How to Embed Matters: Evaluation of EO Embedding Design Choices

Supplementary Material

In the supplementary material, we provide additional results and visualizations that complement the main paper. These analyses cover several experimental axes.

Section A reports full performance accuracy results per method, along with additional performance plots for final-layer GeoFM embeddings, complementing the backbone, SSL objective, and spatial pooling analyses. Section B presents additional visualizations for the concatenation experiments. Finally, Section C provides per-task results for intermediate-layer embeddings, extending the layer-wise analysis.

A. Final-Layer Embeddings: Full Per-Task Results

Overview. This section provides the complete per-task R^2 results for all evaluated GeoFM backbones (final-layer embedding), self-supervised objectives, and pooling strategies: Table 1 reports mean pooling results for SSL4EO pretrained GeoFMs, and similarly Table 2 and Table 3 report max and min pooling, respectively. Methods are sorted by average performance (Avg.) to make consistent cross-task trends explicit and to complement the Q-score–based radar plot shown in the main paper. Additionally, in Table 5 all pooling method results for TerraMind-Small are reported.

R^2 radar plots. Figures 1 and 2 replicate the main-paper radar plots using raw R^2 instead of Q-score. The qualitative trends remain consistent: model families and pooling strategies that rank highly under Q-score also exhibit strong mean predictive performance, confirming that Q-score primarily sharpens separation rather than altering relative rankings. Across both ResNet and ViT backbones, SSL objectives exhibit task-dependent strengths, consistent with the patterns discussed in the main paper. No single objective dominates across all tasks, reinforcing the importance of task-aware embedding selection in embedding-centric workflows.

TerraMind results. For TerraMind ViT-Small, mean pooling clearly dominates min and max variants, confirming that aggregation choice remains critical even for stronger pretrained backbones.

Pooling comparison. Mean pooling consistently yields the strongest performance across backbones. For ViT models, max and min pooling produce similar but consistently lower results. For ResNet models, both max and min pooling substantially degrade performance.

B. Per-Task Concatenation Results

This section provides full per-task R^2 radar plots for the concatenation experiments introduced in the main paper. While the main paper reports per-task ΔR^2 bar plots relative to the stronger baseline, here we show the absolute per-task R^2 values for both individual embeddings and their concatenation.

Concatenation analysis. As shown in Figure 3, concatenation typically preserves the stronger baseline and yields modest, task-dependent gains. In many cases performance remains close to the best single representation, with clearer gains when combining embeddings from different pretraining objectives. Overall, concatenation provides limited but measurable benefits, with selected cases where the joint embedding outperforms the strongest standalone representation on a per-task level.

C. Per-Task Layerwise Results

Overview. This section provides per-task layer-wise performance breakdowns, extending the averaged analysis presented in the main paper. We report both R^2 and Q-score to disentangle predictive accuracy and robustness across splits.

Layer-wise analysis (ViT). Figures 4 and 5 show per-task depth behavior for ViT-Small models. Semantic and land-cover tasks exhibit increasing and saturating trends toward deeper layers, mirroring the average performance curves in the main paper. In contrast, several geophysical tasks saturate earlier or show marginal degradation at the deepest layers, indicating that additional depth does not universally improve performance.

Layer-wise analysis (ResNet). Figures 6 and 7 illustrate a more pronounced depth dependence for ResNet-50 models. While semantic and land-cover tasks benefit from deeper representations, multiple other tasks exhibit a clear performance drop at the final-layer. Intermediate layers therefore remain superior for several targets, supporting the main-paper observation that final-layer embeddings can reduce task-agnostic performance for convolutional backbones.

Use of LLMs

We utilized large language models (LLMs) to refine text and improve readability. All content, including technical material, experimental design, and analyses, was developed by the authors.

Table 1. **Full per-task R^2 scores for tested embedding methods (Mean pooling).** Methods are sorted in ascending order by Avg. For each task, the best-performing method is highlighted in **bold**, and the second-best is underlined.

Method	Biomass Mean	Biomass Std	Crops	Clouds	LC Agri	LC Forest	HI Mean	HI Std	Avg.
ResNet SoftCon (mean)	-0.282	-0.184	0.725	-0.022	0.825	0.806	0.070	-0.561	0.172
ResNet DeCur (mean)	-0.205	-0.127	0.807	-0.042	0.856	0.845	0.198	-0.427	0.238
ResNet MoCo (mean)	-0.139	-0.125	0.798	0.013	0.851	0.838	0.296	-0.332	0.275
ResNet DINO (mean)	0.053	0.005	0.835	-0.203	**0.870**	**0.863**	0.264	-0.282	0.301
ViT DINO (mean)	0.282	0.217	**0.843**	0.334	0.866	**0.863**	0.304	-0.129	0.447
ViT MoCo (mean)	0.375	0.293	0.762	0.338	0.827	0.824	0.471	0.158	0.506
ViT MAE (mean)	0.408	0.335	0.609	0.684	0.800	0.804	0.530	0.145	0.539
ViT FGMAE (mean)	**0.424**	**0.338**	0.630	**0.686**	0.815	0.826	0.531	0.155	0.551
ViT SoftCon (mean)	0.422	0.334	0.763	0.486	0.856	0.851	**0.555**	**0.181**	**0.556**

Table 2. **Full per-task R^2 scores for tested embedding methods (Max pooling).** Methods are sorted in ascending order by Avg. For each task, the best-performing method is highlighted in **bold**, and the second-best is underlined.

Method	Biomass Mean	Biomass Std	Crops	Clouds	LC Agri	LC Forest	HI Mean	HI Std	Avg.
ResNet SoftCon (max)	-1.360	-1.008	0.520	-0.546	0.719	0.703	-1.049	-1.984	-0.501
ResNet MoCo (max)	-0.977	-0.778	0.637	-0.328	0.756	0.754	-0.633	-1.474	-0.255
ResNet DeCur (max)	-0.781	-0.674	0.677	-0.386	0.776	0.770	-0.641	-1.637	-0.237
ResNet DINO (max)	-0.723	-0.589	0.683	-0.498	0.770	0.758	-0.519	-1.336	-0.182
ViT MAE (max)	0.171	0.132	0.433	0.148	0.705	0.696	-0.060	-0.499	0.216
ViT FGMAE (max)	0.155	0.110	0.451	0.101	0.709	0.711	-0.070	-0.390	0.222
ViT DINO (max)	**0.368**	0.240	**0.755**	-0.374	0.796	0.791	-0.121	-0.525	0.241
ViT MoCo (max)	-0.004	0.037	0.670	0.173	0.769	0.762	0.177	-0.215	0.296
ViT SoftCon (max)	0.321	**0.253**	0.731	**0.441**	**0.836**	**0.831**	**0.446**	**0.074**	**0.492**

Table 3. **Full per-task R^2 scores for tested embedding methods (Min pooling).** Methods are sorted in ascending order by Avg. For each task, the best-performing method is highlighted in **bold**, and the second-best is underlined.

Method	Biomass Mean	Biomass Std	Crops	Clouds	LC Agri	LC Forest	HI Mean	HI Std	Avg.
ResNet MoCo (min)	-0.680	-0.647	0.411	-3.071	0.258	0.242	-0.584	-0.801	-0.609
ResNet DeCur (min)	-1.017	-0.873	0.670	-2.105	0.713	0.703	-0.827	-1.531	-0.533
ResNet DINO (min)	-1.016	-0.779	0.718	-1.244	0.779	0.758	-0.833	-1.606	-0.403
ResNet SoftCon (min)	0.043	0.026	0.304	-0.253	0.317	0.238	-0.020	-0.171	0.060
ViT MAE (min)	0.131	0.107	0.460	0.121	0.694	0.673	-0.018	-0.456	0.214
ViT FGMAE (min)	0.175	0.136	0.427	0.109	0.709	0.708	0.021	-0.292	0.249
ViT DINO (min)	**0.370**	0.244	**0.754**	-0.283	0.798	0.787	-0.186	-0.470	0.252
ViT MoCo (min)	-0.014	0.020	0.669	0.181	0.765	0.760	0.173	-0.220	0.292
ViT SoftCon (min)	0.324	**0.251**	0.738	**0.433**	**0.840**	**0.831**	**0.473**	**0.067**	**0.495**

Table 4. **Full per-task R^2 scores for tested embedding methods (CLS token).** Methods are sorted in ascending order by Avg. For each task, the best-performing method is highlighted in **bold**, and the second-best is underlined.

Method	Biomass Mean	Biomass Std	Crops	Clouds	LC Agri	LC Forest	HI Mean	HI Std	Avg.
ViT DINO (CLS)	0.324	0.236	**0.860**	0.134	**0.878**	**0.873**	0.328	-0.129	0.438
ViT SoftCon (CLS)	0.369	0.286	0.769	0.404	0.855	0.850	0.504	0.105	0.518
ViT MAE (CLS)	0.403	0.316	0.584	**0.619**	0.778	0.777	0.516	0.165	0.520
ViT FGMAE (CLS)	**0.413**	**0.331**	0.611	0.610	0.790	0.795	0.527	**0.176**	0.532
ViT MoCo (CLS)	0.386	0.311	0.798	0.431	0.854	0.852	**0.537**	0.131	**0.537**

Table 5. **Full per-task R^2 scores for TerraMind ViT-Small (pooling variants).** Results are reported for min, max, and mean pooling and sorted by Avg.

Method	Biomass Mean	Biomass Std	Crops	Clouds	LC Agri	LC Forest	HI Mean	HI Std	Avg.
TerraMind Small (min)	0.281	0.216	0.730	0.360	0.832	0.829	0.332	-0.081	0.437
TerraMind Small (max)	0.306	0.235	0.738	0.335	0.837	0.829	0.327	-0.099	0.438
TerraMind Small (mean)	0.511	0.384	0.852	0.671	0.900	0.896	0.637	0.239	0.636

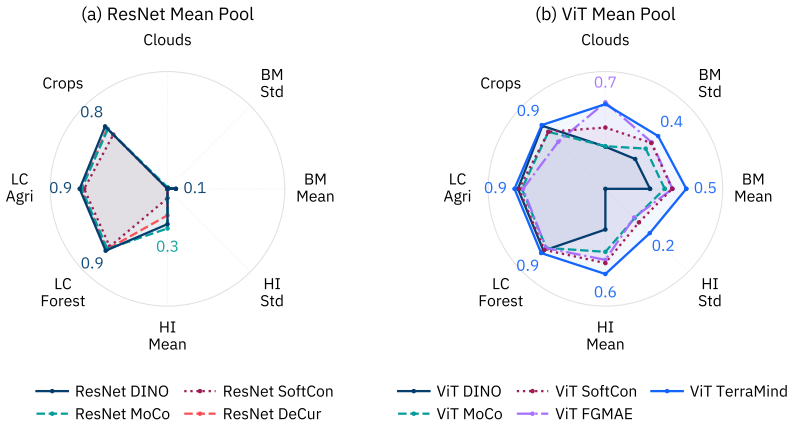

Figure 1. **Per-task R^2 comparison of ResNet-50 (left) and ViT-Small (right) FMs.** Final-layer embeddings with mean pooling are used. In contrast to the main paper's Q-score visualization, this plot reports raw predictive performance (R^2) per task. The overall ranking trends remain consistent: ResNet models perform strongly on semantic/land-cover tasks but show limited transfer beyond them, while ViT models are more balanced across tasks. TerraMind remains the most consistent ViT backbone, DINO is particularly strong on land-cover targets, and FGMAE performs well on cloud-cover and biomass tasks. The radial axis is centered at 0, and the maximum radius is fixed globally with a constant buffer for comparability.

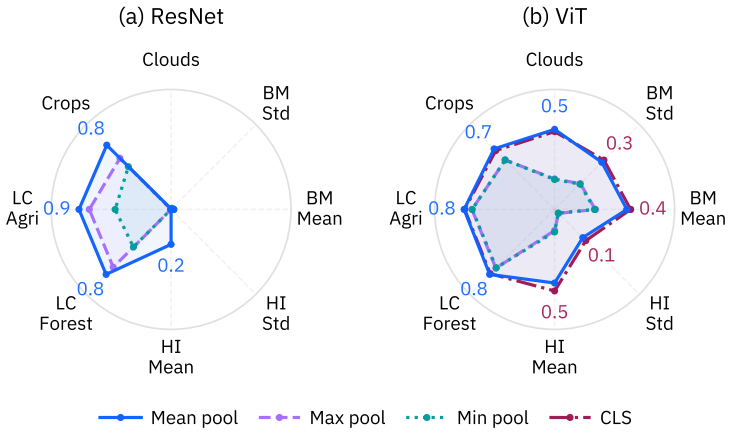

Figure 2. **Per-task R^2 comparison of spatial aggregation methods for ResNet-50 (left) and ViT-Small (right).** Final-layer embeddings are evaluated using mean, min, or max pooling (and the CLS token for ViT), with scores averaged across models. This R^2 view confirms the Q-score trends reported in the main paper: mean pooling consistently yields the strongest performance across tasks and backbones. For ResNet, max pooling generally outperforms min pooling but both degrade performance relative to mean pooling. For ViT, mean pooling again performs best, with CLS comparable on several semantic tasks, while min and max pooling are similar but systematically weaker—especially on non-land-cover targets. The radial axis is centered at 0, with a fixed global maximum radius.

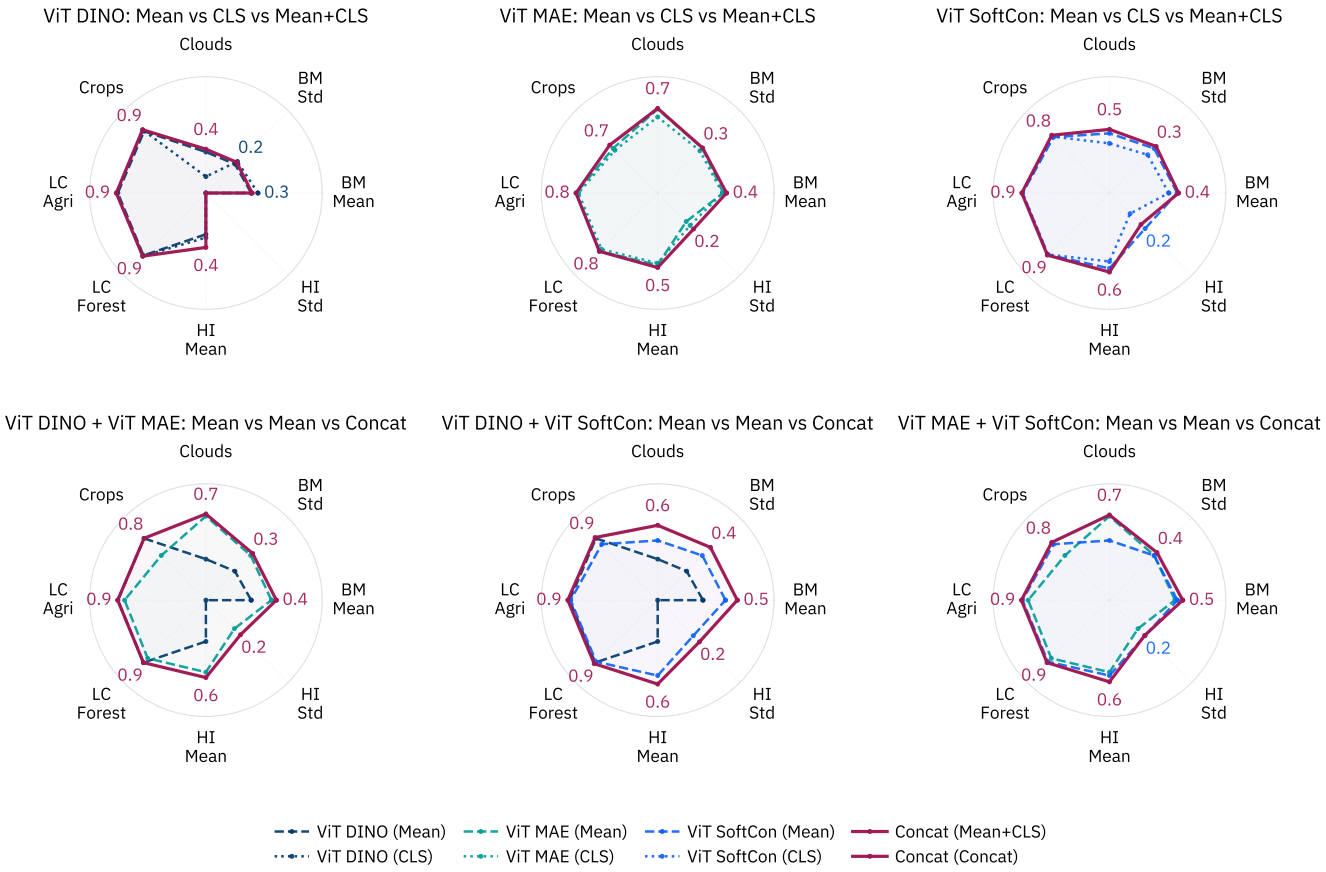

Figure 3. **Per-task R^2 radar plots for embedding concatenation experiments.** We report results for all tested combinations, comparing the two individual baselines with their concatenated representation. The plots illustrate that concatenation typically preserves the stronger baseline and yields modest, task-dependent improvements. In particular, combinations such as SoftCon+DINO most consistently show positive deviations over the individual embeddings, suggesting measurable complementarity between pretraining objectives.

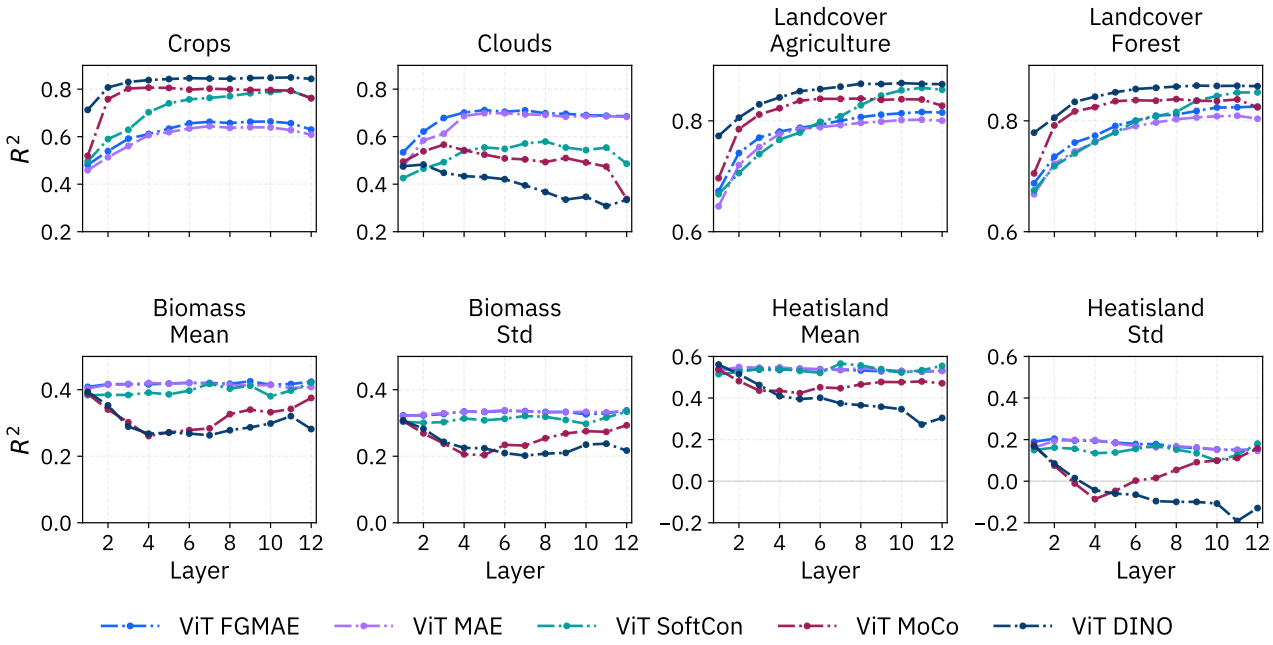

Figure 4. **Layer-wise per-task downstream performance (R^2) for ViT-Small models pretrained on SSL4EO.** Results are shown separately per task across layer depth. Semantic and land-cover targets exhibit increasing and saturating trends toward deeper layers, consistent with the averaged analysis in the main paper. Other tasks show early saturation or slight degradation at greater depth.

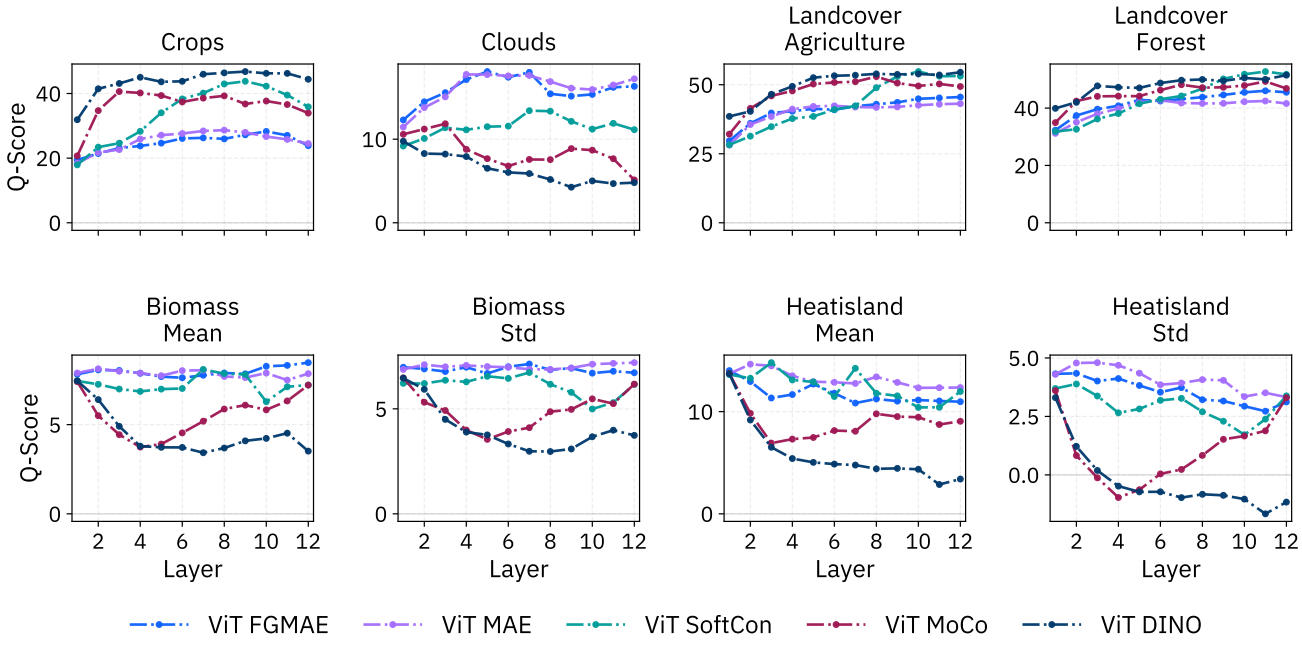

Figure 5. **Layer-wise per-task downstream performance (Q-score) for ViT-Small models pretrained on SSL4EO.** The robustness trends largely mirror the R^2 behavior, confirming that depth-dependent effects are consistent across predictive accuracy and stability metrics.

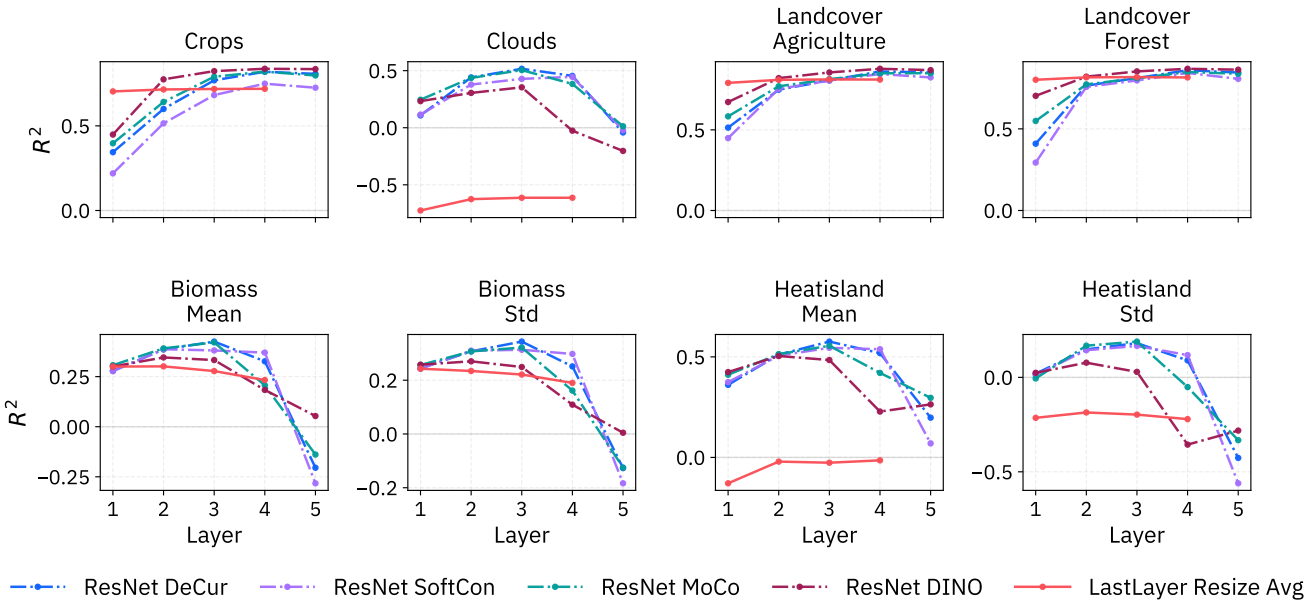

Figure 6. **Layer-wise per-task downstream performance (R^2) for ResNet-50 models pretrained on SSL4EO.** Semantic and land-cover tasks show increasing and saturating trends similar to ViT models. In contrast, several other tasks exhibit a pronounced drop at the final-layer, intermediate layers frequently remain competitive with ViT final-layer embeddings.

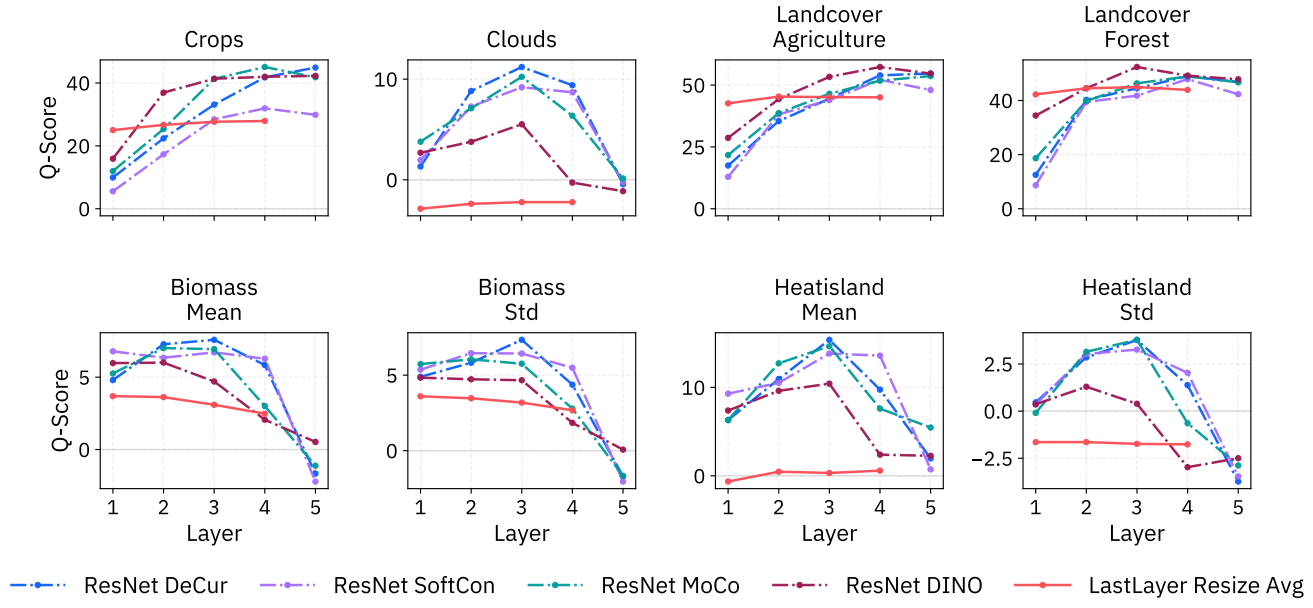

Figure 7. **Layer-wise per-task downstream performance (Q-score) for ResNet-50 models pretrained on SSL4EO.** The robustness metric reinforces the R^2 trends, highlighting stronger depth sensitivity in ResNet compared to ViT backbones.