Efficient Transferability Assessment for Selection of Pre-trained Detectors

Zhao Wang[1] Aoxue Li[2*] Zhenguo Li[2] Qi Dou[1]
[1]The Chinese University of Hong Kong [2]Huawei Noah's Ark Lab
{zwang21@cse., qidou@}cuhk.edu.hk lax@pku.edu.cn li.zhenguo@huawei.com

Abstract

Large-scale pre-training followed by downstream finetuning is an effective solution for transferring deep-learning-based models. Since finetuning all possible pre-trained models is computational costly, we aim to predict the transferability performance of these pre-trained models in a computational efficient manner. Different from previous work that seek out suitable models for downstream classification and segmentation tasks, this paper studies the efficient transferability assessment of pre-trained object detectors. To this end, we build up a detector transferability benchmark which contains a large and diverse zoo of pre-trained detectors with various architectures, source datasets and training schemes. Given this zoo, we adopt 7 target datasets from 5 diverse domains as the downstream target tasks for evaluation. Further, we propose to assess classification and regression sub-tasks simultaneously in a unified framework. Additionally, we design a complementary metric for evaluating tasks with varying objects. Experimental results demonstrate that our method outperforms other state-of-the-art approaches in assessing transferability under different target domains while efficiently reducing wall-clock time 32× and requires a mere 5.2% memory footprint compared to brute-force fine-tuning of all pre-trained detectors. Our assessment code and benchmark will be publicly available.

1. Introduction

Under a paradigm of large-scale model pre-training [6, 8, 16, 20–22, 46, 50] and downstream fine-tuning [7, 17, 52], starting from a good pre-trained model is crucial. Nevertheless, it is too costly to perform selection of pre-trained models by brute-forcibly fine-tuning all available pre-trained models on a given downstream task [26, 57]. Fortunately, existing works have shown the advantages to efficiently evaluate the transferability of pre-trained models with specific design for image classification [13, 29, 34, 36, 42, 58, 59] and semantic segmentation [2, 36] tasks without finetuning. They usually estimate the transferability by measuring the class separation of representations extracted by different pre-trained models [2, 13, 34, 36]. While previous works consider estimating the transferability of classification or segmentation task, this paper aims to rank the transferablity of pretrained models for object detection, as shown in Figure 1. Since object detection methods address both classification and regression sub-tasks together in the same scheme, most assessment methods based on class-separation can hardly be applied, especially for the single-class object detection.

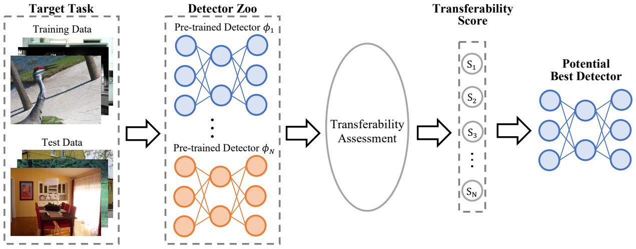

Figure 1. Illustration for selection of pre-trained detectors. The selection is performed by efficient transferability assessment.

To evaluate transferability assessment of object detection, we build up a challenging yet practical experimental setup with a zoo of diverse pretrained detectors. Specifically, we collect 33 large-scale pre-trained object detectors including two-stage [4, 38], single-stage [31, 45], and transformer [11, 64] based detection architectures. Meanwhile, they are equipped with different backbones, varying in ResNets [21], ResNeXts [53], and RegNets [37] and pre-trained with various datasets [12, 28, 32]. Moreover, we adopt 6 downstream tasks from 5 diverse domains under different scenarios including general objects [15], driving [9, 19], dense prediction [41], Unmanned Aerial Vehicle (UAV) [63] and even medical lesions [55]. In contrast, previous classification assementment methods usually simultaneously rank over 10∼20 pretrained models and test on datasets from at most 3 domains [13, 34, 36]. Upon the detection transferability benchmark, we propose to assess classification and regression sub-tasks simultaneously in a

[*]Corresponding author.

unified framework. Moreover, a complementary metric is designed for better assess tasks with various object scales.

Our main contributions are summarized as:

- This paper studies a crucial yet underexplored problem: efficient transferability assessment of pre-trained detectors.
- We build up a challenging detection transferability benchmark, containing 33 pre-trained detectors with different architectures, different source datasets and training schemes, and evaluating on diverse downstream target tasks. A series of metrics are designed to effectively assess these pre-trained detectors.
- Extensive experimental results upon this challenging transferability benchmark show the effectiveness and robustness of the proposed assessment framework compared with previous SOTA methods. Most importantly, it achieves over 32× wall-clock time speedup and only 5.2% memory footprint requirement compared with brute-force fine-tuning.

2. Related Work

2.1. Transferability of Pre-trained Models

Assessing the transferability of pre-trained models is an essential and crucial task. Early works [26, 57] have studied the transferability of the neural networks based on various layers within a single model and various models within a model zoo while the conclusions were drawn from too expensive fine-tuning (see Sec. 5.4), which is not affordable. Further, the transferability of deep knowledge was studied by evaluating the task relatedness [14,47,49,60] and building attribution graphs [43,44]. But these solutions still need costly downstream training and they are not applicable when meeting multi-scale features, multiple sub-tasks, and diverse detection heads, which are the key components in object detection.

Recently, several papers introduced efficient transferability metrics for classification. LEEP [34] proposed to efficiently evaluate the transferability of pre-trained models for supervised classification tasks by calculating the log expectation of the empirical predictor. However, the performance of LEEP degrades when the number of classes within the source data is less than that of the target task. A series of following works [2, 29] further improve LEEP for multi-class classification and pixel-level classification (i.e., semantic segmentation). $\mathcal{N}$-LEEP [29] tried to improve LEEP with Principal Component Analysis on the model outputs. and established a neural checkpoint ranking benchmark for classification. To better assess multi-class classification, Ding *et al.* [13] proposed a series of Cross Entropy (CE) based metrics for pre-trained classification models. Pándy *et al.* [36] proposed to estimate the pairwise Gaussian class separability using the Bhattacharyya coefficient, which could be applicable for both classification and segmentation. In SFDA [42], the authors aimed to leverage the fine-tuning dynamics into transferability measurement in a self-challenging fisher space, which degraded the efficiency.

Different from previous works designed for image-level or pixel-level classification, this paper aims to tackle more challenging pre-trained detector assessment. We thus build up a detector transferability benchmark, which simultaneously ranks 33 pre-trained models over 6 target datasets from 5 diverse domains, comparing with $10 \sim 20$ pre-trained models and 3 target domains in classification scenarios. Moreover, assessing pre-trained detectors should consider both classification and regression sub-tasks together. LogME [58, 59] first extends the transferability assessment from classification to regression and thus can be used as a baseline to assess the transferability of pre-trained detectors. However, it is designed for general regression task, without considering the multi-scale characteristics and inherent relation between coordinates in bounding box regression sub-task. To address these issues, we extend LogME to a series of metrics with special design for object detection.

2.2. Object Detection

Object detection is a practical computer vision task that aims to detect the objects and recognize the corresponding classes from an input image. The object detectors are always trained under supervised [4, 31, 38, 48, 64] and self-supervised [3,11,51,56] schemes on the large-scale datasets [12, 18, 28, 32, 40]. The supervised detectors are trained with ground truth bounding boxes and class labels while the self-supervised ones can only access the training images. Regarding the design of different detection architectures, there are three main streams, including two-stage [4,38,61], single-stage [31, 45, 48], and transformer [3, 11, 64] based detectors. A typical two-stage detector [38] works with 1st-stage proposal generator and 2nd stage bounding box refinement and class recognition while a single-stage detector [31] aims to perform dense predictions for the object location and class. Recent transformer based end-to-end detectors [5, 64] consider the object detection task as a direct set prediction problem optimized with Hungarian algorithm [27], in which lots of human-designed complex components are removed. With these large-scale pre-trained object detectors, how to figure out a good one for a given downstream detection task is crucial but underexplored. In this work, we aim to tackle the efficient transferability assessment for pre-trained detectors, which infers the true fine-tuning performance on a give downstream task.

3. Detector Transferability Benchmark

In this work, we construct a zoo of different detectors pre-trained with various source datasets and and thus build up a transferability benchmark to measure different assess-

Table 1. Pre-training schemes, source datasets and detection architectures used in this work. We include two-stage, single-stage, and transformer based detectors to build a pre-trained detector zoo. These detectors are equipped with different backbones, *e.g.*, ResNets [21], ResNeXts [53], and RegNets [37].

Scheme	Dataset	Type	Detector	Backbone
Supervised	COCO [32]	two-stage	FRCNN [38]	R50 [21] R101 [21] X101-32x4d [21] X101-64x4d [21]
			Cascade RCNN [4]	R50 [21] R101 [21] X101-32x4d [21] X101-64x4d [21]
			Dynamic RCNN [61]	R50 [21]
			RegNet [37]	400MF [37] 800MF [37] 1.6GF [37] 3.2GF [37] 4GF [37]
			DCN [10]	R50 [21] R101 [21] X101-32x4d [21]
		single-stage	FCOS [48]	R50 [21] R101 [21]
			RetinaNet [31]	R18 [21] R50 [21] R101 [21] X101-32x4d [21] X101-64x4d [21]
			Sparse RCNN [45]	R50 [21] R101 [21]
		transformer	DDETR [64]	R50 [21]
	Open Images [28]	two-stage	FRCNN [38]	R50 [21]
		single-stage	RetinaNet [31]	R50 [21]
Self-Supervised	ImageNet [12]	two-stage	SoCo [51] InsLoc [56]	R50 [21] R50 [21]
		transformer	UP-DETR [11] DETReg [3]	R50 [21] R50 [21]

Table 2. Target downstream datasets used in this work, in which they are from 5 diverse domains.

Dataset	Domain	Classes	Images
Pascal VOC [15]	General	20	21K
CityScapes [9]	Driving	8	5K
SODA [19]	Driving	6	20K
CrowdHuman [41]	Dense	1	24K
VisDrone [63]	UAV	11	9K
DeepLesion [55]	Medical	8	10K

ment metrics. In what follows, we will provide details of the detector transferability benchmark.

Problem Setup. With a given downstream detection task and a detection model zoo consisting of N pre-trained models $\{\mathcal{F}_n\}_{n=1}^N$, the aim of model transferability assessment is to produce a transferability score for every pre-trained detector, and then find the best one for further fine-tuning according to the score ranking.

Pre-trained Source Detectors. In the past few years, a great number of detection models arise with very smart design. Typically, these detectors are trained under supervised [38,48,64] or self-supervised [11,51] scenarios. Supervised detectors are always trained and evaluated on large-scale detection datasets, such as COCO [32] and Open Images [28], while self-supervised ones are trained based on ImageNet [12]. In this work, we take three datasets, *i.e.*, COCO, Open Images, and ImageNet, as the source datasets for different pre-training schemes. Upon these source datasets and training schemes, we use 13 detection architectures, including two-stage [4, 10, 37, 38, 51, 56, 61], single-stage [31, 45, 48], and transformer [3, 11, 64] based detectors, for pre-training and obtain a model zoo composed of 33 various pre-trained detectors. These detectors are equipped with different backbones, *e.g.*, ResNets [21], ResNeXts [53], and RegNets [37]. The detailed information is shown in Table 1. With a large variety of pre-trained source detectors, it can be ensured that at least one good model exists for a given target task and our evaluation metric can distinguish between bad and good source models.

Target Datasets. The image domain plays an important role for deep-learning-based models, which directly determines the model performance in a transfer learning scenario [33, 35]. So we cover a wide range of image domains as a challenging but practical setting with diverse scenarios, including general objects [15], driving [9, 19], dense prediction [41], Unmanned Aerial Vehicle (UAV) [63], and even medical lesions [55]. The detailed information of these target datasets are summarized in Table 2.

Evaluation Protocol. Given pre-trained object detectors $\{\mathcal{F}_n\}_{n=1}^N$ and a downstream dataset $\mathcal{D}_t$, a transferability assessment method will produce the transferability scores $\{s_n\}_{n=1}^N$. Following the previous works [2, 13, 29, 36, 42, 58, 59], we take the fine-tuning performance of pre-trained detectors $\{g_n\}_{n=1}^N$ as the ground truth, *i.e.*, mean Average Precision (mAP) as the metric in object detection. Ideally, the transferability scores are positively correlated with true fine-tuning performance. That is, if a pre-trained model $\mathcal{F}_n$ has higher detection mAP than $\mathcal{F}_m$ after fine-tuning ($g_n>g_m$), the transferability score of $\mathcal{F}_n$ is also expected to be larger than $\mathcal{F}_m$ ($s_n>s_m$). So the effectiveness of a transferability metric for assessing the pre-trained detectors is evaluated by the ranking correlation between the ground truth fine-tuning performance $\{g_n\}_{n=1}^N$ and estimated transferability scores $\{s_n\}_{n=1}^N$. We use *Weighted Kendall's* τ_w [23] as the evaluation metric. Larger τ_w indicates better ranking correlation between $\{g_n\}_{n=1}^N$ and $\{s_n\}_{n=1}^N$ and better transferability metric. τ_w is interpreted by

$$\tau_w = \frac{2}{N(N-1)} \sum_{1\leq n<m\leq N} \operatorname{sgn}\left(g_n - g_m\right) \operatorname{sgn}\left(s_n - s_m\right). \tag{1}$$

Here τ_w ranges in $[-1, 1]$, and the probability of $g_n>g_m$ is $\frac{\tau_w+1}{2}$ when $s_n>s_m$. Moreover, we use Top-1 Relative Accuracy (Rel@1) [29] to measure how close model with the highest transferability performs, in terms of fine-tuning

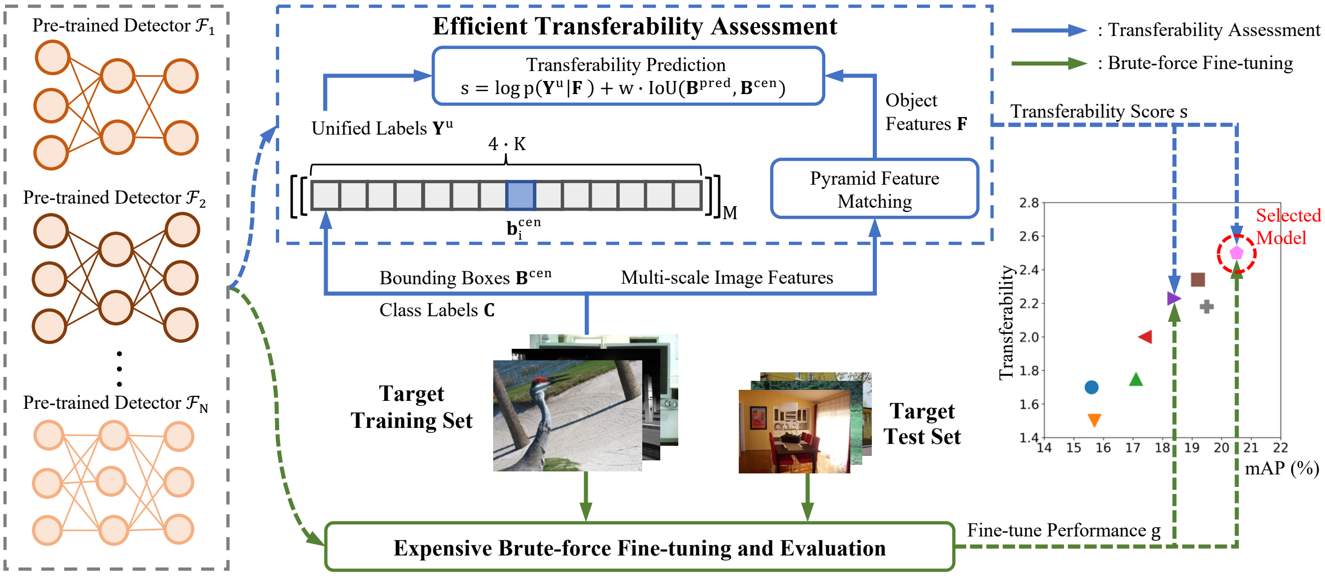
Pre-trained Detector $\mathcal{F}_1$
Pre-trained Detector $\mathcal{F}_2$
Pre-trained Detector $\mathcal{F}_N$
Efficient Transferability Assessment
Transferability Prediction
$s = \log p(\mathbf{Y}^u|\mathbf{F}) + w \cdot \mathrm{IoU}(\mathbf{B}^{pred}, \mathbf{B}^{cen})$
Unified Labels $\mathbf{Y}^u$
$4 \cdot K$
$\mathbf{b}_i^{cen}$
M
Object Features F
Pyramid Feature Matching
Bounding Boxes $\mathbf{B}^{cen}$
Class Labels C
Multi-scale Image Features
Target Training Set
Target Test Set
Expensive Brute-force Fine-tuning and Evaluation
Fine-tune Performance g
: Transferability Assessment
: Brute-force Fine-tuning
Transferability Score s
Selected Model
Transferability
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
14 15 16 17 18 19 20 21 22
mAP (%)

Figure 2. The overview of efficient transferability assessment framework for pre-trained detection models. We build a challenging setting contains various pre-trained object detectors. Based on this challenging setting, we design a pyramid feature matching scheme to handle objects with various sizes and expand the bounding box matrix $\boldsymbol{B}^{cen}$ according to class label matrix $\boldsymbol{C}$ to the unified label matrix $\boldsymbol{Y}^u$ for evaluation. We estimate the maximum evidence $p(\boldsymbol{Y}^u|\boldsymbol{F})$, which indicates the compatibility between the object features $\boldsymbol{F}$ and unified labels $\boldsymbol{Y}^u$. Further, considering IoU as an important metric in object detection, we supply IoU between the predicted bounding boxes and ground truth ones as a complementary term for transferability assessment of detection model.

performance, compared to the highest performing model.

4. Detection Model Transferability Metrics

In this section, we study the problem of efficient transferability assessment for detection model. Given a model zoo with a number of pre-trained detectors and a target downstream task, our goal is to predict their transferability performance on the target task efficiently, without brute-force fine-tuning all pre-trained models. LogME is a classification assessment method designed from the viewpoint of regression [58]. Thus, it can be used to assess transferability of pre-trained detectors which contain both classification and regression subtasks. However, it is designed for general regression and might fail when tackling the bounding box regression task with multi-scale characteristics of inputs and inherent relation between coordinates in outputs. To this end, as shown in Figure 2, we extend LogME to detection scenario by designing a unified framework (U-LogME in Sec. 4.2) which assesses multiple sub-tasks and multi-scale features simultaneously. Furthermore, we propose a complementary metric (i.e., IoU-LogME in Sec. 4.3) for better transferability assessment over objects with varying scales.

4.1. LogME as a Basic Metric

Different from most existing assessment methods that measure class separation of visual features, LogME addresses the problem from the viewpoint of regression. Specifically, it uses a set of Bayesian linear models to fit the features extracted by the pre-trained models and the corresponding labels. The marginalized likelihood of these linear models is used to rank pre-trained models. Since it can address both classification and regression tasks, we can easily extend it to assess object detection framework.

Specifically, we extract multi-scale object-level features of ground-truth bounding boxes by using pre-trained detectors' backbone followed by an ROIAlign layer [38]. In this way, for a given pre-trained detector and a downstream task, we can collect the object-level features of downstream task by using the detector and form a feature matrix $\boldsymbol{F}$, with each row $\boldsymbol{f}_i$ denotes an object-level feature vector. For each $\boldsymbol{f}_i$, we also collect its 4-d coordinates of ground-truth bounding box $\boldsymbol{b}_i$ and class label c_i to form a bounding box matrix $\boldsymbol{B}$ and a class label matrix $\boldsymbol{C}$.

For the bounding box regression sub-task, LogME measures the transferability by using the maximum evidence $p(\boldsymbol{B}|\boldsymbol{F}) = \int p(\boldsymbol{\theta}|\alpha)p(\boldsymbol{B}|\boldsymbol{F}, \beta, \boldsymbol{\theta})d\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the parameter of linear model. α denotes the parameter of prior distribution of $\boldsymbol{\theta}$, and β denotes the parameter of posterior distribution of each observation $p(\boldsymbol{b}_i|\boldsymbol{f}_i, \beta, \boldsymbol{\theta})$. By using the evidence theory [24] and basic principles in graphical models [25], the transferability metric can be formulated as

$$\begin{aligned}\mathrm{LogME} = & \log p(\boldsymbol{B}|\boldsymbol{F}) \\ = & \frac{M}{2}\log\beta + \frac{D}{2}\log\alpha - \frac{M}{2}\log 2\pi \\ & - \frac{\beta}{2}\|\boldsymbol{Fm} - \boldsymbol{B}\|_2^2 - \frac{\alpha}{2}\boldsymbol{m}^T\boldsymbol{m} - \frac{1}{2}\log|A|,\end{aligned} \tag{2}$$

where $\boldsymbol{m}$ is the solution of $\boldsymbol{\theta}$, M is the number of objects, D is the dimension of features, and A is the L_2-norm of $\boldsymbol{F}$.

LogME for classification sub-task can be computed by replacing $\boldsymbol{B}$ in Eq. (2) with converted one-hot class label matrix. By combining the evidences of two sub-tasks together, we obtain a final evidence for ranking the pre-trained detector. However, LogME still struggles to rank the pre-trained detectors accurately due to the following four challenges: 1) The single-scale features of LogME is not compatible with the multi-scale features extracted by pyramid network architecture of pre-trained detector. 2) The huge coordinates variances of different-scale objects make it hard to fit by simple linear model used in LogME. 3) The assessment branch of classification sub-task might fail when the downstream task is single-class detection. 4) The mean squared error (MSE) used in LogME isn't scale-invariant and measures each coordinate separately, which is not suitable for bounding box regression. In what follows, we propose how to address these issues.

4.2. U-LogME

This subsection proposes to unify multi-scale features and multiple tasks into an assessment framework. Here, we propose a pyramid feature mapping scheme to extract suitable features for different scale objects. Meanwhile, we normalize coordinates of bounding boxes and jointly evaluate 4-d coordinates with the same linear network. This can help to reduce coordinate variances and thus benefit ranking. Furthermore, we merge the class labels and normalized bounding box coordinates into a final ground-truth label to joint assess two sub-tasks with the same model. In this way, both single-class and multiple-class downstream detection tasks are unified into one assessment framework. We provide more technical details in the following.

Pyramid Feature Matching. Classical object detectors always include a Feature Pyramid Network (FPN) [30] like architecture with different feature levels. The object features are obtained from different levels according to the corresponding object sizes during training, *e.g.*, very small and large objects will be mapped to bottom-level and top-level image features, respectively. Regarding this, we introduce *pyramid feature matching* to help objects with different sizes find their matched level features. Following FPN [30], we assign an object to the feature pyramid level P_l by the following:

$$l = \left\lfloor l_0 + \log_2(\sqrt{wh}/224) \right\rfloor, \tag{3}$$

where w and h is the width and length of an object on the input image to the network, respectively. Here 224 is the ImageNet [12] training size, and l_0 is the feature level mapped by an object with $w \times h = 224^2$. Inspired by FCOS [48], to better handle too small and large objects, objects satisfying $\max(w,h) < 64$ and $\max(w,h) > 512$ are further forcibly assigned to the lowest and highest level of feature pyramid. Given a bounding box and its P_l, we use an RoI Align layer to crop the P_l-th feature map according to bounding box coordinates and thus obtain its features. Thus, we can extract suitable visual features for multi-scale ground-truth objects.

Improved BBox Evaluation. With these multi-scale object features, our model can predict their coordinates of bounding boxes. In the object detection scenario, bounding box targets can be formulated as corner-wise coordinates $\boldsymbol{b}_i = (x_1, y_1, x_2, y_2)$ or center-wise coordinates $\boldsymbol{b}_i^{cen} = (x_c, y_c, w_c, h_c)$. To avoid coordinate scale issue, each coordinate is rescaled to the range [0,1] by using *bounding box center normalization*. Since the bias in the former one (i.e., $x_1 < x_2, y_1 < y_2$) is hard to be fit by a linear model in LogME, we thus select the latter one. LogME for bounding box regression is obtained by averaging over 4-d coordinates, where each coordinate learns different prior distribution with different α and β. Considering the inherent relation between coordinates, we feed 4-d coordinates of a bounding box as a whole and learn a shared prior distribution for 4 coordinates with the same α and β. More specifically, We expand the dimension of $\boldsymbol{m}$ in Eq. (2) from $\boldsymbol{m} \in \mathbb{R}^D$ to $\boldsymbol{m} \in \mathbb{R}^{D \times 4}$ to match the dimension of $\boldsymbol{B}^{cen}$, resulting in a more efficient and accurate evaluation.

Unified Sub-task Evaluation. Although the correlations among the coordinates are captured by joint evaluation of bounding box coordinates, LogME for regression sub-task still suffers from neglecting object classes information. So how to build the correlation between these sub-tasks? Inspired by the class-aware detection heads of classical detectors [38, 48, 64], we propose *unified sub-task evaluation* for assessing the transferability of pre-trained detectors. To be specific, we combine the bounding box matrix $\boldsymbol{B}^{cen}$ and class label matrix $\boldsymbol{C}$ as the unified label matrix $\boldsymbol{Y}^u$. The bounding box matrix $\boldsymbol{B}^{cen}$ is expanded from $\boldsymbol{B}^{cen} \in \mathbb{R}^{M \times 4}$ to $\boldsymbol{Y}^u \in \mathbb{R}^{M \times (4 \cdot K)}$ according to $\boldsymbol{C}$, where K is the total number of classes. By integrating pseudo bounding boxes filled with 0 coordinates, we obtain unified label matrix $\boldsymbol{Y}^u$ as

$$\boldsymbol{Y}^u = \left[\left[\underbrace{(0,0,0,0)}_{1\,\mathrm{st}}, \ldots, \overbrace{\underbrace{(x_c, y_c, w_c, h_c)}_{c_i\,\text{-th}}}^{\boldsymbol{b}_i^{cen}}, \ldots, \underbrace{(0,0,0,0)}_{K\,\text{-th}} \right]\right]_M . \tag{4}$$

To this end, both bounding boxes and classes information are represented by $\boldsymbol{Y}^u$. Accordingly, $\boldsymbol{m}$ in Eq. (2) is further expanded from $\boldsymbol{m} \in \mathbb{R}^{D \times 4}$ to $\boldsymbol{m} \in \mathbb{R}^{D \times (4 \cdot K)}$ for matching the dimension of $\boldsymbol{Y}^u$. We can take the unified label matrix $\boldsymbol{Y}^u$ as the input for transferability assessment and obtain a unified transferability score for detection as the following:

$$\text{U-LogME} = \log p(\boldsymbol{Y}^u \,|\, \boldsymbol{F}). \tag{5}$$

4.3. IoU-LogME

Although U-LogME addressed multi-scale and multi-task issues of vanilla LogME, the MSE used for fitting fea-

Table 3. Ranking results of of six methods for 1% 33-choose-22 possible source model sets (over 1.9M) on 6 downstream target datasets. Higher τ_w and Rel@1 indicate better ranking and transferability metric. As SFDA is specifically designed for classification task, it is not applicable for the single-class task of CrowdHuman. The results of all three variants of our approach, U-LogME, IoU-LogME, and Det-LogME are reported. The best methods are in red and good ones are in blue.

Measure	*Weighted Kendall*'s tau (τ_w) ↑						Top1 Relative Accuracy (Rel@1) ↑					
Method	KNAS	SFDA	LogME	U-LogME	IoU-LogME	Det-LogME	KNAS	SFDA	LogME	U-LogME	IoU-LogME	Det-LogME
Pascal VOC	0.10±0.18	**0.65±0.13**	0.15±0.22	0.40±0.17	0.58±0.16	**0.78±0.03**	0.94±0.10	**1.00±0.00**	0.91±0.12	0.96±0.05	**1.00±0.00**	**1.00±0.00**
CityScapes	-0.22±0.24	0.45±0.06	0.15±0.20	0.13±0.16	**0.51±0.09**	**0.57±0.08**	0.95±0.06	0.95±0.01	0.95±0.06	0.90±0.04	**0.98±0.02**	**0.98±0.02**
SODA	-0.46±0.09	0.46±0.12	0.13±0.21	0.04±0.17	**0.61±0.09**	**0.61±0.09**	0.88±0.04	0.95±0.02	0.92±0.11	0.87±0.06	**0.98±0.02**	**0.98±0.02**
CrowdHuman	-0.42±0.11	N/A	0.19±0.19	0.21±0.18	**0.34±0.16**	**0.34±0.16**	0.85±0.04	N/A	0.97±0.04	0.97±0.04	**0.98±0.03**	**0.98±0.03**
VisDrone	0.04±0.20	0.53±0.12	0.48±0.17	0.17±0.17	**0.70±0.08**	**0.69±0.08**	0.88±0.17	**1.00±0.00**	0.90±0.15	0.78±0.13	**0.99±0.02**	**0.99±0.02**
DeepLesion	-0.13±0.19	-0.21±0.14	0.08±0.20	**0.52±0.14**	-0.05±0.18	**0.42±0.17**	0.69±0.11	0.65±0.06	0.72±0.17	**0.87±0.28**	0.64±0.07	**0.75±0.38**
Average	-0.18±0.28	0.31±0.10	0.19±0.20	0.24±0.17	**0.45±0.13**	**0.57±0.10**	0.86±0.09	0.91±0.02	0.90±0.11	0.89±0.10	**0.93±0.03**	**0.95±0.08**

tures and bounding box coordinates is not robust to object scales. That is, the larger objects might cause bigger MSE, which makes the model prefers larger objects than smaller ones. Moreover, each coordinate is evaluated separately in mean squared error, without considering the correlations between different coordinates. To overcome this issue, we introduce IoU metric into U-LogME. Specifically, $\boldsymbol{m}$ in Eq. (2) is interpreted as a linear regression model so that $\boldsymbol{Fm}$ can be regarded as the bounding box predictions from a detector naturally. We propose to calculate the IoU between the bounding boxes predictions $\boldsymbol{Fm}$ and the corresponding ground truth bounding boxes $\boldsymbol{B}^{cen}$ as a IoU based transferability measurement. Considering $\boldsymbol{m} \in \mathbb{R}^{D\times(4\cdot K)}$ is computed from the unified label matrix $\boldsymbol{Y}^u \in \mathbb{R}^{M\times(4\cdot K)}$ in Eq. (4), so we downsample $\boldsymbol{m}$ to $\boldsymbol{m}' \in \mathbb{R}^{D\times 4}$ by reserving the values where real coordinates of $\boldsymbol{B}^{cen}$ arise. This IoU-based metric is formulated as:

$$\text{IoU-LogME} = \text{IoU}(\boldsymbol{Fm}', \boldsymbol{B}^{cen}) = \frac{|\boldsymbol{Fm}' \cap \boldsymbol{B}^{cen}|}{|\boldsymbol{Fm}' \cup \boldsymbol{B}^{cen}|}. \tag{6}$$

4.4. Det-LogME

Although IoU-LogME is invariant to objects with different scales, it degrades to 0 when the two inputs have no intersection and fails to measure the absolute difference between them. On the contrary, MSE is good at tackling these cases. Therefore, to take advantage of their strengths, we propose to combine them together and obtain a final detector assessment metric Det-LogME, which is formulated as:

$$\text{Det-LogME} = \text{U-LogME} + \mu \cdot \text{IoU-LogME}, \tag{7}$$

where μ is used for controlling the weight of IoU. U-LogME and IoU-LogME are normalized to $[0, 1]$ upon 33 pre-trained detectors to unify the scale.

5. Experiment

5.1. Experimental Setup

We employ the proposed benchmark in Sec. 3 to conduct experiments. Our method is compared with KNAS [54], SFDA [42], and LogME [58]. KNAS is a gradient-based approach that operates under the assumption that gradients can predict downstream training performance. Therefore, we use it as a point of comparison with our efficient gradient-free approach. SFDA is not applicable for the single-class task of CrowdHuman [41]. Therefore, we present the SFDA results on five other multiple-class datasets. It is worth noting that our proposed pyramid feature matching enhances both LogME and SFDA, facilitating their evaluation. Furthermore, we present results for all three variants of our approach: U-LogME, IoU-LogME, and Det-LogME. We utilize two evaluation measures, namely the ranking correlation *Weighted Kendall*'s tau (τ_w), as defined in Eq. (1), and the Top-1 Relative Accuracy (Rel@1). Larger values of τ_w and Rel@1 indicate better assessment results.

5.2. Main Results

As discussed in [1], transferability metrics can be unstable. To address this issue, we randomly sub-sample 22 models from 33 pre-trained detectors and use 1% of the 33-choose-22 possible source model sets (over 1.9M) for measurement, as shown in Table 3. Our observations indicate that KNAS performs poorly on all 6 target datasets, with negative ranking correlation τ_w. Our method Det-LogME consistently outperforms LogME on 6 downstream tasks. For instance, Det-LogME surpasses LogME by substantial margins of 0.63, 0.48, and 0.34 in terms of ranking correlation τ_w on Pascal VOC, SODA, and DeepLesion, respectively. This validates the high effectiveness of our proposed unified evaluation and complementary IoU measurement. On five multiple-class downstream datasets, Det-LogME still outperforms classification-specific SFDA, particularly on DeepLesion (+0.63 τ_w). This further indicates that classification-specific transferability metrics may not yield optimal results for multi-class detection problems. Furthermore, our Det-LogME achieves better Rel@1 values than all other SOTA methods in average, outperforming the second-best method SFDA by 4% top-1 relative accuracy. Overall, our method is robust and consistently outperforms competing methods over 1.9M sub-sampled model sets.

Regarding the three variants proposed in this paper, U-LogME performs the worst in most cases. We have further observations to make. On one hand, IoU-LogME, which

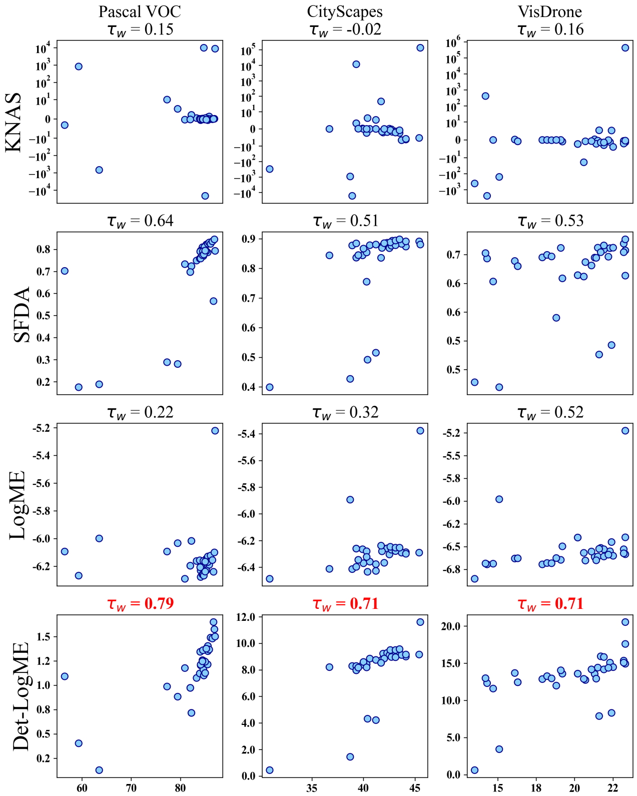

Figure 3. Comparison of ranking scores. The plots illustrate ground-truth fine-tuning performance $\{g_n\}_{n=1}^{N}$ (x-axis), ranking scores (y-axis), and *Weighted Kendall*'s coefficient τ_w for 33 pre-trained detectors on 3 out of 6 target datasets.

uses IoU as a scale-invariant metric, performs better than MSE-based U-LogME in scenarios where the objects vary greatly in scale. This is especially evident in CityScapes [9], SODA [19], and VisDrone [63], where objects captured under driving or UAV scenarios have diverse scales. On the other hand, IoU-based IoU-LogME suffers from the problem of remaining equal to 0 when there is no intersection between the predicted and ground truth bounding boxes, regardless of how far apart they are. In contrast, MSE-based U-LogME can handle this problem with absolute distances. This is particularly evident in DeepLesion [55], where the lesions are very small and difficult to detect. Det-LogME, which combines the advantages of both unified evaluation and IoU measurement, achieves a trade-off and better ranking performance than U-LogME and IoU-LogME. We can also observe that Det-LogME performs better on five multiple-class tasks than the single-class one. This indicates the significant difficulties and challenges involved in evaluating the pre-trained detectors on dense tasks.

Figure 3 illustrates the relationship between predicted transferability and actual fine-tuning performance of 33 pre-trained detectors for our Det-LogME and three SOTA methods on three target datasets. We observe that our Det-LogME consistently shows better positive correlations compared to other SOTA methods in all experiments, which demonstrates the effectiveness of our proposed approach.

Table 4. Effects of different components of Det-LogME.

Method	$\tau_w \uparrow$
Baseline (LogME)	0.22
w/ bbox center norm.	0.33
w/ joint eval. of coord.	0.40
w/ unified sub-task eval.	0.43
w/ IoU (Det-LogME)	**0.79**

Table 5. Effects of different bounding box normalization techniques (border and center) in Det-LogME.

Method	$\tau_w \uparrow$
Baseline (LogME)	0.22
w/ bbox border norm.	0.22
w/ bbox center norm.	**0.33**

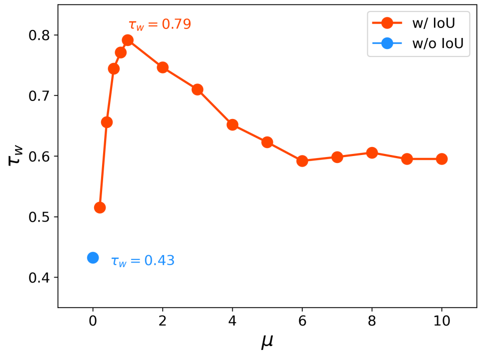

Figure 4. Effects of different weights μ for complementary IoU metric in Det-LogME. The first blue marker indicates Det-LogME without IoU measurement (degrades to U-LogME).

5.3. Ablation Studies

In this subsection, we carefully study our proposed Det-LogME with respect to different components and hyper-parameters by assessing 33 detectors on the Pascal VOC.

Components Analysis of Det-LogME. As shown in Table 4, we can learn that the large variances of different coordinates are eliminated by bounding box center normalization and the ranking correlation τ_w is improved 0.11. By jointly evaluating 4-d coordinates, τ_w improves from 0.33 to 0.40, indicating the inherent correlations among 4 coordinates within a bounding box are captured by our proposed joint evaluation. Under unified sub-task evaluation, τ_w further improves from 0.40 to 0.43, which demonstrates the effectiveness of unifying the supervision information by considering object classes information. Finally, we observe that the inclusion of the complementary IoU metric in our evaluation framework leads to a substantial increase in ranking performance from 0.43 to 0.79 ($+0.36$ τ_w). This finding underscores the importance of IoU measurement in assessing the transferability of pre-trained detectors.

Different BBox Normalization Techniques. To mitigate the detrimental effects of large variances among different coordinates, we introduce bounding box normalization. We experiment with two widely used normalization techniques in detection. Except for center normalization described in

Table 6. Efficiency evaluation of Det-LogME and comparison with brute-force fine-tuning, naive feature extraction, KNAS, SFDA, and LogME. Note that the wall-clock time and memory footprint for SFDA are evaluated on 5 multiple-class downstream detection datasets.

Measure	Wall-clock Time (s) ↓						Memory Footprint (GB) ↓					
Method	Fine-tune (upper bound)	Extract feature (lower bound)	KNAS	SFDA	LogME	Det-LogME	Fine-tune (upper bound)	Extract feature (lower bound)	KNAS	SFDA	LogME	Det-LogME
Pascal VOC	28421.09	128.88	129.99	130.14	130.18	129.89	13.39	0.47	13.86	0.55	0.59	0.66
CityScapes	73973.09	50.50	51.53	52.02	51.80	51.18	22.36	0.99	23.34	1.06	1.08	1.10
SODA	13928.24	50.75	51.89	51.88	51.83	51.19	12.54	0.50	13.04	0.54	0.55	0.55
CrowdHuman	43297.45	170.38	171.48	N/A	178.26	175.06	27.24	0.63	27.87	N/A	1.47	1.46
VisDrone	14067.15	63.38	64.51	69.08	68.16	64.97	16.69	0.65	17.34	0.79	0.80	0.84
DeepLesion	8465.45	37.13	38.24	37.49	37.71	37.34	6.99	0.52	7.51	0.53	0.54	0.54
Average	30358.75	83.50	84.60	68.12	86.32	84.94	16.53	0.63	17.16	0.69	0.84	0.86

Section 4.2, we also try to apply border normalization by dividing the bounding box with the corresponding width and height of the input image to obtain $\boldsymbol{b}_i^{bor} = (x'_1, y'_1, x'_2, y'_2)$. From the results shown in Table 5, we observe that border normalization yields no improvement, and center normalization outperforms it by $+0.1$ τ_w. Center normalization scales the four coordinates of the bounding box to the same range in $[0, 1]$, while border normalization may not work well when significant biases exist among the x-axis and y-axis coordinates, such as when an object is located near the top-right corner of an image where $x_1 \gg y_1$ or $x_2 \gg y_2$.

Weight of Complementary IoU Metric. The weight hyper-parameter μ in Eq. (7) controls the behavior of the complementary IoU metric. Similar to the weights of IoU-related losses [39, 62], the weight μ is crucial to the final transferability score of Det-LogME. We investigate the effects of different weights $\mu \in \{0, 0.2, \dots, 0.8, 1, 2, \dots, 10\}$ of the IoU metric, and the results are presented in Figure 4. We observe that incorporating the IoU metric as a complementary term consistently improves the ranking performance of Det-LogME compared to not including it (the first blue marker, which degrades to U-LogME, with $\tau_w = 0.43$). The best ranking performance of Det-LogME is achieved when the weight is 1, resulting in $\tau_w = 0.79$.

5.4. Efficiency Analysis

In this subsection, we present a comprehensive evaluation of the assessing efficiency of Det-LogME and compare it with brute-force fine-tuning, naive feature extraction, KNAS, SFDA, and LogME from two perspectives: 1) wall-clock time: the average time of fine-tuning (12 epochs) or evaluating (including feature extraction) all 33 pre-trained detectors; 2) memory footprint: the maximum memory required during fine-tuning or evaluating (including feature extraction, the loading of all visual features, and computing transferability metric) all 33 pre-trained detectors. The efficiency of classification-specific SFDA is studied on 5 multiple-class tasks. The results are presented in Table 6.

Wall-clock Time. Over 6 downstream tasks, KNAS has the fastest speed, but according to Table 3, it performs the worst, which is not acceptable in practice. On the other hand, with joint evaluation of bounding box coordinates, Det-LogME runs faster than LogME because Det-LogME does not need to evaluate all 4-d coordinates in a loop. Moreover, Det-LogME runs faster than SFDA, which includes fine-tuning dynamics on all 5 multiple-class datasets. Considering that the selected model will be fine-tuned on the target task, our proposed Det-LogME brings about $(\text{NumSourceModels} - 1)\times$ speedup compared with brute-force fine-tuning, which is 32× in this work.

Memory Footprint. Table 6 indicates that gradient-based KNAS demands a significant amount of memory exceeding 17 GB, rendering it infeasible for most practical applications. In comparison, Det-LogME only marginally increases the memory when compared to LogME and SFDA, making it practical for deployment. Furthermore, Det-LogME demonstrates high memory efficiency by requiring 19× less memory than brute-force fine-tuning, underscoring its potential to save computational resources.

6. Conclusion

In this paper, we aim to address a practical but under-explored problem of efficient transferability assessment for pre-trained detection models. To achieve this, we establish a challenging detector transferability benchmark comprising a large and diverse zoo consisting of 33 detectors with various architectures, source datasets and schemes. Upon this zoo, we adopt 6 downstream tasks spanning 5 diverse domains for evaluation. Further, we propose a simple yet effective framework for assessing the transferability of pre-trained detectors. Extensive experimental results demonstrate the high effectiveness and efficiency of our approach compared with other state-of-the-art methods across a wide range of pre-trained detectors and downstream tasks, notably outperforming brute-force fine-tuning in terms of computational efficiency. We hope our work can inspire further research into the selection of pre-trained models, particularly those with multi-scale and multi-task capabilities.

Acknowledgements. We gratefully acknowledge the support of Mindspore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this work.

References

[1] Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 303–321. Springer, 2022. 6

[2] Andrea Agostinelli, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. Transferability metrics for selecting source model ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7936–7946, 2022. 1, 2, 3

[3] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14605–14615, 2022. 2, 3

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1, 2, 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[7] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020. 1

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 3, 7

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[11] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 1, 2, 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 3, 5

[13] Nan Ding, Xi Chen, Tomer Levinboim, Beer Changpinyo, and Radu Soricut. Pactran: Pac-bayesian metrics for estimating the transferability of pretrained models to classification tasks. *European Conference on Computer Vision*, 2022. 1, 2, 3

[14] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396, 2019. 2

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 3

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1

[17] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4805–4814, 2019. 1

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2

[19] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, et al. Soda10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1, 3, 7

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[23] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 3

[24] Kevin H Knuth, Michael Habeck, Nabin K Malakar, Asim M Mubeen, and Ben Placek. Bayesian evidence and model selection. *Digital Signal Processing*, 47:50–67, 2015. 4

[25] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 4

[26] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 2661–2671, 2019. 1, 2

[27] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2

[28] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1, 2, 3

[29] Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. Ranking neural checkpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2663–2673, 2021. 1, 2, 3

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 2, 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 3

[33] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[34] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305. PMLR, 2020. 1, 2

[35] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 3

[36] Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9172–9182, 2022. 1, 2, 3

[37] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 1, 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2, 3, 4, 5

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 8

[40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[41] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1, 3, 6

[42] 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. *European Conference on Computer Vision*, 2022. 1, 2, 3, 6

[43] Jie Song, Yixin Chen, Xinchao Wang, Chengchao Shen, and Mingli Song. Deep model transferability from attribution maps. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[44] Jie Song, Yixin Chen, Jingwen Ye, Xinchao Wang, Chengchao Shen, Feng Mao, and Mingli Song. Depara: Deep attribution graph for deep knowledge transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3922–3930, 2020. 2

[45] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 1, 2, 3

[46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[47] Yang Tan, Yang Li, and Shao-Lun Huang. Otce: A transferability metric for cross-domain cross-task representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15779–15788, June 2021. 2

[48] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2, 3, 5

[49] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1395–1405, 2019. 2

[50] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1

[51] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level con-

trastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 2, 3

[52] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 1

[53] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1, 3

[54] Jingjing Xu, Liang Zhao, Junyang Lin, Rundong Gao, Xu Sun, and Hongxia Yang. Knas: green neural architecture search. In *International Conference on Machine Learning*, pages 11613–11625. PMLR, 2021. 6

[55] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018. 1, 3, 7

[56] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021. 2, 3

[57] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 1, 2

[58] 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*, pages 12133–12143. PMLR, 2021. 1, 2, 3, 4, 6

[59] Kaichao You, Yong Liu, Ziyang Zhang, Jianmin Wang, Michael I Jordan, and Mingsheng Long. Ranking and tuning pre-trained models: A new paradigm for exploiting model hubs. *Journal of Machine Learning Research*, 23:1–47, 2022. 1, 2, 3

[60] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018. 2

[61] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In *European conference on computer vision*, pages 260–275. Springer, 2020. 2, 3

[62] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020. 8

[63] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 1, 3, 7

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2, 3, 5