This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

How does the Machine Perceive Depth for Indoor Single Images with CNN?

Yihong Wu[1], Yuwen Heng[2], Mahesan Niranjan[1], Hansung Kim[1]
[1]Department of Electronics and Computer Science, University of Southampton
YihongWu@soton.ac.uk, mn@ecs.soton.ac.uk, H.Kim@soton.ac.uk
[2]Bosch Corporate Research, Robert Bosch GmbH
Yuwen.heng@cn.bosch.com

Abstract

Depth estimation from a single image is a challenging problem in computer vision because binocular disparity or motion information is not available in the given input. Whereas impressive performances have been reported in this area recently using end-to-end trained deep neural architectures, as to what cues in the images are being exploited by these black box systems is hard to know. To this end, in this work, we quantify the relative contributions of the known cues of depth in a single-image depth estimation setting using an indoor scene data set. Our work uses feature extraction techniques to relate the single features of shape, texture, colour and saturation, taken in isolation, to predict depth. We find that the shape of objects extracted by edge detection substantially contributes more than others in the indoor setting considered, while the other features also have contributions in varying degrees. These insights will help optimise depth estimation models, boosting their accuracy and robustness. They promise to broaden the practical applications of vision-based depth estimation.

1. Introduction

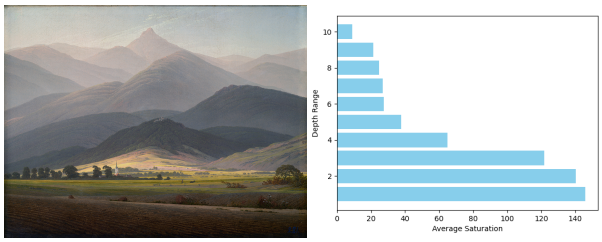
Depth estimation, predicting the distance from an object's surface to the camera, is a key task in the field of computer vision. It plays a crucial role in many applications, such as 3D reconstruction [1], autonomous driving [17, 35], virtual reality (VR) [6], augmented reality (AR) [20], etc. The goal of single-image depth estimation based on deep learning is to infer the depth value of each pixel by analysing the scene information in a single image.

Due to the ill-posed problem of single-image depth estimation, there is a fundamental need to move towards a scene understanding of objects in images so that various characteristics of objects can cue depth information. Recent work on single-image depth estimation using end-to-end trained deep neural network models shows that such cues are collectively learnable and satisfactory depth estimation can be achieved [2, 3, 8, 38]. However, the black-box nature of such models prohibits the understanding of what cues are exploited in single-image depth estimation. The mechanism behind single-image depth estimation based on 2D images in CNNs is still not clearly explained, and the extent to which these models can approximate the human capability of monocular depth perception remains uncertain.

Building on this gap in understanding, and inspired by causality analysis [21], we aimed to investigate the factors that influence depth estimation. This work will pave the way for the development of versatile models applicable to a broader spectrum of depth estimation tasks, moving beyond reliance solely on data-driven approaches. Research has shown that single-image depth cues in 2D images include phenomena such as blurring, shading and brightness [31]. This paper investigated and analysed the factors that influence CNN-based single-image depth estimation using CNNs. To provide a comprehensive understanding, we investigated the roles of the factors relevant to object recognition [11], such as colour and texture, in the context of single-image depth estimation. However, many factors are interrelated and cannot be independently segregated. In the context of scenarios where it is possible to directly and independently extract features from 2D images, we have considered colour, saturation, texture and shape, and each of them holds significant relevance in image processing, exerting varying degrees of impact on the overall outcome.

Colour. Colour is recognised by the perception and interpretation of different wavelengths of light by the eye [15]. The visual information humans gather heavily relies on the presence of colour [23]. Colour helps humans recognise and remember objects faster [12]. Nevertheless, when defining colour, it is critical to recognise that RGB images do not only represent a singular colour but also include various elements in addition to colour, such as shape and texture. To isolate the pure colour information, we utilised a phase scrambling approach [11], which effectively separates the colour from these additional attributes.

Saturation. The second feature of interest is saturation.



(a) Nature Scene Image (b) Saturation Distribution

Figure 1. Saturation Analysis for a Nature Scene

Saturation refers to the purity or intensity of a colour. For instance, high saturation indicates a more vivid and pure colour, while low saturation suggests a lighter or more desaturated colour with a hint of grey. Aerial perspective, within the domain of remote viewing, refers to the impact of the atmosphere on the visual depiction of an object. For instance, in Figure 1a, a nature photograph is displayed. We evenly split images into ten rows, and the average saturation values have been calculated for each row, as depicted in Figure 1b. As the object moves away from the camera, it can be observed that the saturation decreases. Building on this observation, saturation serves as a depth cue for outdoor single-image depth estimation. We aimed to investigate the utility of saturation as a depth cue in indoor scenes.

Texture. In computer vision, texture is defined by repetitive patterns with varying intensities present in an image[33]. Prior research has found that textures are important when influencing a human's perception of distance [28], with specific regions in the brain having been found to be activated when exposed to varying textures [24]. Therefore, we also sought to independently extract the features pertaining to texture and assess their impact on depth estimation.

Shape. A shape is generally considered to be a graphical representation of an object or its external borders, contours or external surfaces. Acquiring precise boundaries of objects in the 3D world based solely on 2D images is challenging. To simplify this process, we defined the shape feature as the edge graph, which corresponds to a greyscale map generated using an edge detection algorithm designed to preserve the object's boundaries. Edges are regarded as one of the primary cues essential for the human visual system [9]. Edge graphs usually represent geometric structures or boundaries between objects. For depth estimation tasks, the geometric features of an object are crucial to inferring its depth. The geometric structure aids depth estimation algorithms in capturing the shapes and relationships between objects [18], with edge maps providing supplementary geometric information. Edge detectors can analyse pixel gradients in different areas of the image, thereby assisting in the estimation of the relative distances between objects.

Our work aims not at improving the performance of any black-box depth estimation architecture but on explaining what such algorithms learn from the features. As such, we offer an analysis of how much each aspect of the input contributes towards the output. Such analysis is important in understanding how CNNs solve depth estimation problems, and ours is the first step in that direction (bulk of the literature is on performance). We believe our research will be a good foundation for other researchers working on explainable AI and developing more efficient depth estimation architectures.

2. Related Work

Interpretability within deep learning is attracting significant and growing interest. Interest in Convolutional Neural Networks (CNNs) and Visual Transformers has been rapidly increasing lately, particularly concerning their interpretability. A study revealed that CNN models trained on ImageNet exhibit a heightened sensitivity to texture information [13]. In addition, recent research has undertaken a comparison of the attributes between CNNs and Transformers across various layers [26] by using Centred Kernel Alignment (CKA) [4, 19]. According to their claims, the transformer allows the early gathering of global information in contrast to CNNs. This results in a robust propagation of features from lower to higher layers in the network. Nevertheless, the primary focus of these enquiries remains centred on model analysis.

In the realm of human depth perception, substantial work has been conducted to investigate cues such as position in the image, texture density and focus blur [5, 14]. Existing works have demonstrated various methods for indoor single-image depth estimation that exhibit good performance [3, 8]. Despite this, an analysis of their operations is still lacking. To the best of our knowledge, there has been no analysis of the contributions of different types of depth cues specifically for deep learning-based depth estimation in indoor single-image scenarios. In the two most relevant prior studies to our work, [16] conducts attribution analysis to identify pixels that contribute most significantly to the final depth map. However, these methods can only offer insights into the low-level workings of CNNs. The analysis in [7] was primarily confined to specific objects situated in outdoor environments, such as animals and vehicles on roadways. In contrast, in our study, we focused on colour, saturation, texture and shape, taking into account that our target application pertains to indoor scenes and requires the extraction of these cues from a single image.

Approaches for emulating the human capacity for gauging depth from an indoor scene still encounter gaps in knowledge. The objective of this paper is to unveil how CNNs extract depth-related information from a single indoor image to attain a more profound comprehension of the disparities between monocular visual depth estimation and

(a) Original RGB (b) Corresponding GT Depth

Figure 2. A Sample from the NYU Dataset

the depth perception exhibited by humans.

Simultaneously, our work offers a foundational framework to facilitate subsequent investigations into assessing the interdependence among pertinent variables in the realm of depth estimation. A prior exploration delved into the causal interplay within 3D reconstruction, deconstructing elements like perspective and depth while also attempting to substantiate the interlinkages amid diverse variables [21]. Nonetheless, the model expounded upon in this enquiry operates on the assumption that the object is symmetric [36]. Through an autoencoder mechanism, it internally dissects the input image into manifold components, as opposed to explicitly extracting a corresponding viewpoint, depth and related insights from the RGB image. In our study, we exclusively extracted various factors from RGB images while carefully investigating the significance of these factors within the context of depth estimation using CNNs. Our work is set to further enable causal analyses in the field of depth estimation, paving the way for future advancements in comprehending the causality of depth estimation.

3. Method

In this section, we consider four cues for single-image depth estimation. In order to compare the appearance of these four different features, we use the same sample in this section. Figure 2 shows the original RGB image and its corresponding ground truth (GT) depth from the NYU dataset [30].

3.1. Colour

Figure 3 illustrates the relationship between the distribution of original RGB three-channel values and the depth maps. The pixels on original RGB images are primarily concentrated between 0 and 100 in the corresponding grey-scale depth maps. The heat map reveals that the values of R, G and B pixels are similarly distributed across a specific depth range. This shows that the factors affecting depth are not significantly related to the distribution of pixels on the RGB channel. More details are shown in the Supplementary Material.

Hue from the hue, saturation and luminance value (HSV) colour space can be an expression of colour. However, hue values represent the projection of the RGB colour space

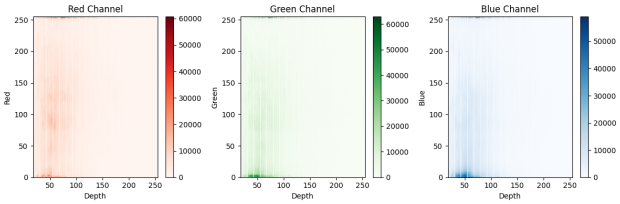


Figure 3. Heat Map of the Relationship between the Distribution of Original RGB Three-channel Values and the Depth Map. The horizontal axis represents the depth range, while the vertical axis corresponds to the pixel count of the R, G and B channels within the respective depth ranges. The colour bar values represent the pixel counts for three respective channels from 500 images randomly selected from the NYU dataset.

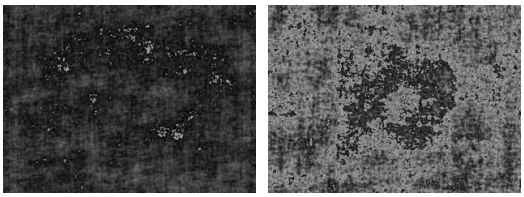
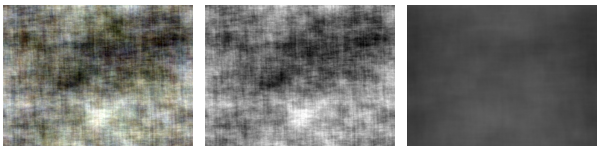


Figure 4. Phase Scrambled H Map and Corresponding Depth Map of Figure 2

onto a non-linear chroma angle [32]. If an output pixel value falls outside the valid range, it necessitates remapping to bring it within the specified range. The chroma angle represents a non-linear trajectory within a continuous, uninterrupted space. Here, starting at 0 degrees is the same as coming full circle to 360 degrees. However, when we apply this idea to an image, like with H maps, the smooth flow is interrupted, creating a series of separated points instead. Figure 4 illustrates the images and corresponding depth maps resulting from the phase scrambling and remapping process applied to the H map from Figure 2, which are mapped back to specific intervals. Some discontinuous blocks can be observed in this figure. Therefore, we did not consider utilising the hue maps as the colour feature.

To examine the contribution of colour to depth, we performed phase scrambling [11] on original RGB images (details shown in Supplementary Material) and their corresponding depth maps to remove influences from shapes, textures and other geometric features. The resulting dataset was labelled "RGB Phase Scrambled". Nevertheless, even after the phase scrambling, the outcome still retains the brightness information, making it not purely a colour feature. Subsequently, these outputs were converted to greyscale, effectively removing the colour information, and the resulting dataset was labelled as "Greyscale Phase Scrambled". To illustrate the role of colour, a comparison of these two phase-scrambled features is conducted.



(a) RGB Image Phase Scrambling (b) Gray-scale Image Phase Scrambling (c) Depth Phase Scrambling

Figure 5. Phase Scrambling Results of Figure 2

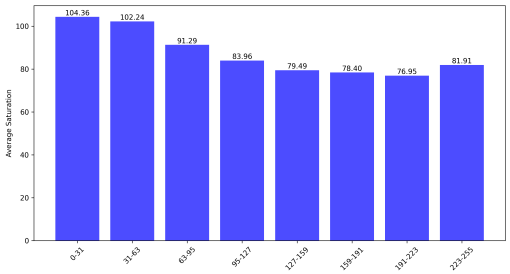


Figure 6. Average Saturation at Different Depth Intervals for Indoor Scenes (NYU dataset)

3.2. Saturation

We investigated whether saturation varies at different depths in indoor scenes. We partitioned this depth range 0-255 in the NYU dataset into eight segments and then calculated the average saturation for each by converting RGB to HSV colour space and extracting the saturation values. Figure 6 shows the average saturation of the NYU dataset in different depth ranges. Based on the observations, it appears that saturation may have less influence on the results for indoor scenes, different from the result for outdoor scenes shown in Figure 1.

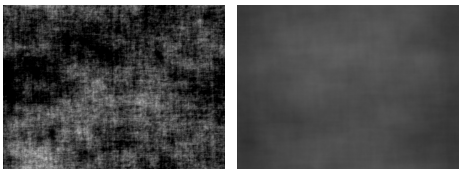
Nevertheless, we intended to further investigate the extent to which this subtle difference can affect depth estimation. In addition, as mentioned above, human depth perception can be influenced by saturation. To assess the specific contribution of saturation, we extracted the saturation feature for experimentation independently.

To extract the features pertaining to saturation, we started by converting the RGB colour space to the HSV colour space and then extracting the saturation maps. Subsequently, these saturation maps are subjected to phase scrambling to eliminate features such as shape, texture and other visual characteristics.

$$\mathrm{V} \leftarrow \max(\mathrm{R}, \mathrm{G}, \mathrm{B}) \tag{1}$$

$$\mathrm{S} \leftarrow \begin{cases} \frac{\mathrm{V}-\min(\mathrm{R},\mathrm{G},\mathrm{B})}{\mathrm{V}} & \text{if } \mathrm{V} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

For each pixel, the V maps are obtained by taking the maximum value (shown in Equation (1)) among the RGB channels. Subsequently, the saturation feature is obtained based on phase scrambling from S maps (shown in Equation (2)). As depicted in Figure 7, Figure 7a illustrates the saturation feature, with Figure 7b displaying its corresponding depth map.



(a) Saturation Feature (b) Saturation Depth

Figure 7. Saturation with Phase Scrambling of Figure 2

(a) Shape with Sobel (b) Shape with Canny (c) Depth

Figure 8. Shape Maps from Figure 2 and Corresponding Ground Truth Depth

3.3. Local Texture

The preference for local textures over global textures stems from the fact that the extraction of global textures includes the consideration of additional factors, including shape and other features. To mitigate the influence of other factors and preserve the texture, the images were segmented into patches and shuffled, thus eliminating global information such as shapes, since this information introduces more features than just textures.

3.4. Shape

The boundaries of an object define its precise outline, marking the separation between the object and its immediate environment. Edge maps are generated through the analysis of gradient variations in image pixel values and identify changes in these values. Although edge maps do not always faithfully represent real object boundaries, when dealing with a single 2D image, they offer an efficient means of simulating object shapes. This feature has been defined as 'shape' for the subsequent experiment.

As shown in Figure 8, we utilised the Canny operator instead of the Sobel operator because the latter will find the gradient in the x and y directions, reflecting the differential changes of pixels [32]. Therefore, not only the shape feature is included when using the Sobel operator, but some texture information may also be introduced.

4. Experiments

As mentioned above, we considered four factors that may contribute to depth estimation: colour, saturation, local texture and shape. All of these features were trained using the baseline model, and the obtained results were analysed.

4.1. Data

We used the NYU dataset [30], which serves as a widely employed dataset in computer vision, particularly for depth estimation research. Comprising images from diverse indoor scenes, it encompasses a variety of objects and furniture. The size of the NYU dataset enhances the representativeness of model training and evaluation. It is derived from 464 scenes in three cities. The resolution of the images is 640×480. 10% of the data is split as the testing set.

4.2. Model

The UNet architecture is preferred for deep learning-based depth estimation due to its comprehensive design, adept at gathering context and integrating features across different scales [2, 3, 8, 38]. This preference is rooted in UNet's feature pyramid structure and efficient reuse of features, enhancing depth estimation by capturing diverse scale information while preserving detail. Our experiments demonstrate that employing ResNet50 as the backbone is sufficient for model convergence on our dataset. Subsequently, we utilised the U-Net network with ResNet50 as the backbone in the following experiment.

Note that we did not compare SOTA models because the focus of our work was not on the performance of the models, but rather on the contribution of various features to indoor single-image depth estimation based on a stable baseline model.

4.3. Evaluation Metrics

We utilised six metrics commonly used in the field of depth estimation [8, 37], which include three accuracy metrics and three error metrics. The accuracy metrics are distinguished by thresholds at 1.25, 1.25^2 and 1.25^3, each reflecting different levels of tolerance for deviation from the true values. Higher values of these accuracy metrics indicate better model performance. For error metrics, the absolute relative error (rel) quantifies the average deviation of predicted values from the actual values. The root mean squared error ($rmse$) can amplify the effect of outliers by taking the square root of the average of the squared deviations from the ground truth, and the logarithmic error (log_{10}) metric mitigates the impact of outliers by applying a logarithmic scale to the error values. Lower values of these error metrics signify superior model performance.

(a) RGB Image (b) GT Depth (c) Estimated Depth

Figure 9. Depth Estimation with a Colour Feature Input. The left and middle images are the original RGB image and the corresponding ground truth depth map, respectively. The image on the right depicts the estimated depth map, which is the result of the model's output after inverse phase scrambling, employing the colour feature as the input.

4.4. Experiments and Analysis

Quantitative results are shown in Table 1, presenting the performance of depth estimation using different input features, evaluated by several metrics (details shown in Sec.4.3). Original RGB images performed the best with high accuracy (a_1, a_2, a_3) and low error (log_{10}, rel, $rmse$). Phase-scrambled RGB and greyscale images, along with saturation inputs, showed significantly worse performance, with greyscale being the least accurate. Inputs of local texture had moderate accuracy and error rates, while shape features performed close to the original RGB images.

4.4.1. Colour

To evaluate the contribution of the colour feature, we trained the model with phase-scrambled RGB images. Figure 9 displays the original RGB image, ground truth depth map and the estimated depth map, the latter of which has been reconstructed from the scrambled image using a pre-stored random matrix. As aligned to the low accuracy indicated in Table 1, it is hard to recognise the original scene structure from the estimated depth.

To simulate scenarios where the model output differs from the ground truth, we added Gaussian noise (mean = 0, std = 25) to the phase scrambled image. Figure 11 shows examples of our phase scrambled image with added Gaussian noise and their corresponding reconstructions. Figure 10 shows the outcomes of introducing Gaussian noise to the phase-scrambled image, followed by its restoration using the pre-stored random matrix. As we can see, despite the introduction of noise through phase scrambling, this noise does not affect the shape and position of objects in the recovered images. The performance in Figure 9c can be attributed to the poor performance of the model when provided with colour phase-scrambled input.

Furthermore, by comparing the respective performances of "RGB Phase Scrambled" and "Grayscale Phase Scrambled" inputs as shown in Table 1, it can be observed that, after excluding the contribution of brightness, colour has a limited impact on depth estimation.

Table 1. Depth Estimation Performance with Different Inputs

Features	$a_1 \uparrow$	$a_2 \uparrow$	$a_3 \uparrow$	$log_{10} \downarrow$	$rel \downarrow$	$rmse \downarrow$
Original RGB Images	98.13 ± 0.0013	99.64 ± 0.0003	99.9 ± 0.0001	0.0176 ± 0.0001	0.0413 ± 0.0008	0.0174 ± 0.0002
RGB Phase Scrambled	43.5 ± 0.0067	72.64 ± 0.0068	87.47 ± 0.0042	0.1498 ± 0.0021	0.4754 ± 0.017	0.113 ± 0.0019
Grayscale Phase Scrambled	36.13 ± 0.0316	64.09 ± 0.041	81.65 ± 0.033	0.1769 ± 0.0143	0.5627 ± 0.0585	0.1364 ± 0.0149
Saturation	36.9 ± 0.0094	65.35 ± 0.0113	82.86 ± 0.0065	0.1718 ± 0.0026	0.5321 ± 0.0208	0.1296 ± 0.0015
Local Texture	49.95 ± 0.0286	77.88 ± 0.0228	90.83 ± 0.0114	0.1276 ± 0.0068	0.3187 ± 0.0215	0.1065 ± 0.0047
Shape	96.46 ± 0.0003	99.12 ± 0.0002	99.71 ± 0.0002	0.0235 ± 0.0001	0.0556 ± 0.0004	0.0224 ± 0.0001

(a) Noise for Whole Image (b) Noise in the Center

Figure 10. Noised with Phase Scrambled Images. Figure 10a illustrates the outcome of applying Gaussian noise to the entire image and subsequently restoring it, while Figure 10b depicts the results of adding noise and restoring only the central area, where both the length and width are half of the original image's dimensions.

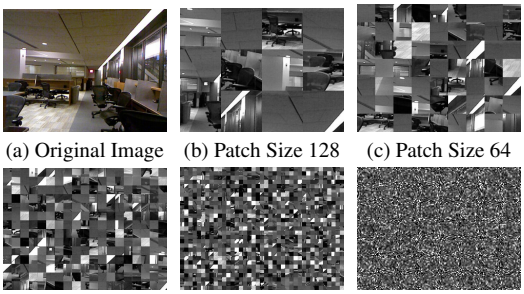
(a) Saturation Map (b) GT Depth (c) Estimated Depth

Figure 11. Depth Estimation with a Saturation Feature Input. The left and middle images are the original RGB image and the corresponding ground truth depth map, respectively. The image on the right depicts the estimated depth map, which is the result of the model's output after inverse phase scrambling, employing a saturation map as the input.

4.4.2. Saturation

We trained the baseline model with saturation maps as the input and evaluated the contribution of the saturation feature. Figure 11 illustrates the saturation map, corresponding ground truth depth and the output from the restoration process. Similarly, due to the poor performance, the restored output lacks discernible features such as object contours.

Table 1 shows that the a_1 is about 37%. Although saturation contributed to estimating the depth of the indoor scene, its contribution was minor. Saturation exhibits lower accuracies compared to other features except for greyscale phase scrambled input. Furthermore, error metrics substantiate this observation. The rel stands at 0.5321, while the root mean square error ($rmse$) is shown as 0.1296. Therefore, using saturation as a measure for depth estimation clearly introduces a significant error compared to the true depth values. Despite its poor performance, saturation still plays a role in assessing depth in indoor scenes. This highlights that saturation can provide some depth cues in certain contexts, although it comes with a higher error margin.



(a) Original Image (b) Patch Size 128 (c) Patch Size 64

(d) Patch Size 32 (e) Patch Size 16 (f) Patch Size 4

Figure 12. Local Texture with Different Patch Sizes of a Random Sample

4.4.3. Local Texture

Variations in the field of view and resolution will impact the size of the patch used to extract local textures. The optimal patch size should align with the specific dataset and scene scale employed. Figure 12 illustrates the paths with varying patch dimensions. As shown in Figure 12b, when we use a large patch size of 128, the texture itself is not isolated because the shape and context information still present in the patches. As the patch size is decreased, the shape of the objects in the image becomes less apparent and, therefore, the textures present in the image are increasingly segregated. As demonstrated in Figure 12e, for the dataset we used, the 16×16 patch size is particularly well-suited for texture extraction while minimising the influence of other features (e.g. shape). This size is large enough to restrict shape details but not so small as to be impractical, unlike the 4×4 patch depicted in Figure 12f.

Table 2 shows the performance of texture inputs (shuffle patches) in different patch sizes. As can be seen in the figure, the accuracy rate gradually increases with the increase of patch sizes in height. This is because a larger

Table 2. Performance with Different Patch Sizes

Size	$a_1 \uparrow$	$a_2 \uparrow$	$a_3 \uparrow$	$log_{10} \downarrow$	$rel \downarrow$	$rmse \downarrow$
4	40.97 ± 0.0037	69.6 ± 0.003	85.99 ± 0.0018	0.1523 ± 0.0008	0.3812 ± 0.0026	0.1239 ± 0.0007
16	49.95 ± 0.0286	77.88 ± 0.0228	90.83 ± 0.0114	0.1276 ± 0.0068	0.3187 ± 0.0215	0.1065 ± 0.0047
32	53.27 ± 0.042	80.64 ± 0.0253	92.46 ± 0.0104	0.1185 ± 0.0087	0.3012 ± 0.0266	0.1009 ± 0.0094
64	74.22 ± 0.0166	92.48 ± 0.0122	97.51 ± 0.0057	0.0747 ± 0.0039	0.1885 ± 0.0199	0.0629 ± 0.002
128	93.12 ± 0.0066	98.47 ± 0.0019	99.5 ± 0.0007	0.0358 ± 0.0016	0.0863 ± 0.0039	0.0338 ± 0.0015

(a) Gray-scale Image (b) GT Depth (c) Estimated Depth

Figure 13. Depth Estimation with a Local Texture Input with Patch Size 16×16. The left and middle images are the original RGB image and the corresponding ground truth depth map, respectively. The image on the right depicts the estimated depth map, which is the result of the model's output after inverse shuffle by using a pre-stored random matrix, employing a local texture feature as input.

(a) Shape Map (b) GT Depth (c) Estimated Depth

Figure 14. Depth Estimation with a Shape Feature Input. The left and middle images are the original RGB image and corresponding ground truth depth map. The right is the estimated depth map from the model trained with shape maps.

patch contains more information besides the texture, such as the shape of the object.

A sample of the original greyscale image, along with the corresponding depth map and the estimated depth map, is displayed in Figure 13, demonstrating the results of the model trained with local texture inputs. To focus on local textures during training, Figure 13a and Figure 13b are split into 16×16 patches and these patches are shuffled using the same random matrix to eliminate global scene information, such as object shapes. As shown in Figure 13c, the estimated depth map only provides a coarse approximation of the scene's depth, distinguishing between nearer and farther areas but failing to capture the precise depth details.

The local texture appears as a minor factor in depth estimation, yielding an a_1 accuracy of a mere 50% in Table 1. Error metrics also show this trend although they are slightly better than the colour and saturation features. This issue happens because the position changes to the patches weaken their connections, thereby making it harder for the model to understand objects and context. This proposition finds corroboration in the robust performance observed upon deploying the shape feature as the input data source in Sec 4.4.4.

4.4.4. Shape

As we noted in Table 1, the shape comes across as the most dominant feature in these experiments, significantly outperforming other cues taken in isolation. We suggest this is because the dataset contains indoor scenes of objects such as furniture with accurately extractable edges whose relative orientations and geometric forms can serve as powerful cues as seen in Figure 14 (b and c).

The outcome aligns with the finding presented in [16], suggesting that CNNs are capable of deducing the depth map using merely a limited subset of pixels from the input image. This hypothesis aligns with human perceptual abilities, which allow for the extraction of approximate distance assessments from images that depict geometric shapes.

4.4.5. Generalisation

In light of the fact that models using shape maps as input exhibit performance approximating that of models employing original RGB images as the input, we have assessed the generalisation capacity of shape models trained with shape maps on the NYU dataset. We applied it to a diverse set of indoor environments from a different dataset [25] that includes kitchens, bedrooms, bathrooms and various other scenes. The performance of the shape model is depicted in Figure 15, illustrating its ability to predict depth maps even for scenes from a different domain, and the performance is similar to that of the original RGB model. However, shape maps, as input for depth estimation, still have their limitations. For instance, in the fourth-row images in Figure 15, the sink only has partial edges, leading to poor depth prediction. Additional results are presented in the Supplementary Material.

Shape maps require significantly less memory storage compared to original RGB images, while still providing comparable performance. In a similar vein, event cameras are designed to only detect rapid changes in pixel intensity [27, 29] which often occur at the edges of objects or where there is texture variation, which is similar to a shape map. Moreover, event cameras have previously been applied in the field of depth estimation [10]. Given these considerations, our research may offer supporting evidence for the application of event cameras in single-image depth estima-

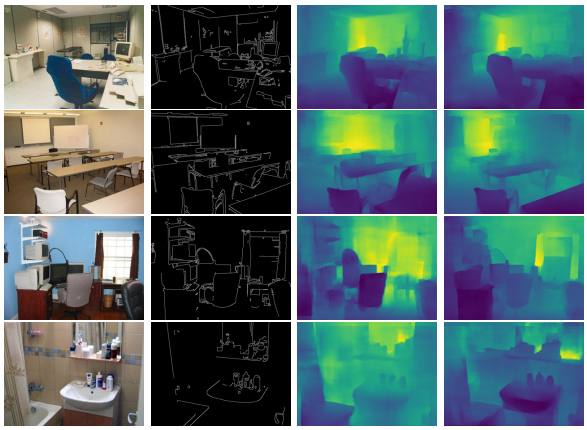


Figure 15. Performance of Shape ONLY Model with New Indoor Scenes from other Domains. The left column displays original RGB scene images, the second column presents corresponding edge maps and the third column showcases the results generated by the pre-trained shape-input model. The right column exhibits the outcomes produced by the pre-trained original-RGB-image-input model.

tion.

In addition to this, our work will also contribute to 3D sketch reconstruction. A sketch input is a hand-drawn image that is primarily used to capture the basic shape and key features of an object [22, 34]. It shows the overall outline and main features of the object, and this contour information is linked to the concept of 'shape' discussed in our work. For example, [22] convert sketches into multi-view 2D images that capture the surface depth and normals, then fuse these into a 3D point cloud and eventually convert them into a polygonal mesh, and finally generate detailed 3D models from simple line sketches. Therefore, our study may also provide supporting evidence for 3D sketch reconstruction.

4.4.6. Discussion

Different types of input data have varying effects on the performance of depth estimation. Comparative analysis of diverse evaluation metrics clearly highlights the superior role of shape information in the depth estimation task. Colour, saturation and local texture collectively enhance the indoor scene depth estimation, although the influence of colour and saturation appears relatively circumscribed.

For phase-scrambled and local texture inputs, human vision finds it difficult to interpret images when their phase information is scrambled or only shuffled patches are present. In contrast, machines are adept at using these inputs to predict depth maps. Given that the models can output corresponding depth maps when employing these as inputs, their performance, albeit not optimal, is still noteworthy compared to the ability of humans.

5. Limitations

Throughout our study, we sought to isolate each feature we were evaluating. However, it is difficult to entirely isolate individual features. For instance, during the extraction of shape features, the edge detector might inadvertently capture some texture information. Additionally, our reliance on the NYU dataset, which is tailored to indoor scenes, and the use of a CNN-based model may limit the generalisability of our findings to other environments or architectures.

These challenges present exciting avenues for future research. Extending this work to outdoor scenes and exploring a broader range of models, such as Vision Transformers, could enhance the robustness and applicability of our insights. Investigating the interaction effects of multiple depth cues—such as colour combined with texture—and refining feature isolation techniques could further deepen our understanding of depth estimation, paving the way for more comprehensive and practical advancements.

6. Conclusion

In this work, we have sought to decouple and quantify the relative contributions of various depth cues in single-image depth estimation. Whereas good results have been demonstrated in the literature by the end-to-end training of deep neural network models to achieve this task, ours is the first attempt to understand the degree to which some known cues of depth contribute when taken in isolation. Our results show that, in a dataset of indoor scenes, shape extracted by edge detection is relatively the most significant contributor, while other cues (colour, saturation and texture) also play a role. In achieving these conclusions, this work sought to carefully design feature extraction techniques that aimed to isolate a single feature from the other known ones, which is non-trivial. We speculate (and this is the subject of our current research) that, on different depth inference problems (e.g. outdoor scenes), the relative contributions of texture and saturation are likely to play a greater role. This kind of decomposition which we have extracted can serve to shift research more in the direction of understanding and explaining how powerful models, such as deep neural networks, work in scene understanding as opposed to simply offering estimation performance as black-box function approximators.

Acknowledgement

This work was supported 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)

References

[1] Mona Alawadh, Yihong Wu, Yuwen Heng, Luca Remaggi, Mahesan Niranjan, and Hansung Kim. Room acoustic properties estimation from a single 360° photo. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 857–861. IEEE, 2022. 1

[2] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018. 1, 5

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1, 2, 5

[4] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012. 2

[5] James E Cutting and Peter M Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pages 69–117. Elsevier, 1995. 2

[6] Anthony Dickson, Alistair Knott, and Stefanie Zollmann. Benchmarking monocular depth estimation models for vr content creation from a user perspective. In *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2021. 1

[7] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2183–2191, 2019. 2

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27, 2014. 1, 2, 5

[9] Muhammad Shahid Farid, Maurizio Lucenteforte, and Marco Grangetto. Edges shape enforcement for visual enhancement of depth image based rendering. In *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, pages 406–411. IEEE, 2013. 2

[10] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020. 7

[11] Yunhao Ge, Yao Xiao, Zhi Xu, Xingrui Wang, and Laurent Itti. Contributions of shape, texture, and color in visual recognition. In *European Conference on Computer Vision*, pages 369–386. Springer, 2022. 1, 3

[12] Karl R Gegenfurtner and Jochem Rieger. Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, 10(13):805–808, 2000. 1

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[14] James J Gibson. *The perception of the visual world.* Houghton Mifflin, 1950. 2

[15] Andrzej Grzybowski and Konrad Kupidura-Majewski. What is color and how it is perceived? *Clinics in dermatology*, 37 (5):392–401, 2019. 1

[16] Junjie Hu, Yan Zhang, and Takayuki Okatani. Visualization of convolutional neural networks for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3869–3878, 2019. 2, 7

[17] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. 1

[18] Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, and Shenghua Gao. Geometric structure based and regularized depth estimation from 360 indoor imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2020. 2

[19] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 2

[20] Wonwoo Lee, Nohyoung Park, and Woontack Woo. Depth-assisted real-time 3d object detection for augmented reality. In *ICAT*, pages 126–132, 2011. 1

[21] Weiyang Liu, Zhen Liu, Liam Paull, Adrian Weller, and Bernhard Schölkopf. Structural causal 3d reconstruction. In *European Conference on Computer Vision*, pages 140–159. Springer, 2022. 1, 3

[22] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. 3d shape reconstruction from sketches via multi-view convolutional networks. In *2017 International Conference on 3D Vision (3DV)*, pages 67–77. IEEE, 2017. 8

[23] Maureen Neitz and Jay Neitz. Molecular genetics of color vision and color vision defects. *Archives of ophthalmology*, 118(5):691–700, 2000. 1

[24] Aina Puce, Truett Allison, Maryam Asgari, John C Gore, and Gregory McCarthy. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of Neuroscience*, 16 (16):5205–5215, 1996. 2

[25] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009. 7

[26] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 2

[27] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018. 7

[28] James H Rowland. *The effects of texture on distance estimation in synthetic environments*. PhD thesis, Monterey, California; Naval Postgraduate School, 1999. 2

[29] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 156–163, 2020. 7

[30] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 3, 5

[31] Cassandra T Swain. Integration of monocular cues to create depth effect. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2745–2748. IEEE, 1997. 1

[32] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 3, 4

[33] Mihran Tuceryan and Anil K Jain. Texture analysis. *Handbook of Pattern Recognition and Computer Vision*, pages 235–276, 1993. 2

[34] Fei Wang, Yu Yang, Baoquan Zhao, Junkun Jiang, Teng Zhou, Dazi Jiang, and Tie Cai. Deep 3d shape reconstruction from single-view sketch image. In *2020 8th International Conference on Digital Home (ICDH)*, pages 184–189. IEEE, 2020. 8

[35] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[36] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 3

[37] Yihong Wu, Yuwen Heng, Mahesan Niranjan, and Hansung Kim. Depth estimation from a single omnidirectional image using domain adaptation. In *European Conference on Visual Media Production (CVMP)*, pages 1–9, 2021. 5

[38] Yihong Wu, Yuwen Heng, Mahesan Niranjan, and Hansung Kim. Depth estimation for a single omnidirectional image with reversed-gradient warming-up thresholds discriminator. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 5