Self-Supervised Domain Adaptation with DINOv2-HRNet for Multi-Task Ultrasound Landmark Detection

Anna Panagiotakopoulou[1], Alexandros Barmperis[1,2], Vasileios E. Katsigiannis[1], and George K. Matsopoulos[1]

[1] Biomedical Engineering Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece
[2] Archimedes Research Unit, Athena Research Center, Athens, Greece

Supplementary Material

S1. Complete Phase-1 Adaptation-Duration Sweep

Section 4.2 of the main paper (Figure 4) reports the frozen-probe adaptation-duration sweep, macro-averaged over all nine tasks; the text states the key values (no-SSL baseline, the ep20 bulk-only peak, and the selected ep104 checkpoint). Table S1 below reports the complete sweep, including the intermediate tail-length and full-resolution variants omitted from the main text for space.

Table S1. Frozen-probe SSL-duration sweep (fold-0 internal validation). Challenge-blend selection metric and mean radial error (MRE, px).

Segment	Encoder / SSL duration	blend ↓	MRE (px) ↓
—	No SSL (off-the-shelf DINOv2)	0.0890	31.27
	SSL, 10 ep (bulk only)	0.0872	26.88
	SSL, 20 ep (bulk only)	0.0782	24.70
	SSL, 60 ep (bulk only)	0.0836	26.43
224px bulk + tail	SSL, 60 ep + 5-ep 518px tail	0.0748	25.23
	SSL, 60 ep + 10-ep 518px tail	0.0755	25.50
	SSL, 100 ep (bulk only)	0.0974	25.95
	SSL, 100 ep + 4-ep 518px tail	**0.0747**	25.09
	SSL, 10 ep	0.0929	25.79
518px full-res	SSL, 20 ep	0.0901	25.96
	SSL, 30 ep	0.0934	26.92

S2. Task-Specific Sensitivity to Phase-1 Adaptation Duration

Section 4.2 of the main paper reports frozen-encoder probe results macro-averaged over all nine tasks (Table S1 above). Here we report the same frozen-probe checkpoints broken down per task, to check whether the preferred Phase-1 adaptation duration is uniform across tasks.

Table S2. Task-specific sensitivity to Phase-1 adaptation duration. For each task, the best-scoring frozen-probe checkpoint (mean radial error, MRE, orig. px, fold-0 internal validation) and how the selected ep104 checkpoint compares to that task-specific optimum.

Task	Best checkpoint	Best MRE (px)	ep104 MRE (px)	Rel. diff. from best
AOP	ep20	4.47	4.82	+7.8%
FA	ep60	18.05	18.27	+1.2%
FUGC	ep100	9.64	9.94	+3.1%
HC	ep104	25.68	25.68	0.0%
fetal_femur	ep104	17.53	17.53	0.0%
A4C	ep60	54.01	55.01	+1.9%
PLAX	ep60	19.35	20.56	+6.3%
PSAX	ep20	45.30	46.21	+2.0%
IVC	ep20	17.21	27.88	+62.0%

The preferred adaptation duration varies across tasks, with task-specific optima spanning ep20–ep104. Nevertheless, the selected ep104 checkpoint remains within 8% MRE of the task-specific optimum for eight of nine tasks. The larger deviation for IVC should be interpreted cautiously given the small sample size of this task. Adaptation-duration preference is thus task dependent, but ep104 provides a reasonable single shared checkpoint for the multi-task setting.

S3. Compute Cost, Throughput, and Memory Footprint

Full breakdown behind the headline numbers in Section 3.2 of the main paper. All figures measured on a single NVIDIA A100-SXM4-80GB GPU.

Table S3. Parameter counts, throughput, and peak memory for both training phases. Phase-1 GFLOPs are per optimization step (with gradient accumulation); Phase-2 GFLOPs are per forward pass at batch size 1.

Stage	Configuration	Throughput	Peak memory	GFLOPs
Phase 1	224px bulk (bs32×8 accum)	64.4 img/s	33.16 GB	27,641 / step
Phase 1	518px tail (bs16×16 accum)	15.9 img/s	55.03 GB	68,579 / step
Phase 2	Training, bf16 (bs64)	84.4 img/s	29.24 GB	1051.8 / fwd
Phase 2	Inference, fp32 (bs64)	17.8 img/s	6.06 GB	—
Phase 2	Inference, bf16 (bs64)	127.3 img/s	5.65 GB	—

Phase 1 trains the full encoder (351.7M parameters, all trainable) over 191K unlabeled frames; the complete 100-bulk+4-tail schedule totals ∼96 GPU-hours, making Phase 1 the more expensive stage overall despite Phase 2's larger per-step batch throughput. Phase 2 has 312.2M parameters (58.2M trainable: the last four unfrozen encoder blocks, neck, and nine task heads). Inference currently runs in fp32 (the precision used by `predict.py`/`evaluate.py` at submission time); bf16 autocast inference is ∼7× faster at lower memory and is a straightforward future optimization, not yet enabled in the submitted pipeline.

S4. Phase-2 Decoder Ablation: Per-Fold Breakdown

Section 4.3 of the main paper (Table 2) reports the five-fold mean ± SD comparing the simple decoder against the HRNet-style neck, fixed epoch-104 Phase-1 checkpoint and upgraded recipe. Table S4 below reports the complete per-fold breakdown.

Table S4. Phase-2 decoder ablation, per fold. Challenge-blend selection metric and mean radial error (MRE, px), fixed epoch-104 Phase-1 checkpoint and upgraded recipe.

Fold	Simple decoder		HRNet decoder	
	blend ↓	MRE (px) ↓	blend ↓	MRE (px) ↓
0	0.0717	24.61	0.0763	27.37
1	0.0846	23.64	0.0680	24.25
2	0.0658	23.54	0.0648	23.24
3	0.0736	26.40	0.0612	23.02
4	0.0745	27.99	0.0830	30.55
Mean	0.0740	25.24	0.0707	25.69

S5. Qualitative PCA Visualizations of Adapted Features

Extended version of the qualitative feature analysis in Section 2.1 of the main paper, showing the progression of DINOv2 patch-token features over the course of Phase-1 SSL adaptation.

S6. Reviewer Response Summary

S6.1 Reviewer jWeV

1.1 Agree.
1.2 Agree.
1.3 Agree — citations added (Introduction).

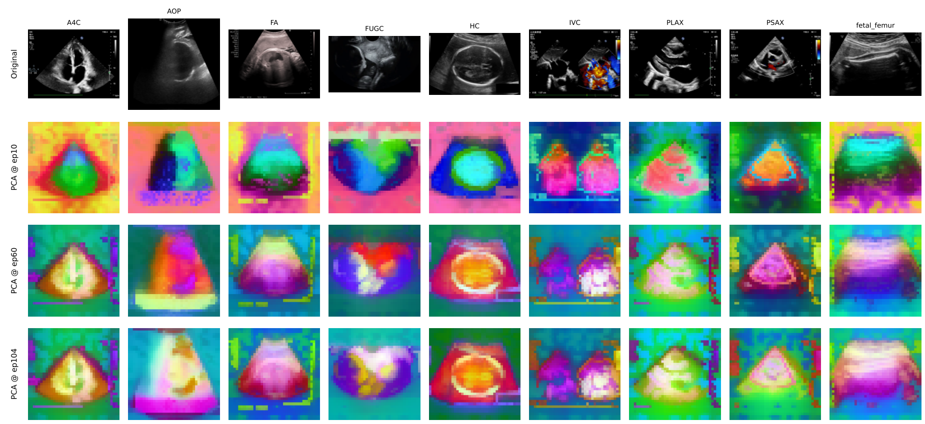

Fig. S1. Representative PCA visualizations of DINOv2 patch-token features during ultrasound-domain adaptation, at epoch-10, epoch-60, and epoch-104 of the SSL recipe described in Section 2.1 of the main paper.

S6.2 Reviewer ZxK3

2.1 Agree — Abstract; Section 4.1.
2.2 Agree — Section 4.2, Figure 4; Tables S1–S2.
2.3 Agree — Section 2.2, Figure 2.
2.4 Agree — Section 4.2.
2.5 Agree — Section 4.2, Figure 4 (epoch 0).

S6.3 Reviewer KKV9

3.1 Agree — Section 4.1.
3.2 Agree — Section 4.2; Table 2.
3.3 Agree — Section 3.2; Table S3.
3.4 Agree — Section 4.2; Tables S1–S2.
3.5 Agree — Section 4.1.