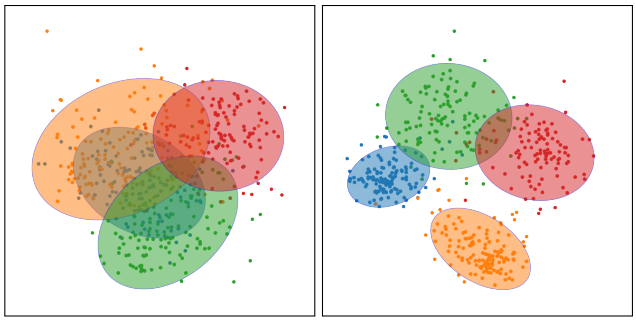

Figure 2. Illustration of the intuition: when images of the target classes overlap less in the embedding space of a source model, then it is a more suitable source to transfer from (left: a poor source; right: a better one). We use the Bhattacharyya coefficient to estimate the overlap between classes, each modeled as a Gaussian.

3.2. Class separability

The key idea behind our method is that if the target images are class-wise separable in the source model feature space $f_s(\cdot)$, then this source model allows for good classification for the target task. This intuition is shown in Fig. 2, where we show two embeddings of 4 target classes. We argue that the left dataset is more difficult to transfer to than the right dataset because the amount of class overlap is higher. We posit that the class overlap is proportional to the error of a sufficiently expressive fine-tuned classifier on the target dataset, and hence is proportional with the transferability of the source model to the target data. Our approach measures the amount of class separability of the target dataset in the source model feature space and uses that as transferability score $\mathcal{S}_{s\to t}$.

Bhattacharyya coefficient. The Bhattacharyya coefficient [8] (BC) is a measure of the amount of overlap between two distributions; in our case we want to measure the overlap between the probability densities of two target classes p_{c_i}, p_{c_j}:

$$\mathrm{BC}(p_{c_i}, p_{c_j}) = \int \sqrt{p_{c_i}(x)\, p_{c_j}(x)}\, \mathrm{d}x \tag{1}$$

Per-class Gaussian distributions. In order to compute the Bhattacharyya coefficient, we need to define the probabilistic model p_c for the target classes. We chose to model each class distribution with a Gaussian in the source embedding space: $p_c = \mathcal{N}(\mu_c, \Sigma_c)$, with:

$$\mu_c = \frac{1}{N_c} \sum_i [\![y_i = c]\!]\, f_s(x_i)$$

$$\Sigma_c = \frac{1}{N_c - 1} \sum_i [\![y_i = c]\!]\, (f_s(x_i) - \mu_c)\, (f_s(x_i) - \mu_c)^\top$$

where $N_c = \sum_i [\![y_i = c]\!]$, the number of images in this class.

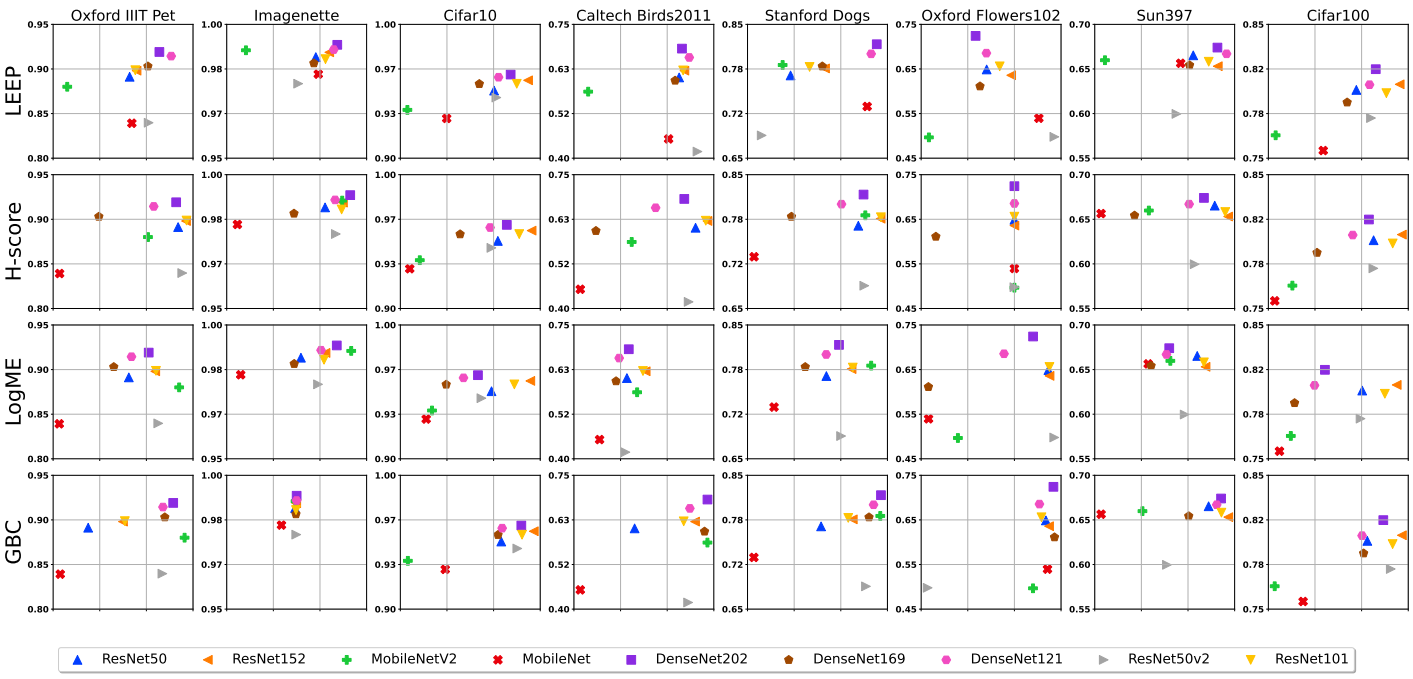

Figure 3. Overview of source selection experiments. For eight different target datasets we show the correlation between the accuracy of the model ($\mathcal{A}$, Y-axis) and the transferability scores ($\mathcal{S}$,X-axis) of LEEP, H-score, LogMe and GBC. See text for details.

While we do not suggest that per-class distributions are necessarily (multivariate) Gaussians, the advantage of using this model is that the Bhattacharyya coefficient can be computed in closed form from the class means and covariance matrices, using the Bhattacharyya distance D_B:

$$D_B(c_i, c_j) = \frac{1}{8}(\mu_{c_i} - \mu_{c_j})^\top \Sigma^{-1}(\mu_{c_i} - \mu_{c_j}) + \frac{1}{2}\ln\left(\frac{|\Sigma|}{\sqrt{|\Sigma_{c_i}||\Sigma_{c_j}|}}\right) \quad (2)$$

where $\Sigma = \frac{1}{2}(\Sigma_{c_i} + \Sigma_{c_j})$, the Bhattacharyya coefficient is then $\mathrm{BC}(\cdot,\cdot) = \exp -D_B(\cdot,\cdot)$.

Gaussian Bhattacharyya coefficient. Our final transferability score estimates the overlap of all classes by taking the sum of the pairwise coefficients:

$$\mathrm{GBC}_{s\to t} = -\sum_{i,j} [\![i \neq j]\!] \, \mathrm{BC}(c_i, c_j) \quad (3)$$

The final score uses the *negative* sum, because higher Bhattacharyya coefficients correspond to more overlap between the classes and therefore less transferability.

Theoretical guarantee. A nice property of GBC is that it provides some theoretical guarantee: when a classification head is fine-tuned on top of a fixed feature backbone and the per-class Gaussian assumption holds, then GBC is equivalent to an upper-bound on the optimal Bayes classification error [21, 40]. However, when the full model is fine-tuned, it is difficult to draw such a strong guarantee, as for all transferability metrics. For this, we rely on strong empirical results to demonstrate that GBC works well also in this general case.

3.3. Practical considerations

PCA dimensionality reduction. In practice, we transform the source embedding using the PCA projection into a fixed dimensional feature space of 64 dimensions. The reason for doing so is that different source architectures produce features with different number of dimensions and the Bhattacharyya coefficient is affected by the dimensionality due to its use of the determinant of the covariance matrix (Eq. (2)), which would make GBC scores difficult to compare. Moreover, reducing the number of dimensions allows to better estimate the Gaussian model.

Covariance estimation. To compute GBC, we need to estimate the per-class covariance matrices for all target classes. However, estimating the full covariance is infeasible, since the number of samples in a class can be very low, for example in the Caltech-USCD Birds [63] dataset, on average there are only 30 samples per class. Therefore, we experiment with both diagonal covariance matrices and spherical ones.

	Pets	Imagenette	CIFAR-10	CUB'11	Dogs	Flowers102	SUN	CIFAR-100	**Average**
LogMe	-0.06	0.58	0.25	0.2	0.08	0.00	-0.19	0.34	0.15
H-score	0.06	0.59	0.45	0.16	-0.01	0.09	0.09	0.34	0.22
LEEP	**0.63**	**0.65**	**0.52**	0.25	0.59	-0.46	**0.40**	**0.55**	0.39
GBC	0.55	0.63	0.46	**0.43**	**0.80**	**0.23**	0.32	0.35	**0.47**

Table 1. Overview of results for transferability for source selection in image classification. We depict for eight different target datasets the weighted Kendall τ_w between the accuracy of the fine-tuned model and the transferability scores from LEEP, LogMe, H-score and the proposed GBC. Our proposed method obtains the highest average τ_w across the different datasets.

Time complexity. In order to compute GBC, we first extract source model features $f_x(\cdot)$ for all images in the target data, the corresponding complexity if $O(NF)$, for N images, and where F denotes the complexity of extracting features. First, for PCA estimation using SVD, it costs $O(ND^2)$ to obtain the projection matrix. Then, to project samples and obtain their per-class means and covariance estimates in the reduced d-dimensional space is $O(NDd)$ and $O(NDd^2)$, respectively. Finally, computing the Bhattacharyya distance (2) between two classes in the reduced space with diagonal covariance matrices costs $O(d)$, so estimating our transferability metric (3) is $O(C^2d)$. In practice, the total run time largely depends on the cost of extracting features for the target dataset: $O(NF)$.

4. Experiments

In this section, we evaluate our proposed GBC transferability metric. We consider various transfer learning tasks to compare our proposed method against related work.

4.1. Classification: architecture transferability

Experimental setup. We consider several different source model architectures pre-trained on ImageNet. We want to identify which architecture would perform best on a given target dataset. To this end, we follow the experimental setup in [72] and evaluate our method using 8 different target datasets and 9 commonly utilized network architectures. Concretely, we fix the target dataset and we compute $\mathcal{A}_{s\to t}$ and $\mathcal{S}_{s\to t}$ for each architecture. Then, we measure weighted Kendall rank correlation τ_w (as proposed by You *et al*. [72]) between the reference $\mathcal{A}_{s\to t}$ and predicted $\mathcal{S}_{s\to t}$ across all the architectures, and report the results. We repeat this experiment for every target dataset.

We use the following target datasets: CIFAR-10 & 100 [36], Imagenette [28], Oxford IIIT Pets [51], Caltech-USCD Birds 2011 (CUB'11) [63], Stanford Dogs [32], Oxford Flowers 102 [48], and SUN-397 [70].

As source architectures we use ResNet-50, ResNet-101 & ResNet-152 [25], ResNetV2-50 [26], DenseNet-101, DenseNet-169 & DenseNet-201 [29], MobileNet [27], and MobileNetV2 [56] from the Keras library [13].

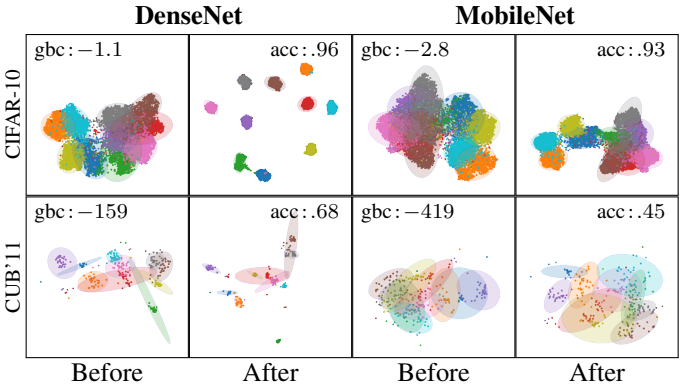

Figure 4. Feature distribution of CIFAR-10 (top) and 10 (randomly selected) classes of CUB'11 (bottom), visualized with UMAP.

The target accuracy $\mathcal{A}_{s\to t}$ is computed by evaluating the target model after fine-tuning each architecture on each target dataset for 100 epochs (with SGD with Momentum, using a batch size of 64 and learning rate of 10^{-4}).

We compare our method to three competitive baselines: H-score [4], LEEP [47] and LogME [72][1].

Main results and comparisons. We present the results of the full source architecture transferability experiment in Tab. 1. Notably, GBC achieves the highest average rank correlation τ_w of .47 over all the target datasets. Moreover, GBC is the only method to exhibit positive rank correlations for all target datasets. GBC determines the single best performing architecture for 3 target datasets, while LEEP and H-score for 2, and LogME for none. Furthermore, the best architecture is among the top-3 suggested models by GBC in 7 datasets, while only in 6 for LEEP, in 3 for H-score, and in 1 for LogME.

Fig. 3 presents the scatter plots of the accuracies $\mathcal{A}$ and estimated transferability scores $\mathcal{S}$ obtained by each method on each dataset. GBC showcases increasing trends across all datasets. These results demonstrate that GBC outperforms previous work for source architecture selection.

Fig. 4 shows the feature distributions before and after fine-tuning for two models with different GBC transferabil-

[1]LEEP & LogMe: github.com/thuml/LogME; H-score: git.io/J1WOr

ity scores (DenseNet and MobilNet). Each row shows a separate experiment on a different target dataset (CIFAR-10, CUB'11). In both cases, MobileNet has lower GBC scores than DenseNet and also results in lower accuracy after fine-tuning, demonstrating that our method works as intended.

Influence of regularization. We evaluated the influence of GBC's regularization parameters (Sect. 3.3). We used CIFAR-10 as the target dataset and transferred from the 9 source architectures listed above. For PCA we considered {16,32,64,128}-dimensional projections, and for the covariance estimation the regularization variants: {full, diagonal, spherical}. From the results we conclude that spherical regularization with 64-dimensional PCA projections delivers the best performance. Please see supplementary material for full details. Hence, in all classification experiments we use these settings (Sect. 4.1 & Sect. 4.2).

We want to highlight that using these settings the covariance estimation is robust, even in a low data regime: (1) On the smallest dataset we consider (CUB'11, 29 samples per class), GBC outperforms all previous methods in Tab. 1 and is on-par with the best in Tab. 2; (2) We computed the Pearson correlation (ρ) between GBC's performance and the number of samples per class. They are essentially uncorrelated ($\rho = -0.048$), suggesting that GBC does not perform worse with fewer samples per class.

Computational cost. To provide an indicative reference, we compare here the run times of several transferability metrics on CIFAR-100 (on a single CPU). After the feature extraction stage (shared by all metrics), GBC runs in 7.8s, vs 12.0s for LogME, 6.1s for H-score, and 0.2s for LEEP.

4.2. Classification: dataset transferability

Experimental setup. Good transferability metrics should correlate with a model's performance on target test data, as mentioned in Sect. 3.1. To evaluate this, we follow the setup from [47]: Given a fixed source model, the goal is to rank target datasets according to the actual performance of the source model after fine-tuning it on the target training set.

For this set of experiments all our source models have a ResNet-50 [25] architecture. Our first source model is trained on ImageNet [55]. This ImageNet [55] source model also acts as initialisation for the other 5 source models, trained on the following datasets: CIFAR-10 & CIFAR-100 [36], Fashion-MNIST [69], SUN397 [70] and Caltech-USCD Birds2011 [63]. This results in 6 source models. We use the same datasets as targets except for ImageNet, resulting in 5 target datasets. This results in 25 source-target pairs used as experiments.

For each of these 25 experiments, we use a single source model and a single *main* target dataset. Following [47], we construct a set of 100 *subsampled* target datasets from this main target dataset. Each subsampled target dataset is obtained by sampling uniformly between 2% and a 100% of the target classes and using all available images for these classes. For example, when the CIFAR-100 dataset is used as main target dataset, 100 subsampled datasets are created, each containing the CIFAR-100 images for 2–100 (randomly selected) classes.

For each of these subsampled target datasets, the trans-

Source	LEEP [47]	LogME [72]	H-score [4]	GBC *Ours*
		CIFAR-10		
CUB'11	0.68	0.71	0.69	**0.72**
CIFAR-100	**0.75**	**0.75**	0.73	0.69
F-MNIST	0.68	0.70	**0.72**	0.68
SUN	0.73	**0.75**	0.72	0.67
ImageNet	0.68	0.68	0.69	**0.71**
		CIFAR-100		
CUB'11	**0.90**	0.29	0.59	**0.90**
CIFAR-10	**0.92**	0.29	0.88	**0.92**
F-MNIST	**0.88**	0.24	0.26	**0.88**
SUN	**0.90**	0.30	0.88	**0.90**
ImageNet	0.91	0.25	0.88	**0.92**
		Fashion-MNIST		
CUB'11	**0.72**	0.71	0.71	0.71
CIFAR-10	0.72	**0.73**	0.69	0.69
CIFAR-100	**0.71**	**0.71**	0.70	0.69
SUN	**0.71**	**0.71**	0.69	**0.71**
ImageNet	**0.72**	0.71	0.69	0.70
		Caltech-USCD Birds 2011		
CIFAR-10	**0.87**	-0.59	0.83	0.86
CIFAR-100	**0.87**	-0.58	0.80	**0.87**
F-MNIST	**0.70**	-0.50	0.51	0.69
SUN	**0.88**	-0.60	0.80	**0.88**
ImageNet	**0.89**	-0.59	0.73	0.88
		SUN-397		
CUB'11	**0.95**	0.87	0.54	**0.95**
CIFAR-10	**0.95**	0.87	0.12	**0.95**
CIFAR-100	**0.95**	0.88	0.51	**0.95**
F-MNIST	**0.95**	0.86	0.54	**0.95**
ImageNet	**0.96**	0.87	0.55	**0.96**
Average	**0.82**	0.40	0.66	**0.82**

Table 2. Overview of results for transferability for target selection in image classification, where the transferability of subsampled target datasets are estimated (following the setup in [47]). From the results we observe that the proposed GBC method performs on par with the current state-of-the-art LEEP method.

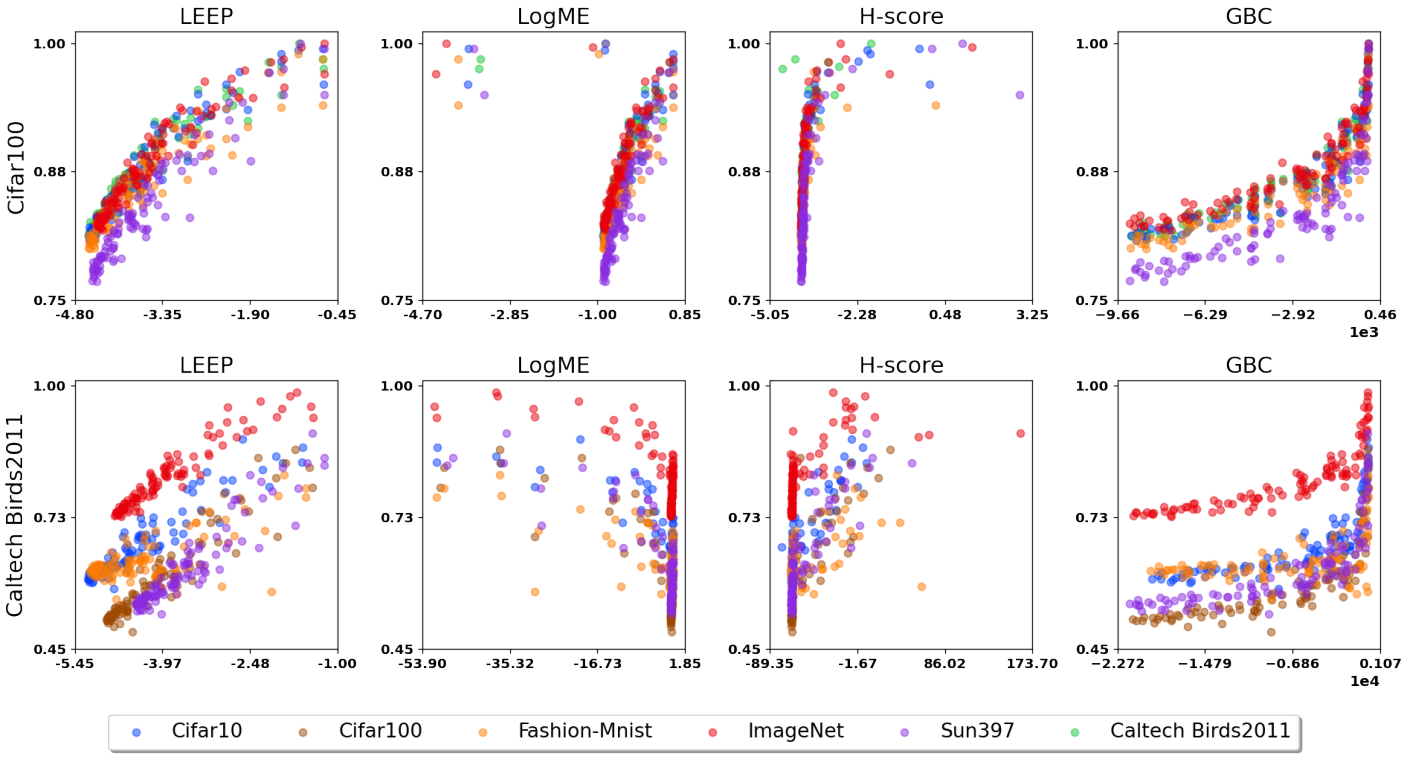

Figure 5. This figure illustrates the scatter plots of LEEP, LogME, H-score, and GBC for CIFAR-100 and CUB'11 as target datasets. In each figure, the transferability score $\mathcal{S}_{s\to t}$ of the method is on the *X-axis*, with the corresponding $\mathcal{A}_{s\to t}$ of each fine-tuned model on the *Y-axis*. From the plots we observe that while LogME and H-score tend to struggle to differentiate between target datasets, both GBC and LEEP showcase increasing trends.

ferability score of the source model is estimated. To obtain the target model $m_{s\to t}$, we fully fine-tuned the source model for each subsampled target dataset, for 100 epochs (using SGD with Momentum, with a batch size of 10 and learning rate of 10^{-3}, these hyper-parameters are the same as in LEEP [47]).

Evaluation. The accuracy $\mathcal{A}_{s\to t}$ is obtained by evaluating the final target models on the target test set (removing labels not sampled for this particular target task). We measure correlation between the transferability metric $\mathcal{S}_{s\to t}$ and the accuracy $\mathcal{A}_{s\to t}$ using the weighted Kendall rank correlation τ_w. The baselines we use are H-score, LogME, and LEEP. For fair comparisons, each method is evaluated on the same set of 100 random target datasets in all experiments.

Results. We present the quantitative results in Tab. 2. We observe that our proposed GBC has the top performance on 15 (out of 25) experiments, LEEP has the top performance on 19 experiments, LogME has the top performance in 5 cases, and H-score has the top performance in 1 case, where we include ties in our counting.

GBC and LEEP achieve an average τ_w score of .82, much higher than H-score (.66) and LogME (.40). Further, both GBC and LEEP consistently showcased high τ_w values ($\geq$ .67) across all experiments, while both H-score (.12) and LogME underperform for certain target datasets (*e.g.* CUB'11 for LogME). These results confirm that the proposed GBC method outperforms LogME and H-score, and is on par with LEEP in this setting.

We illustrate the correlation between $\mathcal{A}_{s\to t}$ and $\mathcal{S}_{s\to t}$ on CIFAR-100 (*top*) and CUB'11 (*bottom*) in Fig. 5. We observe that LogME and H-score fail to distinguish well between certain target datasets, *i.e.* assign near identical transferability scores despite the differences in target accuracies. On the other hand, both LEEP and the proposed GBC distinguish between the target datasets well, with persistent monotonically increasing trends.

4.3. Segmentation: dataset transferability

Experimental setup. We now turn to a transfer learning scenario for semantic segmentation, following the setup of [42]: 17 datasets spanning very different image domains (consumer photos, autonomous driving, aerial imagery, underwater, indoor scenes, synthetic, close-ups) containing 6–150 classes each: ADE20K [74], BDD [73], CamVid [9], CityScapes [15], COCO [11, 34, 39], IDD [62], iSAID [65,

68], ISPRS [54], KITTI [1], Mappilary [44], Pascal Context [43], Pascal VOC [20], ScanNet [17], SUIM [31], SUN RGB-D [59], vGallery [66], and vKITTI2 [10, 22]. While [42] used their setup to investigate what factors are important for good transfer learning, they did not aim to predict transferability. Nevertheless, we interpret one of their measurements as a transferability metric.

We use the low-shot target training regime of [42], which is arguably the most interesting scenario for transfer learning. The target training set is limited to 150 images for all datasets, except COCO and ADE20k, where the limit is set to 1000 images since they contain a large number of classes.

We use a HRNetV2-W48 backbone [64] with a linear classifier on top. This model offers excellent performance for semantic segmentation [64] and was also used in [37, 42]. We train a source model on each dataset.

We consider all $17 \times 16 = 272$ valid $\langle$source model, target dataset$\rangle$ pairs (for each target dataset we do not consider its corresponding source model trained on the full training set). For each pair, we compute the transferability metrics. We also compute the actual mean Intersection-over-Union performance by fine-tuning the source model on the target training set, and then evaluate on the target test set.

We evaluate in two scenarios like before: (1) given a fixed source model, we rank all valid target datasets; and (2) given a fixed target dataset, we rank all valid source models. For each scenario we measure the correlation with τ_w and also the top-1 selection accuracy: for scenario (1) the percentage of targets where the source with the highest predicted transferability score also has the highest actual performance, and for scenario (2) the same, however with the role of source and target reversed.

GBC estimation. For semantic segmentation, instead of one label per image, we have predictions at the pixel level. To estimate the transferability metrics, we consider each pixel x_i and its ground truth label y_i as a separate observation. Since using all observations for all metrics is too computationally expensive, we subsample 1000 pixels as observations per training image. We subsample using a class-balanced sampling strategy (*i.e*. sample inverse-proportionally to the label frequency), which we found to improve results for all metrics. To make the comparison completely fair, we always use the exact same subsampled pixels for each image to calculate all transferability metrics.

For semantic segmentation, even after subsampling, we have generally many more observations than for image classification. Therefore, instead of modelling spherical Gaussians, we model Gaussians with a diagonal covariance matrix, which offer a greater modeling capacity. This improved results for our method.

Image Domain Similarity. In [42] they demonstrated that transfer learning performance was reasonably correlated with image domain similarity (IDS) between the source and target dataset. In this paper we interpret IDS as a transferability metric. IDS was established as follows [42]: First a multi-task model (trained on multiple sources) was applied to 1000 randomly sampled images of each dataset, resulting in a single embedding vector per image. Then each target image embedding was matched to its closest source image embedding. Finally, IDS is the average euclidean distance between these matched embeddings. We obtained all IDS metrics from the authors in personal correspondence.

		IDS [42]	LEEP [47]	LogME [72]	GBC *Ours*
fixed target	top-1	0.41	0.47	0.59	**0.65**
	τ_w	0.45	0.53	**0.63**	0.59
fixed source	top-1	0.41	0.24	0.00	**0.76**
	τ_w	0.36	0.62	0.08	**0.69**

Table 3. Overview of results for transferability estimation for semantic segmentation. GBC outperforms all transferability methods in terms of top-1 accuracy, which is the most important measure in a practical transfer learning application.

Results. Tab. 3 presents the results. For scenario (1), when choosing a source model for a fixed target dataset, our method has the highest top-1 accuracy: it outperforms all other methods in choosing the best source, which is the main goal in a practical application. When looking at the weighted Kendall τ_w, which measures overall ranking correctness, LogME is best, and our method is second. In scenario (2), determining for a fixed source model to which target dataset it transfers best, LogME completely fails. While IDS and LEEP perform better, they are still significantly below our method in both top-1 accuracy and τ_w. We conclude that our proposed GBC transferability metric is the overall best transferability metric for semantic segmentation.

5. Conclusion

In this paper, we introduce the Gaussian Bhattacharyya coefficient (GBC), a novel transferability metric which measures the amount of overlap between target classes (each modelled as a Gaussian) in the source feature space. The societal impact is that it reduces the need for heavy training procedures in transfer learning by selecting good models to transfer from in an efficient manner. We compare our method against state-of-the-art transferability metrics: LogME [72], LEEP [47], and H-score [5] and show that GBC outperforms them (or is on par) on most evaluation criteria. A key limitation of GBC is that it is designed for classification tasks only (not regression).

References

[1] Hassan Alhaija, Siva Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision : Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 2018. 8

[2] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. In *NeurIPS*, pages 21428–21439, 2020. 2, 3

[3] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *TPAMI*, 2015. 1

[4] 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 Int. Conf. on Image Processing*, pages 2309–2313, 2019. 1, 2, 3, 5, 6

[5] 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 Int. Conf. on Image Processing*, 2019. 8

[6] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *NeurIPS*, 2007. 2

[7] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, 2007. 2

[8] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Indian Journal of Statistics*, 1946. 3

[9] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Patt. Rec. Letters*, 30(2):88–97, 2009. 7

[10] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv*, 2020. 8

[11] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. 7

[12] L-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018. 1

[13] François Chollet et al. Keras. https://keras.io/api/applications/, 2015. 5

[14] Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. Best practices for fine-tuning visual classifiers to new domains. In *ECCV*, 2016. 1

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 7

[16] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 2018. 2

[17] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 8

[18] Hal Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2009. 2

[19] Jie Ding, Vahid Tarokh, and Yuhong Yang. Model selection techniques: An overview. *IEEE SPM*, 2018. 2

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012. 8

[21] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013. 4

[22] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 8

[23] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 1

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 3, 5, 6

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 5

[27] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. Technical report, arXiv, 2017. 5

[28] J. Howard. Imagenette. github.com/fastai/imagenette/. 5

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 5

[30] M. Huh, P. Agrawal, and A.A. Efros. What makes imagenet good for transfer learning? In *NeurIPS LSCVS workshop*, 2016. 1

[31] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *IROS*, 2020. 8

[32] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR Workshops*, 2011. 5

[33] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, 2004. 2

[34] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *CVPR*, 2019. 7

[35] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, 2019. 1

[36] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5, 6

[37] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020. 8

[38] Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. Ranking neural checkpoints. In *CVPR*, pages 2663–2673, 2021. 1, 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva

Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 7

[40] Brian Mak and Etienne Barnard. Phone clustering using the bhattacharyya distance. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2005–2008. IEEE, 1996. 4

[41] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009. 2

[42] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *arXiv*, 2021. 1, 2, 7, 8

[43] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 8

[44] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 8

[45] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1

[46] Jiquan Ngiam, Daiyi Peng, Vijay Vasudevan, Simon Kornblith, Quoc V. Le, and Ruoming Pang. Domain adaptive transfer learning with specialist models. Technical report, arXiv, 2018. 1

[47] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *ICML*, 2020. 1, 2, 3, 5, 6, 7, 8

[48] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conf. on CVGIP*, 2008. 5

[49] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Trans. on Neural Networks*, pages 199–210, 2010. 2

[50] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. KDE*, 2009. 2

[51] O. M. Parkhi, A. Vedaldi, A. Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5

[52] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. Technical report, arXiv, 2018. 2

[53] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[54] Franz Rottensteiner, Gunho Sohn, Markus Gerke, Jan Dirk Wegner, Uwe Breitkopf, and Jaewook Jung. Results of the ISPRS benchmark on urban object detection and 3d building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2014. 8

[55] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 1, 3, 6

[56] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottleneck. In *CVPR*, 2018. 5

[57] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2016. 1

[58] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[59] S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. 8

[60] Yang Tan, Yang Li, and Shao-Lun Huang. Otce: A transferability metric for cross-domain cross-task representations. In *CVPR*, 2021. 1, 2, 3

[61] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, 2019. 1, 2

[62] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and C.V. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *Proc. WACV*, 2019. 7

[63] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 4, 5, 6

[64] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 8

[65] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *CVPR Workshops*, 2019. 7

[66] Philippe Weinzaepfel, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. Visual localization by learning objects-of-interest dense match regression. In *CVPR*, 2019. 8

[67] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016. 2

[68] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 7

[69] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. Technical report, ArXiV, 2017. 6

[70] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from Abbey to Zoo. In *CVPR*, 2010. 5, 6

[71] Xi Yan, David Acuna, and Sanja Fidler. Neural data server: A large-scale search engine for transfer learning data. In *CVPR*, 2020. 1

[72] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021. 1, 2, 3, 5, 6, 8

[73] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 7

[74] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 7

[75] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv*, 2019. 1