CraterBench-R: Instance-Level Crater Retrieval for Planetary Scale

Supplementary Material

7. Complete Baseline Results

Table 5 reports retrieval performance for all 30 frozen backbones evaluated on Curated-5K, using the best pooling strategy per model. This extends Table 2 in the main paper, which shows a representative subset.

8. Pooling Ablation

Table 6 reports performance for every ViT backbone under all four pooling strategies. Table 7 reports CNN results with GAP and GeM where applicable.

Pooling preferences vary by pretraining objective. CLS pooling is strongest for DINO v1 backbones, where the self-supervised objective explicitly trains the CLS token. DINOv2 and DINOv3 favor max pooling, likely because their training distributes discriminative information more uniformly across patch tokens. MarsDINO is mixed: ViT-S prefers GeM while ViT-B is best with CLS.

GeM as a robust default. GeM with $p{=}3$ is within 2 points of the best strategy for nearly all backbones, making it a practical default when model-specific tuning is not feasible.

9. Token Selection Strategy Comparison

Table 8 compares seven token selection strategies at $K{=}64$ for both ViT-S/16 backbones. Attention-based strategies (attention, norm$\times$attention) consistently rank first. On DINO, the top three strategies (norm$\times$attention, attention, norm) perform within 1% mAP of each other, while on MarsDINO the attention advantage is larger ($+9$ mAP over norm). Random selection is a surprisingly strong baseline, achieving 95% of the optimal mAP on DINO and 90% on MarsDINO. CLS-distance (selecting tokens *farthest* from CLS) is consistently worst, confirming that salient—not diverse—tokens drive retrieval quality.

10. Instance-Token Aggregation Ablation

We ablate the three design axes of the instance-token aggregation pipeline introduced in Sec. 4: seed selection, token-to-seed assignment, and matching strategy. All experiments use late interaction unless otherwise noted.

Assignment strategy. Tables 9 and 10 compare four assignment strategies at each K, using the best seed method per model. Hard nearest-neighbor (`hard_top1`) is the strongest or within 1% mAP of the best for nearly all K on both models. Soft assignment (`soft_top2/4`) is competitive but adds no consistent benefit. Dense attention-weighted assignment (`group_dense`) underperforms at $K{\geq}16$, likely because spreading mass across all seeds dilutes cluster coherence. The rightmost column shows the Phase 2 raw-attention baseline (no aggregation); aggregation improves over raw at every K by +1.3 to +17.9 mAP.

Seed selection. Table 11 compares attention and FPS seeding. FPS dominates on DINO at $K{\geq}8$, reaching mAP = .760 vs. .734 for attention at $K{=}64$—FPS produces spatially diverse seeds that better partition the token space. On MarsDINO, attention seeds are stronger at every K except 16 (where FPS edges ahead by 0.4%), suggesting that MarsDINO's attention heads already identify instance-relevant regions.

Matching strategy. Table 13 shows that late interaction is essential for multi-token retrieval: pooling instance tokens into a single vector (mean or max) recovers only 40–60% of the late-interaction mAP.

Descriptor compression. Table 12 evaluates token quantization for storage-constrained deployment. FP16 and INT8 (per-vector symmetric scalar quantization) are *lossless*: mAP changes by $\leq$0.02 points on both backbones at $K{=}32$. Product quantization with $m{=}96$ sub-vectors (16$\times$ compression) loses only 1.0–1.2 mAP points; aggressive PQ-48 (32$\times$) loses 6.2–7.3 points. These results show that INT8 storage ($K{=}32$: 12.4 KB/image) is a practical default, and PQ-96 (3.1 KB/image) is viable when storage is severely constrained.

However, pooled instance tokens still outperform the single-vector baselines from Table 2 when K is large enough ($K{\geq}32$), confirming that aggregation concentrates discriminative signal.

Table 5. Complete frozen-backbone retrieval results on Curated-5K (best pooling per model). Best per column in **bold**, second-best underlined.

Model	Pretraining	Params	Pool	R@1	R@5	R@10	mAP
CNNs (ImageNet-1k supervised)							
ResNet-18	Sup. IN-1k	11 M	GAP	.102	.155	.182	.179
ResNet-50	Sup. IN-1k	24 M	GeM	.142	.217	.248	.244
ResNet-101	Sup. IN-1k	43 M	GAP	.103	.170	.200	.184
ConvNeXt-T	Sup. IN-1k	28 M	GeM	.075	.121	.146	.134
EfficientNet-B0	Sup. IN-1k	4 M	GAP	.150	.214	.250	.248
DenseNet-121	Sup. IN-1k	7 M	GAP	.110	.170	.202	.189
VGG-16	Sup. IN-1k	134 M	GAP	.068	.108	.134	.124
MobileNetV3-L	Sup. IN-1k	4 M	GAP	.133	.201	.235	.230
Supervised ViTs							
ViT-S/16 AugReg	Sup. IN-21k→1k	22 M	Mean	.115	.187	.219	.200
ViT-B/16 Orig	Sup. IN-21k→1k	86 M	Mean	.091	.142	.171	.158
DeiT-S/16	Sup. IN-1k	22 M	GeM	.137	.202	.232	.229
DeiT-B/16	Sup. IN-1k	86 M	Max	.187	.267	.303	.303
DeiT3-S/16	Sup. IN-1k	22 M	Mean	.064	.109	.129	.116
DeiT3-B/16	Sup. IN-1k	86 M	GeM	.148	.210	.235	.239
Self-supervised ViTs (natural images)							
ViT-S/16 DINO	SSL (IN-1k)	22 M	CLS	.273	.360	.402	.420
ViT-S/8 DINO	SSL (IN-1k)	22 M	CLS	.296	.383	.419	.454
ViT-B/16 DINO	SSL (IN-1k)	86 M	CLS	.295	.387	.425	.454
ViT-B/8 DINO	SSL (IN-1k)	86 M	GeM	.304	.379	.409	.461
ViT-S/14 DINOv2	SSL (LVD-142M)	22 M	Max	.226	.302	.336	.355
ViT-B/14 DINOv2	SSL (LVD-142M)	87 M	Max	.240	.323	.360	.377
ViT-S/16 DINOv3	SSL (LVD-1.7B)	22 M	CLS	.258	.354	.395	.406
ViT-B/16 DINOv3	SSL (LVD-1.7B)	86 M	CLS	.218	.321	.363	.353
ViT-L/16 DINOv3$_{\text{sat}}$	SSL (Sat-493M)	303 M	Max	.208	.287	.321	.337
ViT-L/16 DINOv3$_{\text{lvd}}$	SSL (LVD-1.7B)	303 M	CLS	.162	.233	.274	.259
ViT-7B/16 DINOv3$_{\text{lvd}}$	SSL (LVD-1.7B)	6.7 B	Max	.247	.340	.377	.393
ViT-7B/16 DINOv3$_{\text{sat}}$	SSL (Sat-493M)	6.7 B	Max	<u>.330</u>	<u>.416</u>	<u>.450</u>	<u>.505</u>
Other pretraining objectives							
ViT-B/16 MAE	Recon. (IN-1k)	86 M	GeM	.022	.042	.052	.043
ViT-B/16 CLIP	Lang.-img (WIT)	86 M	GeM	.058	.091	.109	.107
Domain-specific (Mars orbital imagery)							
ViT-S/16 MarsDINO	DINO (Mars)	22 M	GeM	.269	.356	.391	.412
ViT-B/16 MarsDINO	DINO (Mars)	85 M	CLS	**.374**	**.472**	**.503**	**.553**

Table 6. Effect of token pooling on ViT backbones (R@1 / mAP). Best pooling per backbone in **bold**.

	CLS		**Mean**		**Max**		**GeM** (p=3)	
Backbone	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
DINO v1								
ViT-S/16 DINO	**.273**	**.420**	.177	.279	.126	.210	.258	.392
ViT-S/8 DINO	**.296**	**.454**	.162	.257	.110	.185	.185	.286
ViT-B/16 DINO	**.295**	**.454**	.226	.351	.194	.313	.262	.399
ViT-B/8 DINO	.303	.465	.263	.402	.226	.356	**.304**	**.461**
DINOv2								
ViT-S/14 DINOv2	.200	.327	.177	.287	**.226**	**.355**	.195	.315
ViT-B/14 DINOv2	.157	.262	.188	.306	**.240**	**.377**	.221	.348
DINOv3								
ViT-S/16 DINOv3	**.258**	**.406**	.226	.366	.256	.405	.225	.366
ViT-B/16 DINOv3	**.218**	**.353**	.188	.307	.207	.329	.194	.313
ViT-L/16 DINOv3$_{\text{sat}}$	.188	.314	.181	.303	**.208**	**.337**	.204	.334
ViT-L/16 DINOv3$_{\text{lvd}}$	**.162**	**.259**	.121	.205	.137	.226	.119	.204
ViT-7B/16 DINOv3$_{\text{lvd}}$	.215	.353	.139	.244	**.247**	**.393**	.204	.337
ViT-7B/16 DINOv3$_{\text{sat}}$	.132	.235	.261	.412	**.330**	**.505**	.298	.460
MarsDINO (domain-specific)								
ViT-S/16 MarsDINO	.241	.368	.249	.381	.225	.355	**.269**	**.412**
ViT-B/16 MarsDINO	**.374**	**.553**	.369	.540	.344	.511	.369	.544
Other pretraining								
ViT-B/16 MAE	.013	.022	.013	.022	.008	.019	**.022**	**.043**
Supervised ViTs								
ViT-S/16 AugReg	.104	.181	**.115**	**.200**	.085	.150	.109	.187
ViT-B/16 Orig	.085	.146	**.091**	**.158**	.080	.135	.082	.143
ViT-B/16 CLIP	.019	.043	.056	.103	.049	.090	**.058**	**.107**
DeiT-S/16	.108	.188	.131	.224	.100	.175	**.137**	**.229**
DeiT-B/16	.159	.260	.182	.292	**.187**	**.303**	.181	.290
DeiT3-S/16	.059	.111	**.064**	**.116**	.011	.026	.051	.092
DeiT3-B/16	.120	.203	.141	.229	.104	.177	**.148**	**.239**

Table 7. CNN baseline retrieval results. Models with both GAP and GeM pooling are shown; best per model in **bold**.

Model	Pool	R@1	R@5	R@10	mAP
ResNet-18	GAP	.102	.155	.182	.179
ResNet-50	GAP	.103	.159	.191	.182
ResNet-50	GeM	**.142**	**.217**	**.248**	**.244**
ResNet-101	GAP	.103	.170	.200	.184
ConvNeXt-T	GAP	.072	.120	.143	.130
ConvNeXt-T	GeM	**.075**	**.121**	**.146**	**.134**
EfficientNet-B0	GAP	**.150**	**.214**	**.250**	**.248**
EfficientNet-B0	GeM	.141	.207	.235	.239
DenseNet-121	GAP	.110	.170	.202	.189
VGG-16	GAP	.068	.108	.134	.124
MobileNetV3-L	GAP	.133	.201	.235	.230

Table 8. Token selection strategy comparison at K=64 (ViT-S/16). Best per model in **bold**.

	DINO		MarsDINO	
Strategy	R@1	mAP	R@1	mAP
Attention	.507	.716	**.453**	**.642**
Norm×Attn	**.507**	**.717**	.450	.641
Norm	.505	.714	.374	.549
Random	.477	.681	.395	.576
Spatial grid	.457	.659	.390	.566
CLS similarity	.439	.626	.352	.512
CLS distance	.384	.561	.349	.515

Table 9. Assignment comparison — ViT-S/16 DINO (late interaction, best seed per K). Raw = Phase 2 attention selection without aggregation.

	hard_top1		soft_top2		soft_top4		group_dense		Raw
K	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	mAP
4	.108	.180	.104	.174	.100	.168	**.109**	**.181**	.168
8	**.231**	**.378**	.219	.358	.189	.304	.192	.303	—
16	**.425**	**.623**	.416	.612	.394	.582	.296	.457	.444
32	**.514**	**.726**	.511	.724	.497	.708	.432	.628	—
64	.541	.759	**.543**	**.760**	.536	.753	.492	.698	.716

Table 10. Assignment comparison — ViT-S/16 MarsDINO (late interaction, best seed per K).

	hard_top1		soft_top2		soft_top4		gtp_dense		Raw
K	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	mAP
4	.163	.267	.166	.270	.169	.275	**.171**	**.276**	.167
8	**.258**	**.400**	.257	.399	.255	.394	.251	.388	—
16	**.351**	**.528**	.343	.517	.341	.508	.322	.484	.446
32	**.415**	**.602**	.407	.591	.395	.577	.366	.536	—
64	**.457**	**.650**	.449	.639	.439	.626	.388	.561	.642

Table 11. Seed selection comparison (late interaction, best assignment per seed). Bold = best per model and K.

	DINO				MarsDINO			
	Attention		FPS		Attention		FPS	
K	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
4	**.109**	**.181**	.084	.152	**.171**	**.276**	.113	.201
8	.203	.326	**.231**	**.378**	**.258**	**.400**	.227	.371
16	.336	.504	**.425**	**.623**	.353	.524	**.351**	**.528**
32	.449	.647	**.514**	**.726**	**.415**	**.602**	.414	.598
64	.521	.734	**.543**	**.760**	**.457**	**.650**	.444	.632

Table 12. Effect of descriptor quantization on exhaustive late-interaction mAP (K=32). Compression is relative to FP32 (1536 B/token).

Method	B/tok	Comp.	DINO	MarsDINO
FP32	1536	1×	.726	.602
FP16	768	2×	.726	.602
INT8	388	4×	.726	.602
PQ (m=96)	96	16×	.715	.590
PQ (m=48)	48	32×	.664	.529

Table 13. Matching strategy comparison (attention seeds, best assignment). Descriptor size shows bytes per image (384-dim, float32).

	DINO			MarsDINO		
K	Late	Mean	Max	Late	Mean	Max
4	.181	.144	.107	.276	**.281**	.245
8	.326	.207	.141	.400	.343	.311
16	.504	.267	.193	.524	.376	.362
32	.647	.323	.239	.602	.391	.390
64	.734	.322	.268	.650	.401	.404