NLE's DUNE

Distilling a UNiversal Encoder from Heterogeneous 2D and 3D Teachers

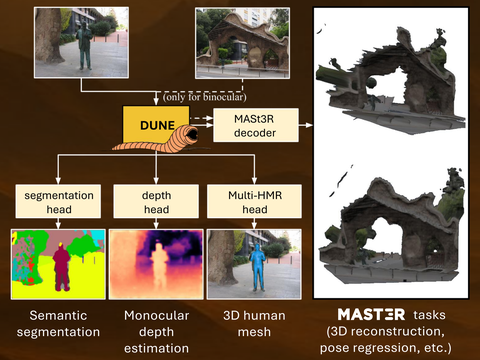
Mert Bulent SARIYILDIZ
Philippe WEINZAEPFEL
Thomas LUCAS
Pau DE JORGE
Diane LARLUS
Yannis KALANTIDIS

NAVER LABS Europe

https://github.com/naver/dune

Goal

A single encoder for diverse 2D & 3D tasks

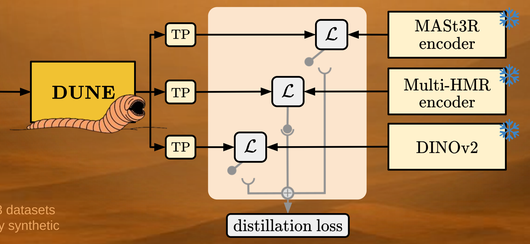


HETEROGENEOUS DATA

DINO-v2: 18M generic images (subset of original training set) from 3 datasets
Multi-HMR: 500k images from 4 datasets, centered on humans, mainly synthetic
MASt3R: 2M real and synthetic images from 12 datasets (indoor scenes, outdoor landmarks, objects, etc.)

Multi-teacher distillation

- ✓ Transformer projectors
- ✓ Teacher dropping [UNIC]
- ✓ Feature standardization



+ decoder fine-tuning

HETEROGENEOUS TEACHERS

MASTER
3D Foundation model

Multi-HMR
Human Perception model

DINOv2
2D Foundation model

Ablations

Data Sharing	ADE20K (mIoU ↑)	NYUd (RMSE ↓)	MapFree (AUC ↑)	BEDLAM (PA-PVE ↓)
No data sharing	41.6	0.426	93.2	68.7
Generic data sharing	40.1	0.416	92.7	71.7
Full data sharing	**44.9**	**0.377**	**93.7**	**68.3**

Data sharing among teachers

Distil. Data	Proj. Design	ADE20K (mIoU ↑)	NYUd (RMSE ↓)	MapFree (AUC ↑)	BEDLAM (PA-PVE ↓)
IN-19K	LP	42.4	0.446	91.4	83.9
IN-19K	TP	**44.9**	0.433	93.6	73.5
All	SP	42.3	0.413	92.2	73.1
All	LP	44.7	0.384	91.5	78.2
All	TP	**44.9**	**0.377**	**93.7**	**68.3**

Distillation data and projector design

Other results

Method	Encoder	Co3Dv2 ↑ RRA@15	Co3Dv2 ↑ RTA@15	Co3Dv2 ↑ mAA(30)	RealEstate10K ↑ mAA(30)
DUSt3R	ViT-Large	93.3	88.4	77.2	61.2
MASt3R	ViT-Large	**94.6**	**91.9**	**81.8**	76.4
DUNE	ViT-Base	92.2	90.7	78.8	**79.9**

Multi-view pose regression

Model	Cityscapes (mIoU ↑)	NYUv2 (mIoU ↑)	ScanNet (mIoU ↑)	Avg. (mIoU ↑)
Pri3D	56.3	54.8	61.7	57.6
MASt3R	58.9	60.2	57.0	58.7
DUNE (no proj.)	65.6	66.1	61.2	64.3
DUNE	**70.6**	**68.2**	**65.2**	**68.0**

Semantic segmentation

Teachers are heterogenous

PCA visualization of encoder outputs

Results with a universal encoder

Model	Encoder Arch.	Training Data	Training Res.	2D vision ADE20k (mIoU ↑)	2D vision NYUd (RMSE ↓)	3D vision BEDLAM (F1-score ↑)	3D vision BEDLAM (PA-PVE ↓)	3D vision MapFree (AUC ↑)
Teacher models								
DINO-v2	ViT-Large	LVD-142M	518	47.7	0.384	-	-	-
Multi-HMR	ViT-Large	HMR-500K	672	-	-	95	36.9	-
MASt3R	ViT-Large	MASt3R-1.7M	512	-	-	-	-	91.2
State-of-the-art ViT-Base encoders								
DINO-v2	ViT-Base	LVD-142M	518	47.3	0.399	86	76.5	89.6
AM-RADIO-v2.5	ViT-Base	DataComp-1B	512	**50.0**	0.718	89	83.2	93.1
DUNE	ViT-Base	DUNE-20.7M	336	44.9	0.377	91	68.3	93.7
DUNE	ViT-Base	DUNE-20.7M	448	45.6	0.358	**94**	**56.0**	94.7

A ViT-Base encoder for MASTER

- ✓ Comparable or improved performance
- ✓ Smaller encoder (Vit-B vs. ViT-L originally)
- ✓ State-of-the-art Visual Relocalization (Map-free)

Map-free Reloc. Leaderboard

Method	AUC (VCRE < 45px)	AUC (VCRE < 90px)	Precision (VCRE < 45px)
DUNE + MAST3R	0.840	0.943	64.4%
MAST3R (Ess.Mat + D.Scale)	0.817	0.933	63.0%
interp_metric3d_loftr_3d2d	0.681	0.796	39.9%

[UNIC] UNIC: Universal Classification Models via Multi-teacher Distillation. Sarıyıldız *et al.* ECCV'24.
[MASt3R] Grounding image matching in 3D with MASt3R. Leroy *et al. ECCV'24.*
[Multi-HMR] Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. Baradel *et al.* ECCV'24.
[DINOv2] DINOv2: Learning robust visual features without supervision. Oquab *et al.* TMLR'24.

DUNE RISE font by Fontswan
Background image created with ChatGPT-4o