ECCV 2026 · Poster · #11927

Towards Robust Driving Perception:
A Flexible Scale-Driven Family for Self-Supervised Monocular Depth Estimation

Zhaowen Zhu[1,3]*, **Li Zhang**[1]†, **Yujie Chen**[1]*, Tian Zhang[1,4], Yingjie Wang[2], Mingxia Zhan[1]

[1]Hefei University of Technology, Hefei, China [2]Nanjing University of Posts and Telecommunications, Nanjing, China
[3]Chengpin Home Tech, Nanjing, China [4]Differential Robotics, Hangzhou, China
** Equal contribution* *† Corresponding author: lizhang@hfut.edu.cn*

Hefei University of Technology

Nanjing Univ. of Posts & Telecommunications

Project Page & Demo Videos
startnew.github.io/projects/flexdepth
Code: github.com/startnew/flexdepth

Motivation

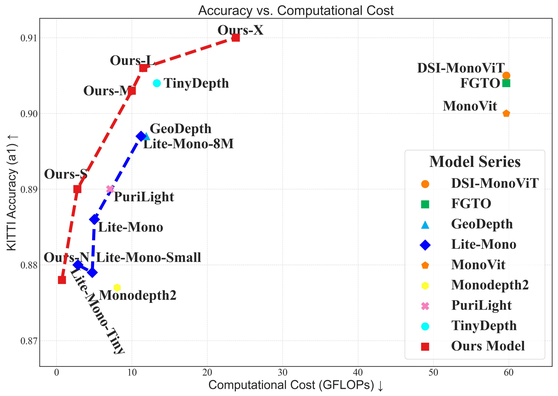
- Self-supervised monocular depth models are **locked to a single scale** and degrade sharply in dynamic driving scenes.
- Dynamic-scene-specialized networks are **too heavy for edge devices** in vehicles.

→ **One model family: every scale, robust to dynamics.**

Contributions

- **FlexDepth** — a scale-driven lightweight self-supervised depth **family** (Nano → X-Large) studying the accuracy–efficiency trade-off.
- **Two-stage static–dynamic decoupled training** — independent confidence evaluation for static background and dynamic objects.
- **Scale-Driven Decoder (SDD)** — efficient inference + accurate depth, elastically configurable per device.

Accuracy vs. Compute — KITTI



FlexDepth (red) dominates the accuracy–compute frontier at every scale, from 0.7 GFLOPs (Nano) to 24.6 GFLOPs (X-Large).

Model Family & Efficiency

Model	Params	GFLOPs	4090	V100	SD8Elite
Flex-Nano	**1.5M**	**0.7**	**1583**	**774**	**37.6**
Flex-Small	**6.1M**	**2.8**	**1215**	**606**	**18.6**
Flex-Medium	**12.7M**	**10.0**	**510**	**231**	**5.8**
Flex-Large	**15.2M**	**11.5**	**467**	**216**	**5.2**
Flex-X-Large	**32.3M**	**24.6**	**321**	**130**	**3.0**
Lite-Mono-Tiny	2.1M	2.8	1048	312	15.9
MonoViT	33.6M	59.7	265	126	1.9

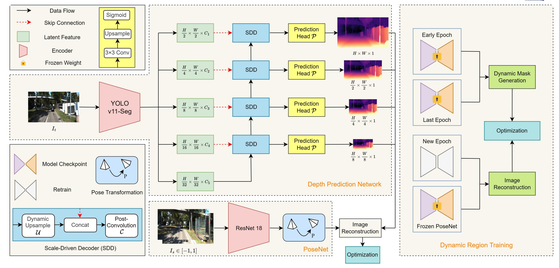
640×192 input; FPS, GPU: batch 16; Snapdragon 8 Elite: batch 1, performance mode.

Zero-Shot Generalization — Make3D

Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓
Monodepth2	0.322	3.589	7.417	0.163
GeoDepth	0.296	2.750	6.735	0.153
Flex-X-Large (Ours)	**0.279**	**2.799**	**6.599**	**0.144**

KITTI-trained models tested directly on Make3D — no fine-tuning.

FlexDepth Overview — Two-Stage Training



Stage 1: depth network (YOLOv11-Seg encoder + SDD) and PoseNet are trained jointly with photometric self-supervision. Stage 2: motion-induced inconsistency between early- and late-epoch predictions yields dynamic-object masks; with a frozen PoseNet the depth network is retrained — zero extra inference cost.

Scale-Driven Decoder (SDD)

- **No pre-convolution** — global semantics are preserved before spatial expansion (removes the bottleneck of classic upsample–concat pipelines).
- **Dynamic upsampling** — learned content-aware offsets + Pixel Shuffle + grid sample give sharp, non-uniform upsampling and crisp object boundaries.
- **Two post-conv bottlenecks** — HEB (dense, short gradient paths) for small models; HPB (CSP-style, fine-grained) for large ones.
- **Inverted prediction head** (upsample → predict) for larger models; classic head keeps Nano/Small real-time.

Decoder Configuration per Family Member

Family member	Post-Conv C	Upsampling U	Head P
Nano / Small	HEB	Dynamic	Classic
Medium / Large / X-Large	HPB	Dynamic	Inverted

Static–Dynamic Decoupled Training

- **Stage 1 —** photometric loss (SSIM + L1) with minimum reprojection & auto-masking, edge-aware smoothness; depth net + PoseNet (ResNet-18, 6-DoF) trained jointly.
- **Stage 2 —** dynamic regions betray themselves: early- vs late-epoch disparity disagrees. Mask M uses an **adaptive per-pixel threshold**; retrain with masked normal-distribution loss + geometric depth smoothness.
- **No auxiliary signals** — no semantics, no optical flow, no pseudo depth; single frame at test time.

$$\mathbf{M} = [\,|\mathrm{disp}_t^{\beta} - \mathrm{disp}_{t'}^{\beta}| < g(\varphi(f_{S_t}, f_{S_{t'}}))\,]$$

β = 0.7; g(φ(·)): adaptive per-pixel threshold. Static where early/late-epoch predictions agree; dynamic otherwise.

Training Recipe

- **NAdam optimizer.** Epochs (Stage 1 + Stage 2): KITTI 30 + 20, Cityscapes 30 + 10.
- **Batch 12** (Nano/Small; encoder lr 5e-5, decoder 1e-4), **batch 6** (Medium and above; lr 5e-5).
- Flex-X-Large trains on KITTI in **~10 h on a single RTX 4090.**
- **Monocular videos only** — KITTI Eigen split & Cityscapes; no extra labels.

Results — KITTI (Eigen Split, 640×192)

Method	Abs Rel↓	Sq Rel↓	RMSE↓	δ<1.25↑	GFLOPs↓
Monodepth2 ('19)	0.115	0.903	4.863	0.877	8.0
Lite-Mono-Tiny ('23)	0.110	0.837	4.710	0.880	2.8
PuriLight ('26)	0.106	0.747	4.536	0.890	7.1
Flex-Nano (Ours)	**0.110**	**0.794**	**4.678**	**0.878**	**0.7**
Flex-Small (Ours)	**0.104**	**0.713**	**4.458**	**0.890**	**2.8**
GeoDepth ('25)	0.100	0.694	4.381	0.897	11.9
TinyDepth ('24)	0.096	0.665	4.249	0.904	13.3
Flex-Medium (Ours)	**0.096**	**0.639**	**4.253**	**0.903**	**10.0**
Flex-Large (Ours)	**0.095**	**0.642**	**4.199**	**0.906**	**11.5**
MonoViT ('22)	0.099	0.708	4.372	0.900	59.7
DSI-MonoViT ('25)	0.096	0.711	4.321	0.905	59.7
Self-Distillation ('25)	0.095	0.619	4.156	0.905	64.5
Flex-X-Large (Ours)	**0.093**	**0.605**	**4.114**	**0.910**	**24.6**

Results — Cityscapes (416×128)

Method	Abs Rel↓	Sq Rel↓	RMSE↓	δ<1.25↑	GFLOPs↓
ManyDepth ('21) 2f	0.114	1.193	6.223	0.875	12.1
DynamicDepth ('22) 2f+Se	0.103	1.000	5.867	0.895	18.5+
PuriLight ('26)	0.099	1.014	5.892	0.901	5.7
Flex-Nano (Ours)	**0.107**	**1.261**	**6.133**	**0.893**	**0.6**
Flex-Small (Ours)	**0.100**	**1.078**	**5.813**	**0.904**	**2.2**
Flex-Medium (Ours)	**0.089**	**0.885**	**5.358**	**0.917**	**8.0**
Flex-Large (Ours)	**0.087**	**0.911**	**5.310**	**0.924**	**9.2**
ProDepth ('24) 2f	0.095	0.876	5.531	0.908	35.5
DSI-MonoViT ('25)	0.088	0.872	5.410	0.918	47.7
Flex-X-Large (Ours)	**0.086**	**0.877**	**5.268**	**0.926**	**19.6**

M: monocular video only. 2f: two test frames. +Se: extra semantic segmentation.

Dynamic-Region Evaluation — Cityscapes

Method	Aux.	Dyn AbsRel↓	Dyn δ↑	Sta AbsRel↓	Sta δ↑	GFLOPs↓
ProDepth	–	0.115	0.884	0.092	0.911	35.5
Mining-BrNet	✓	0.119	0.872	0.080	0.922	27.0
DSI-MonoViT	–	0.097	0.918	0.086	0.919	47.7
Flex-X-Large (Ours)	**–**	**0.091**	**0.935**	**0.085**	**0.925**	**19.6**

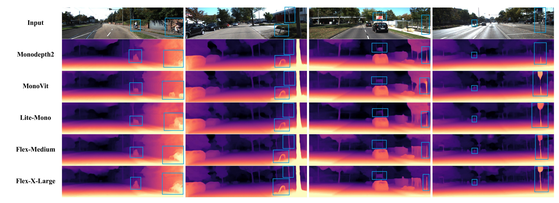
Aux.: pixel motion / pseudo depth. Best dynamic-region accuracy with the lowest compute — using no auxiliary cues.

vs. Foundation Models (Depth Anything V2)

Method	Type	Params	GFLOPs	AbsRel↓	δ<1.25↑
DA2 ViT-L @1722×518	Zero-shot	335M	1947	0.070	0.956
DA2 ViT-S @1722×518	Zero-shot	25M	137	0.077	0.944
DA2 ViT-L @644×196	Zero-shot	335M	276	0.092	0.915
DA2 ViT-S @644×196	Zero-shot	25M	19	0.110	0.881
Flex-X-Large† (Ours)	**Self-sup.**	**32M**	**25**	**0.063**	**0.952**

KITTI, LS alignment + improved GT. † ~1/10 params & ~1/78 FLOPs of DA2 ViT-L, yet lower Abs Rel (0.063 vs 0.070).

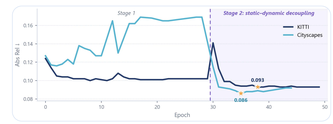
Qualitative — KITTI



vs. Monodepth2 / MonoViT / Lite-Mono: sharper thin structures and more complete objects (blue boxes).

Dynamic Objects Across Training

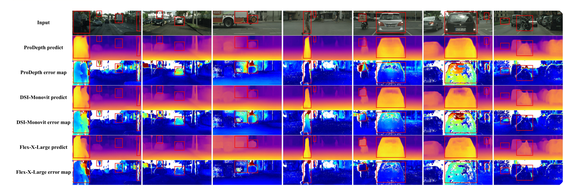
Dynamic-object masks emerge from early/late-epoch disparity inconsistency and are refined as training progresses (black = dynamic).



PLACEHOLDER
Per-epoch depth predictions of dynamic objects — drop your figure here.
占位：各 epoch 动态物体深度预测对比图

Stage-2 static–dynamic decoupling (shaded) further reduces Abs Rel on both datasets — KITTI 0.102 → 0.093, Cityscapes 0.115 → 0.086.

Qualitative — Cityscapes (Dynamic Scenes)



Depth & error maps (blue = small, red = large error) vs. ProDepth and DSI-MonoViT. Flex-X-Large is cleanest on dynamic objects (red boxes).

Ablation Highlights (Flex-X-Large, KITTI)

- **Full model: 0.093 Abs Rel / 0.910 δ** — best overall.
- **w/o Stage-2 decoupling: 0.100 / 0.900** — largest drop.
- **Gain ≠ more training:** Stage-1 + fine-tune = 0.102 vs. Stage-1+2 = 0.093.
- **Adaptive mask > fixed threshold:** Sq Rel 0.605 vs. 0.629.

Conclusion

- **FlexDepth** — a flexible scale-driven self-supervised depth family for robust driving perception.
- **SOTA at every scale** on KITTI & Cityscapes — no auxiliary info, single frame.
- **Strong on dynamics** (CS dynamic δ = 0.935) and zero-shot (Make3D Abs Rel 0.279).
- **Edge-ready:** Flex-Nano 0.7 GFLOPs → 37.6 FPS on Snapdragon 8 Elite.

Acknowledgement & Links

Supported by the National Natural Science Foundation of China (Grant 62332016).
Code: github.com/startnew/flexdepth
Project & demo videos: startnew.github.io/projects/flexdepth