Cofinanțat de Uniunea Europeană

HUB Român de Inteligență Artificială
cod SMIS: 351416

Metric Monocular Depth Estimation for UAVs

Horatiu Florea
PhD Supervisor: Prof. Dr. Eng. Sergiu Nedevschi

Technical University of Cluj-Napoca
Image Processing and Pattern Recognition Research Center

11/6/25

Workshop Hub Român Inteligență Artificială
Intelligent and Autonomous Systems

Monocular Depth Estimation

- Enabler for many key tasks in perception systems for autonomous vehicles:
  3D scene reconstruction, object localization, etc.
- Camera-**only** based depth estimation, although ill-posed problem, presents advantages
- Data related challenges remain: quantity, diversity, quality **– especially for UAV data**

- Recent trends:
  - focus on zero-shot generalization capabilities
  - training on massive, diverse datasets, adaptation to camera model, conditions
  - wider (re)adoption of synthetic data in supervised training
  - alternatively, self-supervised training is sought for greater training flexibility
  - new directions: sequence consistency (video), diffusion-based architectures

Demand for Metric Scale Recovery

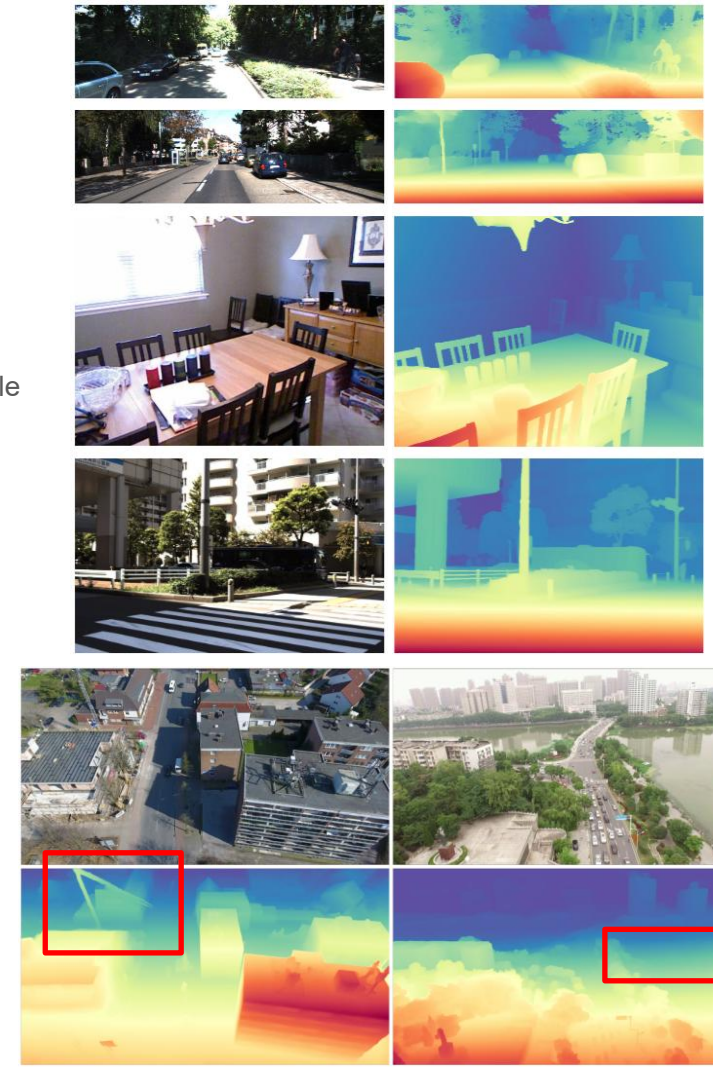
- Themes in MDE: Reconstruct Anything
  - 2024: Breakthroughs in zero-shot generalization: Depth Anything, UniDepth
- Depth Anything: Supervised Relative MDE foundation model

  Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J. and Zhao, H., 2024. **Depth anything**: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

  - Impressive results at the expense of providing scale
  - Idea: train over vast (62M) unlabeled data with heavy augmentations
  - Diverse depth ranges can hinder learning across datasets

  ➔ MiDaS: Scale and shift invariant loss: zero median, unit scale

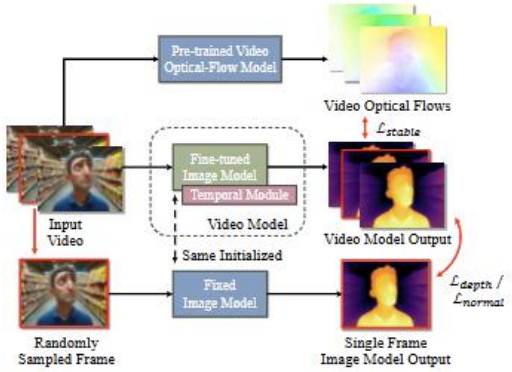
- Critical enabler for adopting Relative MDEs: ability to scale relative results to metric values
  - Either as a post-processing operation or through network modifications / training



Generalization of Depth Anything V2 (Relative model)

2025 Themes in MDE

- 📉 Reduced focus on self-supervised, depth completion
- 📊 Continued work on universal monocular metric depth [UniDepth 2, UniK3D]
- NEW New avenues on *unified* perception – e.g. Depth + Optical Flow + 2D/3D Tracking [L4P]
- 📈 Increased attention: **Video Depth**

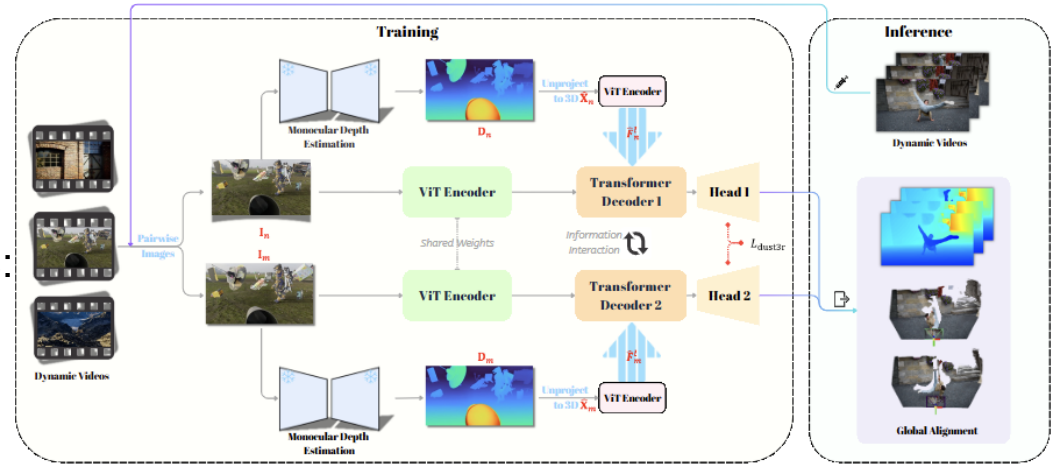


Buffer Anytime: Zero-Shot Video Depth and Normal from Image Priors

- Pretrained MDE with added temporal layers, trained without depth/normal data
- Training supervision: auxiliary Optical Flow net for consistency,
  MDE net for maintaining alignment
- Up to 110 frames, producing relative estimations

Align3R: Aligned MDE for Dynamic Videos

- Based on DUSt3R (Multi-View ➔ 2D-3D mappings for 3D vision)
- Injects encoded features from auxiliary depth estimations
- Adapts DUSt3R's multi-frame global optimization into two phases:
  global (keyframe level) and local (remaining frames)
- Relative estimations



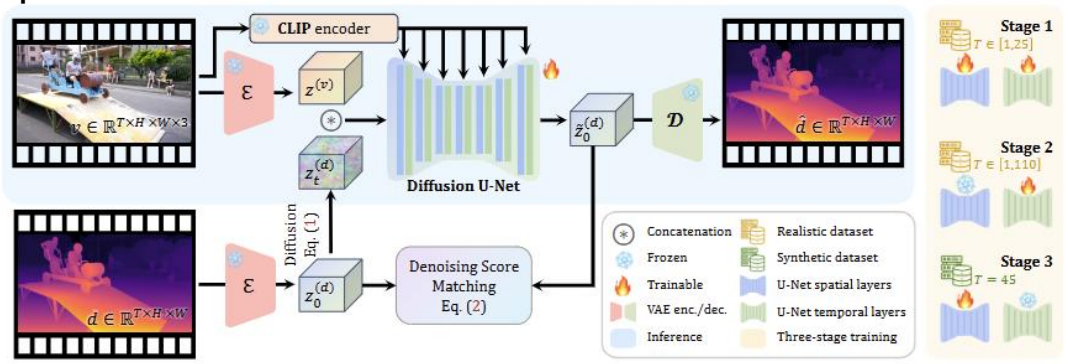
Video methods selected from CVPR 2025 accepted papers

2025 Themes in MDE

- 📈 Increased attention: **Video Depth**

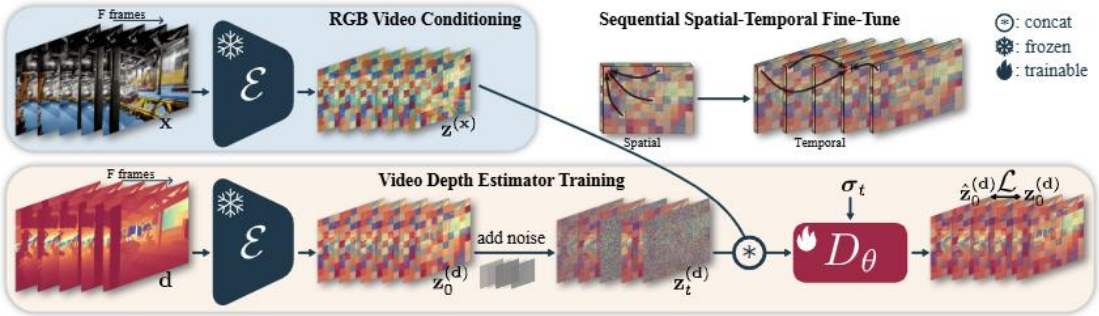
DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos

- Repurposes [SVD] image-to-video latent diffusion for video-to-depth
- 3-stage training: spatial+temporal ➔ temporal ➔ spatial layers
- Up to 110 frames, relative estimations
- Inference strategy for stitching long videos: latent noise is initialized using previous results with added noise in overlapping regions



ChronoDepth: Learning Temporally Consistent Video Depth from Video Diffusion Priors

- Similar general approach to DepthCrafter, only 2 train stages (spatial ➔ temporal), trains decoder as well
- Improved inference strategy aimed at streaming applications:
  On overlapping frames, direct conditioning using previous depths ⇔ for this, training features independently sampled noise for each frame
- Relative output, improved results over DepthCrafter



Video methods selected from CVPR 2025 accepted papers

Baseline Metric Scaling Methods

- Offline Evaluation – scale each prediction to match Ground Truth d*

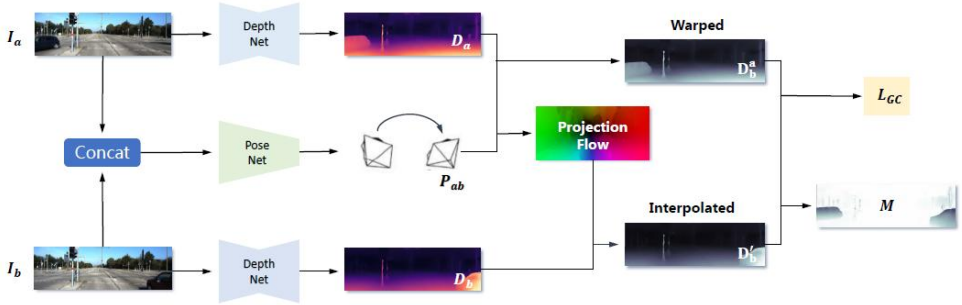
$$(s,t) = \arg\min_{s,t} \sum_{i=1}^{M} (s\mathbf{d}_i + t - \mathbf{d}_i^*)^2 \qquad \hat{\mathbf{d}} = s\mathbf{d} + t$$

  - Not applicable to online, real-world use as GT is not available

- Naive online method – Fixed Factor Scaling
  - Evaluate model on representative validation set with GT and store median (s, t) for reuse
  - In practice, this leads to poor performance
    - Output scale varies across domains, scenes and even between consecutive frames
    - For self-supervised training, methods for ensuring scale consistency are an active research topic

Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M. M., & Reid, I. (2019). **Unsupervised scale-consistent depth and ego-motion learning from monocular video**. Advances in neural information processing systems, 32, 35-45.

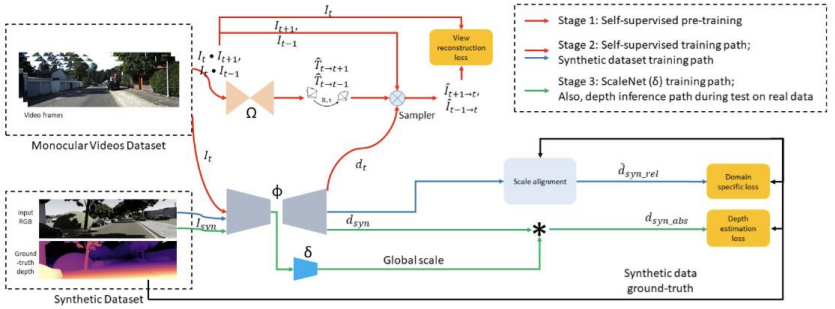
- Estimate depth for consecutive frames and force consistency between them (after pose alignment)

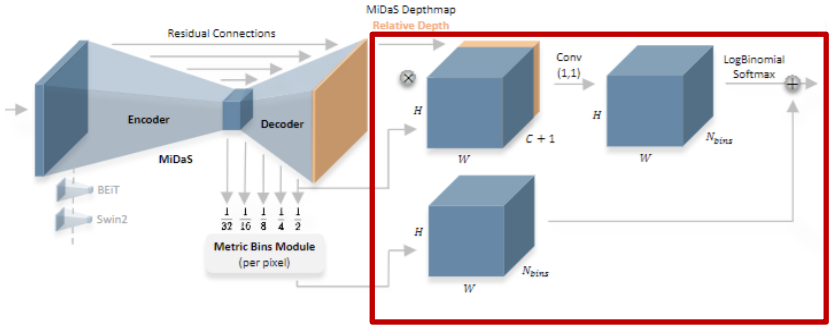


Learning Metric Scale from Metric GT

Leveraging Synthetic Data - Train a third, scale estimation network trained using only synthetic data

Swami, K., Muduli, A., Gurram, U. and Bajpai, P., 2022. **Do what you can, with what you have**: Scale-aware and high quality monocular depth estimation without real world labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 988-997).





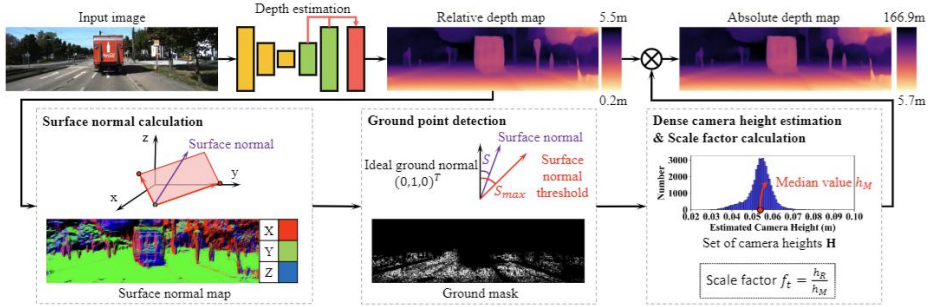
Scaling Head

Additional Metric Scaling Heads - When metric data is available, a metric scaling head(s) can be trained, using intermediary decoder outputs

Bhat, S.F., Birkl, R., Wofk, D., Wonka, P. and Müller, M., 2023. **Zoedepth**: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288.

Camera Height Scaling as Post-Processing - Use ground points in relative estimation to determine camera height and scale it to match known height

Xue, F., Zhuo, G., Huang, Z., Fu, W., Wu, Z. and Ang, M.H., 2020, October. **Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications**. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2330-2337). IEEE.

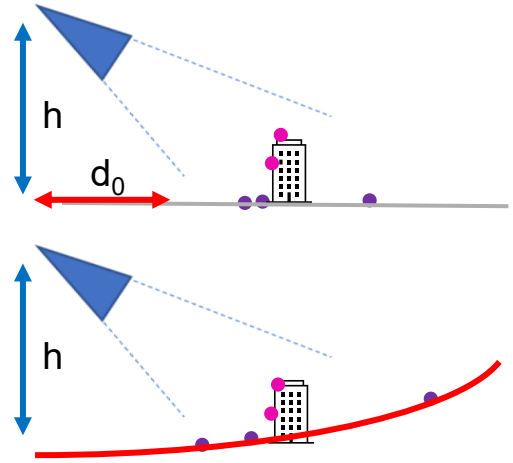


AGL-based Metric Scaling for UAVs

- Indoor UAV scenarios: use downward Time-of-Flight sensor to determine height and use it for scaling

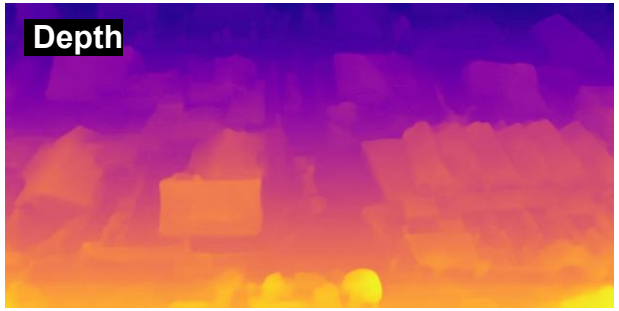
Pan, Y., Liu, B., Liu, Z., Shen, H., Xu, J., Fu, W. and Yang, T., 2024. MoNA Bench: A Benchmark for Monocular Depth Estimation in Navigation of Autonomous Unmanned Aircraft System. Drones, 8(2), p.66.

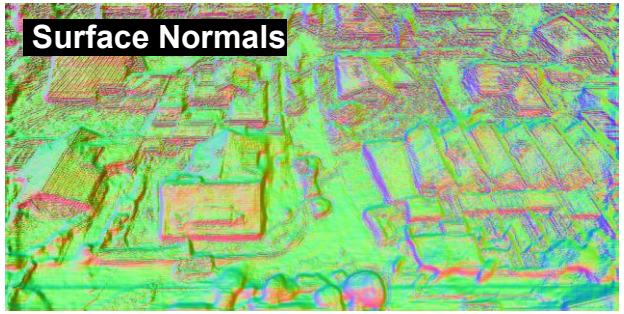
- Our initial task: use known height (Above Ground Level - AGL) for scaling outdoors
  - Ground selection based on surface normals and semantic segmentation
  - Scaling factor is searched to align computed height with AGL

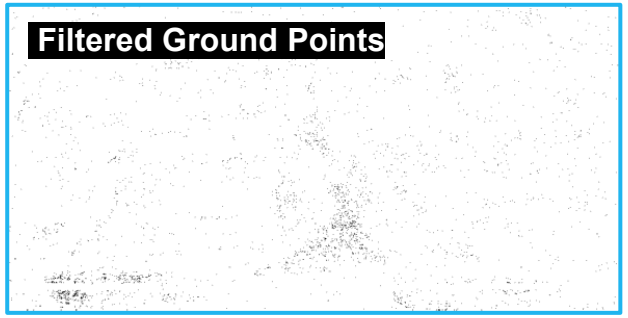


- Requires AGL information – additional Radar altimeter or preloaded data
- Can be affected by distance to observed scene and terrain variations



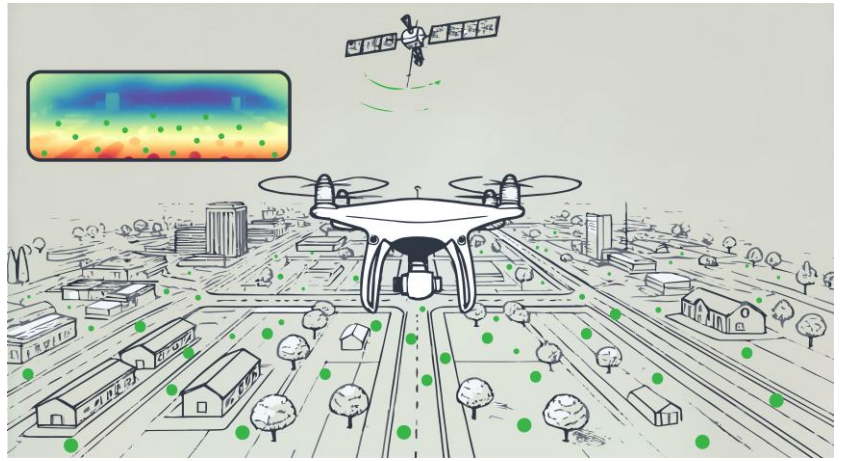






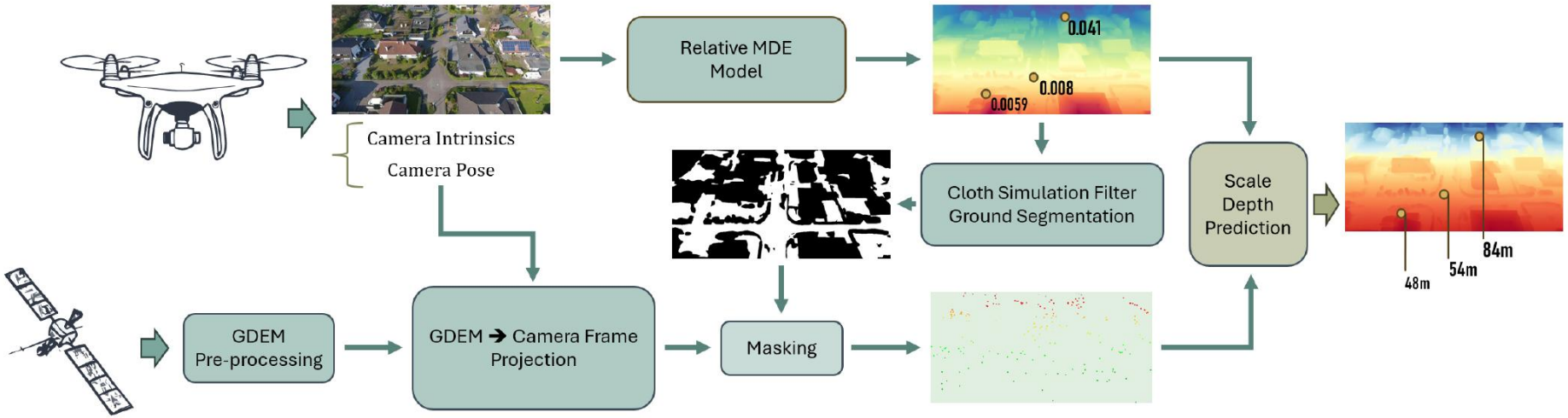
TanDepth: Leveraging Global DEMs for Metric Depth

Florea, H. and Nedevschi, S., 2025. **TanDepth: Leveraging global DEMs for metric monocular depth estimation in UAVs**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.



- General scale recovery method tailored for aerial scenes based on Global Digital Elevation Models (GDEMs)
  - Once projected in the image, the (densified) GDEM acts as a sparse Ground Truth against which scaling can take place

- Designed for *all types of Relative MDE* methods: self-supervised or supervised relative depth estimators
  - Scaling by two parameters: scale and shift (MiDaS)



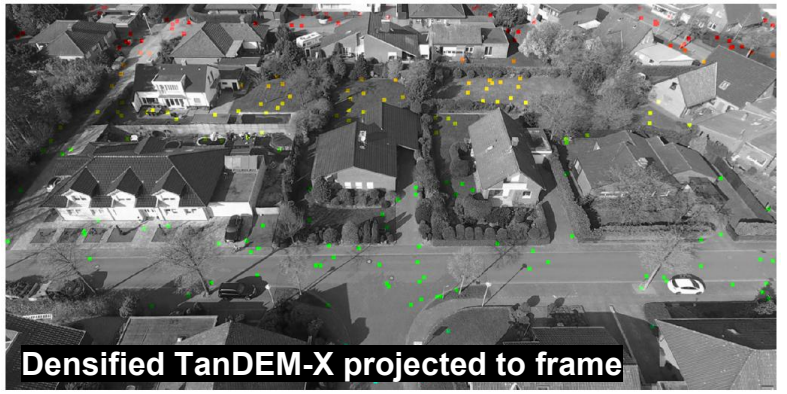
TanDepth: Leveraging Global DEMs for Metric Depth

- Filtering out non-ground points improves scaling performance –typically requires segmentation net.
  - ➢ Developed adaptation for Cloth Simulation Filter (CSF) ground segmentation for direct use on 3D point clouds derived from Relative MDEs



- Typical GDEM data: 30m resolution – can be too low at times
  - Risk of having few/none ground GDEM points visible in frame

➢ Prior to flight, GDEM is densified by sampling an interpolated 3D mesh computed on the original points

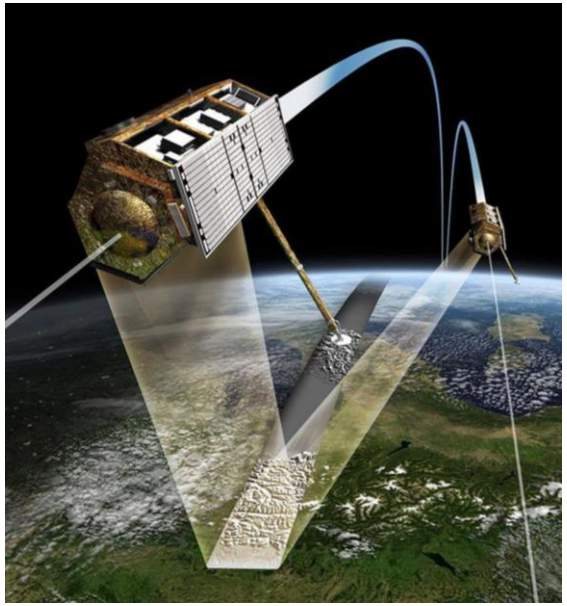


TanDepth: Leveraging Global DEMs for Metric Depth

- Tested Depth Anything V1 paired with DLR's TanDEM-X GDEM data
  - Public release (Nov. 2023): 30m resolution; 12m commercially available

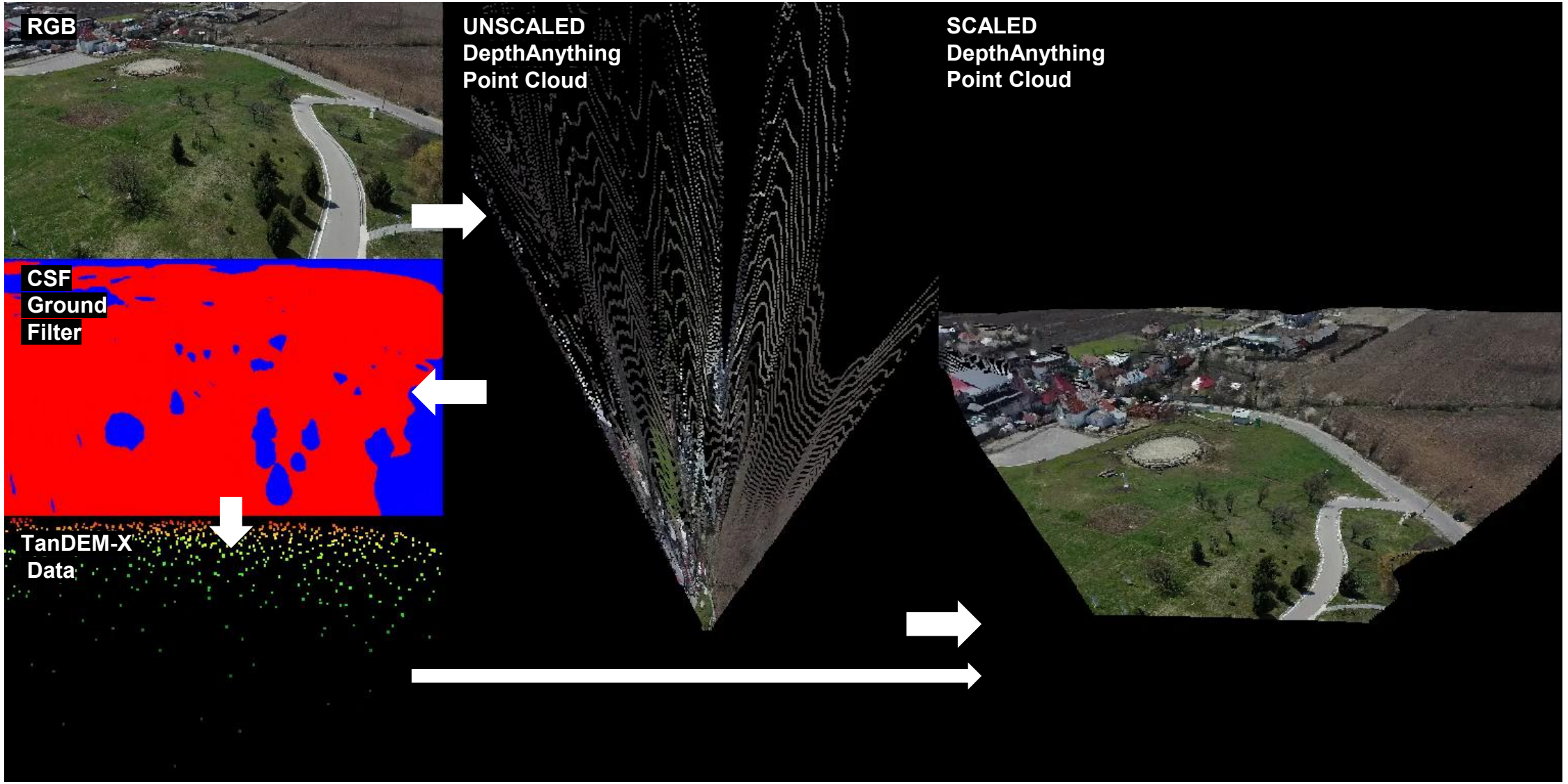
- TanDepth outperforms other online scaling methods, approaching the results obtained with offline Ground Truth scaling
- Scaling performance generalizes well across a variety of scenes, in multiple datasets with varying terrain profiles

Scene	Scaling	Lower is better ↓				Higher is better ↑					
		AbsRel	SqRel	RMSE	LogRMSE	δ_1	δ_2	δ_3	$\bar{\delta}_1$	$\bar{\delta}_2$	$\bar{\delta}_3$
UAVid Germany	Fixed	0.062	0.424	4.693	0.072	0.983	0.999	1.000	0.265	0.494	0.714
	Cam. Height	0.039	0.262	3.854	0.055	0.988	0.999	1.000	0.519	0.763	0.878
	ZoeDepth*	0.041	0.218	3.659	0.054	0.992	0.999	1.000	0.358	0.720	0.899
	TanDepth	**0.031**	**0.143**	**2.800**	**0.044**	**0.993**	**1.000**	**1.000**	**0.523**	**0.863**	**0.944**
	GT	*0.027*	*0.133*	*2.720*	*0.041*	*0.994*	*1.000*	*1.000*	*0.654*	*0.879*	*0.938*
UFO Depth Oveselu	Fixed	0.218	6.966	25.147	0.290	0.485	0.800	0.972	0.076	0.148	0.213
	Cam. Height	0.112	2.097	12.233	0.124	0.885	0.996	1.000	0.184	0.336	0.462
	ZoeDepth*	0.171	2.933	14.866	0.185	0.707	0.998	1.000	0.069	0.140	0.212
	TanDepth	**0.048**	**0.344**	**5.498**	**0.061**	**0.997**	**1.000**	**1.000**	**0.330**	**0.592**	**0.784**
	GT	*0.040*	*0.285*	*5.057*	*0.053*	*0.997*	*1.000*	*1.000*	*0.430*	*0.701*	*0.849*
UFO Depth Chilia	Fixed	0.425	28.214	67.528	0.614	0.036	0.295	0.680	0.001	0.002	0.004
	Cam. Height	0.052	0.675	9.078	0.066	0.991	0.999	1.000	0.347	0.602	0.774
	ZoeDepth*	0.116	3.072	23.232	0.151	0.850	0.998	1.000	0.095	0.201	0.319
	TanDepth	**0.040**	**0.442**	**7.520**	**0.055**	**0.994**	**0.999**	**1.000**	**0.427**	**0.715**	**0.872**
	GT	*0.035*	*0.362*	*7.064*	*0.049*	*0.995*	*0.999*	*1.000*	*0.496*	*0.779*	*0.904*
WildUAV	Fixed	0.345	7.052	18.334	0.304	0.295	0.913	0.997	0.007	0.015	0.026
	Cam. Height	0.207	9.483	10.981	0.178	0.828	0.959	0.966	0.116	0.230	0.345
	ZoeDepth*	0.502	14.876	27.551	0.415	0.062	0.673	0.977	0.001	0.001	0.003
	TanDepth	**0.072**	**1.269**	**4.106**	**0.085**	**0.962**	**0.996**	**0.997**	**0.322**	**0.531**	**0.670**
	GT	*0.041*	*0.157*	*2.560*	*0.052*	*0.995*	*1.000*	*1.000*	*0.448*	*0.708*	*0.846*



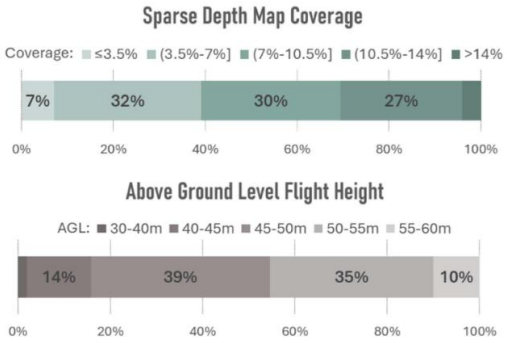
Images: DLR https://tandemx-science.dlr.de/

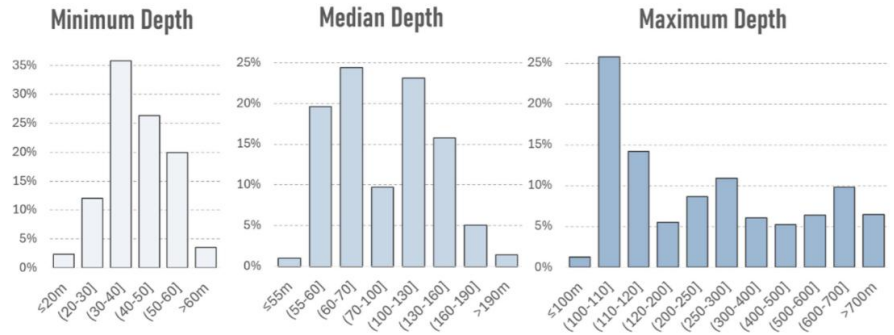
TanDepth: Demo



TanDepth: UAVid-3D-Scenes Dataset

- Depth-focused extension of the UAVid dataset for real-world UAV depth estimation
- Built using COLMAP: ~5500 RGB–depth pairs + intrinsics and poses
- Includes both dense and sparse reconstructions from 12 urban/suburban scenes in Germany and China.
- Metric scale through alignment with aerial LiDAR (Germany) and landmarks (China)
- Enables research on metric monocular depth estimation and joint depth–semantic learning in aerial imagery.









Future Prospects & Applications

- Main computational bottleneck: CSF segmentation – using a coarser mesh speeds up CSF, at the cost of segmentation stability
  - Better: lower the resolution of the input depth map ➔ **2.8x** speedup, **0.04** ↗ RMSE
- Enhancing performance by integrating metric priors further upstream in the network (from post-processing step)
- Leveraging latest approaches for video consistency could also improve results

Use cases:

- ➢ Land Monitoring (landslides, deforestation, etc.)
- ➢ Flood Extent Modelling/Damage Assessment
- ➢ Infrastructure Deformation Monitoring
- ➢ Solar Power Potential Estimation (shadow prediction, roof surface measurements
- ➢ Terrain Traversability Analysis for Path Planning



Thank you!

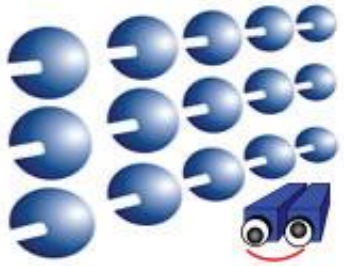


Image **P**rocessing and **P**attern **R**ecognition
Research **C**enter
horatiu.florea@cs.utcluj.ro


HUB ROMÂN
DE INTELIGENȚĂ ARTIFICIALĂ