Towards Explainability in Monocular Depth Estimation

Vasileios Arampatzakis[1,2], George Pavlidis[2], Kyriakos Pantoglou[2], Nikolaos Mitianoudis[1], and Nikos Papamarkos[1]

[1]Democritus University of Thrace, Xanthi, Greece
[2]Athena Research Center, Xanthi, Greece

Monocular Depth Estimation

- Inferring depth information from 2D images
- Crusial for: Robotics, Autonomous driving, Augmented reality

- An inherently ill-posed problem
- Ambiguities caused by the projection of the 3D world to 2D images
- Significance: Enhances scene understanding and 3D perception

- Deep Learning-based methods outperform traditional approaches
- CNNs, Vision Transformers capture complex patterns in images

Typical Explainability Methods

- Unveiling the rationale behind model predictions
- Crusial for: Transparency & trust, Model improvement, Bias detection, User understanding, Regulatory compliance

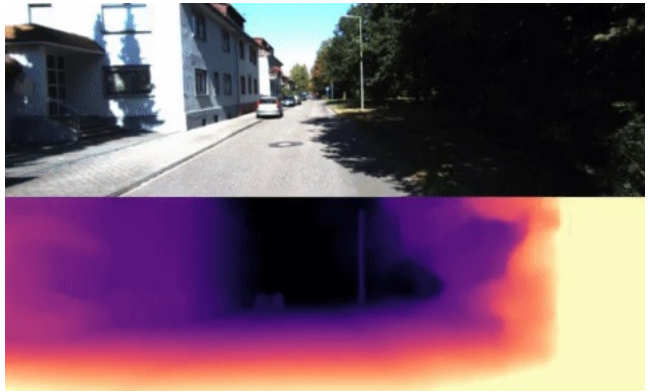
- Methods: Feature visualization, Saliency maps, Attention mechanisms, LIME (Local Interpretable Model-agnostic Explanations), SHAP (Shapley Additive exPlanations), Grad-CAM (Gradient-weighted Class Activation Mapping)

- Trade-offs between simplicity and accuracy in explanation methods
- Some methods are model-specific, while others are model-agnostic
- The need to validate explanations and ensure they reflect true model behavior

Unique Approach: Connecting with Human Perception

- **Objective**: Enhance explainability by aligning model predictions with how humans perceive depth
- **Key idea**: *As a dataset is limited to provide only a single cue, the accuracy of the methods indirectly reflects their success in learning the selected depth cue.*

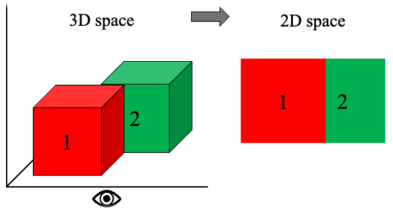
Input: RGB image
(multiple visual depth cues)



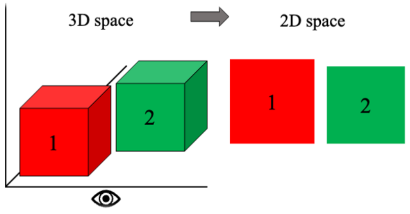
Output: depth map

Cutting & Vishton (1995): Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth

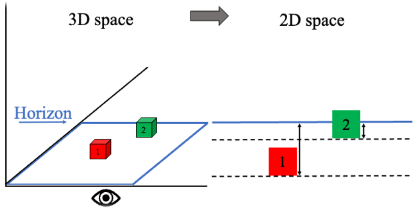
Visual Depth Cues



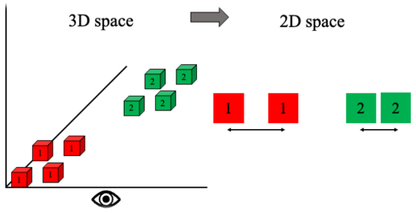
Occlusion



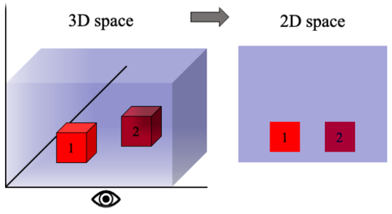
Relative Size



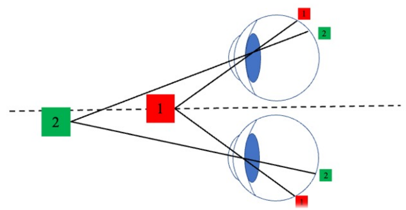
Height in visual field



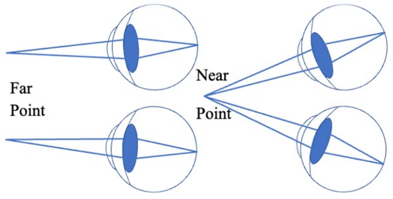
Relative Density



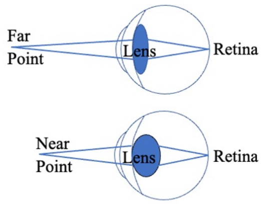
Aerial Perspective



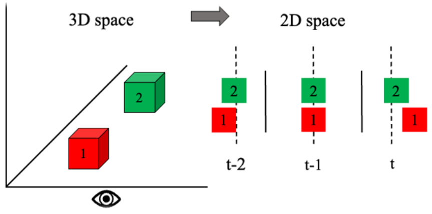
Binocular Disparity



Convergence



Accommodation

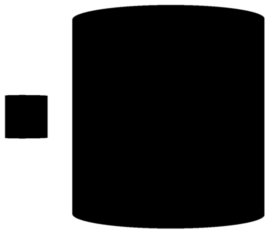


Motion Perspective

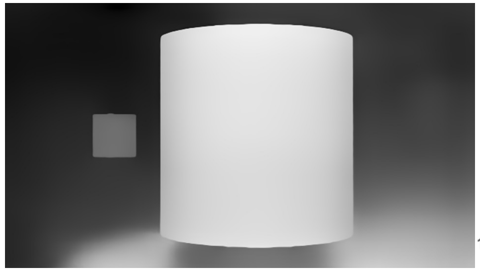
Nagata (1987): *How to reinforce perception of depth in single two-dimensional pictures*

Visual Depth Cue (VDC) Dataset

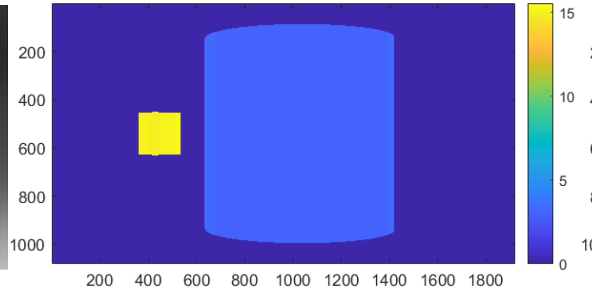
- A synthetic resource for monocular depth estimation inspired by human perception (Cutting & Vishton, Nagata)
- Exclusive depth cue representation in each image
- Relative Size (≈ 23800 images):
  - 2D images of black cylindrical objects at various distances against a white background, created through perspective projections of the corresponding virtual 3D scenes



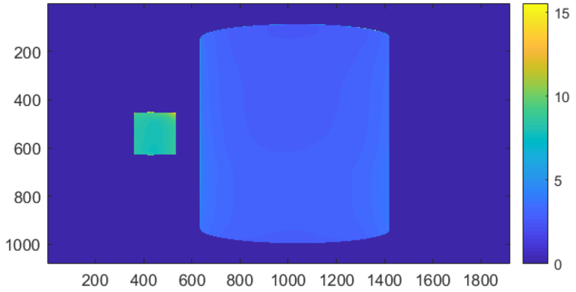
original image



predicted depth



pseudo-coloured groundtruth depth



pseudo-coloured masked predicted depth

Metrics

d_i: groundtruth depth value
$\hat{d}_i$: predicted depth value
N: number of samples

- **Absolute Relative Error:**

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|d_i - \hat{d}_i\right|}{d_i}$$

- **Squared Relative Error:**

$$SqRel = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(d_i - \hat{d}_i\right)^2}{d_i}$$

- **Linear Root Mean Squared Error:**

$$RMSE(lin) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_i - \hat{d}_i\right)^2}$$

Metrics

d_i: groundtruth depth value
$\hat{d}_i$: predicted depth value
N: number of samples

- **Logarithmic Root Mean Squared Error:**

$$RMSE(log) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log d_i - \log \hat{d}_i\right)^2}$$

- **Scale-invariant Mean Squared Error** [*Eigen*]:

$$sRMSE(log) = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\left(\log d_i - \log \hat{d}_i + a(d_i, \hat{d}_i)\right)^2},$$

where $a\left(d_i, \hat{d}_i\right) = \frac{1}{N}\sum_i(\log \hat{d}_i - \log d_i)$

- **Accuracy with threshold** (δ_x):

(%) of d_i such that $\max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) = \delta < thr,$

where $thr = 1.25, 1.25^2, 1.25^3$

Models

- 12 Pretrained state-of-the-art models:
  - *MiDaS* (4 variations)
  - *Monodepth2* (6 variations)
  - *DenseDepth* (2 variations)

Table: Evaluation on KITTI dataset, using the Eigen split

Model	**Year**	**Citations**	**Version**	***AbsRel***	***SqRel***	***RMSE***	***$RMSE_{log}$***	**δ_1**	**δ_2**	**δ_3**
MiDaS	2020	721	*dpt_hybrid*	0.062	0.222	2.575	0.092	0.959	0.995	0.999
Monodepth2	2019	1708	*mono_640x192*	0.115	0.903	4.863	0.193	0.877	0.959	0.981
DenseDepth	2018	435	*kitti*	0.093	0.589	4.170	0.171	0.886	0.965	0.986
Eigen	2014	3782	*(baseline)*	0.190	1.511	7.156	0.270	0.692	0.899	0.967

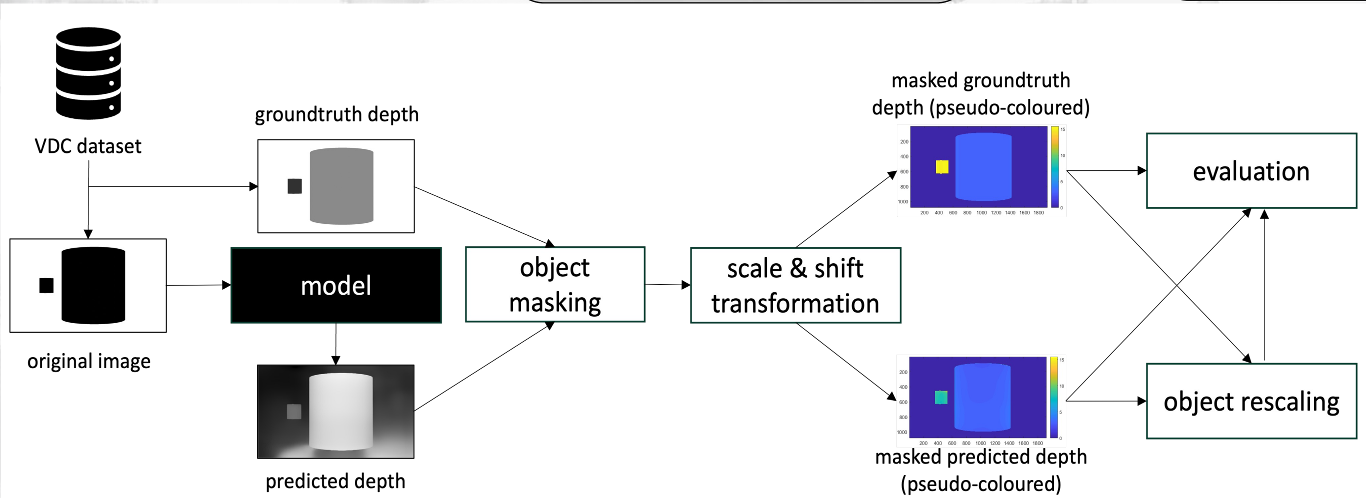
**Ranftl et al. (2020): Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer*

Experiment Pipeline Overview

Dataset integration
- Applying VDC to pretrained models

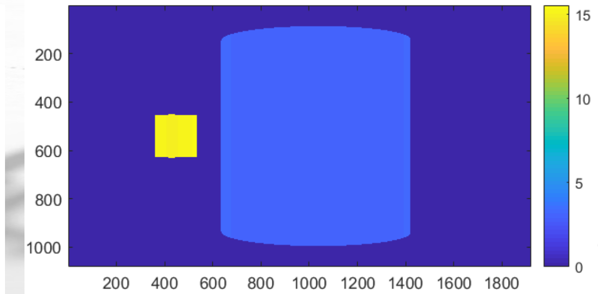
Masks & Transformations
- Applying binary objects' mask
- Applying scale and shift alignment transformations*

Metric evaluation
- Calculating typical metrics
- Calculating metrics after object rescaling

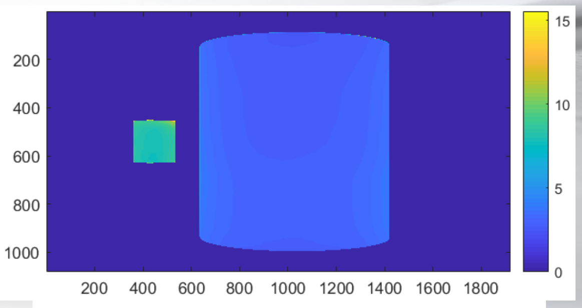


Results

An error in the estimated depth of the far object (depicted significantly smaller) will have a negligible impact on the metrics, while the estimate should be balanced



groundtruth depth



masked predicted depth

Table: Evaluation on VDC dataset, mean values

		Typical metric results								***Rescaled object metric results***							
Model	**Version**	***AbsRel***	***SqRel***	***RMSE***	***$RMSE_{log}$***	***sRMSE***	***δ_1***	***δ_2***	***δ_3***	***AbsRel***	***SqRel***	***RMSE***	***$RMSE_{log}$***	***sRMSE***	***δ_1***	***δ_2***	***δ_3***
MiDaS	*midas_v21*	0.056	1.718	10.097	0.040	0.003	0.964	0.987	0.993	0.111	7.562	21.852	0.088	0.014	0.853	0.908	0.935
Monodepth2	*stereo1024x320*	0.104	57.259	21.226	0.078	0.008	0.903	0.957	0.978	0.202	228.261	42.372	0.165	0.031	0.708	0.799	0.856
DenseDepth	*kitti*	0.103	3.472	20.153	0.083	0.009	0.902	0.951	0.972	0.207	16.657	40.576	0.178	0.036	0.704	0.785	0.837

Results

- *MiDaS* demostrate superior performance (rescaled $\delta_1 \approx 0.85$):

  Potential for partial learning of the relative size cue

- *Densedepth* (*nyu*) also exhibit enhanced accuracy:

  Pivotal role of training datasets: *Densedepth* (*kitti*) displays weaker results

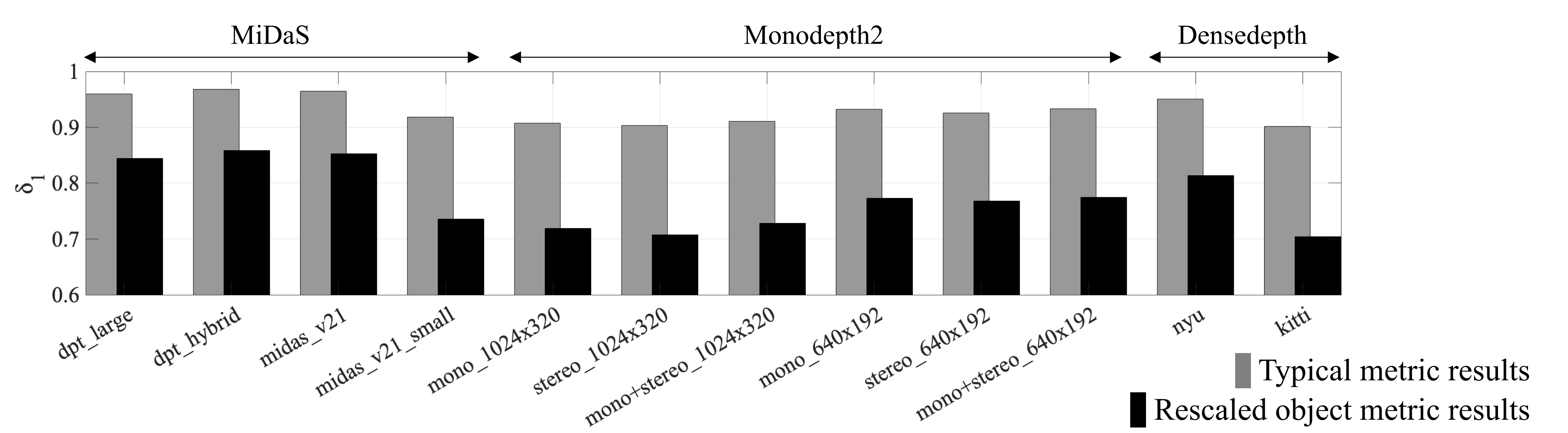


Figure: Accuracy with threshold δ_1 of the pertained models tested on our dataset

VDC RelSize:

Conclusion

- Preliminary study: New explainability concept for monocular depth estimation
- Creating a novel dataset:
  Introducing the Visual Depth Cue Dataset (VDC)
- Testing pretrained methods on a single visual depth cue:
  Exploring relative size
- Assessing indirect success:
  Metrics unveil monocular depth estimation performance
- Balancing metrics:
  The role of rescaled object assessments
- Future directions:
  - Expanding VDC: Incorporating other visual depth cues → benchmark dataset
  - Evaluating state-of-the-art method efficiency
  - Bridging Deep Learning with Human Perception
  - New depth estimation models aligned with human perception