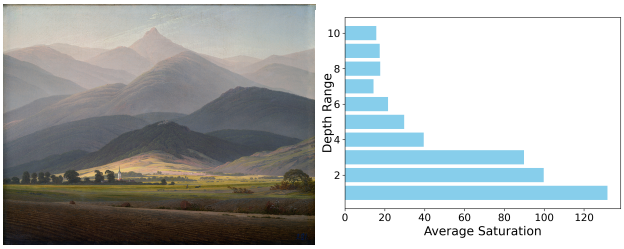
# APPENDIX: HOW DOES THE MACHINE PERCEIVE DEPTH FOR INDOOR SINGLE IMAGES?




(a) Nature Scene Image (b) Saturation Distribution

**Fig. 1**: Saturation Analysis for a Nature Scene

## 1. INTRODUCTION

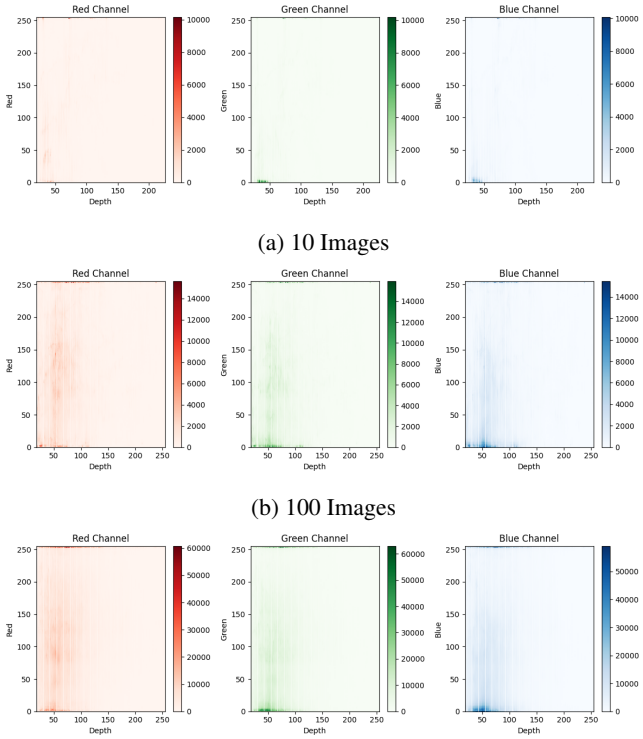
### Saturation Aerial Perspective

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.

## 3. METHODOLOGY

### 3.1. Colour

Figure 2 represents the results of the accumulated values obtained from datasets of 50, 100, and 500 randomly sampled images. The depth range is depicted on the horizontal axis, while the vertical axis indicates the number of pixels in the R, G, and B channels corresponding to the specific depth ranges. These images show that the factors affecting depth are not significantly related to the distribution of pixels on the RGB channel.




(a) 10 Images

(b) 100 Images

(c) 500 Images

**Fig. 2**: Heatmap of the relationship between the distribution of 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 different numbers of images randomly selected from the NYU dataset.

Listing 1: Pseudocode of Phase Scrambling [1]

```
imFourier = fft2(input)
Amp = abs(imFourier)
Phase = angle(imFourier)
Phase = Phase + RandomPhase
imScrambled = ifft2(Amp * exp(1j * Phase))
imScrambled = GetRealPart(imScrambled)
```

List 1 presents the pseudo-code for the phase scrambling process.

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 onto a non-linear chroma angle [2]. 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 3 illustrates the images and corresponding depth maps resulting from the phase scrambling and remapping process applied to the H map from Figure 5, 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

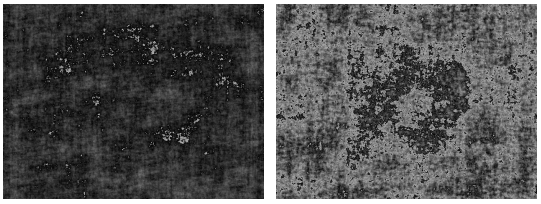


**Fig. 3**: Phase Scrambled H Map and Corresponding Depth Map of Figure 5

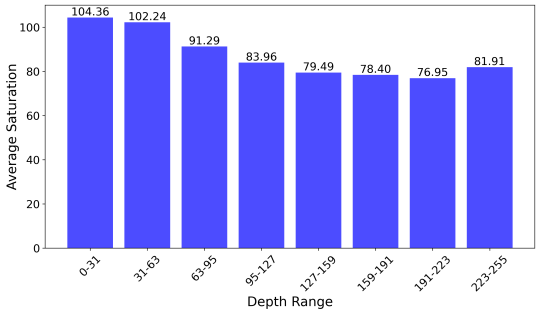



**Fig. 4**: Average Saturation at Different Depth Intervals for Indoor Scenes (NYU data set)

as the colour feature.

## 3.2. Saturation

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

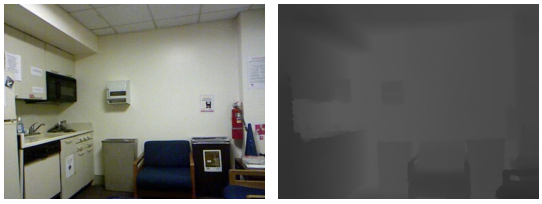
$$\mathrm{S} \leftarrow \begin{cases} \frac{\mathrm{V}-\min(\mathrm{R},\mathrm{G},\mathrm{B})}{\mathrm{V}} & \text{if } \mathrm{V} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

For each pixel, the V maps are obtained by taking the maximum value (Eq.1) among the RGB channels. Subsequently, the saturation feature is obtained based on phase scrambling from S maps (Eq.2).

We investigated whether saturation varies at different depths in indoor scenes. We partitioned this depth range 0-255 in the NYU data set into eight segments and then calculated the average saturation for each by converting RGB to HSV colour space and extracting the saturation values. Figure 4 shows the average saturation of the NYU data set 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 4.

# 4. EXPERIMENTS

Figure 5 shows the original RGB image and its corresponding ground truth (GT) depth from the NYU data set [3].



(a) Original RGB (b) GT Depth

**Fig. 5**: A Sample from NYU Dataset

### 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 [4, 5, 6, 7]. 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.

### Evaluation Metrics

We utilised six metrics commonly used in the field of depth estimation, 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.

### 4.1. Colour

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 6 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.

(a) Noise for Whole Image (b) Noise in Central Region

**Fig. 6**: Noised with Phase Scrambled Images. Figure 6a illustrates the outcome of applying Gaussian noise to the entire image and subsequently restoring it, while Figure 6b 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.

### 4.5. Generalisation

Figure 7 illustrates the RGB channels alongside their respective shape maps, as well as the output depth maps generated by the trained model using these inputs. Despite the substantial disparity in information content between the RGB images and shape maps, their contributions to depth estimation appear to be similar.

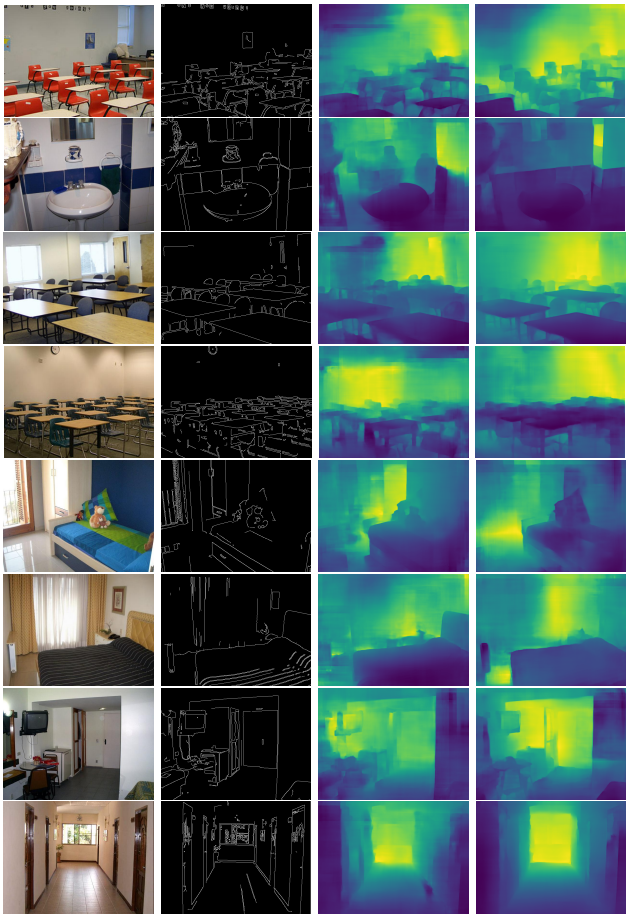


**Fig. 7**: Performance of Shape ONLY model with New Indoor Scenes from other Domains. The left column displays RGB scene images, the second column presents corresponding edge maps, and the third column showcases the results generated by the pre-trained shape-input model. The right column exhibits the outcomes produced by the pre-trained RGB-input model.

## SUPPLEMENT MATERIALS

### Saturation

Figure 8 illustrates saturation maps with different saturation values, while Figure 9 displays various RGB images with different saturation values alongside their corresponding model performance. It can be observed that the model's performance does not exhibit a strong sensitivity to different saturation values. As the saturation values increase, there is a slight decline in performance. We hypothesised that this decline is due to the presence of more noise in images with high saturation values, as depicted in Figure 8, which negatively impacts the model's performance.

Please note that during training, the channel order is BGR. However, for the sake of convenience in checking, the images have been converted to RGB channel order.

### Contrast

Due to the inclusion of shape, shading, and other information, Contrast cannot be extracted independently. The adopted method involves utilising a trained model and incrementally adjusting the contrast of the test set images during the reasoning process. This enables observation of the performance of the model's depth estimation and facilitates analysis.

Figure 10 shows images with different contrast values ranging from 0.2 to 5.

Figure 11 illustrates that when the contrast remains relatively stable compared to the original image, such as within the range of 0.6-1.6, we observed minimal changes in performance. This observation leads us to suspect that the narrower depth range typically found in indoor scenes may contribute to this phenomenon, as the variations within this small range might not be noticeable.

Considering the contrast formula, output = saturate $(src * alpha + beta)$, excessive or insufficient contrast values can result in a loss of picture details, leading to a significant decline in performance.

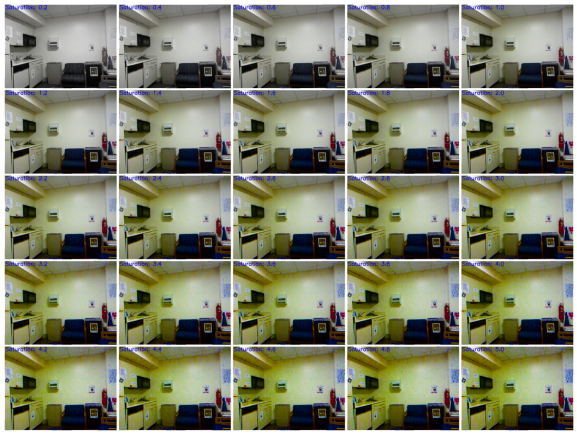


**Fig. 8**: Saturation Maps with Different Saturation Values

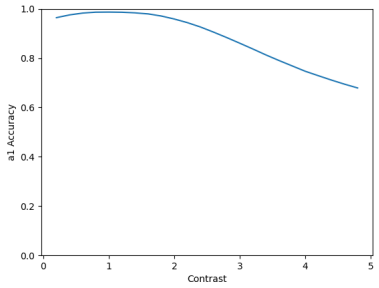



**Fig. 9**: Different Saturation RGB Images and Model Performance

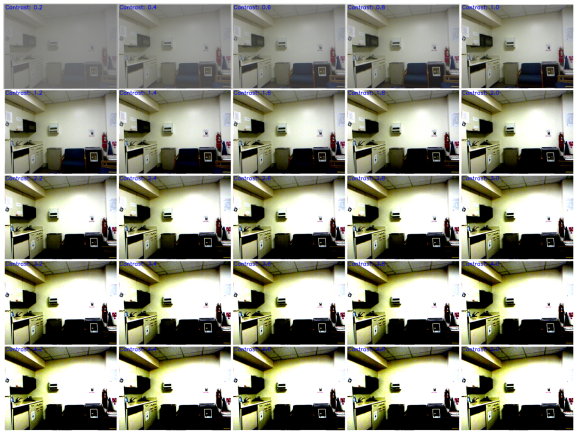


**Fig. 10**: Contrast Maps with Different Contrast Values

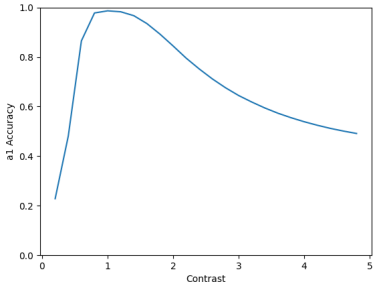



**Fig. 11**: Different Contrast RGB Images and Model Performance

## Discussion

However, Figure 8 and Figure 10, show that these approaches merely appeared to mirror the acquired knowledge of the data-driven model. The model attained its optimal performance when presented with input data characterised by the same levels of original saturation and contrast as those found in the training dataset. Therefore, this method is not suitable for the study of saturation and contrast contributions.

### 1. REFERENCES

[1] Yunhao Ge, Yao Xiao, Zhi Xu, Xingrui Wang, and Laurent Itti, "Contributions of shape, texture, and color in visual recognition," in *Proc. ECCV*. Springer, 2022, pp. 369–386.

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

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. ECCV*. Springer, 2012, pp. 746–760.

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka, "Adabins: Depth estimation using adaptive bins," in *Proc. CVPR*, 2021, pp. 4009–4018.

[5] Yihong Wu, Yuwen Heng, Mahesan Niranjan, and Hansung Kim, "Depth estimation for a single omnidirectional image with reversed-gradient warming-up thresholds discriminator," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

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

[7] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Proc. NeurIPS*, vol. 27, 2014.