Structure Guided Monocular Depth Estimation

Aditya Shourya,(i6353515)[1]

Abstract

This work presents a modest method for comprehending the structure of the monocular depth estimation problem. We can devise a regression-based method to address the issue of depth estimation in situations when disparity calculations could only provide an estimate of actual depth by comparing structures in the local manifold.Our goal is not to enhance the precision or computational efficiency of any current model. Additionally, the study looks into the effects and capabilities of several structure-guided loss approaches for depth estimation. For our trials, we employ a feature fusion model using a standard Resnet backbone and the ReDWeb dataset. The paper offers suggestions on how to combine and choose geometrical losses in a way that improves depth map estimation while minimizing redundancy.Our results demonstrate how such an approach might provide a more profound comprehension of the mutual benefit between structural guiding losses and regression based models.

Keywords: Depth Estimation 1, Structure Guidance 2, Regression 3, ReDWeb 5

1. Introduction

Monocular depth estimation, although useful, is still a difficult and severely underconstrained problem to solve while still holding the interpretibility of the model. It requires the use of numerous, occasionally subtle visual clues, distant context, and past knowledge to solve them. We require training data that reflects the diversity of the visual world and is equally varied in subject settings in order to develop models that perform well in a range of scenarios.
The range and operating conditions of sensors that offer dense ground-truth depth in dynamic scenes, like time-of-flight or structured light, are constrained.In Stereo cameras we have left-right consistency which helps in calculating disparity of the subject by calculating pixel shift between 2 images . But a monocular dataset is more readily available and practical to hold . Thus quite a lot of work has gone into estimating relative depth in monocular images in the recent few years as researched by Xian et al. (2018) .

2. Related Work

A crucial component of computer vision is depth perception, which is handled using both monocular and stereo methods. Saxena et al. (2009) provide an example of monocular methods, which concentrate on obtaining depth information from a single image by using machine learning algorithms to determine relative distance. In contrast, disparities between corresponding points in two images are used by stereo approaches,i.e true depth can be calculated by seeing the pixel shift in left to right images.Such Disparity calculations like the work of Hirschmuller and Scharstein (2005), to compute depth. While monocular techniques are useful in situations where stereo information is not available, stereo techniques provide higher accuracy by taking advantage of binocular differences. These techniques are particularly useful in applications that require accurate depth perception, like robotic navigation and 3D reconstruction.In this paper we would only concern ourselves with Monocular based approach. But similar arguments could be made on a broad spectrum about stereo based approach.

2.1. Stereo Based Approach

Zbontar and LeCun (2016) present a novel method of stereo matching with convolutional neural networks (CNNs) in this seminal work. In order to achieve effective stereo matching, the authors suggest training the CNN to compare image patches. The model shows enhanced performance in matching disparities, a crucial component of precise depth perception, by utilizing deep learning techniques. This work, which was published in the Journal of Machine Learning Research, makes a substantial contribution to the use of neural networks in stereo vision.

The ”Displets” method is a technique that was developed by Vogel et al. (2015) to address stereo ambiguities by incorporating object knowledge. The paper, which was published in the Proceedings of Xian et al. (2018), focuses on improving stereo image depth perception. By utilizing object-level data, Displets advances the state-of-the-art in stereo vision by facilitating more precise and dependable stereo matching.

A paper on improving the effectiveness of deep learning techniques for stereo matching is presented by Pang et al. (2017). The authors of the paper, which was published in the Proceedings of CVPR, provide methods for enhancing the stereo matching algorithms' computational effectiveness and speed. Because of this work, deep learning algorithms will be more useful for tasks requiring quick depth perception in stereo pictures. It is crucial for real-time applications.

2.2. Monocular Based Approach

Learning depth information from individual monocular images is a tough topic that is addressed in the study by Saxena et al. (2009). The authors suggest a machine learning-based method to infer depth signals in the lack of stereo pairs, which are frequently used for depth perception. The model learns to correlate visual patterns with appropriate depth information by utilizing a wide range of characteristics. When stereo information is absent or unfeasible, the use of monocular images is especially pertinent. Through trials that provide correct depth estimates, the authors illustrate the efficacy of their system. By going beyond conventional stereo setups, this work makes a substantial contribution to the larger field of computer vision.

Laina et al. (2016) have published a paper outlining a unique method for unsupervised depth prediction from monocular videos. The scientists take advantage of the underlying structure in the visual data to create a model that can anticipate depth information in scenarios where explicit depth sensors are not present. The model can learn depth cues without labeled ground truth thanks to the unsupervised learning paradigm, which makes it useful in a variety of real-world situations. The suggested technique shows promise for applications like autonomous navigation and picture comprehension as it performs well in estimating depth from monocular movies. This work makes a significant contribution to the field of computer vision by extending the possibilities for depth perception in monocular environments.

3. Dataset : ReDWeb

Xian et al. (2018) investigates the issue of monocular relative depth perception in the outdoors and is examined in this paper. We present an easy-to-use yet efficient technique for automatically producing dense relative depth annotations from web stereo images. Additionally, we suggest a new dataset with a variety of images and dense relative depth maps that correlate. In addition, an enhanced ranking loss is included to address unbalanced ordinal relations, compelling the network to concentrate on a subset of challenging pairs. Test findings show that our suggested method improves additional dense per-pixel prediction tasks, such as semantic segmentation and metric depth estimation, in addition to achieving state-of-the-art accuracy in relative depth perception in the wild.

A long-standing task in computer vision is monocular depth estimation, which has numerous applications, including robotics, 3D modeling, and 2D-to-3D conversion. Despite notable advancements, in recent times owing to the triumph of deep convolutional networks (ConvNets), depth estimate from monocular pictures continues to be difficult, particularly for photos captured in the wild. The majority of cutting-edge techniques developed for one dataset frequently perform worse on another. For instance, models (like NYUDv2) that were trained on indoor datasets are unable to accurately predict depth in outside scenarios. In line with the Robust Vision Challenge 20181, our objective is to utilize a single model to forecast relative depth in a variety of scenarios.

In reality, many applications—like depth-of-field and 2D-to-3D conversion only require relative depth. Chen et al. (2016) presented a "Depth in the Wild" (DIW) dataset of 495k web photos, where each image was manually annotated with two points of ordinal connection (near '¡' and further '¿') in order to recover relative depth for monocular images in the wild. To obtain satisfactory predictions, training with a single pair of ordinal relations is insufficient (refer Figure 1). In light of the aforementioned observations, the following query emerges: how may diversified photos and matching dense relative depth maps be obtained at a low cost?

Xian et al. (2018) describes an efficient approach to automatically build disparity maps from web stereo images, since a disparity map depicts the relative depth of a scene. Since the horizontal component of a correspondence map can be viewed as a disparity map and web stereo picture pairings are not always well-calibrated, we choose not to use stereo matching but rather a cutting-edge optical flow method to compute correspondence maps. Thus, we present a new dataset called "Relative Depth from Web" that includes matching relative depth maps and 3600 scene-diverse photos.

Training with many pairs of supervision employing a ranking loss can yield good results, as suggested by Chen et al. (2016). In a similar manner, we train a ConvNet to predict relative depth. We turn to online sampling to investigate the diversity of sampled point pairs rather than training with fixed point pairs [6]. The issue of imbalanced ordinal relations, or the fact that there are considerably fewer equal relations than other two relations (closer and further), arises when sampling at random. We create an enhanced ranking loss to mitigate the issue posed by unbalanced ordinal relations and enhance model capacity. Specifically, we sort the loss of each unequal pair at each iteration to prevent the difference between two uneven depth values from being too high.

	Monocular Images	Ground Truth
count	3600	3600
Online Sampling Pairs	300	.

Table 1: For our experiments we will keep online sampling to a humble limit of 300 for our experiments.Note that Higher sampling almost always result in better depth estimations

4. Loss Functions

In computer vision, geometrical loss functions are essential for improving depth estimates. They enhance the accuracy of models by drawing attention to the spatial relationships present in an image. The fidelity of depth estimates is increased by these loss functions, which help the model more accurately represent the geometric structure of a scene. Geometric loss functions help produce more accurate and significant findings when determining depth from images. By ensuring that the depth maps predicted by these methods closely match the actual three-dimensional layout of the scene, they improve the

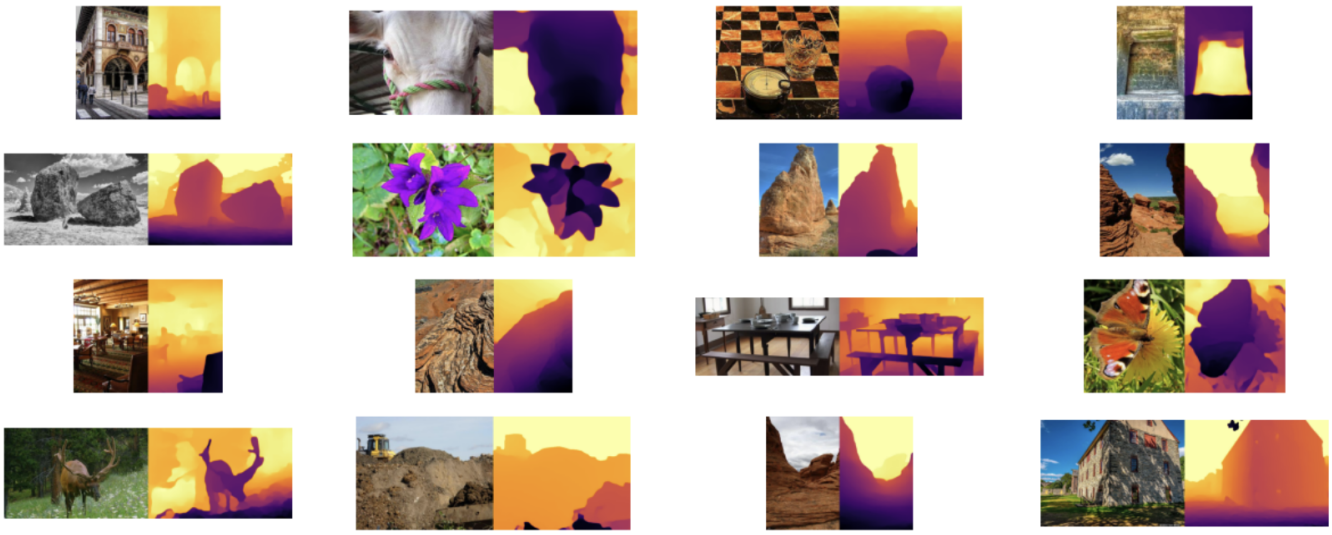

Figure 1: "Relative depth from web" ReDweb Dataset Examples Xian et al. (2018)

overall performance of depth estimation models in a variety of applications, including augmented reality, robotics, and scene understanding.

4.1. Structural Similarity Index

A typical metric to measure the similarity between two photographs is the Structural Similarity Index (SSI) loss. It is particularly helpful when evaluating an image's perceptual quality, especially when depth prediction is involved. The structural similarity between a predicted image (P) and a ground truth image (G) is measured by the SSI loss.

$$SSI(P,G) = \frac{2\mu_P\mu_G + c_1}{\mu_P^2 + \mu_G^2 + c_1} \cdot \frac{2\sigma_{PG} + c_2}{\sigma_P^2 + \sigma_G^2 + c_2} \tag{1}$$

Here, the terms are defined as follows:

- μ_P and μ_G are the means of P and G, respectively.
- σ_P and σ_G are the standard deviations of P and G, respectively.
- σ_{PG} is the covariance of P and G.
- c_1 and c_2 are constants to stabilize the division with a weak denominator.

A metric for measuring the similarity between two photographs is the Structural Similarity Index loss. It gauges how closely a predicted image (P) resembles a ground truth image (G) in terms of both brightness and structural information when used in depth prediction or image processing.

Instead of evaluating pixel-by-pixel differences, the Structural Similarity Index compares local patterns of pixel intensities in P and G.The total similarity is computed by combining these elements.

Here, is a brief exploration of the components:

- **Luminance (l):** Represents the average pixel intensity. It measures the overall brightness of the images.
- **Contrast (c):** Captures the standard deviation of pixel intensities, reflecting the variation in brightness. High contrast indicates a wide range of intensities.
- **Structure (s):** Describes the covariance of pixel intensities between P and G. It assesses how patterns in pixel intensities are related between the two images.

Figure 2: This figure shows how this example is less than ideal for SSI Loss . Although there is a lot of structure that needs preserving the luminance remains almost the same as its background .A good example pair would be to use them when the prominent subject has higher luminescence and is well separated with contrasting edges.

These three elements are then multiplied to generate the Structural Similarity Index, or SSI. For stability, the formula incorporates constants (c1 and c2) to prevent division by ex-

tremely small numbers. The index is a number between -1 and 1, where 1 denotes perfect resemblance.

In conclusion, the SSI loss evaluates the local structure, brightness, and contrast in the images in addition to pixel-wise variations. Because of this, it becomes a more perceptually meaningful metric for assessing the quality and similarity of images, particularly in applications like depth prediction where precise judgments of scene similarity depend on capturing the structural features.

4.2. Inverse Depth Smoothness loss

A regularization term that is frequently employed in computer vision applications, especially in depth estimate issues, is Depth Inverse Smoothness Loss. This loss function encourages neighboring pixels to have similar depth values in order to smooth down the anticipated depth map. In order to produce visually cohesive depth maps and avoid artifacts like abrupt depth shifts, this is essential.

Typically, the Depth Inverse Smoothness Loss is expressed as follows:

$$L_{\text{depth\_smooth}}(D) = \sum_{i,j} \left| \frac{1}{D_i - D_j} \right| \cdot w_{ij}$$

- D represents the predicted depth map.
- i and j denote pixel indices.
- D_i and D_j are the predicted depths at pixels i and j, respectively.
- w_{ij} is a weight that can be used to emphasize or deemphasize certain pixel pairs. It is often derived based on image gradients or other considerations.

The inclusion of term $\frac{1}{D_i - D_j}$ in the absolute value guarantees that regions with significant variations in anticipated depth values between adjacent pixels are penalized by the loss function. The loss promotes smoothness in the anticipated depth map by encouraging neighboring pixels to have comparable depth values by including the inverse of the depth difference.

The loss is applied throughout the full anticipated depth map, and the summation is carried out over every pair of adjacent pixels.

A key element in depth estimation jobs, Depth Inverse Smoothness Loss enhances the predicted depth maps' visual quality and overall accuracy. It is frequently combined with additional losses to produce a thorough objective function that is used to train depth estimate algorithms.

Figure 3: A Pair of example sets which can help us understand the use of depth loss. The coherence of a crowd (Right) remains sound in the depth map even though the subjects are continguous. And well sepration on the subjects improves the depth recording of the big fan on the left

4.3. Mini Batch Sampling

In Mini Batch sampling as per Xian et al. (2018) We employ online sampling, or sample pairings within each mini-batch, to examine the diversity of samples rather than training with fixed point pairs from every image . We randomly select N point pairs (i,j) for each input image I. N is the total number of point pairs, and i and j stand for the first and second points' locations, respectively. We first acquire depth values (gi, gj) from the associated ground-truth depth map in order to label the ordinal relation lij between each pair of points. We then define the ground-truth ordinal relation lij as follows:

$$l_{ij} = \begin{cases} 1 & \text{if } \frac{g_i}{g_j} > 1 + \sigma \\ -1 & \text{if } \frac{g_i}{g_j} < 1 - \sigma \\ 0 & \text{otherwise} \end{cases}$$

- i_k and j_k represent the locations of the first and second points in the k-th pair.
- $l_k \in \{+1, -1, 0\}$ is the corresponding ground-truth ordinal relationship between i_k and j_k, indicating further $(+1)$, closer (-1), and equal (0).
- Note that there exists the problem of imbalanced ordinal relations, i.e., the number of equal relations is far less than the other two relations.

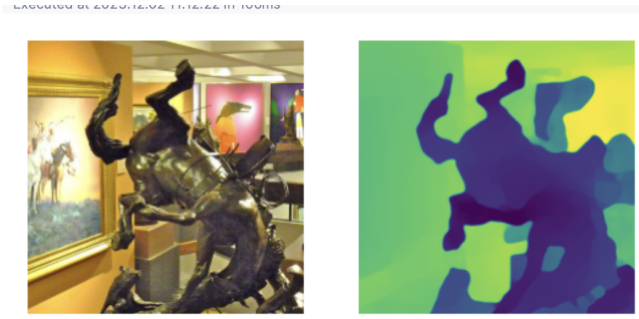

Figure 4: Consider this example of a horse in a museum along side its depth map . to calculate ordinal relations between different manifolds of the image we choose many areas at random and refer to its ground truth to caculate ordinal relations

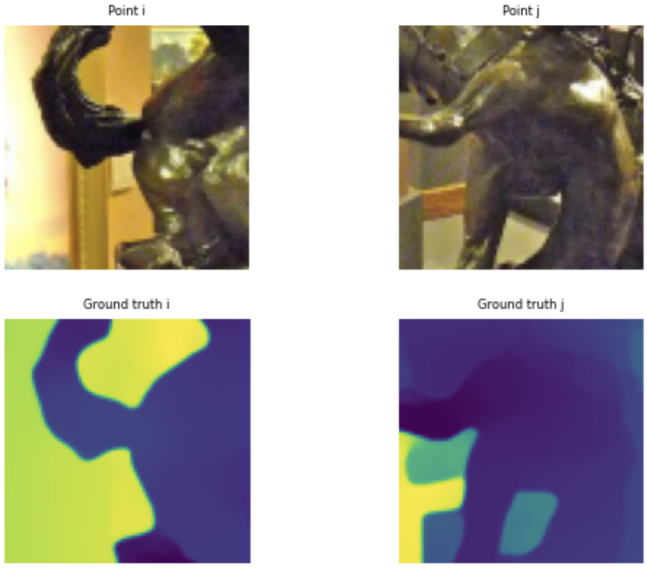

Figure 5: Here we focus on one such sampling pair to get our ordinal relation. here the ordinal relation is the difference between the average pixel value with it's corresponding ground truth.This essentially converts our problem into solving a regression equation

To enable our ConvNet to be trained with imbalanced ordinal relations, an appropriate loss function is needed. In this paper, we design an improved ranking loss $L(I, G, z)$, which can be formulated as follows:

$$L(I, G, z) = \sum_{k=1}^{N} \omega_k \phi(I, i_k, j_k, l_k, z)$$

where z is the estimated relative depth map, ω_k and $\phi(I, i_k, j_k, l_k, z)$ are the weight and loss of the k-th point pair, respectively. Note that ω_k can only be 0 or 1 in our experiments. $\phi(I, i_k, j_k, l_k, z)$ takes the form:

$$\phi = \begin{cases} \log(1 + \exp[(-z_{ik} + z_{jk})l_k]), & \text{if } l_k \neq 0 \\ (z_{ik} - z_{jk})^2, & \text{if } l_k = 0 \end{cases}$$

We initialize all ω_k as 1; then, the loss can be seen as a ranking loss. To avoid the difference of two unequal depth values being too large and ease the problem of imbalanced ordinal relations, we first sort the loss of unequal pairs at each iteration, and then ignore the smallest part by setting corresponding ω_k to 0. More specifically, we empirically set the smallest 25% of ω_k to 0. Therefore, the ratio of equal relation would be increased, so that the problem of imbalanced ordinal relations can be alleviated. In addition, the ConvNet is thus enforced to focus on a set of hard pairs during training.

5. Combining Geometrical losses with Edge Guiding

A thorough methodology for training a Convolutional Neural Network (ConvNet) for monocular depth estimation with skewed ordinal relations and complex scenes as per Xian et al. (2018). Along with an edge-guided ranking loss includes functionality for online and edge-guided sampling.

The purpose of the online sampling function is to select point pairs at random from the network predictions. Next, corresponding pairings are found in the ground truth, with an emphasis on incorporating only legitimate pairs that exhibit consistent masks. This feature is essential for adding diversity to the training process so that the ConvNet can pick up skills from a variety of difficult samples.

In contrast, the edge-guided sampling expression makes use of the edge data present in the input image. To create point pairs, it chooses anchor points from edges and calculates the coordinates surrounding each anchor. By introducing an organized sampling strategy that treats edge points as anchors, this method improves the network's capacity to capture information at object borders.

In order to calculate a ranking loss, the edge guided ranking loss criterion combines these sampling techniques, which addresses the uneven ordinal relations seen in monocular depth estimation tasks. To improve the model's capacity to collect subtle depth information, a regularization term that uses a multi scale gradient matching technique is added to the loss.Masks are necessary to ensure that only valid point pairs are computed, that masked regions are taken into account, and that consistency with the ground truth is maintained.

Such expression enables the addition of different geometrical losses to increase the depth estimation model's usefulness and resilience. Because they penalize variations between the

expected and ground truth depth values, these losses are essential to improving the predictions. In the projected depth maps, geometrical losses like gradient loss and structural similarity index help to capture subtle edges and details. Researchers and practitioners can experiment with different combinations of geometrical losses and weights to customize the loss function and adjust the training process to the unique features of the depth estimation problem at hand.For further understanding its use cases refer DepthSense (2023).

6. Resnet Backbone with Feature fusion in Depth Perception

One of the core tasks in computer vision is depth perception, which is essential to augmented reality, robotics, and driver less cars. Complex neural network topologies that can extract fine information from visual input are needed to estimate depth accurately. In this regard, the advancement of depth perception models has been greatly aided by feature fusion techniques and the ResNet (Residual Network) architecture, which is renowned for its deep learning capabilities.

The difficulty of training very deep networks is addressed by He et al. (2016) with the introduction of ResNet. The vanishing gradient issue affects traditional deep networks, which makes it challenging to learn extraordinarily deep designs. ResNet addresses this problem by presenting the idea of residual learning. ResNet learns residual functions, which are the differences between a block's input and output, as opposed to attempting to directly learn the intended underlying mapping.

The residual block, which contains skip connections, is the fundamental building unit of ResNet. The network can learn residual features thanks to these connections, which let input data to move directly to deeper layers.Deep networks can be trained more easily with this topology since layers may concentrate on learning the residual, which makes optimization simpler.

6.1. Feature Fusion

The integration of multi-scale information is typically beneficial to depth perception. Different layers of a neural network capture features at different levels of abstraction in computer vision applications, such as depth estimation. By combining these features, feature fusion strategies seek to improve the model's comprehension of both tiny details and global context.

Skip connections are commonly used to produce feature fusion in ResNet-based depth perception models. These links allow information from prior layers to be directly merged into later ones by skipping one or more layers. The goal is to improve depth predictions by capturing both high-level semantic information and low-level features, such as textures and edges.

ResNet's capacity to collect hierarchical information is very useful in the context of depth perception. The architecture is very good at acquiring sophisticated representations of depth signals, such as boundaries between objects or simple geometric forms. As a feature extractor, each residual block in ResNet gradually learns and refines features that are essential for precise depth estimation.

A Brief understanding of what Resnet Backbone with feature fusion would work in its forward pass :

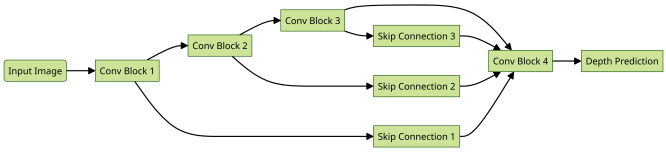

Figure 6: A Block Diagram representing how skip connections (feature fusion) works in Neural Networks

- Initially, a number of convolutional layers process the RGB or depth picture input. Basic characteristics like borders and textures are captured by these layers.
- Following processing, the input goes through a series of residual blocks. The network can capture increasingly abstract representations of depth information as each block learns the residuals and refines the features.
- In feature fusion, skip connections—also referred to as shortcut connections—are essential. They make it easier to integrate features from several levels by connecting the output of one layer to a deeper layer.
- The fusing of characteristics from lower-level blocks to higher-level blocks is made possible by the skip connections. This fusion enhances the model's ability to perceive depth by allowing it to make better use of both local information and global context.
- The fused features from the final residual block are used to construct the final depth prediction. The complex hierarchical representations that the ResNet architecture has learnt are advantageous for this prediction.

7. Evaluation

A crucial step in creating and evaluating depth perception models is validation. Building accurate, resilient, and reliable models in the field of computer vision, especially for tasks like depth estimation, requires an awareness of the significance of validation. Here, we explore the importance of validation and how it affects the performance of depth perception models.

	loss_train	loss_val	ssim_train	ssim_val	mse_train	mse_val
0	0.327318	0.2071	0.133579	0.425158	0.327318	0.20583
1	0.050247	0.042086	0.58262	0.608989	0.050247	0.041778
2	0.054153	0.047912	0.647623	0.622574	0.054153	0.047796
3	0.047921	0.039307	0.680513	0.654263	0.047921	0.039069
4	0.043474	0.038338	0.688893	0.655723	0.043474	0.038063

Figure 7: Training Log through epochs

- Ensuring Generalization: A model's capacity to apply its learnt representations to previously unobserved data is

measured through validation. In order to perceive depth, the model must correctly estimate depths for both new, untrained images and the training set. To assess how well the model generalizes to various settings, lighting situations, and item configurations, a well-designed validation set is essential.

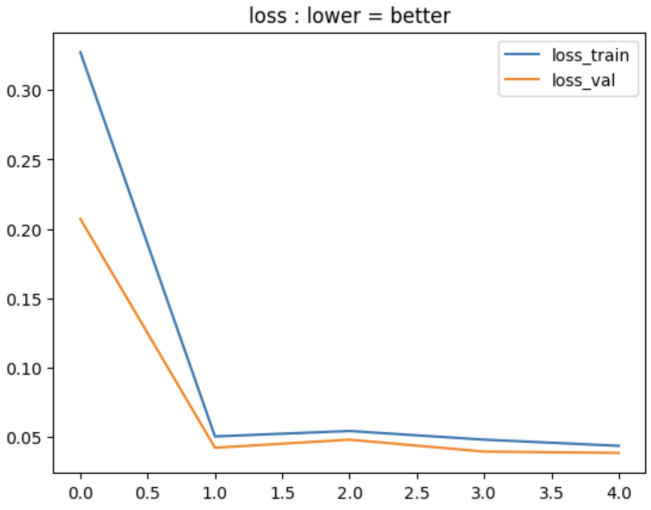

Figure 8: Best Epoch output. $\text{loss}_{fn} = EdgeGyudideLoss()$

- Hyper Parameter tuning : When choosing the optimal model and adjusting hyperparameters, validation is essential. It is possible to evaluate several model architectures and hyperparameter combinations during the training phase. In order to compare these models and determine which configuration is most likely to perform well on unseen data, one can use the validation set as a benchmark.

- Overcoming Overfitting : When a model performs remarkably well on the training data but is unable to generalize to new data, this is known as overfitting. Validation helps identify overfitting by assessing how well the model performs on newly-introduced data. An ideal balance can be found by keeping an eye on how well the model performs on both the training and validation sets. This will guarantee that the model recognizes pertinent patterns without learning noise from the training set.

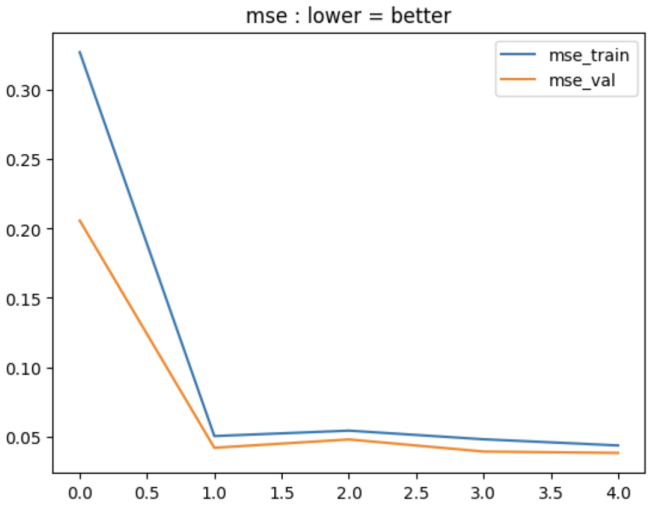

Figure 9: Mean Squared Error Performance

- Model Robustness . Evaluating the Robustness of the Model: Models for depth perception need to be resilient to changes in the input data. Validation aids in evaluating a model's ability to adapt to various environmental factors, such as variations in weather, lighting, or views. A strong model performs consistently in a variety of circumstances, and validation sheds light on how well the model manages difficulties that arise in the actual world.

- Model Bias Finding Data Biases: Validation plays a crucial role in identifying biases that exist within the dataset. Model performance can be greatly impacted by biases, such as imbalances in object distributions or discrepancies in illumination conditions. Through the assessment of the model on an independent validation set, scholars and professionals can detect and rectify biases, guaranteeing that the model's forecasts are impartial and fair.

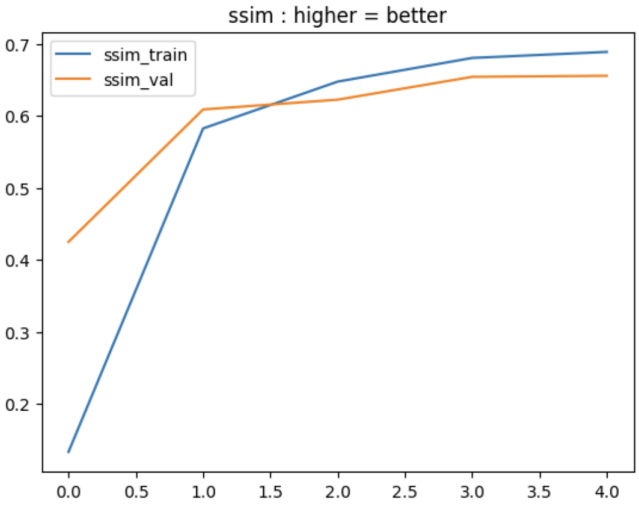

Figure 10: Structural Similarity Index performance

8. Evaluation from the wild web

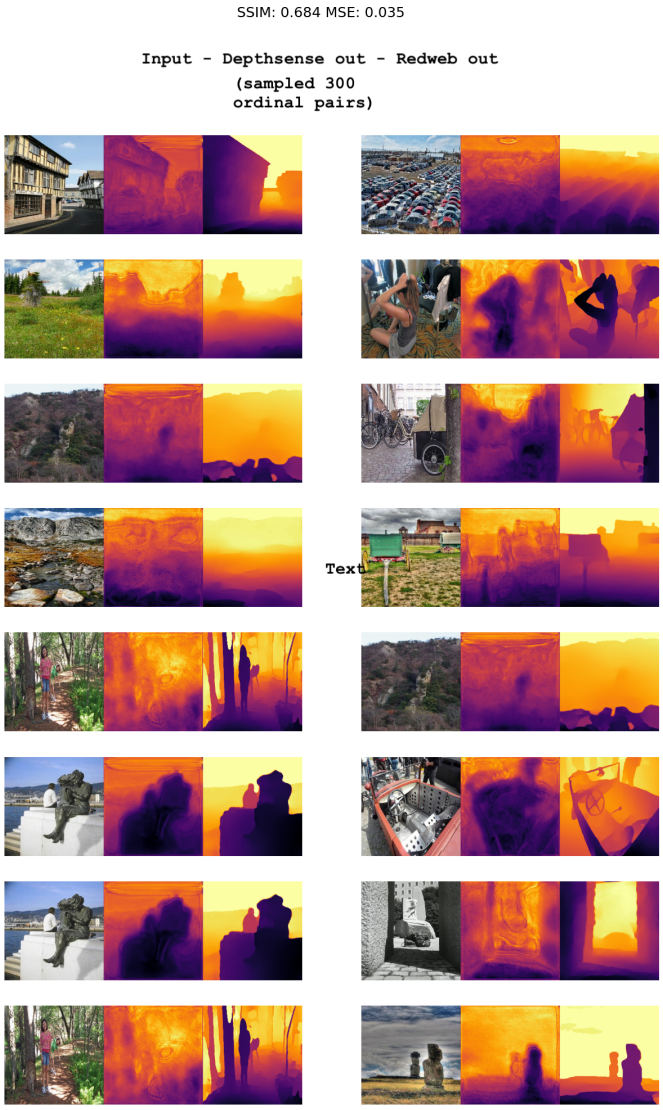

Figure 11: The output of the model on the validation set is shown in this picture. Specifically, when the distances are too small or when there is insufficient contrast, the model performs less well than ideal. But it performs relatively well on easier tasks, like segmenting the sky,or if the image has a fewer prominent subjects if we keep in mind that the model was restricted to sample of just 300 ordinal distances per image for saving on computational cost

When evaluating the capacity of machine learning models for generalization, the validation set is essential. But when you take into account a validation set that has a different distribution than the training set, it becomes much more important. This situation, which is sometimes called a "domain shift," simulates the difficulties that depth perception models could face in the real world when used in various situations.

A validation set with a varied distribution in the context of depth perception includes differences in settings, lighting, and item configurations that might not be adequately reflected in the training data. This deliberate change enables practitioners and academics to assess a model's ability to adjust to new situations and make sure that the learnt features are reliable and transferable outside of the training environments.

The validation set is crucial for assessing how well machine learning models can generalize. However, its significance increases when you consider a validation set whose distribution differs from that of the training set. This scenario, commonly referred to as a "domain shift," mimics the challenges that real-world depth perception models could encounter in a variety of contexts.

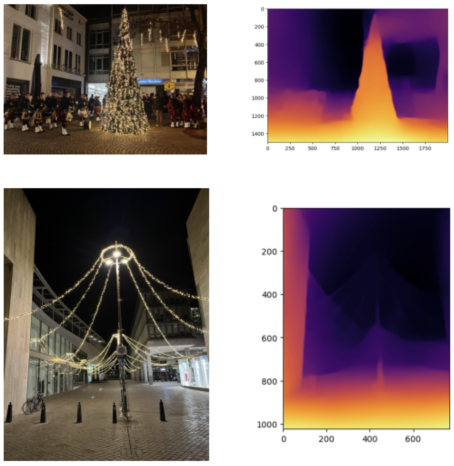

Figure 12: **Top image**: you can see that model does enough to detect the christmas tree and maintains a balance in depth smoothness when we look closely at the band surrounding the tree . which hints at the claim that judicially combining geometrical losses can improve on the quality and usability of depth perception . **Bottom Image** : since the main subject is too thin the chances of genrating ordinal distances from the lampost decreases quite a lot and almost fails to capture its complete structure

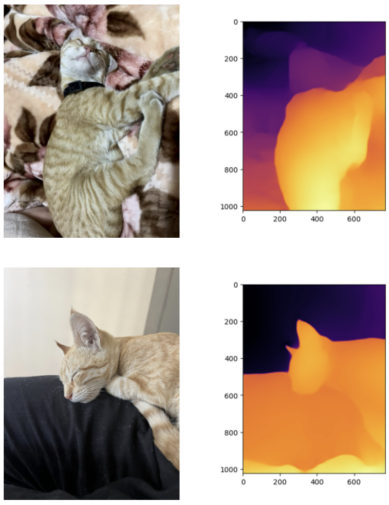

Figure 13: **Top Image** : Since the subject is close enough to the camera and does not inherit a lot of structure the model performs well on preserving smoothness. **Bottom Image** : A Cherry-picked image of my cat shows the ideal scenario when we can get by using a smaller parameter model/ high inference for tasks that demands less of the structure

In the context of depth perception, a validation set with a variable distribution contains variations in illumination, set-

tings, and item configurations that may not be sufficiently reflected in the training data. With this intentional alteration, researchers and practitioners can evaluate a model's adaptability to novel circumstances and confirm that the acquired characteristics are dependable and applicable outside training contexts.

References

Chen, W., Fu, Z., Yang, D., Deng, J., 2016. Single-image depth perception in the wild, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).

DepthSense, 2023. Depthsense: A github repository for depth estimation. `https://github.com/adishourya/DepthSense`.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Hirschmuller, H., Scharstein, D., 2005. Real-time stereo matching using adaptive window-based dynamic programming, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Pang, J., Sun, W., Yan, J., et al., 2017. Efficient deep learning for stereo matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Saxena, A., Sun, M., Ng, A.Y., 2009. Learning depth from single monocular images, in: Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS).

Vogel, C., Schindler, K., Roth, S., 2015. Displets: Resolving stereo ambiguities using object knowledge, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z., 2018. Monocular relative depth perception with web stereo data supervision, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zbontar, J., LeCun, Y., 2016. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research .