ZipDepth

Lightweight zero-shot monocular depth, anywhere, on any device

Fabio Tosi • Luca Bartolomei • Matteo Poggi • Stefano Mattoccia | University of Bologna

Project Page | Code & Weights

ECCV 2026 | MALMÖ | SEPT 8–12

Foundation-model generalization, 6.1 M parameters.

55× fewer parameters (6.1M vs. 335M teacher)

217× fewer GMACs (3.0 vs. 652.7 teacher)

137× less energy/frame (397 mJ vs. 54.5 J)

+1.6–4.6 AbsRel gap to the teacher – the honest cost

Why this is hard

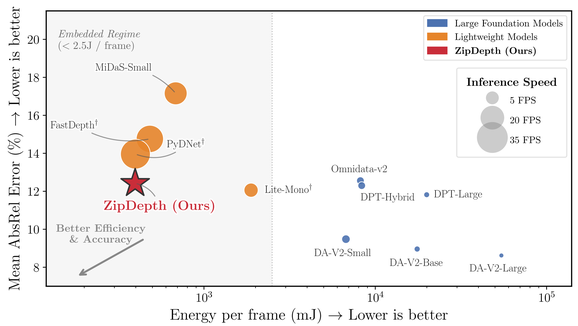
Foundation models: accurate, not deployable

Foundation models made monocular depth work *everywhere*: indoor, outdoor, synthetic, in the wild. They were never meant to run anywhere.

Lightweight models: deployable, not accurate

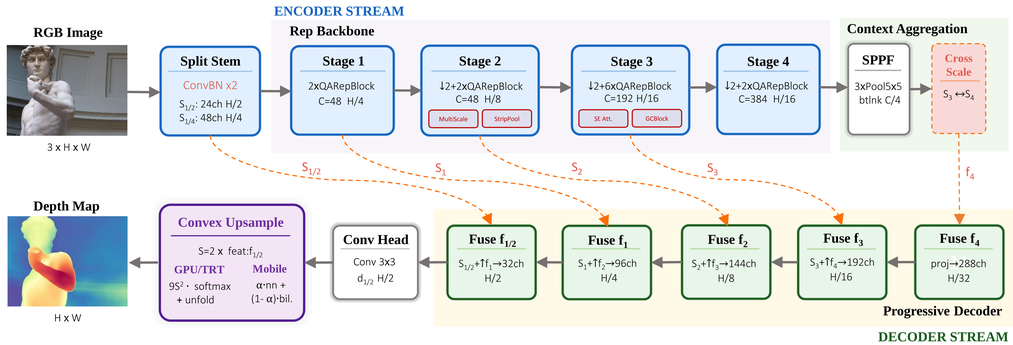
They run on device, but learn from self-supervision on *one* domain (typically KITTI). Off that domain they degrade silently.

Efficiency vs. zero-shot accuracy: still mutually exclusive.



Energy per frame vs. mean AbsRel, Jetson Orin NX @ 15 W. Bubble size = FPS.

The network



01 Split stem

The $\frac{1}{2}$-res activation is kept as a skip, for boundaries at full resolution.

Remove it: NYUv2 AbsRel $8.4 \rightarrow 9.1$.

02 Reparameterized

Three branches while training, one 3×3 conv at inference.

Exact fusion: $6.79 \rightarrow 6.14$ M parameters.

03 Linear-cost context

Dilated depthwise, strip pooling, SE, GC, SPPF. No self-attention.

Remove SPPF: NYUv2 AbsRel $8.4 \rightarrow 9.0$.

04 Dual upsampling

Convex on GPU, Conv2D-only gate on NPU, one per class of hardware.

Boundary F1 $+36\%$ over bilinear.

Distilled from **Depth Anything v2-Large** over **14.1 M images** across **17 domains** • trained on **two RTX 3090s** • exported natively to TensorRT, CoreML, ONNX and TFLite, with no custom operators.

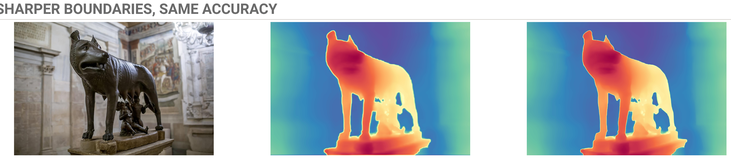
Results

ZERO-SHOT ACCURACY AND COST PER FRAME • JETSON ORIN NX @ 15 W

	EFFICIENCY				NYUv2		KITTI		ETH3D		ScanNet		DIODE	
	Par. (M)	MACs (G)	FPS	J/frame	AR↓	δ_1 ↑	AR↓	δ_1 ↑	AR↓	δ_1 ↑	AR↓	δ_1 ↑	AR↓	δ_1 ↑
LARGE PRETRAINED MODELS														
Marigold	899.2	6205.0	—	—	6.6	95.3	11.5	87.0	6.4	96.1	6.8	94.9	25.7	79.7
DPT-Large	343.0	258.5	0.7	19.9	9.1	91.9	11.0	88.2	9.0	92.8	8.4	93.1	21.6	74.4
DPT-Hybrid	122.4	128.0	1.7	8.19	9.3	91.5	13.2	84.2	9.6	92.6	8.9	92.5	21.8	73.7
DA-V2-Small	24.8	57.8	2.2	6.74	6.1	96.2	8.4	93.4	6.3	96.7	5.2	97.2	21.4	75.9
DA-V2-Large *teacher*	335.0	652.7	0.3	54.5	5.1	97.1	7.7	94.9	5.1	97.8	4.2	98.0	21.0	76.7
LIGHTWEIGHT MODELS, RETRAINED ON OUR DATA														
Lite-Mono†	8.3	13.5	7.3	1.88	8.7	92.7	10.9	89.5	8.7	93.6	9.6	90.2	22.4	72.9
MiDaS-Small	21.3	4.6	20.1	0.68	11.5	86.8	25.9	54.3	13.4	83.7	10.8	87.9	24.2	71.1
GuideDepth†	5.8	5.1	16.6	0.86	9.2	91.8	14.9	80.6	10.2	91.9	9.6	90.5	22.7	72.4
FastDepth†	4.0	2.2	28.9	0.48	10.3	89.8	16.8	74.3	12.3	87.2	11.0	87.6	23.4	71.0
PyDNet†	1.9	5.5	34.8	0.40	11.4	87.1	12.9	84.9	10.4	90.6	11.7	85.8	23.4	70.8
ZipDepth	**6.1**	**3.0**	**34.4**	**0.40**	**8.4**	**93.3**	**12.3**	**86.4**	**10.0**	**92.2**	**8.8**	**92.1**	**22.6**	**73.9**

1st 2nd 3rd within each group, ranked separately. † Lightweight baselines retrained from scratch on our 14.1 M images, so the comparison isolates architecture from data.

SHARPER BOUNDARIES, SAME ACCURACY



RGB | Bilinear | Convex, ours

Boundary F1 on UnrealStereo4K: **0.088** bilinear → 0.105 NPU path → **0.120** GPU path, $+36\%$, while NYUv2 AbsRel moves only from 8.3 to 8.4.

One model, any scene

no fine-tuning, no per-domain weights

Motion blur | Underwater | Reflective | Aerial | Adverse weather

FPS, fused FP32 → best backend 389 → **1317** RTX 3090 • 351 → **773** RTX 3070 • 118 → **196** Orin 50 W • 34 → **77** Orin 15 W • 50 → **75** server CPU • 40 → **45** laptop CPU • 360 → **715** iPad M4 • 240 → **375** iPhone 12 • **16** Android GPU