APPENDIX
Building a Winning Team: Selecting Source Model Ensembles using a Submodular Transferability Estimation Approach

Vimal K B*[1], Saketh Bachu*[1,3], Tanmay Garg[1], Niveditha Lakshmi Narasimhan[2], Raghavan Konuru[2], and Vineeth N Balasubramanian[1]

[1]Indian Institute of Technology, Hyderabad [2]KLA [3]University of California, Riverside
* Equal Contribution. Corresponding author: vimalkb96@gmail.com

In this appendix, we provide additional details which we could not include in the main paper due to space constraints, including additional results, details and analysis that provide more insights into the proposed method. In particular, We discuss the following:

Table of Contents

A1. Comparison against OTCE

In this section, we compare OSBORN with the OTCE metric. OTCE is limited by its ability to estimate transferability for a single source model; however, we naively add the OTCE scores of the individual models present in the ensemble to make it a multi-source variant. The results in terms of various correlations are shown in Tab. A1. OSBORN outperforms OTCE by 131.76% in terms of WKT, 235.59% in terms of KT and 513.33% in terms of PCC.

Target Dataset	Weighted Kendall's τ		Kendall's τ		Pearson	
	OTCE	Ours	OTCE	Ours	OTCE	Ours
Oxford102Flowers	0.406	**0.616**	0.118	**0.400**	0.086	**0.456**
OxfordIIITPets	0.186	**0.558**	0.075	**0.453**	0.109	**0.666**
StanfordDogs	0.093	**0.477**	0.05	**0.427**	0.088	**0.604**
Caltech101	0.179	**0.565**	0.223	**0.335**	0.068	**0.486**
StanfordCars	0.300	**0.486**	0.123	**0.368**	0.100	**0.549**
Average	0.233	**0.540**	0.118	**0.396**	0.090	**0.552**

Table A1: OTCE vs OSBORN (Ours)

A2. Modified Baselines

In this section, we understand the effect of adding the model cohesion term W_C to our baselines i.e. MS-LEEP and E-LEEP. Table A2 shows the results. While it expectedly improves correlations of these baselines (further corroborating the usefulness of our proposed cohesiveness term), OSBORN still achieves higher correlations than these modified baselines.

A3. Additional Experiments

In this section, we present the results of additional experiments we conducted on tasks like multi-domain/domain adaptation and semantic segmentation. We could not include details about these in the main paper due to space constraints. We start by describing the datasets used, models trained and then report the performance of OSBORN and other baselines on these tasks.

Multi-domain/Domain Adaptation Dataset: DomainNet. We use the DomainNet [A14] dataset to test OSBORN in a challenging multi-domain source pool setting. DomainNet consists of 6 domains (styles) namely, Clipart (C), Infograph (I), Painting (P), Quickdraw (Q), Real (R) and Sketch (S), each covering 345 common object categories. Out of

Target Dataset	**Weighted Kendall's τ**				**Kendall's τ**				**Pearson**			
	MS	**E**	W_C **+ MS**	W_C **+ E**	**MS**	**E**	W_C **+ MS**	W_C **+ E**	**MS**	**E**	W_C **+ MS**	W_C **+ E**
Oxford102Flowers	0.086	-0.019	0.413	**0.459**	0.138	0.0739	0.315	**0.330**	0.23	0.164	**0.401**	0.385
OxfordIIITPets	0.414	0.393	**0.540**	0.522	0.346	0.326	0.473	**0.475**	0.504	0.5	0.666	**0.676**
Caltech101	**0.435**	0.409	0.314	0.385	0.240	0.231	**0.242**	0.236	0.353	0.341	0.315	**0.354**
StanfordDogs	0.326	-0.472	0.348	**0.384**	0.244	-0.236	0.269	**0.326**	0.398	-0.154	0.496	**0.571**
StanfordCars	0.115	0.018	0.066	**0.147**	0.137	0.071	0.144	**0.185**	0.256	0.163	0.360	**0.434**
Average	0.275	0.097	0.265	**0.301**	0.221	0.110	0.246	**0.259**	0.348	0.222	0.383	**0.407**

Table A2: Comparison of baselines and modified baselines. Note: MS: MS-LEEP, E: E-LEEP, W_C: Model Cohesion term

Target Domain	**Weighted Kendall's τ**			**Kendall's τ**			**Pearson**		
	MS	**E**	**Ours**	**MS**	**E**	**Ours**	**MS**	**E**	**Ours**
Real	0.057	0.026	**0.576**	0.016	-0.011	**0.415**	0.010	-0.033	**0.518**
Infograph	0.165	0.163	**0.298**	0.046	0.048	**0.230**	0.076	0.057	**0.308**
Clipart	0.003	-0.076	**0.040**	0.115	0.078	**0.161**	**0.248**	0.193	0.179
Average	0.075	0.038	**0.305**	0.059	0.038	**0.269**	0.111	0.072	**0.335**

Table A3: Comparison of different ensemble transferability estimation metrics for classification tasks on the DomainNet dataset. Averaged across 3 domains, OSBORN achieves the best results under all the correlation values. MS: MS-LEEP, and E: E-LEEP.

these 6 domains, we evaluate the performance of OSBORN on 3 domains, that are Real (R), Infograph (I) and Clipart (C).

Semantic Segmentation Datasets. For conducting experiments on the semantic segmentation tasks, we choose 10 popularly used segmentation datasets, Pascal Context [A12], Pascal VOC [A4], COCO [A11], CamVid [A2], CityScapes [A3], India Driving Dataset (IDD) [A18], Berkeley Deep Drive (BDD) [A19], Mapillary Vistas [A13], SUIM [A9], and SUN RGB-D [A17]. Out of these 10 datasets, we evaluate and compare the performance of OSBORN with baselines on 3 target datasets, namely Camvid [A2], CityScapes [A3], and SUIM [A9].

Model Architectures (DomainNet). For building the source pool for the multi-domain experiments, we use the same models as we used in the fully-supervised pre-training setting i.e ResNet-101 [A6] and DenseNet-201 [A8]. Initially, both models are initialized with the fully-supervised ImageNet weights [A10], and we then train them on 6 domains of the DomainNet dataset.

Model Architectures (Semantic Segmentation). For semantic segmentation, we employ a FCN [A16] with ResNet-101 [A6] backbone, and a Lite R-ASPP with MobileNetv3 backbone [A7] as our source model architectures. The capacity of the former is much higher than the latter thus bringing in diversity. We initialize these models with the COCO pre-trained weights [A11] and then train them on the 10 datasets to include them in our source pool [1].The rest of the experimental setup is the same as in Section 5 of the main paper.

Results on DomainNet. We compare OSBORN with the baseline metrics, i.e. MS-LEEP and E-LEEP, in terms of WKT, KT, and PCC. The correlation values are reported in Tab. A3, averaged across three target domains.

Results on Semantic Segmentation. Apart from MS-LEEP and E-LEEP, the paper [A1] also proposes two additional metrics for predicting transferability on semantic segmentation tasks, which are namely IoU-EEP and SoftIoU-EEP. In this section, we compare the performance of OSBORN with these two metrics as well. We present the experimental results for the semantic segmentation tasks in Tab. A4. As seen in the table, OSBORN improves transferability estimation when compared to previous works.

A4. Implementation Details

Here, we describe miscellaneous details pertaining to the experiments reported in Section 5 of the main paper.

Optimal Transport Computation. We use the Python Optimal Transport Library (POT) to conduct our experiments. To keep the computational cost in check, we use a stratified representative set of 5000 samples from the train sets to calculate the Wasserstein distance (since it involves extracting the source and target latent). This makes our method tractable and practical. We perform stratified sampling to follow a class-balanced approach, i.e. we sample the images inversely proportional to their class frequencies in the

[1] Our baselines MS-LEEP and E-LEEP use custom proprietary model architectures that are not publicly available. We hence followed the authors' code and obtained guidelines from them in using their method on the models used in our work, and picked the best-performing hyperparameters for the results corresponding to their baselines shown in this work.

Target Dataset	Weighted Kendall's τ					Kendall's τ					Pearson				
	MS	**E**	**IoU**	**sIoU**	**Ours**	**MS**	**E**	**IoU**	**sIoU**	**Ours**	**MS**	**E**	**IoU**	**sIoU**	**Ours**
Camvid	-0.173	-0.279	0.175	-0.074	**0.190**	-0.006	-0.108	0.030	-0.050	**0.114**	0.088	-0.050	0.071	-0.024	**0.091**
Cityscapes	-0.356	-0.390	-0.306	-0.153	**0.056**	-0.166	-0.188	-0.115	-0.090	**0.108**	-0.263	-0.241	-0.191	-0.154	**0.216**
SUIM	0.052	0.051	0.191	0.097	**0.237**	-0.014	-0.016	**0.084**	0.075	0.078	-0.024	-0.028	**0.230**	0.164	0.112
Average	-0.159	-0.053	0.020	-0.043	**0.161**	-0.062	-0.104	0.0003	-0.022	**0.1**	-0.066	-0.106	0.037	-0.005	**0.140**

Table A4: Comparison of different ensemble transferability estimation metrics for semantic segmentation tasks. On average, we beat all the previously proposed methods for estimating transferability for semantic segmentation in terms of correlations. Note, MS: MS-LEEP, E: E-LEEP, IoU: IoU-EEP, sIoU: SoftIoU-EEP.

Target Dataset	Weighted Kendall's τ		Kendall's τ		Pearson	
	Standard	**Frobenius**	**Standard**	**Frobenius**	**Standard**	**Frobenius**
Oxford102Flowers	**0.616**	0.614	**0.400**	0.390	**0.456**	0.463
OxfordIIITPets	**0.558**	0.539	**0.453**	0.446	**0.666**	0.660
Caltech101	**0.565**	0.557	**0.335**	0.329	**0.486**	0.483
StanfordDogs	0.477	**0.581**	0.427	**0.508**	0.604	**0.628**
StanfordCars	**0.486**	0.445	**0.368**	0.361	**0.549**	0.544
Average	0.540	**0.547**	0.397	**0.407**	0.552	**0.556**

Table A5: In this table, we report the change in correlations obtained using a Frobenius norm based regularizer rather than a standard (non-regularized) method for the fully-supervised pre-trained models (classification tasks).

train set. Also, we standardize all three terms in OSBORN to avoid the dominance of any term on the others.

Input Data. In the case of classification tasks, we resize the input images to 224 × 224, and in the case of semantic segmentation, we resize them to 256 × 256 (for computational feasibility). Since semantic segmentation is a dense prediction task with a high computational cost, we follow the strategy mentioned in [A1] and sample 1000 pixels from an image. Considering class imbalances in semantic segmentation datasets, we sample pixels inversely proportionally to the frequency of their class categories in the target dataset, similar to what MS-LEEP performed in their experiments.

A5. Visualization of Results

In Fig A2, we show t-SNE plots for data points of different classes in StanfordCars when passed through ensembles selected using various methods. We see that the ensemble selected by our method is better at segregating classes and closer to the Optimal as compared to MS-LEEP.

A6. Results with Frobenius Norm Regularizer

As mentioned in Section 3 of the paper, there is an option to use a regularizer to solve the OT problem. In this section, we investigate the usage of a Frobenius norm regularizer [A15],[A5] in the experiments for image classification tasks (both fully-supervised and self-supervised pre-training settings). In Tab. A5, we show the results of OSBORN with the use of a Frobenius norm regularizer (column: Frobenius) and without any regularizer (column: Standard) for the fully-supervised pre-training setting. We observe that both variations give comparable results on an average. In Tab. A6, we report the results for a self-supervised pre-training setting. In contrast to Tab. A5, we observe that a Frobenius norm regularizer improves the performance substantially in this case. We hypothesize that self-supervised pre-training may make a model more conducive to the source datasets, which a Frobenius norm regularizer offsets while performing optimal transport computations by making them much easier and structured.

A7. Weighted version of OSBORN

While our results in the main paper showed that OSBORN outperforms existing state-of-the-art as is in its simple form, we conducted additional experiments to study the influence of weighting each component of OSBORN. Our studies showed that this can vary for different target datasets. Fig. A1 shows these results for the Oxford102Flowers dataset. For target datasets such as OxfordIIITPets and Oxford102Flowers, we observe that when we give more weightage to W_D and subsequently to W_T, as compared to W_C, we achieve higher correlations. We believe this is because these datasets have some fine-grained characteristics in each class, which need more attention for classification. We believe that such a trend holds for transfer from coarse-grained to fine-grained datasets in general, while we observed a higher weightage to W_T to provide more favorable results in other settings. As stated earlier, while not using any weighted coefficients for the terms in OSBORN is by itself beneficial, carefully picking weights

Target Dataset	Weighted Kendall's τ		Kendall's τ		Pearson	
	Standard	Frobenius	Standard	Frobenius	Standard	Frobenius
Oxford102Flowers	0.492	**0.549**	0.293	**0.336**	0.272	**0.306**
OxfordIIITPets	0.316	**0.357**	0.123	**0.139**	0.193	**0.232**
StanfordDogs	0.140	**0.170**	0.074	**0.110**	0.210	**0.236**
Caltech101	0.484	**0.488**	0.279	**0.308**	0.345	**0.374**
StanfordCars	0.207	**0.260**	0.100	**0.139**	0.198	**0.232**
Average	0.328	**0.365**	0.174	**0.206**	0.244	**0.276**

Table A6: In this table, we understand the difference in correlations obtained using a Frobenius norm-based regularizer rather than a standard (non-regularized) method for the self-supervised pre-trained models (classification tasks).

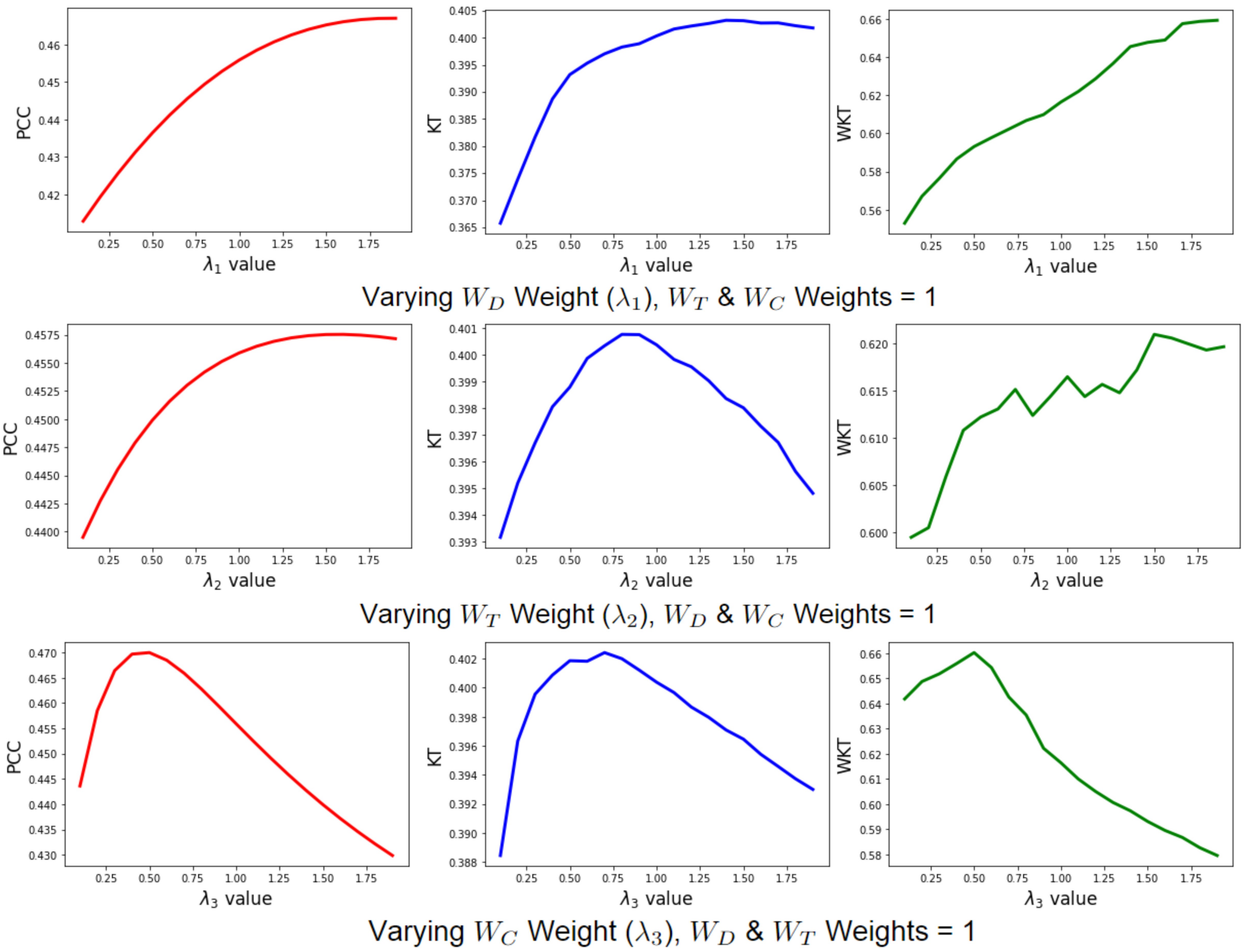

Figure A1: Relation between weighted coefficient values for terms in OSBORN and corresponding correlation scores for Oxford102Flower

for a specific target dataset can further improve performance. Learning these weighting coefficients would be an interesting direction for future work.

A8. Balancing Three Components of OSBORN

To study further on importance of each component of OSBORN, we conducted experiments by completely removing one of the terms and reporting the resulting correlations/results in Table A7. The analysis demonstrates, inter-

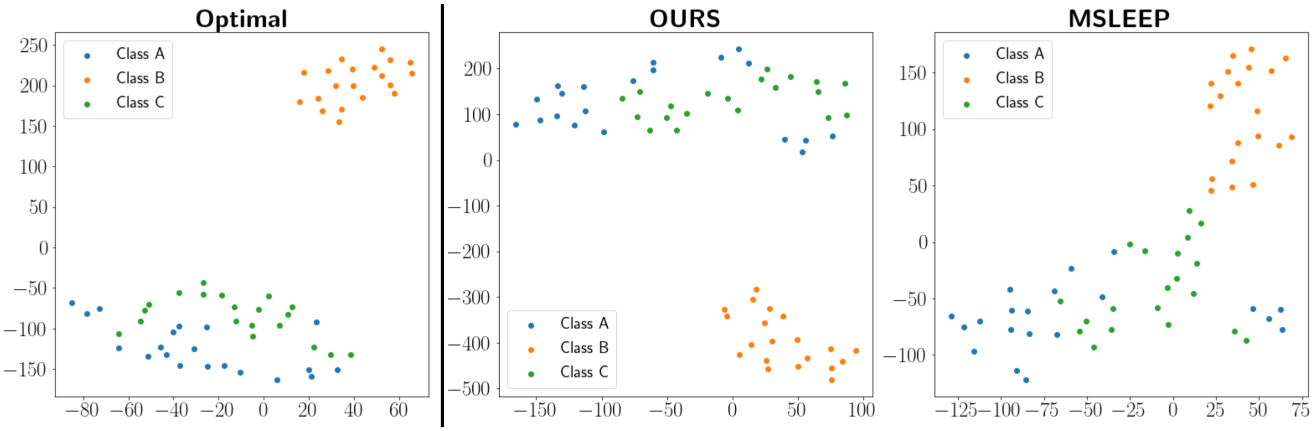

Figure A2: t-SNE plots of features learned by corresponding method's ensembles on StanfordCars dataset. '*Optimal*' chooses best ensemble with exhaustive search

Target Dataset	$W_D + W_T + W_C$	$W_D + W_T$	$W_D + W_C$	$W_T + W_C$
OxfordIIITPets	**0.666**	0.539	<u>0.657</u>	0.622
Oxford102Flowers	**0.455**	0.418	<u>0.435</u>	0.405
StanfordCars	**0.548**	0.524	<u>0.526</u>	0.512
StanfordDogs	<u>0.604</u>	0.496	**0.643**	0.563
Caltech101	0.486	<u>0.501</u>	**0.517**	0.309
Average	<u>0.552</u>	0.496	**0.556**	0.482

Table A7: Comparison of pearson corr. scores. **Bold** represents highest score, <u>Underline</u> represents second highest score.

estingly, that the inclusion of the W_C term significantly improves correlation scores. Our metric includes domain difference (W_D) and task difference (W_T) besides the model cohesiveness term (W_C). While selecting models from the source pool, our objective is not just minimizing the model disagreement via (W_C) but the entire metric. Through the interplay and equilibrium of these three components, model collapse is prevented.

References

[A1] Andrea Agostinelli, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. Transferability metrics for selecting source model ensembles. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7926–7936, 2022. 13, 14

[A2] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. Video-based Object and Event Analysis. 13

[A3] 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*, June 2016. 13

[A4] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W.H. Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5017–5026, October 2021. 13

[A5] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014. 14

[A6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 13

[A7] Andrew G. Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 13

[A8] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017. 13

[A9] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE/RSJ, 2020. 13

[A10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 13

[A11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12,* *2014, Proceedings, Part V*, volume 8693, pages 740–755, 2014. 13

[A12] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2014. 13

[A13] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. 13

[A14] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 12

[A15] Alain Rakotomamonjy, Rémi Flamary, and Nicolas Courty. Generalized conditional gradient: analysis of convergence and applications. *CoRR*, abs/1510.06567, 2015. 14

[A16] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 13

[A17] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 13

[A18] 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 *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 1743–1751, 2019. 13

[A19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 13