Explaining Monocular Depth Estimation - Diving into Regional Differences in the Prediction

Sabine Fischer[1], Øyvind Ødegård[2,3], Frank Lindseth[1], Steven Yves Le Moan[1], and Gabriel Hansen Kiss[1]

[1] Department of Computer Science, NTNU, Holtermannsvegen 2, 7030 Trondheim
{sabine.fischer, frankl, steven.lemoan, gabriel.kiss}@ntnu.no
[2] Department of Archaeology and Cultural History, NTNU, Erling Skakkes gt 47B, 7491 Trondheim
oyvind.odegard@ntnu.no
[3] NTNU University Museum, Erling Skakkes gate 47B, 7491 Trondheim

Abstract. Recent foundational depth estimation models achieve impressive accuracy on various scenes. However, due to their black box nature, we lack knowledge about how they utilize the input for their predictions, and hence about their applicability to domains sparsely covered by labeled datasets. This paper, therefore, applies occlusion to investigate the influence of the input on different parts of the predicted depth maps for underwater scenes. Our results show that foundational depth estimation models combine global and local features to estimate relative distance, and that their influence differs significantly between the background and the rest of the scene. This suggests that extending post-hoc explanations to consider relationships between multiple input and output features can enrich our understanding of monocular depth estimation models and potentially help to gauge their applicability to new domains.

Keywords: Computer Vision, Monocular Depth Estimation, Deep Learning, Explainable AI

1 Introduction

Image-based depth estimation is utilized for various tasks, like image dehazing and novel view synthesis. Foundational deep-learning models for monocular depth estimation have recently gained popularity due to their ability to infer distances from a single input frame. In contrast to earlier deep-learning models, they are trained on a mixture of images from multiple domains, aiming to generalize performance [10,12,13]. However, like other deep-learning models, their large and complex architecture makes them non-interpretable (black-box nature) [11]. The resulting lack of knowledge about how they utilize input images to generate predictions and hence how inputs whose features differ from those in the training and evaluation sets affect the predictions impedes their application to safety critical tasks, like robotic control, and to domains only sparsely covered by available labeled data sets, like underwater scenes. Explainability methods

aim to make AI systems' reasoning, model, or evidence for a decision understandable to humans [11]. Their application could, therefore, address the challenges introduced by the black-box nature of monocular depth models by exposing how input features influence depth estimation and allowing users to analyze to what degree emerging patterns in this mapping apply to a target domain.

However, the only paper generating post-hoc explanations for such a foundational model studies the effects of a single depth cue in a simple scene [3]. Other work on explainable depth estimation focuses on the effects of the input on a single target, either the prediction as a whole [4,7] or a single inserted object [6]. In contrast, we investigate how spatial input features influence multiple different regions in the predictions of foundational monocular depth estimation models. We study this on images of underwater scenes, which differ in both content and appearance from terrestrial images [8], and are only sparsely covered by labeled training data. We apply occlusion, a perturbation-based explainability method, to quantify the influence of the input on multiple target regions for two such transformer-based depth estimation models, Depth Anything v1 [12] and v2 [13] (DA-V1, DA-V2). We show that the importance of input features differs between target regions in the prediction, especially between the background and parts that depict objects. Hence, extending post-hoc explanations to multiple targets in the prediction can enrich our understanding of depth estimation models.

2 Related work

Work on explainable AI largely focuses on regression, classification, and segmentation [11]. The outputs of these tasks are generally single scalars or discrete labels. In contrast, dense depth estimation predicts a depth map for each input, typically a 2D array of continuous values. This expands the space of possible mappings between the input and output, allowing for more complex relationships. We aim to generate new insights into this relationship for depth estimation by generating explanations that capture the influence of the input features on multiple target regions. In contrast, previous work explaining monocular depth models analyzes how the input influences a single target, the predicted depth map as a whole [4,7] or the predicted distance to a specific object [3,6].

In this context, occlusion has been applied to learn a mask that simultaneously minimizes the difference between predictions for the occluded and original image, and maximizes sparseness of the non-occluding weights [7]. By visualizing the mask, the authors identified edges, areas inside some objects, and the region around the vanishing point as especially influential for indoor and driving scenes [7]. While this provides fine-grained information about the magnitude of influence of input features on the overall prediction, it does not capture how the depth map changes. Another paper studied how manipulating features of a single object overlayed on a driving scene affects the predicted distance to this object [6]. They found that CNN-based models rely on the size and position of its shadow [6]. They describe how these input features influence a specific part of the prediction, but do not consider the potential influence of the scene context.

A more recent paper that applied style transfer, and compared the resulting prediction errors, showed that CNN-based depth prediction models rely more on textures and transformer-based models more on shape [4]. Since earlier work was conducted on CNN-based models [7,6], while foundational depth estimation models like DA-V1 and DA-V2 integrate transformers in their architecture[12,13], this indicates, the earlier results might not transfer to current models, especially in light of the associated performance gains. Only the most recent paper included a transformer-based foundational model in its analysis [3]. Using synthetic data depicting cylindrical objects of different sizes, the authors conclude that this model utilizes relative size as a depth cue [3]. However, they do not describe how the presence of a second non-manipulated object influences the prediction. This further highlights the need for research on how current models utilize their input for monocular depth prediction, which we study in this paper.

3 Methods

To gain insight into how spatial features of the RGB input images influence foundational depth estimation models, we apply occlusion. This perturbation-based explainability method is model-agnostic and does not utilize gradients, making it suitable for pre-trained models and easy to extend to multiple targets.

We split the input frame into m regions $R_{in} \in \mathbb{R}^2$ and measure the difference between the depth map predicted for the original image P^* and for the image after occluding a region R_{occl} in the input $P^{(R_{occl})}$, i.e. setting the RGB values inside R_{occl} to $[0, 0, 0]$. Following Equation 1, we calculate attribution values $A_{i,j,k,l} \in \mathbb{R}^4$ for each pixel in the input $I_{i,j}$ on each pixel in the output $P_{k,l}$, with $n_R(i, j)$ as the number of input regions R_{in} that contain a pixel $I_{i,j}$.

$$A_{i,j,k,l} = \frac{1}{n_R(i,j)} \sum_{R_{in} \in I} P^*_{i,j,k,l} \text{ - } P^{(R_{occl})}_{i,j,k,l} \tag{1}$$

To quantify the input's influence on a target region R_{out} in the predicted depth map, we consider three aggregation functions f_a. To capture the full magnitude of influence, we average the absolute attribution values, i.e. $f_{abs}(A_{i,j,k,l}) = \frac{1}{|R_{out}|} \sum_{P_{k,l} \in R_{out}} |A_{i,j,k,l}|$. We compute additional, separate means of just the positive (f_{pos}) and just the negative (f_{neg}) attribution values. These describe how much the features in the manipulated region push the predicted distance to a target region towards or away from the camera, respectively.

Inspired by distance-dependent appearance shifts in underwater images and the varying degrees of resemblance between underwater and terrestrial objects, we group pixels by their semantic class or predicted distance to form target regions and input regions that are occluded together. The semantic classes are given by preexisting annotations. The distance-based depth bins are generated by clustering pixels according to their prediction value in P^* using Multi-Otsu thresholding [9]. In order of increasing distance, we call them "front", "mf", "mid", "mb", and "back". We generate additional attribution maps by occluding

regions inside a sliding window of size $s_{window} \in \mathbb{N}^2$ moving it across the input with stride $s_{stride} \in \mathbb{N}$. To balance the resolution of the attribution maps in terms of input and output, predictions are down-sampled to size $\frac{h_{in} \times w_{in}}{s^2_{stride}}$.

We quantify the influence of local features, i.e. features of the input region covering the same 2D region as the target in the output, compared to global features, i.e. parts of the input outside this region, using the Jaccard dissimilarity between a semantic region or depth bin M_{class} and the region $M_{top}(A)$ that covers the highest n_{top} absolute values within an attribution map A (see Equations 2 and 3). We generate $M_{top}(A)$ for two choices of n_{top}, 10% of the size of the attribution map ($n_{top,10}$) and the highest amount of attribution values that can fit within the mask M_{class} ($n_{top,fit}$).

$$M_{top,i,j}(f_a(A)) = \begin{cases} 1 & \text{if } a_{i,j} >= a_{sorted}[n_{top}] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$d_{Jaccard}(M_{class}, M_{top}) = 1 - \frac{|M_{class} \cap M_{attr}|}{|M_{class} \cup M_{top}|} \tag{3}$$

Datasets and Models All experiments are conducted on the TEST split of SUIM dataset [8]. It provides 110 pairs of images of underwater scenes and associated masks with human annotations assigning one of eight classes to each pixel: background water body (water), human diver (diver), aquatic plants and seagrass (plant), wrecks or ruins (ruin), robots including AUVs/ROVs/instruments (robot), reefs and invertebrates (reef), fish and vertebrates (fish), or seafloor and rocks (rock). We study two recent transformer-based models, DA-V1 [12] and DA-V2[13]. They mainly differ in the training data for their teacher models, real and synthetic images for DA-V1, and only synthetic images for DA-V2 [12,13]. Neither selects SUIM as part of their training datasets [12,13]. Since most of the included labeled data sets target scenarios that exclude underwater images [12,13], images from this domain make up a small fraction of the training data.

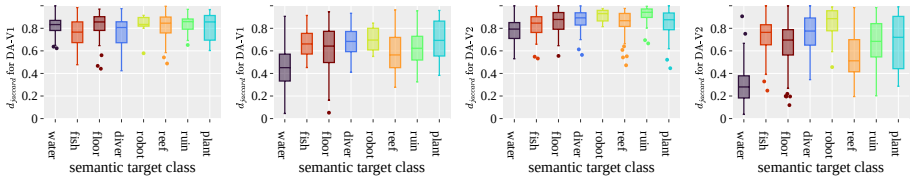
Implementation details All experiments were implemented with PyTorch and Captum. We utilize pre-trained models from Hugging Face [2,1].For the sliding window approach, we use $s_{window} = (16, 16)$ and $s_{stride} = 8$. Images with more than 800000 pixels were reduced to that size using bilinear interpolation.

4 Results

An initial analysis of the cosine similarity of attribution maps generated with the sliding window shows that DA-V1 and DA-V2 use different input regions to predict distances for different target regions (see Table 1). The mean cosine similarities ranging from 0.80 to 0.93 between $f_{abs}(A)$ maps indicate a high but not complete overlap between the most influential input regions on predictions for the whole image and the distance/semantic class based target regions. There is a higher variation in this overlap for $f_{pos}(A)$ or $f_{neg}(A)$. A one-way ANOVA test on these cosine similarities shows that the differences in attribution maps

Table 1: Summary and ANOVA test statistics describing cosine similarities between attribution maps $f_a(A)$ targeting full predictions and regions of interest, with H_0: "The mean similarities are the same for each ROI of the same type".

		DA-V1				DA-V2			
		mean	std	p	η^2	mean	std	p	η^2
semantic regions	f_{abs}	0.84	0.11	$\mathbf{1{\cdot}10^{-7}}$	0.11	0.80	0.12	$\mathbf{2{\cdot}10^{-12}}$	0.17
	f_{neg}	0.64	0.19	$\mathbf{8{\cdot}10^{-7}}$	0.11	0.49	0.24	$\mathbf{5{\cdot}10^{-13}}$	0.17
	f_{pos}	0.81	0.12	$\mathbf{7{\cdot}10^{-13}}$	0.051	0.81	0.44	0.19	0.026
depth bins	f_{abs}	0.91	0.093	$\mathbf{1{\cdot}10^{-7}}$	0.11	0.93	0.072	$\mathbf{7{\cdot}10^{-44}}$	0.32
	f_{neg}	0.69	0.23	$\mathbf{8{\cdot}10^{-7}}$	0.11	0.47	0.31	$\mathbf{2{\cdot}10^{-7}}$	0.11
	f_{pos}	0.80	0.20	$\mathbf{7{\cdot}10^{-13}}$	0.051	0.89	0.14	$\mathbf{1{\cdot}10^{-19}}$	0.15



(a) $n_{top,10}$ (b) $n_{top,fit}$ (c) $n_{top,10}$ (d) $n_{top,fit}$

Fig. 1: Jaccard distances between the n_{top} highest attributions $M_{top}(f_{abs}(A))$ and the semantic regions M_{class} for DA-V1 (left) and DA-V2 (right).

between different semantic regions and depth bins are statistically significant ($p < 0.05$) in all but one case. The moderate to large effect sizes η indicate that differences between these classes explain parts of the variance in the influence different regions in the input have on different regions in the models' output.

Semantic Classes A one-way ANOVA test comparing the Jaccard distances between semantic target regions and the associated most influential input regions shows significant differences in the importance of local compared to global features between the semantic classes for both models and for all three aggregation functions. According to post-hoc t-tests on $f_{abs}(A)$, for DA-V2, these differences are statistically significant when comparing "water" to any other semantic class (see Table 2b). Except for the pair ("water", "plant"), this applies independent of the choice of n_{top}. For DA-V1 differences in these Jaccard distances are again significant for all pairs that include "water" (see Table 2a) $n_{top,fit}$. However, the pair ("water", "fish") is the only one for which we find a statistically significant difference for both choices of n_{top}. Overall, the distributions of the Jaccard distances indicate that a large proportion of the $n_{top,10}$ most influential input pixels lie outside each semantic target region (see Figure 1). The distances are considerably lower for $n_{top,fit}$. Notably, for this threshold, both models utilize more local features for "water" than for any other semantic class as indicated by the especially low Jaccard distances.

Table 2: p-values of a t-test testing H_0: "The mean Jaccard distances between class regions M_{class} and the associated most influential input regions $M_{top}(f_{abs}(A))$ are the same for each semantic class".

(a) DA-V1

	$n_{top,10}$							$n_{top,fit}$						
	water	diver	fish	floor	plant	reef	robot	water	diver	fish	floor	plant	reef	robot
diver	0.37							$\mathbf{1{\cdot}10^{-12}}$						
fish	**0.023**	1.0						$\mathbf{2{\cdot}10^{-15}}$	1.0					
floor	1.0	0.55	0.098					$\mathbf{2{\cdot}10^{-6}}$	1.0	1.0				
plant	1.0	1.0	1.0	1.0				$\mathbf{3{\cdot}10^{-4}}$	1.0	1.0	1.0			
reef	1.0	0.46	0.072	1.0	1.0			$\mathbf{6{\cdot}10^{-5}}$	0.19	0.34	1.0	1.0		
robot	1.0	1.0	0.86	1.0	1.0	1.0		$\mathbf{2{\cdot}10^{-5}}$	1.0	1.0	1.0	1.0	0.35	
ruin	1.0	0.12	**0.015**	1.0	1.0	1.0	1.0	$\mathbf{5{\cdot}10^{-4}}$	1.0	1.0	1.0	1.0	1.0	1.0

(b) DA-V2

	$n_{top,10}$							$n_{top,fit}$						
	water	diver	fish	floor	plant	reef	robot	water	diver	fish	floor	plant	reef	robot
diver	$\mathbf{7{\cdot}10^{-5}}$							$\mathbf{7{\cdot}10^{-22}}$						
fish	**0.026**	1.0						$\mathbf{2{\cdot}10^{-36}}$	1.0					
floor	$\mathbf{2{\cdot}10^{-4}}$	1.0	1.0					$\mathbf{3{\cdot}10^{-18}}$	0.18	0.32				
plant	1.0	1.0	1.0	1.0				$\mathbf{4{\cdot}10^{-5}}$	1.0	1.0	1.0			
reef	**0.014**	1.0	1.0	1.0	1.0			$\mathbf{1{\cdot}10^{-11}}$	$\mathbf{2{\cdot}10^{-5}}$	$\mathbf{8{\cdot}10^{-6}}$	0.39	1.0		
robot	$\mathbf{3{\cdot}10^{-5}}$	1.0	**0.018**	0.37	1.0	0.16		$\mathbf{1{\cdot}10^{-6}}$	1.0	1.0	0.10	1.0	$\mathbf{2{\cdot}10^{-3}}$	
ruin	$\mathbf{6{\cdot}10^{-8}}$	1.0	$\mathbf{2{\cdot}10^{-3}}$	0.11	1.0	0.05	1.0	$\mathbf{3{\cdot}10^{-9}}$	1.0	1.0	1.0	1.0	0.38	0.51

We also analyze the influence of occluding all input pixels associated with each semantic class at once. According to the resulting mean relative attribution values normalized by the number of occluded pixels, the prediction for any region is influenced most by occluding the region itself (see Figure 2). While the meaningfulness of analyzing inhomogeneous pairs is generally limited by the number of images showing both classes, most images depict the water column. For both models, the influence relationship between "water" and "fish" or "diver" regions is noticeably asymmetric. The influence of occluding the background water column on the predictions for "fish" or "diver" regions is higher than the influence of occluding those regions on "water".

Predicted Distance According to a one-way ANOVA test, the Jaccard distances between depth bins and the associated top attribution regions differ significantly between the depth bins. The test indicates statistical significance for all models and aggregation functions independent of the choice for n_{top}, except when selecting the $n_{top,10}$ top values in $f_{pos}(A)$ for DA-V1. Because the depth bins are ordered by distance, we conduct post-hoc t-tests on the direction-sensitive at-

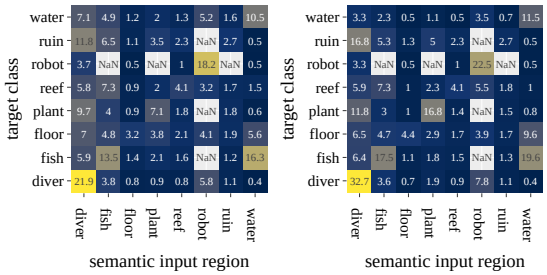
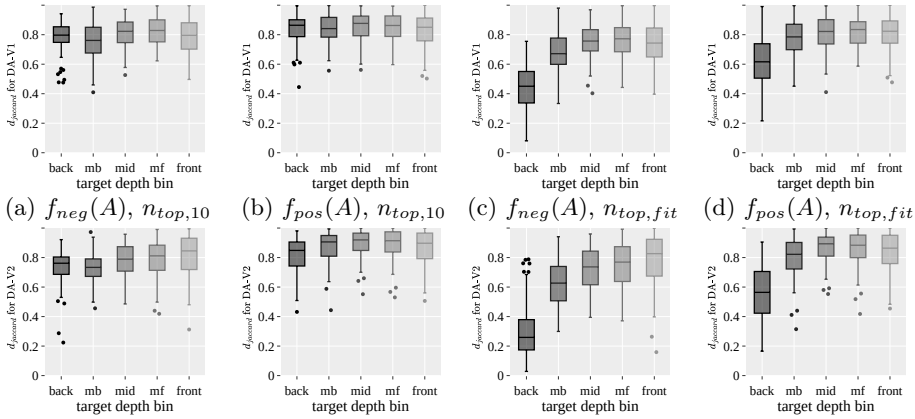


Fig. 2: Means of relative attributions using $f_{abs}(A)$ when occluding full semantic regions for DA-V1 (left) & DA-V2 (right), normalized by the occluded area. Values in the heatmap were scaled by factor 10^5.



(a) $f_{neg}(A)$, $n_{top,10}$ (b) $f_{pos}(A)$, $n_{top,10}$ (c) $f_{neg}(A)$, $n_{top,fit}$ (d) $f_{pos}(A)$, $n_{top,fit}$

(e) $f_{neg}(A)$, $n_{top,10}$ (f) $f_{pos}(A)$, $n_{top,10}$ (g) $f_{neg}(A)$, $n_{top,fit}$ (h) $f_{pos}(A)$, $n_{top,fit}$

Fig. 3: Jaccard distance between the depth bins M_{class} and the top associated attribution values $M_{top}(f_a(A))$ for DA-V1 (top) and DA-V2 (bottom).

tribution maps $f_{pos}(A)$ and $f_{neg}(A)$ (see Table 3), skipping the combination for which the ANOVA test showed no significance. Save for one exception for DA-V2, there are significant differences in these Jaccard distances between "back" and all other depth bins for $f_{neg(A)}$. Furthermore, all pairs showing significant differences in these distances include "back" or "mb". Generally, the distribution of the Jaccard distances indicates that a large proportion of the most relevant features for each depth bin are located outside it (see Figure 3). For the $n_{top,fit}$ top attribution values, the median distances for "back" and "mb" decrease, indicating a higher influence of local features for these depth bins.

When occluding whole depth bins, occluding the target bin itself is associated with the highest absolute mean relative values in $f_{neg}(A)$ (see Figures 4a and 4c). This indicates that the features that most strongly indicating a larger distance to any target bin to either model are located within the same depth bin. The absolute values calculated using f_{neg} are lower for pairs that are farther

Table 3: p-values of a t-test testing H_0: "The mean Jaccard distances between depth bins M_{class} and the associated most influential input regions $M_{top}(f_a(A))$ are the same for each depth bin".

(a) DA-V1

	$n_{top,10}$				$n_{top,fit}$							
	$f_{neg}(A)$				$f_{neg}(A)$				$f_{pos}(A)$			
	back	mb	mid	mf	back	mb	mid	mf	back	mb	mid	mf
mb	0.39				$\mathbf{1{\cdot}10^{-25}}$				$\mathbf{2{\cdot}10^{-15}}$			
mid	0.44	$\mathbf{1.7{\cdot}10^{-4}}$			$\mathbf{1{\cdot}10^{-41}}$	$\mathbf{9{\cdot}10^{-5}}$			$\mathbf{9{\cdot}10^{-21}}$	0.57		
mf	**0.041**	$\mathbf{6.7{\cdot}10^{-5}}$	1.0		$\mathbf{3{\cdot}10^{-34}}$	$\mathbf{4{\cdot}10^{-6}}$	1.0		$\mathbf{9{\cdot}10^{-22}}$	0.41	1.0	
front	1.0	0.78	0.58	0.083	$\mathbf{5{\cdot}10^{-35}}$	**0.019**	1.0	0.62	$\mathbf{5{\cdot}10^{-17}}$	1.0	1.0	1.0

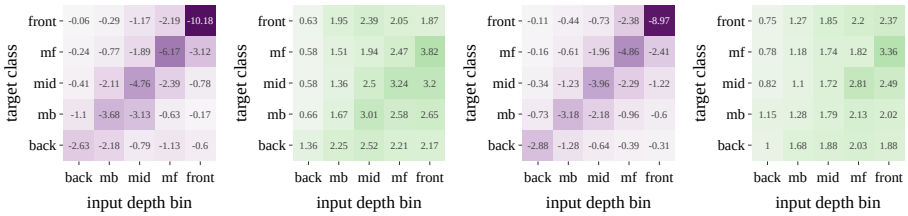
(b) DA-V2

		$f_{neg}(A)$				$f_{pos}(A)$			
		back	mb	mid	mf	back	mb	mid	mf
$n_{top,10}$	mb	1.0				$\mathbf{4{\cdot}10^{-4}}$			
	mid	**0.023**	$\mathbf{2{\cdot}10^{-3}}$			$\mathbf{8{\cdot}10^{-8}}$	0.69		
	mf	$\mathbf{6{\cdot}10^{-4}}$	$\mathbf{6{\cdot}10^{-4}}$	1.0		$\mathbf{4{\cdot}10^{-6}}$	1.0	1.0	
	front	$\mathbf{4{\cdot}10^{-4}}$	$\mathbf{4{\cdot}10^{-5}}$	1.0	1.0	$\mathbf{2{\cdot}10^{-3}}$	1.0	0.85	1.0
$n_{top,fit}$	mb	$\mathbf{2{\cdot}10^{-30}}$				$\mathbf{6{\cdot}10^{-24}}$			
	mid	$\mathbf{1{\cdot}10^{-47}}$	$\mathbf{3{\cdot}10^{-6}}$			$\mathbf{3{\cdot}10^{-37}}$	$\mathbf{7{\cdot}10^{-4}}$		
	mf	$\mathbf{4{\cdot}10^{-48}}$	$\mathbf{9{\cdot}10^{-8}}$	1.0		$\mathbf{1{\cdot}10^{-34}}$	$\mathbf{4{\cdot}10^{-3}}$	1.0	
	front	$\mathbf{3{\cdot}10^{-50}}$	$\mathbf{3{\cdot}10^{-11}}$	0.085	0.81	$\mathbf{5{\cdot}10^{-29}}$	0.31	1.0	1.0

away from each other, indicating a closer link between regions at more similar depths. Furthermore, we find that the influence between inhomogeneous pairs is asymmetric for both aggregation functions (see Figure 4). With only two exceptions, the absolute values of attributions associated with occluding a closer region and the prediction for any farther away region as the target are higher than those associated with the opposing occluded and target regions.

5 Discussion & Conclusions

Overall, our results show that DA-V1 and DA-V2 utilize a combination of features within and outside each target region to produce predictions for that region. The high Jaccard distances between target regions and the input regions with the highest influence on them indicate that DA-V1's and DA-V2's predictions strongly depend on the overall context of the scene. At the same time, we would expect the influence of occluding a full region relative to its size to be evenly distributed among present target classes if it reflected only the information it provides about the overall scene. Hence, the high influence of occluding a region on its own prediction indicates that the models also utilize local features, likely



(a) f_{neg}(A), DA-V1 (b) $f_{pos}(A)$, DA-V1 (c) $f_{neg}(A)$, DA-V2 (d) $f_{pos}(A)$, DA-V2

Fig. 4: Means of relative attribution values when occluding full depth bins, normalized by the occluded area. Values in the heatmaps were scaled by factor 10^5.

the region's texture or color, since the shape and position of the black silhouette resulting from its occlusion match those of the target region. The high similarity in patterns found for DA-V1 and DA-V2 indicates that the generated attribution maps are consistent between functionally similar models, which is one of the characteristics of good explanations [11]. Combining context and local features also aligns well with depth cues used by humans such as relative texture, and changes in color and saturation related to aerial perspective [5]. This alignment further supports the potential of the attribution maps as post-hoc explanations. While more research is needed to verify whether the mechanisms of foundational depth estimation models align with human cognition, this also suggests that such models can identify and combine relevant depth cues in underwater scenes.

Furthermore, we show that the influence of input regions differs significantly between different regions in the predicted depth maps. This is illustrated by the asymmetric influence relationships between the depth bins that indicates that closer regions have a stronger effect on the predictions. Both the results for semantic regions and the depth bins show that the relative importance of local and global features particularly differs between the background ("water", "back", "mb") and the rest of the scene. Although the background contains fewer edges, contrast, and objects whose presence is linked to influential features identified by earlier work [3,4,6,7], a higher proportion of influential features are located within the background. However, the smoothness of the background might be a relevant cue in itself. At some distance from the camera, any details blur and become hazy making the background appear flat. Although this happens at closer distances underwater, a similar effect can be observed in terrestrial scenes.

In conclusion, although our work is limited by the pre-selection of target regions and potential biases due to the introduction of additional edges during occlusion, we show that extending post-hoc explanations to multiple target regions can enrich our understanding of the influence of the input on 2D predictions. Promising directions for future work include investigating which features of predicted regions lead to differences in the importance of input regions or higher-level input features, and adapting gradient-based explainability methods to this task. In addition, extending the experiments to additional models or datasets, especially ones that depict other domains or include ground truth, could help

to distinguish to what degree the identified patterns generalize between models and domains, or are related to a domain gap between underwater images and training data.

Disclosure of Interests. Sabine Fischer received funding from the PERSEUS project, a European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101034240.

References

1. Depth-Anything-V2-Large. https://huggingface.co/depth-anything/Depth-Anything-V2-Large
2. Depth Anything (large-sized model, Transformers version). https://huggingface.co/LiheYoung/depth-anything-large-hf (2024)
3. Arampatzakis, V., Pavlidis, G., Pantoglou, K., Mitianoudis, N., Papamarkos, N.: Towards Explainability in Monocular Depth Estimation. In: Meo, R., Silvestri, F. (eds.) Machine Learning and Principles and Practice of Knowledge Discovery in Databases. vol. 2134, pp. 412–419. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-74627-7_34
4. Bae, J., Moon, S., Im, S.: Deep Digging into the Generalization of Self-Supervised Monocular Depth Estimation. In: AAAI Conference on Artificial Intelligence. vol. 37, pp. 187–196 (Jun 2023). https://doi.org/10.1609/aaai.v37i1.25090
5. Cutting, J., Vishton, P.: Perceiving layout and knowing distances: The interaction, relative potency, and contextual use of different information about depth. In: Perception of Space and Motion, vol. 5, pp. 69–177 (Jan 1995)
6. van Dijk, T., de Croon, G.: How Do Neural Networks See Depth in Single Images? In: IEEE/CVF International Conference on Computer Vision. pp. 2183–2191 (2019)
7. Hu, J., Zhang, Y., Okatani, T.: Visualization of Convolutional Neural Networks for Monocular Depth Estimation. In: IEEE/CVF International Conference on Computer Vision. pp. 3868–3877. IEEE, Seoul, Korea (South) (2019). https://doi.org/10.1109/ICCV.2019.00397
8. Islam, M.J., et al.: Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1769–1776 (2020). https://doi.org/10.1109/IROS45743.2020.9340821
9. Liao, P.S., Chen, T.S., Chung, P.C.: A fast algorithm for multilevel thresholding. J. Comput. Inf. Sci. Eng. **17**(5) (2001)
10. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision Transformers for Dense Prediction. In: IEEE/CVF International Conference on Computer Vision. pp. 12159–12168. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.01196
11. Schwalbe, G., Finzel, B.: A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Min Knowl Disc **38**(5), 3043–3101 (2024). https://doi.org/10.1007/s10618-022-00867-8
12. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024)
13. Yang, L., et al.: Depth Anything V2. Adv. Neural Inf. Process. Syst. **37**, 21875–21911 (2024)