Feature Contribution in Monocular Depth Estimation

Hui Yu Lau[1], Srinandan Dasmahapatra[1], and Hansung Kim[1]

University of Southampton, Southampton SO17 1BJ, United Kingdom

Abstract. Monocular Depth Estimation (MDE) is an inherently ill-posed problem due to the lack of binocular depth cues, despite this there have been significant research done in this field in recent years. In an attempt to bridge understanding between human and machine perception, this paper investigates learned concepts from the general-purpose model Depth Anything, focusing on features that are known to be present in the human visual system. We perform interventions on different image features within the KITTI and NYUv2 dataset, evaluating performance on these intervened inputs. This led to interesting insights on how and how much each of these features influence depth perception. These insights contribute to bridging understanding of how humans and machines perform MDE respectively, and we also hope it provides a new way for future work to devise more robust methods of training neural networks for MDE.

Keywords: computer vision tasks · monocular depth estimation · understandable artificial intelligence · human visual system

1 Introduction

Monocular Depth Estimation (MDE) aims to predict a dense depth map for a given input image. Research in this field cover a wide range, from engineering better performance and generalization to understandability and tackling specific problems. It is a field under active research due to its applications in autonomous driving [22], VR scene reconstruction [1], and robotics [7].

Given its nature of predicting depth from single two-dimensional images, MDE is an inherently ill-posed problem. In recent years, the field has seen improvement in various forms: from powerful general-purpose models that aim to produce large general models via mixing dataset in training [19], to a recent work that aims to utilize large amounts of unlabelled data during training for the same purpose [24]; as well as improvements in zero-shot scaled capabilities through methods such as inclusion of camera intrinsics [10]. These have shown that MDE in machines using depth cues only found in single images is practical. However despite leaps of improvement in performance, the specific mechanisms used by deep learning models to perform MDE and how these mechanisms compare to the human visual system are still poorly understood.

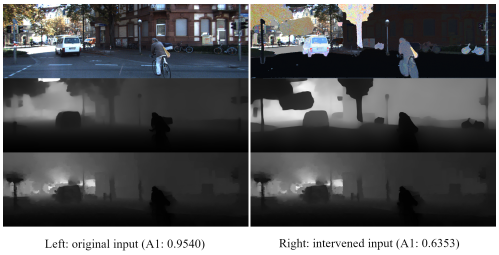

Fig. 1: An example of how destroying intensity information affects performance. Top: Input, Middle: Prediction, Bottom: Ground truth dense map using infill

To address this gap in knowledge, we are inspired by works of causality in computer vision [15,26], the human visual system [5,8], and existing attempts at understanding MDE [6,25]. We build on these works to investigate how features in images contribute to estimating depth. When performing analysis between causal features it is important to choose the right features [26]. Existing works that analyse how MDE can be better understood have placed focus on different feature levels, ranging from placement of shapes [6] to image features [8] and to latent representations within models [25]. In this work we choose to focus on image features since it matches best with how the human visual system is understood to work [8].

Following conclusions from [6, 23], we accept the importance of global shape and object placement to be significant and its removal to be devastating. We instead focus on how texture, hue, saturation, and intensity impacts MDE when global shape is maintained. An example of intervention and its effects is shown in (Fig. 1). While [23] trains separate models on images intervened to retain only single features, our work extends that work and is distinct in that we investigate learned relationships in an existing state of the art model, and crucially retains global shape and object placement, which is destroyed in the mentioned paper.

Texture in computer vision refers to the repetitive change in intensities in parts of an image. It has been shown that the human visual system contains explicit neural pathways to recognize texture, and that it contributes to depth perception in humans. [5, 18] This can be understood intuitively as humans - given similar texture, the further away an object the denser the texture should appear on the image. Following our goal of comparing machine to human perception, we choose it as one of our features.

Colour is the result of light perceived by a vision system, activated in the human visual system in the visible light spectrum. Humans rely heavily on colour for our visual understanding of the world [20]. In this paper, we consider colour as the mixture of light received at a point. When considering colour, an appropriate way of representing colour has to be chosen.

We choose to use the HSV colourspace, since we deemed it to represent attributes closer to physical attributes and human understanding than other representations. HSV is a colourspace designed to allow for more intuitive user

interaction, proposed to capture colour in terms of its most noticeable features: hue, saturation etc. This idea and the process for converting between the RGB and HSV colourspaces are well known, and described in [12].

Hue is the dominant wavelength of a colour considered in the HSV colour space. It contains information of the object through its natural colouration, but is also affected by environmental lighting. At the beginning of this research it was unclear how hue could affect depth perception, but in later experiments we found its ability to help define boundaries of objects to be powerful.

Saturation is determined by the purity of colour. A singular wavelength makes for the most saturated colour, while both mixing different wavelengths of light leads to reduced saturation. It was originally conceived that saturation would play a larger role in outdoor scenes than in indoor ones due to particulates in the air diffusing colour, but this was not confirmed by quantitative results. However, qualitative results suggest that large changes in saturation between neighbouring objects tend to cause a difference in depth prediction.

Value is the brightness of colour, but to avoid confusion we shall refer to it as "intensity" from hence on. In the HSV colour space brightness ranges from black at 0 to the brightest possible colour at maximum. In experiments we find that intensity carries the most contribution out of the three colour features, especially in outdoor scenes where changes in intensity often causes the model to predict a large change in depth.

2 Related Works

MDE is an increasingly popular research topic, and there have been various studies that aim to understand inner workings of MDE models. A study looking at MonoDepth [9] finds the model mainly estimates depth based on the height of objects within the image, suggesting models learn shortcuts to estimate depth. However what other shortcuts exists remains a gap in literature. This highlights the importance of understanding what a neural network is using to perform its predictions.

Causal reasoning is a field of study that has recently seen applications in computer vision. These methods aim to disentangle the causal relationships that exist within computer vision tasks. [15] introduces causal reasoning into 3D scene reconstruction and saw an increase in performance. 3D scene reconstruction being an ill-posed problem where spurious links traditionally required strong regularization. This highlights the importance of encouraging models to learn reliable relationships instead of spurious ones. A review on the use of causal reasoning on computer vision also highlighted the importance of choosing nodes on a structural causal model as features appropriate to the task at hand [4].

Extensive work has been done to understand the importance of global shape and placement on monocular estimation, highlighting how these features are used by deep learning models to make predictions. In a recent study also investigating the roles of visual cues on MDE, [23] found that when shape is not preserved, model performance drops severely. Similarly, a study studies the effect of chang-

ing shape and positions of obstacles on a general purpose model [6]. Finally, it is well known that the design of CNNs with their convolution architecture commonly used in computer vision tasks [11] is to capture spatial and shape relationships across scales. These reasons make it clear to us the importance of shape in MDE, and thus we do not investigate is effect relative to the other lesser understood features.

2.1 Monocular Depth Cues in the Human Visual System

Depth prediction performed in the human visual system is well studied. [5] analysed different visual cues that contributes to depth prediction, the ones which are not applicable to MDE have been removed from this list. They are defined as follows:

- Occlusion: when close objects occlude parts of those further.
- Relative size and density: difference in size of shapes and textures that should be of similar sizes.
- Height in the visual field: the height of an object when viewed, relative to a vanishing point.
- Aerial perspective: the effect of atmospheric particulates on perceived colour.
- Motion perspective: the difference of motion between close and distant objects.

2.2 Current MDE methods

CNN based models have been widely used in MDE [9], with various modifications such as residual connections [11] and encoder-decoder networks with skip connections [2]. Lately, vision transformers have become increasing popular, acting as the backbones of a number of larger general purpose approaches. The power of these models has led to researchers providing larger amounts of data in training, leading to [19, 24]. In [24], the authors train a student model from a teacher model with extra unlabelled data, which are perturbed, creating a more challenging training set that requires the model to learn more further cues. This inspired us to investigate other cues in MDE tasks which can similarly make more challenging training, perhaps even in a more meaningful way.

2.3 Causal Reasoning

Causal reasoning is the investigation of causality, the relationship of cause and effect through a mathematical framework [17]. Given observations, this field of study aims to discover the relationships between phenomena, whether one is the cause of another, and how information flows through this network of causal links. Spurious links are those that contain statistical correlation between two nodes, but one is not a causal ancestor of the other. Unfortunately, spurious links are abundant in data sets used in computer vision in form of data bias [26]. While

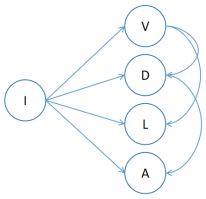

Fig. 2: A structural causal graph directly adapted from [15].

there are methods that debias models once known [3, 13], the discovery of these correlations is still a challenging task.

For example, in the above mentioned mentioned [15], authors tackle 3D scene reconstruction, which's ill-posed nature traditionally called for strong regularization. They find that by posing internal representations of the image (I) as a structural causal graph (Fig. 2) between view point (V), depth (D), lighting (L), and albedo (A), they were able to introduce a form of implicit regularization that improved performance. This shows the potential of tackling complex computer vision tasks as a combination of different feature cues.

3 Methodology

To analyse the effect of different image features on performance, we establish a definition of feature contribution. The contribution of a feature is the percentage drop in performance the model sees after an intervention has been performed to scramble information on said feature. For example, if destroying hue information via randomization causes performance to drop by a large amount, then we would say hue has a high contribution for MDE.

To test the effect of destroying feature information while still maintaining general shape information of individual objects and general spacial relationship, we apply phase-scrambling [8] (for texture) and randomizing of hue, saturation, and intensity (for colour) to individual objects within images. This means that the outlines of images are maintained, but features between objects that might tell the model how they are related in depth are randomized.

In order to obtain a working set of data, we propose an automatic pipeline to intervene on images (Fig. 3). This pipeline consists of the following steps: a general purpose segmentation method first segments images into individual objects, overlapping object maps are removed, and lastly independent intervention operations are performed on a per-object basis based on the segmentation maps.

After we obtain a working dataset, we perform evaluation of the dataset using Depth Anything [24]. We compare the prediction against its label and calculate contribution. Depth Anything is a general purpose, transformer-based model. We chose this model for the following reasons. First, its aim at being general purpose means it claims to have learnt certain invariant feature links

Fig. 3: The steps we take to create different intervened images. The images on the right are from top to bottom: texture, hue, saturation, and intensity interventions.

applicable throughout various datasets. Second, its performance at its time of publishing makes it an ideal candidate for testing what state-of-the-art models learn. Third, its training method already includes image perturbations, and we are interested if these perturbations in training have provided resistance towards the destruction of certain features: texture and colour.

More details of our implementation are given below.

3.1 Segmentation

Segmentation of the image is performed via the general-purpose model Segment Anything [14]. Due to the intervention operations to be performed downstream, our method requires independent objects to be segmented away from each other.

Segment Anything returns an object masks based on a starting point. To perform segmentation on the entire image a grid of points of interest are applied, each returning an object map. In some cases, two or more points of interest are placed on a single object, making their maps overlap. In these cases, we calculate the area of overlap, and if the overlap is significant define the masks to be of the same object, and the larger object mask is taken as the final object mask.

3.2 Intervention

Destruction of texture is done through phase-scrambling [8], which is a technique commonly used in signal processing and psychology. This method randomizes the phase of an image in the Fourier transform space, destroying texture information but retaining colour information. We opted for this method instead of averaging colour because averaging colour introduces texture information in the form of a perfectly equal gradient.

Destruction of hue, saturation, and intensity is done through randomizing each of these features for each individual object. This retains texture within object since each object maintains a similar change in colour, but between objects any depth cue given by these features would be randomized.

Table 1: Comparison of results between original image, interventions of texture, hue, saturation, and intensity across the whole dataset.

Image type	A1	A2	A3	AbsRel	log10
KITTI original	0.9545	0.9896	0.9974	0.0796	0.0342
KITTI texture	0.8107	0.9388	0.9750	0.1468	0.0660
KITTI hue	0.9485	0.9901	0.9973	0.0833	0.0358
KITTI saturation	0.9449	0.9882	0.9967	0.0845	0.0363
KITTI intensity	0.8249	0.9526	0.9826	0.1332	0.0602
NYUv2 original	0.9714	0.9957	0.9990	0.0528	0.0229
NYUv2 texture	0.8751	0.9767	0.9946	0.1102	0.0477
NYUv2 hue	0.9703	0.9951	0.9988	0.0540	0.0234
NYUv2 saturation	0.9646	0.9937	0.9987	0.0588	0.0255
NYUv2 intensity	0.9265	0.9872	0.9966	0.0834	0.0362

3.3 Evaluation

Evaluation of results will be done using five performance metrics. Accuracy scores, which measure the fraction of pixels that falls within an acceptable threshold from ground truth at 1.25, squared and cubed; absolute relative error, which measures the absolute error between ground truth and prediction, divided by ground truth to obtain an error measured proportionally to ground truth; and log10 error, which measures the absolute error between the base 10 logs of prediction and ground truth. We chose these metrics due to their widespread use in existing work [4].

Once we have calculated loss metrics, we can use them to compare reduction in performance on intervened images. The difference in performance tells us how much the feature contributed to the model's prediction.

3.4 Data

We used the KITTI [21] and NYUv2 [16] datasets, since they have been widely used in outdoors and indoors studies respectively and can provide a good context to compare with other studies. Since Depth Anything claims not to be pretrained with either dataset, this also makes them good for a blind baseline test.

4 Results

Tab. 1 shows a comparison of results of the raw set of images and its interventions. A clear trend can be seen where hue sees the least drop in performance, followed by saturation, and large losses can be seen in intensity and texture. Analysis of these results follow.

Fig. 4: Example of texture intervention.
Top: Original, Middle: Prediction, Bottom: Ground truth dense map using infill

4.1 Texture

To test the contribution of texture on MDE, we performed inference on the phase-scrambled images of both datasets. Fig. 4 and Fig. 5 show examples of a comparison of raw images with their destroyed texture, and resulting depth maps. Texture appears consistently as the feature with the highest contribution. This can be explained by looking at texture as containing local shape information since the importance of shape has been shown in earlier works, and that it was identified as a key depth cue in human vision in [5].

Looking at Fig. 4, it is interesting to note the modes of failure. It can be seen that an image contains a section of road with destroyed texture. When inferenced using ground truth, the model is successful at predicting the road as an object with varying depth, but predicts the entire road as being of a single depth when texture is destroyed. This also applies for other images, for example with the backs and seats of chairs (Fig. 5).

It can also be seen that by destroying natural texture, the relationships between neighbouring objects are destroyed. Previously well predicted neighbouring objects now see jumps in depth, sometimes dramatically.

Results above show the importance of texture in depth prediction, both as depth cue and as differentiation between object boundaries.

4.2 Hue

To test the contribution of hue on MDE, we performed inference on the hue-randomized set of images for both datasets. Hue appears consistently as the least contributing feature within the four. This is unsurprising since the hue of

Figure-a

Left: original (A1: 0.9897) Right: intervened (A1: 0.9651)

Figure-b

Left: original (A1: 0.9719) Right: intervened (A1: 0.9200)

Fig. 5: Examples of texture interventions. With image curves adjusted for visibility. Top: Original, Bottom: Prediction

an object is rarely dramatically changed under natural and common artificial lighting situations. The hue of an object thus contains information about the nature of the object rather than their depth.

Looking at Fig. 6 (left) however, we can clearly see instances where hue affects the model's decisions on object boundaries, especially on reflective surfaces. It is especially notable that depending on how the segmentation model segments objects, the depth model draws different boundaries for the reflection of chairs. This suggests that while hue in and on itself does not strongly contribute to depth performance, its role in determining object boundaries might bring new insight on how to improve this challenging aspect of MDE.

4.3 Saturation

To test the contribution of saturation on MDE, we performed inference on the saturation-randomized set of images for both datasets. Saturation was the second most contributing colour feature for both KITTI and NYUv2. We found that in certain cases, intervening on an object and its neighbour's saturation result in changes to depth prediction, however this effect is not as strong as that seen in intensity. Saturation was hypothesized to be related to shadows and aerial perspective [5] in outdoor scenes.

There is no direct relationship between saturation and where depth prediction increased or decreased. Looking at Fig. 6 (right), we can see cases where decrease in distance prediction coming from both increased and decreased saturation. This is interesting, since it suggests that the model might have learnt to use saturation as a depth cue as a comparison tool between objects, and not just within a single object.

Fig. 6: Left: example of hue intervention. Right: example of saturation intervention.

4.4 Intensity

To test the contribution of intensity on MDE, we performed inference on the intensity-randomized set of images for both datasets. Out of the three colour features, intensity is the most contributing, although it contributes much more to the outdoors KITTI dataset than it does in NYUv2, which matches our intuition of strong shadows being used as depth cues in outdoor scenes.

Looking at qualitative results, it is found that in KITTI, intensity contributed greatly to object dissection and placement, especially in scenes where a strong vanishing point was unavailable (Fig. 7). A strong change in intensity here resulted in neighbouring objects to be projected to vastly different depths. This suggests that the model correlates a sudden change in intensity to a sudden change in depth.

A similar trend can be seen in the NYUv2 dataset, where sudden shifts of intensity are interpreted as a change in object occupying different depths (Fig. 8 (left)). However we also see cases where clusters of small objects changing intensity within themselves to not induce this behaviour, instead confounding the depth of the cluster(Fig. 8 (right)).

5 Conclusion

5.1 summary

In this study, we built upon existing work looking at MDE through a human-inspired angle, incorporating insights from various studies that aimed to improve interpretability in other computer vision tasks. We analysed the effects of four

different image features on MDE, comparing performance change and qualitatively analyzed the mechanics of change. We identified texture as the most contributing features within our tested features, and have shown a stable performance drop through the other features over both indoors and outdoors datasets, covering questions left unanswered by existing work. We quantitatively identified different contributing features in indoors and outdoors scenes, and made hypotheses to explain such differences, which can be extended for further studies.

Our work provides evidence that certain visual cues used by humans to identify depth are also used by machines. Such as texture density and boundary defined by intensity and saturation. We thus provide insight on how human and machine perception relate to each other.

By identifying the contributing features to learned relationships and their modes of failure, our work also provide a starting point to identifying robust features in MDE that models could be encouraged to focused on by various learning methods.

5.2 Limitations and future work

Current work has extended to looking at surface level features such as hue, saturation, and intensity, but as mentioned above it is important to use meaningful features to represent tasks. It is yet unclear what other features other than texture, shape, and HSV contribute to MDE. Thus, work going forward could be to extend a similar approach to features proposed by other studies, or to investigate inner model workings to look for other possible image features or latent features available for intervention.

The current method relies mainly on comparing overall performance with a few intervention examples, complemented by qualitative insights. A future direction would be to apply more robust statistical methods to further justify current qualitative insights by categorizing the types of failures that appear through intervention, such as by clustering data points by change of loss, or investigating change of deviation for different interventions.

The approach above can be reinforced by or taken from existing literature on causal discovery [17], where statistical methods can be used to identify whether features are direct descendants of each other, or whether they share confounding ancestors or are colliders.

Acknowledgements

This work was partially supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction (EP/V03538X/1) and partially by Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government (24ZC1200, Research on hyper-realistic interaction technology for five senses and emotional experience)

Left: original image (A1: 0.9588) Right: intervened image (A1: 0.8668)

Left: original image (A1: 0.9162) Right: intervened image (A1: 0.6039)

Fig. 7: Examples of intensity interventions outdoors. Note the difference in feature contribution between the two images, one with a clear vanishing point and the other without.

Left: Raw imput (0.9420)
Right: Intervened image (0.7408)

Left: Raw imput (0.9564)
Right: Intervened image (0.7948)

Fig. 8: Examples of intensity interventions indoors.

References

1. Alawadh, M., Wu, Y., Heng, Y., Remaggi, L., Niranjan, M., Kim, H.: Room acoustic properties estimation from a single 360° photo. In: 2022 30th European Signal Processing Conference (EUSIPCO). pp. 857–861 (2022). https://doi.org/10.23919/EUSIPCO55093.2022.9909598
2. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning (arxiv) (2019), https://arxiv.org/abs/1812.11941
3. Alvi, M., Zisserman, A., Nellaaker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (September 2018)
4. Arampatzakis, V., Pavlidis, G., Mitianoudis, N., Papamarkos, N.: Monocular depth estimation: A thorough review. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(4), 2396–2414 (2024). https://doi.org/10.1109/TPAMI.2023.3330944
5. Cutting, J.E., Vishton, P.M.: Chapter 3 - perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth*. In: Epstein, W., Rogers, S. (eds.) Perception of Space and Motion, pp. 69–117. Handbook of Perception and Cognition, Academic Press, San Diego (1995). https://doi.org/https://doi.org/10.1016/B978-012240530-3/50005-5, https://www.sciencedirect.com/science/article/pii/B9780122405303500055
6. Dijk, T.v., Croon, G.d.: How do neural networks see depth in single images? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
7. Dong, X., Garratt, M.A., Anavatti, S.G., Abbass, H.A.: Towards real-time monocular depth estimation for robotics: A survey. IEEE Transactions on Intelligent Transportation Systems **23**(10), 16940–16961 (2022). https://doi.org/10.1109/TITS.2022.3160741
8. Ge, Y., Xiao, Y., Xu, Z., Wang, X., Itti, L.: Contributions of shape, texture, and color in visual recognition. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 369–386. Springer Nature Switzerland, Cham (2022)
9. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
10. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9233–9243 (October 2023)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
12. Joblove, G.H., Greenberg, D.: Color spaces for computer graphics. In: Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques. p. 20–25. SIGGRAPH '78, Association for Computing Machinery, New York, NY, USA (1978). https://doi.org/10.1145/800248.807362, https://doi.org/10.1145/800248.807362
13. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
15. Liu, W., Liu, Z., Paull, L., Weller, A., Schölkopf, B.: Structural causal 3d reconstruction. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 140–159. Springer Nature Switzerland, Cham (2022)
16. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
17. Pearl, J.: Causality. Cambridge University Press, 2 edn. (2009)
18. Peuskens, H., Claeys, K.G., Todd, J.T., Norman, J.F., Hecke, P.V., Orban, G.A.: Attention to 3-d shape, 3-d motion, and texture in 3-d structure from motion displays. Journal of Cognitive Neuroscience **16**, 665–682 (2004), https://api.semanticscholar.org/CorpusID:14985355
19. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1623–1637 (2022). https://doi.org/10.1109/TPAMI.2020.3019967
20. Tuceryan, M., Jain, A.K.: Texture analysis, p. 235–276. World Scientific Publishing Co., Inc., USA (1993)
21. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: International Conference on 3D Vision (3DV) (2017)
22. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: CVPR (2019)
23. Wu, Y., Heng, Y., Niranjan, M., Kim, H.: Depth insight – contribution of different features to indoor single-image depth estimation (arxiv) (2023), https://arxiv.org/abs/2311.10042
24. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10371–10381 (June 2024)
25. You, Z., Tsai, Y.H., Chiu, W.C., Li, G.: Towards interpretable deep networks for monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12879–12888 (October 2021)
26. Zhang, K., Sun, Q., Zhao, C., Tang, Y.: Causal reasoning in typical computer vision tasks. Science China Technological Sciences **67**(1), 105–120 (Jan 2024)