Self-Supervised Learning in Multi-Task Graphs through Iterative Consensus Shift – Supplementary Material –

Emanuela Haller*[1,2]
haller.emanuela@gmail.com
Elena Burceanu*[1,3]
eburceanu@bitdefender.com
Marius Leordeanu[1,2,4]
leordeanu@gmail.com

[1] Bitdefender, Romania
[2] University Politehnica of Bucharest, Romania
[3] University of Bucharest, Romania
[4] Institute of Mathematics of the Romanian Academy, Romania

In the following supplementary material, we provide additional details and experimental results that emphasize our main contributions. Sec. 1 analyzes the distribution gap between the source domains of the expert models and our target domains. Sec. 2 provides additional details regarding the considered expert models. Supplementary qualitative results are attached to current material and a brief description of the results is presented in Sec. 3.

1 Out-of-distribution experts adaptation

CShift requires no human-annotated data for the target domain. We take advantage of existing state-of-the-art expert models that distill research years and valuable expertise and provide reliable pseudo-labels for each of the considered tasks. When applied to novel domains, the weakness of these experts is that they are trained on different distributions. We first transfer their knowledge in our graph edges. Then our learning method, by exploiting and enforcing the overall consensus among all tasks, allows the graph to adapt by itself to the target domain, thus overcoming the domain gap, as shown in the following.

To emphasize the domain adaptation capabilities of CShift, we employ the Maximum Mean Discrepancy [3] (MMD) method for measuring the domain dissimilarity between our target domain and the expert source domains. MMD is a strong and widely used [5, 6, 13] non-parametric metric for comparing the distributions of two datasets. We follow the methodology in [3] and compute the unbiased empirical estimate of squared MMD. Our experiments show (Tab. 1) that there is a large distributional shift between our target domain and the domains of the original expert models. In conjunction with the ones presented in the Experimental Analysis Section of our main paper, these results prove our method's unsupervised domain adaptation capabilities.

We further analyze the gap between the source domain of *depth* and *normals* experts and one of our testing datasets: Replica. The experts [16] are originally trained on Taskonomy dataset, which is a real-world dataset, while Replica is a synthetic dataset. We will compute the discrepancy in distribution using MMD as mentioned above. Considering that the obtained discrepancy is not an absolute measure, we will also use the synthetic Hypersim dataset to perform a relative comparison. The analysis is performed both for the input level and the expert's mid-level features. For computing MMD, we average over multiple runs, each

	rgb	*depth*	*normals*
MMD(replica$_{part1}$, replica$_{part2}$)	5.4	17.8	17.4
MMD(replica$_{part1}$, hypersim)	3.4	20.1	20.6
MMD(replica$_{part1}$, taskonomy)	13.1	23.3	20.2

Table 1: We report the MMD between one of our target domains (Replica dataset) and the source domain of the *depth* and *normals* expert models (Taskonomy dataset), considering both *rgb* input and mid-level embeddings of the experts. Compared to another synthetic dataset (Hypersim), we observe a smaller distribution shift than for Taskonomy, which contains real-world samples. We also validate our assumptions by comparing two different splits of Replica. For readability, we report MMD $\times$100.

containing 100-1600 samples per dataset. The results in Tab. 1 show that there is a significant domain shift in the input for the pre-trained experts on Taskonomy, both at the *rgb* level but also through the eyes of the experts (*depth* and *normals* columns). Notice that the Hypersim dataset is closer to Replica (compared with Taskonomy) since both use synthetic data.

2 Expert models

Our graph contains a total of 13 task nodes, including the *rgb* one, thus we consider 12 experts ranging from trivial color-space transformations to heavily trained deep nets: **1)** halftone computed using python-halftone; **2)** grayscale and **3)** hsv computed with direct color-space transformations; **4)** depth and **5)** surface normals obtained from the XTC [16] experts; **6, 7, 8)** small, medium and large scale edges extracted using a Sobel-Feldman filter [2], and more complex **9)** edges extracted using the DexiNed [9] expert; **10)** super-pixel maps extracted using SpixelNet [14]; **11)** cartoonization got from WBCartoon [12] and **12)** semantic segmentation maps computed with HRNet [11]. The deep nets expert models are trained on a large variety of datasets: **4)** and **5)** Taskonomy [15], **9)** BIPED [8], **10)** SceneFlow [7] + BSDS500 [1], **11)** FFHQ [4], **12)** ADE20k [17]. Note that these datasets are built for a different purpose, on a different distribution than ours.

3 Additional qualitative results

We attached to the archive two videos showing additional qualitative results. In **no_ground-truth.mp4** video we present examples for the $super-pixel$, $edges$, and $cartoon$ tasks, for which we do not have ground-truth annotations in Replica [10] dataset. We start with the RGB input, which is the first input of all the edges reaching the ensemble, and the expert models' input generating the initial pseudo-labels. The next columns show the results of the Expert model and CShift results. We see on $super-pixel$ and $cartoon$ that CShift removes a large amount of noise and hallucinations from the Expert, improving the surfaces. For edges, it removes noisy structure coming from the texture rather than being real edges.

In the second video, **with_ground-truth.mp4**, we show *depth* and *normals* tasks for which we have access to the ground-truth labels, and we take advantage of this to show more insights on the performance. Except for the before mentioned RGB input, Expert, and CShift columns, we also add a column containing the ground truth. To better visualize the differences, in the last column, we draw a map where green represents pixels where CShift outperforms the Expert and red indicates pixels where the Expert is better. We highlight that the green areas are predominant. We also see how new elements (objects in the scene) from both domains start to become visible, even though they are missing in the Experts.

References

[1] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 2011.

[2] Jerome A. Feldman, Gary M. Feldman, Gilbert Falk, Gunnar Grape, J. Pearlman, Irwin Sobel, and Jay M. Tenenbaum. The stanford hand-eye project. In *IJCAI*, 1969.

[3] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In *NIPS*, 2006.

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[5] Wouter M. Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE TPAMI*, 2021.

[6] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

[7] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[8] Xavier Soria, Edgar Riba, and Angel Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *WACV*, 2020.

[9] Xavier Soria, Edgar Riba, and Angel Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *WACV*, 2020.

[10] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv*, 2019.

[11] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020.

[12] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *CVPR*, 2020.

[13] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, 2017.

[14] Fengting Yang, Qian Sun, Hailin Jin, and Zihan Zhou. Superpixel segmentation with fully convolutional networks. In *CVPR*, 2020.

[15] Amir Roshan Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018.

[16] Amir Roshan Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J. Guibas. Robust learning through cross-task consistency. In *CVPR*, 2020.

[17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.