Supplementary Material
Transferability Estimation using Bhattacharyya Class Separability

Michal Pándy* Andrea Agostinelli Jasper Uijlings Vittorio Ferrari Thomas Mensink

Google Research

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

In this supplementary material, we discuss the main limitations of our work with possibilities for extensions, analyze some of GBC's covariance regularization approaches, and provide further detail on the datasets used in the semantic segmentation experiments. We also provide additional experimental insights by visualizing the scatter plots across all of the experiments from Section 4.2, and by evaluating GBC using Pearson's correlation and the un-weighted Kendall's rank correlation in classification settings.

1. Limitations and Future work

Here we reflect on some of the key limitations of the proposed GBC method.

Network Architectures. Firstly, the selected source architectures play a key role in evaluating the proposed method. It is plausible that different architectures yield different class distributions in the embedding space, which could impact the per-class Gaussian approximation of GBC. Further, all the architectures are sensitive to training hyper-parameter choices, which introduces further complications in estimating ground truth transferability scores. To allow for fair comparison, we have used the same network architectures as in [23] as much as possible, and verified that our networks are trained until convergence. However, a different set of used architectures may influence the results.

Classification networks. GBC measures pairwise class overlap and hence is suitable for transfer in a classification setting. However, many interesting transfer learning problems are in regression, or even unsupervised learning and reinforcement learning. In terms of regression, it would be useful to extend GBC, for example by binning the regression variable and replace the *class* overlap with an overlap between the used bins. We believe all of these directions present promising avenues for future work.

*Currently at Waymo.

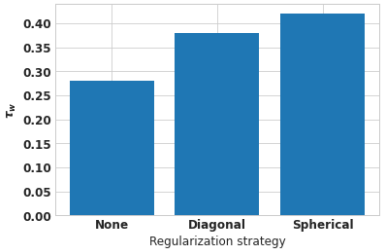

Figure 1. This figure demonstrates the superiority of spherical regularization with respect to other strategies in terms of weighted Kendall's rank correlation.

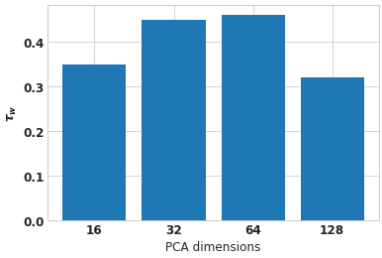

Figure 2. This figure demonstrates the sensitivity of GBC to PCA dimensionality on Cifar10 in terms of weighted Kendall's rank correlation.

2. Empirical analysis of design choices

In this experiment, we evaluate the influence of GBC's design choices introduced in our method section. We use CIFAR10 as a target dataset and transfer from the 9 source architectures pre-trained on ImageNet as described in our experiments.

Effects of regularization strategies We compare three Gaussian covariance regularization strategies: none, diagonal, and spherical. As we can observe in Figure 1, adding regularization improves our results, with the best τ_w in the case of spherical regularization.

Effects of PCA dimensions In this experiment, we compare the performance of GBC across multiple PCA dimensions: 16, 32, 64, and 128. We discover performance improvements up to 64 dimensions (see Figure 2), after which we observe a decline in τ_w. Hence, it can be beneficial to carefully select the appropriate PCA dimensions for the given use case.

3. Detailed Results for Semantic Segmentation

Here we provide additional results for semantic segmentation. In particular Table 1 shows per-target results for the source selection task (ranking source datasets for a particular fixed target dataset), while Table 2 shows per-source results for the target selection task (ranking target datasets for a fixed source dataset). For each table we provide the Weighted Kendall Tau correlation metric for every dataset, and we indicate whether a transferability metric correctly predicts the top-1 source (in source selection) or the top-1 target (in target selection).

4. Additional Results for Image Classification

In this section we provide additional experimental results for the image classification experiments.

- For the source selection experiments (Section 4.1 / Table 1 in the main paper), we include different correlation metrics: Weighted Kendall Tau (Table 3a, also in the main paper), Kendall Tau (Table 3b), and Pearson's r coefficient (Table 3c).
- For the dataset transferability experiments (Section 4.2 / Table 2 in the main paper), we include also the Weighted Kendall Tau (Table 4a, used in the main paper), Kendall Tau (Table 4b), and Pearson's r coefficient (Table 4c).
- Moreover, we provide scatter plots for all target datasets used in Figure 3, extending Figure 4 in the main paper.

Target Dataset	τ_w				Top-1			
	IDS [13]	**LEEP** [16]	**LogME** [23]	**GBC** *Ours*	**IDS** [13]	**LEEP** [16]	**LogME** [23]	**GBC** *Ours*
Pascal Context [14]	0.48	0.64	**0.82**	0.73	-	✓	✓	✓
Pascal VOC [8]	0.25	0.29	**0.52**	0.44	-	-	✓	✓
ADE20K [25]	0.21	0.42	**0.43**	0.41	✓	✓	✓	✓
COCO [5, 11, 12]	-0.12	-0.15	**0.20**	-0.14	-	-	-	-
KITTI [1]	0.64	0.77	0.82	**0.84**	-	-	-	-
CamVid [3]	0.62	**0.76**	0.75	0.73	✓	✓	✓	✓
CityScapes [6]	0.62	0.88	0.92	**0.93**	-	✓	✓	✓
IDD [19]	0.57	0.80	0.87	**0.90**	✓	✓	✓	✓
BDD [24]	0.77	0.85	**0.93**	0.90	✓	✓	✓	✓
MVD [15]	0.58	0.65	**0.72**	0.66	-	-	-	-
ISPRS [17]	0.26	0.08	0.52	**0.66**	✓	-	-	✓
iSAID [20, 22]	0.35	**0.36**	0.13	0.27	-	-	✓	✓
SUN RGB-D [18]	0.53	0.52	**0.61**	0.57	✓	-	-	-
ScanNet [7]	0.43	**0.71**	0.65	0.61	✓	-	-	-
SUIM [10]	0.39	0.27	**0.64**	0.58	-	✓	✓	✓
vKITTI2 [4, 9]	0.70	0.65	**0.73**	0.64	-	✓	✓	✓
vGallery [21]	0.40	0.44	**0.50**	0.27	-	-	-	-
Average	0.45	0.53	**0.63**	0.59	0.41	0.47	0.59	**0.65**

Table 1. Per-target results for segmentation source selection.

Source Dataset	τ_w				Top-1			
	IDS [13]	**LEEP** [16]	**LogME** [23]	**GBC** *Ours*	**IDS** [13]	**LEEP** [16]	**LogME** [23]	**GBC** *Ours*
Pascal Context [14]	0.53	**0.78**	0.01	0.72	✓	-	-	-
Pascal VOC [8]	0.34	0.52	0.06	**0.69**	✓	-	-	✓
ADE20K [25]	0.48	**0.74**	0.01	0.69	-	✓	-	✓
COCO [5, 11, 12]	0.12	0.66	0.15	**0.73**	-	-	-	-
KITTI [1]	0.48	0.60	0.09	**0.70**	-	-	-	✓
CamVid [3]	0.61	0.57	0.01	**0.72**	-	-	-	✓
CityScapes [6]	0.38	0.57	0.17	**0.74**	-	-	-	✓
IDD [19]	0.48	0.59	0.24	**0.70**	-	-	-	✓
BDD [24]	0.55	0.69	0.12	**0.74**	✓	-	-	✓
MVD [15]	0.58	0.63	0.21	**0.68**	-	-	-	-
ISPRS [17]	-0.10	0.59	0.09	**0.69**	-	-	-	✓
iSAID [20, 22]	0.30	0.73	0.02	**0.79**	✓	✓	-	✓
SUN RGB-D [18]	0.40	0.54	0.06	**0.62**	✓	✓	-	✓
ScanNet [7]	0.40	0.59	0.05	**0.60**	✓	✓	-	✓
SUIM [10]	0.32	0.61	0.11	**0.72**	-	-	-	✓
vKITTI2 [4, 9]	0.61	0.62	0.17	**0.71**	✓	-	-	✓
vGallery [21]	-0.01	**0.52**	-0.19	0.51	-	-	-	-
Average	0.36	0.62	0.08	**0.69**	0.41	0.24	0.00	**0.76**

Table 2. Per-source results for segmentation target selection.

	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**

(a) Metric: Weighted Kendall Tau (τ_w)

	Pets	Imagenette	CIFAR-10	CUB'11	Dogs	Flowers102	SUN	CIFAR-100	**Average**
LogME	-0.06	**0.56**	0.44	**0.28**	0.17	0.0	-0.17	0.50	0.22
H-score	0.17	0.54	**0.56**	**0.28**	0.06	0.13	0.0	0.50	0.28
LEEP	**0.5**	0.5	**0.56**	**0.28**	0.39	-0.28	**0.28**	**0.56**	0.35
GBC	0.44	0.39	0.50	0.22	**0.67**	**0.17**	0.17	0.39	**0.37**

(b) Metric: Kendall Tau (τ)

	Pets	Imagenette	CIFAR-10	CUB'11	Dogs	Flowers102	SUN	CIFAR-100	**Average**
LogME	0.33	**0.58**	0.62	0.45	0.28	0.37	-0.14	0.54	0.38
H-score	0.37	0.52	**0.83**	0.35	0.25	-0.0	-0.01	0.78	0.39
LEEP	0.28	0.12	0.81	0.03	0.48	-0.2	0.21	**0.75**	0.31
GBC	**0.40**	0.45	0.81	**0.49**	**0.44**	**0.57**	**0.27**	**0.77**	**0.52**

(c) Metric: Pearson correlation coefficient (ρ)

Table 3. Overview of results for transferability for source selection in image classification.

Source	**LEEP** [16]	**LogME** [23]	**H-score** [2]	**GBC** *Ours*
		CIFAR-10		
CUB	0.68	0.71	0.69	**0.72**
C-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	**0.90**	0.29	0.59	**0.90**
C-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	**0.72**	0.71	0.71	0.71
C-10	0.72	**0.73**	0.69	0.69
C-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		
C-10	**0.87**	-0.59	0.83	0.86
C-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	**0.95**	0.87	0.54	**0.95**
C-10	**0.95**	0.87	0.12	**0.95**
C-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**

(a) Weighted Kendall Tau (τ_w)

LEEP [16]	**LogME** [23]	**H-score** [2]	**GBC** *Ours*
	CIFAR-10		
0.51	0.53	0.54	**0.55**
0.58	0.57	0.56	0.55
0.49	0.48	**0.50**	0.49
0.53	0.55	**0.56**	**0.56**
0.53	**0.53**	0.50	0.49
	CIFAR-100		
0.84	0.74	0.65	**0.84**
0.85	0.74	0.73	**0.85**
0.81	0.69	0.66	**0.81**
0.83	0.72	0.74	0.82
0.86	0.74	0.77	**0.86**
	Fashion-MNIST		
0.48	0.46	**0.49**	0.48
0.47	**0.49**	0.46	0.46
0.47	**0.48**	0.46	0.46
0.49	0.47	0.47	0.47
0.49	0.48	0.47	**0.50**
	Caltech-USCD Birds 2011		
0.79	-0.04	0.69	**0.79**
0.79	-0.02	0.67	**0.80**
0.33	-0.03	0.29	**0.33**
0.79	-0.04	0.67	0.78
0.82	-0.02	0.65	**0.82**
	SUN-397		
0.92	0.91	0.27	**0.92**
0.92	0.90	0.22	0.91
0.90	0.89	0.28	**0.90**
0.91	0.90	0.26	**0.91**
0.92	0.91	0.26	**0.92**
0.69	0.52	0.51	**0.69**

(b) Kendall Tau (τ)

LEEP [16]	**LogME** [23]	**H-score** [2]	**GBC** *Ours*
	CIFAR-10		
0.69	**0.70**	0.65	**0.70**
0.75	**0.77**	0.70	**0.77**
0.63	0.65	0.58	**0.66**
0.71	**0.73**	0.66	0.72
0.69	**0.71**	0.64	0.70
	CIFAR-100		
0.94	0.29	0.56	0.87
0.95	0.29	0.61	0.86
0.92	0.21	-0.29	0.85
0.95	0.19	0.56	0.87
0.95	0.25	0.58	0.82
	Fashion-MNIST		
0.63	**0.64**	0.60	0.61
0.61	**0.62**	0.57	0.59
0.61	**0.62**	0.56	0.59
0.63	**0.63**	0.59	0.60
0.62	**0.63**	0.59	0.60
	Caltech-USCD Birds 2011		
0.94	-0.75	0.87	0.77
0.94	-0.77	0.81	0.76
0.66	-0.68	0.24	0.43
0.95	-0.72	0.56	0.83
0.94	-0.77	0.76	0.75
	SUN-397		
0.92	0.91	0.27	**0.92**
0.92	0.90	0.22	0.91
0.90	0.89	0.28	**0.90**
0.91	0.90	0.26	**0.91**
0.92	0.91	0.26	**0.92**
0.82	0.36	0.54	0.75

(c) Pearson correlation coefficient (ρ)

Table 4. Overview of results for transferability for target dataset transferability.

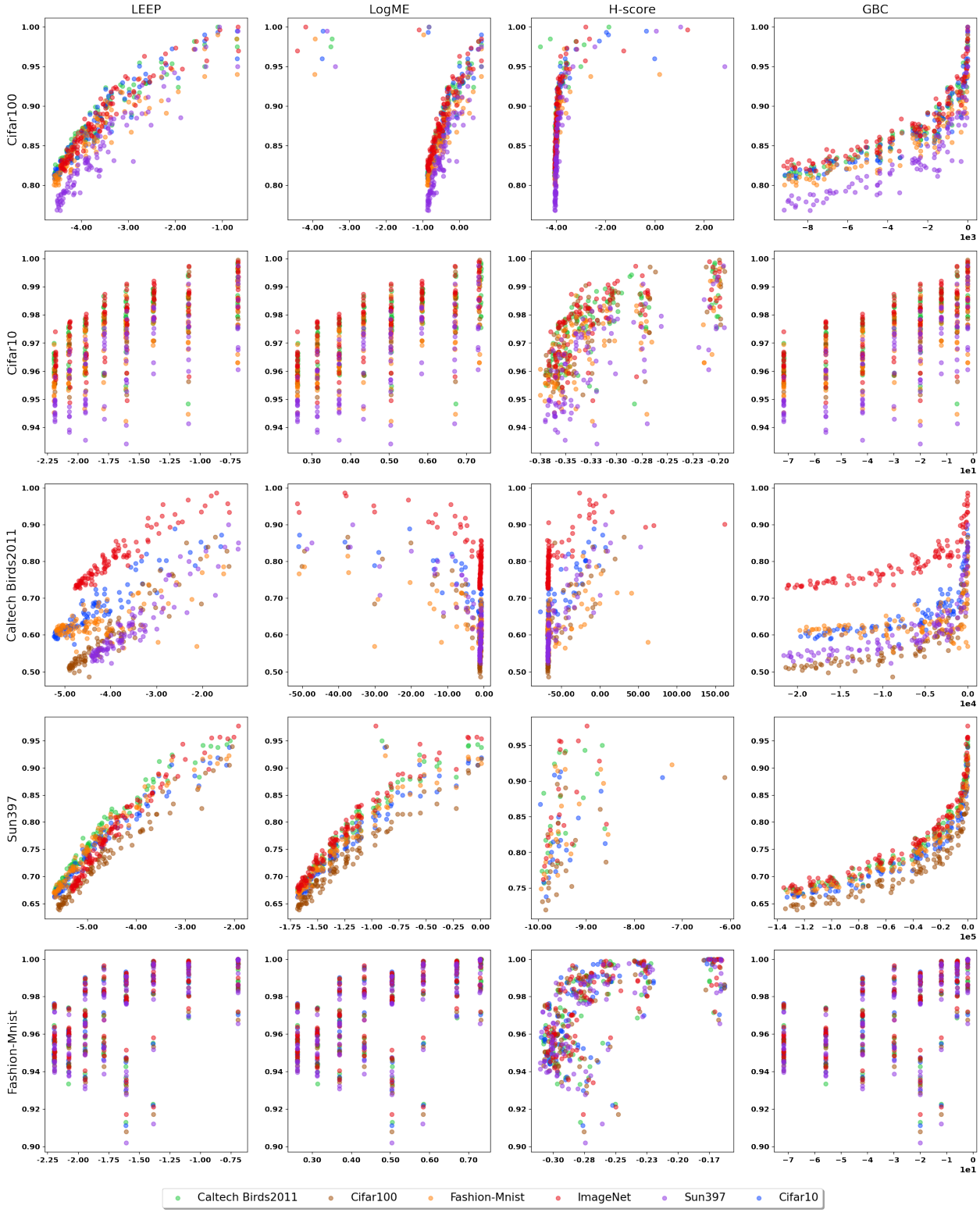

Figure 3. This figure illustrates the scatter plots of LEEP, LogME, H-score, and GBC for all datasets used in the dataset transferability experiment (see Section 4.2 and Fig 4 of the main paper). In each plane, the transferability score $\mathcal{S}_{s\to t}$ of the method is on the X-axis, with the 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 some of the target datasets, both GBC and LEEP showcase increasing trends.

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. 3

[2] 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. 4

[3] 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. 3

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv*, 2020. 3

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. 3

[6] 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. 3

[7] 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. 3

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012. 3

[9] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 3

[10] 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. 3

[11] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *CVPR*, 2019. 3

[12] 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. 3

[13] 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. 3

[14] 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. 3

[15] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 3

[16] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *ICML*, 2020. 3, 4

[17] 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. 3

[18] S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. 3

[19] 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. 3

[20] 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. 3

[21] Philippe Weinzaepfel, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. Visual localization by learning objects-of-interest dense match regression. In *CVPR*, 2019. 3

[22] 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. 3

[23] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021. 1, 3, 4

[24] 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. 3

[25] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 3