Supplementary Material for Bridging Generalization Gaps in High Content Imaging Through Online Self-Supervised Domain Adaptation

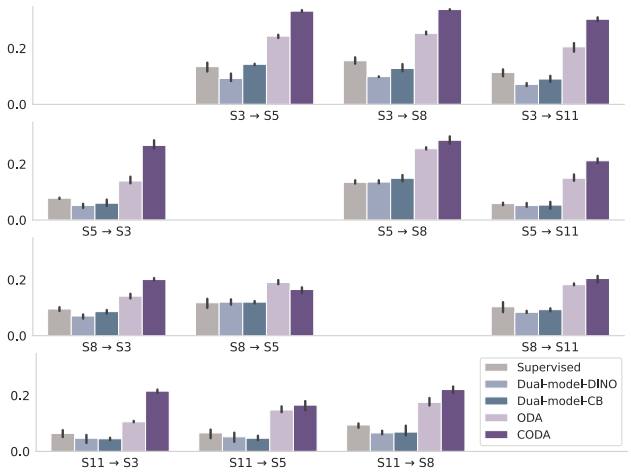

Figure 7. *Out-of-domain test performance (Acc.).* We compare the various learning setups trained on data from one source and applied to some target data, without access to labels from the target.

A. Appendix overview

Here, we provide further details and results from the experiments carried out in this work. In Section A.1 we present supplementary results and figures derived from the experiments discussed in the main text and in Section A.2 we provide additional information regarding the datasets and sources used in this study.

A.1. Additional results

Here we provide auxiliary information, results and figures from the experiments run and data used in this work.

In Table 4, re report results from the same experiments, discussed in the main text and reported in Table 2. We report F1 scores, corroborating the results reported in the main text when using Accuracy. The main results reported for in Table 2 (excluding TTT), are also reported in bar-chart format in Figure 1 – clearly visualizing the performance boosts of ODA and CODA compared to the baseline. In Table 5 additional experiments containing the performance of the DUAL-Model-MAE and TTT are reported between each of the sources used.

Table 4. Generalization performance across target sources (*F1.*)

Source	Target S3	S5	S8	S11	S7	S10	Model type (Set trained)
S3	-	10.9 ± 1.9	11.0 ± 0.5	8.1 ± 0.3	15.3 ± 0.5	9.0 ± 1.1	Supervised
	-	6.2 ± 0.7	6.5 ± 0.5	4.9 ± 0.5	10.2 ± 0.4	6.4 ± 0.2	Dual-model-DINO
	-	10.6 ± 0.5	8.3 ± 0.8	6.0 ± 0.5	13.2 ± 0.7	7.9 ± 0.8	Dual-model-CB
	-	6.4 ± 1.3	5.3 ± 1.9	5.6 ± 2.7	7.6 ± 3.5	5.2 ± 2.2	TTT
	-	22.2 ± 0.7	23.3 ± 0.1	19.1 ± 1.7	23.5 ± 0.6	15.6 ± 0.7	ODA
	-	30.7 ± 0.6	31.9 ± 0.5	28.6 ± 1.3	32.8 ± 0.4	24.4 ± 0.3	CODA
S5	5.5 ± 0.8	-	9.4 ± 0.4	2.5 ± 0.3	7.1 ± 0.9	9.4 ± 0.0	Supervised
	3.6 ± 0.3	-	10.9 ± 0.5	2.6 ± 0.2	5.6 ± 0.3	8.9 ± 0.7	Dual-model-DINO
	3.5 ± 0.6	-	11.0 ± 0.8	2.9 ± 0.1	5.2 ± 0.1	9.9 ± 0.6	Dual-model-CB
	4.1 ± 1.2	-	7.2 ± 1.6	3.6 ± 0.5	6.0 ± 1.7	6.6 ± 1.4	TTT
	13.4 ± 0.7	-	22.0 ± 0.6	13.1 ± 0.5	19.5 ± 0.9	18.0 ± 0.5	ODA
	23.3 ± 1.0	-	25.0 ± 1.1	18.0 ± 0.9	23.1 ± 0.4	24.0 ± 0.5	CODA
S8	5.6 ± 0.2	9.1 ± 0.6	-	6.2 ± 1.3	14.1 ± 0.3	9.6 ± 0.3	Supervised
	3.4 ± 0.2	8.8 ± 0.3	-	4.7 ± 0.3	9.2 ± 0.3	8.1 ± 0.1	Dual-model-DINO
	4.8 ± 0.4	8.9 ± 0.3	-	5.8 ± 0.2	9.2 ± 0.2	8.5 ± 0.5	Dual-model-CB
	2.4 ± 0.1	1.4 ± 0.1	-	3.0 ± 0.1	1.5 ± 0.0	2.3 ± 0.1	TTT
	10.1 ± 1.0	13.4 ± 0.3	-	15.0 ± 0.6	16.2 ± 1.5	13.5 ± 0.6	ODA
	16.1 ± 0.5	13.9 ± 1.3	-	17.3 ± 1.0	19.3 ± 1.1	16.7 ± 0.5	CODA
S11	4.1 ± 0.5	3.7 ± 0.2	6.3 ± 0.4	-	5.0 ± 0.3	4.2 ± 0.3	Supervised
	1.8 ± 0.3	1.6 ± 0.5	3.5 ± 0.1	-	2.3 ± 0.3	1.8 ± 0.3	Dual-model-DINO
	2.1 ± 0.4	2.1 ± 0.3	4.3 ± 0.5	-	2.8 ± 0.6	2.4 ± 0.4	Dual-model-CB
	2.4 ± 0.6	2.4 ± 0.8	4.4 ± 1.5	-	3.8 ± 0.8	2.9 ± 0.8	TTT
	7.9 ± 0.5	10.4 ± 0.7	13.8 ± 1.2	-	13.2 ± 0.7	6.8 ± 0.1	ODA
	18.2 ± 0.9	13.2 ± 0.9	19.4 ± 0.7	-	19.7 ± 1.1	8.7 ± 0.6	CODA

Table 5. Comparison of MAE based dual-model and Test-Time Training performance (accuracy).

Source	Target S3	S5	S8	S11	S7	S10	Model type (Set trained)
S3	-	11.9 ± 3.5	12.8 ± 3.3	12.1 ± 1.0	14.9 ± 2.6	9.4 ± 1.5	Dual-model-MAE
	-	8.2 ± 1.4	9.0 ± 1.0	9.0 ± 2.8	10.8 ± 3.3	7.6 ± 1.6	TTT
S5	7.6 ± 2.2	-	13.9 ± 0.4	6.8 ± 1.3	10.2 ± 0.8	12.4 ± 1.3	Dual-model-MAE
	6.6 ± 1.6	-	10.4 ± 1.8	5.9 ± 1.0	9.2 ± 2.2	9.0 ± 1.2	TTT
S8	9.1 ± 0.4	11.0 ± 0.1	-	10.7 ± 0.1	14.4 ± 0.4	10.5 ± 0.4	DUAL-model-MAE
	5.1 ± 0.6	4.6 ± 0.5	-	5.2 ± 0.5	3.3 ± 0.2	4.7 ± 0.6	TTT
S11	7.9 ± 0.6	8.2 ± 0.1	10.6 ± 1.1	-	9.4 ± 0.2	7.7 ± 0.2	DUAL-model-MAE
	4.7 ± 0.2	4.9 ± 0.3	7.8 ± 1.4	-	7.0 ± 0.0	5.6 ± 0.6	TTT

A.2. Detailed Data Description

As described in section 4 in the main text, the primary experiments focus on a subset of the data from four (anonymized) partners within the JUMP-CP consortium [3]. The data of those sources along with two additional test sources are described in Table 1. Here we include further information about these sources. Starting with the four primary sources that were selected, based on containing the largest subsets of data from each of the different microscope types used, thus, providing the most diverse set of data sources:

- **S3** contains 25 plates, totaling 9,600 unique wells and 85,409 images in total, belonging to 13 distinct experimental batches. These were captured using the Opera Phoenix microscope in widefield mode, using laser excitation and a 20X/1 NA objective.

- **S5** contains 24 plates, totaling 9,216 unique wells and 82,256 images in total, belonging to 23 distinct experimental batches. These were captured using the CV8000 confocal microscope, using laser excitation and a 20X/0.75 NA objective.
- **S8** contains 4 plates, totaling 1,536 unique wells and 13,824 images in total, belonging to 4 distinct experimental batches. These were captured using the ImageExpress Micro confocal microscope, using LED excitation and a 20X/0.75 NA objective.
- **S11** contains 7 plates, totaling 2,688 unique wells and 23,373 images in total, belonging to 4 distinct experimental batches. These were captured using the Operetta widefield microscope, using LED excitation and a 20X/1 NA objective.

Two additional sources were also used for auxiliary testing. Both use similar microscope setups to that used by S5, allowing comparison of generalization performance between models trained and tested in sources with similar imaging setups.

- **S7** contains 7 plates, totaling 2,688 unique wells and 24,192 images in total, belonging to 7 distinct experimental batches. These were captured using the CV7000 confocal microscope, using laser excitation and a 20X/0.75 NA objective.
- **S10** contains 6 plates, totaling 2,304 unique wells and 13,812 images in total, belonging to 6 distinct experimental batches. These were captured using the CV8000 confocal microscope, using laser excitation and a 20X/0.75 NA objective.