Semester Project — 360° Stereo Depth Estimation with Learnable Cost Volume on Real-World Data

Salim Cherkaoui (salim.cherkaoui@epfl.ch)
Supervisors : Charles Corbière and Alexandre Alahi

Abstract

In this project, we developed an innovative system for creating a dataset tailored for 360-degree stereo imaging. The setup comprises two 3D cameras positioned one above the other, with a LiDAR sensor situated between them. This arrangement captures sequences of images simultaneously with LiDAR data, we then enable the precise mapping of LiDAR point clouds onto the upper camera's images. In the second phase of the project, we trained a specialized neural network model for depth estimation, named 360SD-Net, using this newly compiled dataset. Notably, the 360SD-Net model is a stereo model specifically adapted for 360-degree images. Prior to this project, it had only been trained on synthetic data, as a real-world dataset of this nature did not exist. The outcomes of our study, including results and implementation details, can be accessed at our project repository: https://github.com/scherkao31/vita_360stereo.

1. Introduction

The advancement of stereo imaging technology, especially in the realm of 360-degree vision, presents unique challenges and opportunities in the field of computer vision. While the creation of specialized datasets for stereo imaging has been a cornerstone in advancing research and applications, there has been a notable gap in datasets specifically designed for 360-degree stereo vision.

In response to this need, our project introduces a novel dataset specifically tailored for 360-degree stereo imaging. This dataset is a pioneering effort in providing real-world, diverse, and comprehensive data, enabling the detailed capture and analysis of urban and natural environments in 360 degrees. The dataset is crafted using a setup comprising dual 3D cameras and a LiDAR sensor. This arrangement not only captures detailed visual data but also provides accurate depth information, crucial for stereo vision applications.

Building on this, the project further explores the application of this unique dataset in training and enhancing the 360SD-Net model (Ning-Hsu Wang, 2020). Previously, the 360SD-Net, a model designed for depth estimation from 360-degree images, was primarily trained on synthetic data due to the absence of a suitable real-world dataset. Our work addresses this limitation by providing the model with rich, real-world training data, thereby significantly advancing its potential for accurate depth perception in 360-degree real world imagery.

This report details the development of this dataset, its characteristics, and the process of integrating it with the 360SD-Net model.

2. Related work.

In the field of computer vision, the creation of datasets and the development of models for stereo depth estimation have been pivotal. Our work draws inspiration and builds upon significant prior research in these areas.

2.1. DrivingStereo and KITTI Datasets

The DrivingStereo (Guorun Yang, 2019) and KITTI datasets (Yiyi Liao, 2021) have been fundamental in advancing stereo depth estimation. The KITTI dataset, with its extensive real-world driving scenes, has been a benchmark in the field, offering a diverse range of scenarios for stereo, optical flow, and 3D object detection. Similarly, the DrivingStereo dataset, with its large-scale and diverse stereo image pairs collected under various driving conditions, provides a rich resource for training and evaluating stereo vision algorithms. Both datasets have been instrumental in fostering advancements in stereo depth estimation, providing valuable real-world data for researchers and practitioners. However, it is important to note that these datasets focus on conventional images and do not include equirectangular (360-degree) imagery, which presents unique challenges and opportunities in stereo vision.

2.2. MP3D and SF3D: Synthetic Data for Stereo Estimation

The MP3D (Matterport3D) and SF3D datasets (A. Chang, 2017) represent a different approach, where synthetic environments are used to generate data. MP3D, derived from 3D scans of indoor spaces, offers a detailed and varied set of environments for training stereo vision models. SF3D, on the other hand, provides a synthetic framework for generating stereo imagery, allowing for controlled experimentation and testing. The use of synthetic data addresses the limitations of real-world data collection and provides a scalable way to generate diverse and challenging scenarios for stereo depth estimation. These datasets have been particularly crucial in enabling the initial development of models and methodologies for 360-degree depth estimation. While capturing 360-degree images in the real world is feasible, acquiring corresponding depth maps for these images poses a significant challenge, making synthetic datasets like MP3D and SF3D invaluable in this area of research.

2.3. PSM-Net and 360SD-Net

PSM-Net (Pyramid Stereo Matching Network) (Jia-Ren Chang, 2018) represents a significant baseline in the field, known for its effectiveness in estimating depth from stereo images. To adapt to the unique challenges of equirectangular images, PSM-Net was enriched and evolved into 360SD-Net (Ning-Hsu Wang, 2020). This adaptation specifically targeted the deformations characteristic of 360-degree imagery. However, the training and testing of 360SD-Net were primarily conducted on the MP3D and SF3D synthetic datasets, due to the absence of a real-world dataset comprising 360-degree images with corresponding depth maps. This limitation highlighted the need for a dataset like the one we developed, which provides real-world 360-degree stereo images with accurate depth information.

3. Stereo360 : an in-house 360° Dataset

Stereo360 is an in-house 360° dataset under development. As stated, the dataset is crafted using a setup comprising dual 3D cameras and a LiDAR sensor. Currently, it contains 5615 panoramic pairs divided in three sequences with sparse ground-truth depth maps obtained via LiDAR. The training and testing pairs have a resolution of 1920x3840 pixels initially. We have cropped the images to their centers (512x3840 pixels) only to avoid parts of the images where we could not get depth data using our LiDAR setup.

3.1. General Approach for Mapping LiDAR Points onto Equirectangular Images

The main challenge was to be able to aligne LiDAR point clouds with 360-degree camera images. It involved a series of transformations and optimizations, designed to ensure precise mapping. Here's an overview of the approach (as developed in Algorithm 1):

Setting Up Source and Destination Points: We begin by establishing a set of 3D points measured by the LiDAR as our source points (Line 3 : Algorithm 1). Corresponding points on the equirectangular images are identified as destination points. This step is crucial for creating a reference for aligning the LiDAR data with the camera images.

Applying Rotation and Translation: The next step involves applying a rotation and translation to the source points (Line 7 : Algorithm 1). The rotation aligns the orientation of the LiDAR data with that of the camera, while the translation adjusts for any positional differences between the LiDAR sensor and the camera.

Converting to Spherical Coordinates: Once the points are transformed, they are converted into spherical coordinates (Line 8 : Algorithm 1). This is a key step, as equirectangular images inherently represent data in a spherical coordinate system (because they are a transformation of fisheye images, see Figure 1).

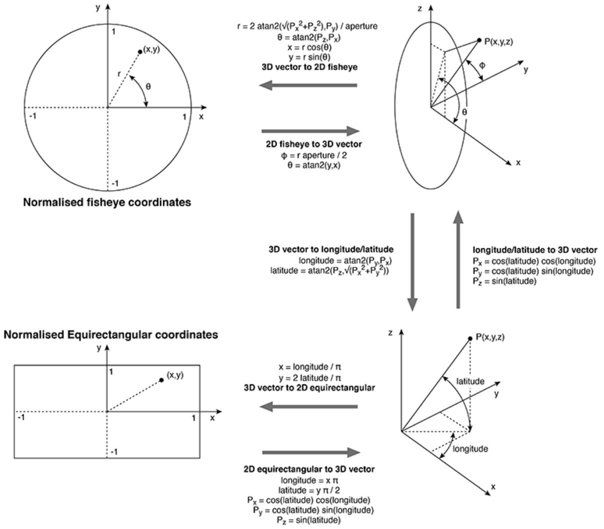
y
1
(x,y)
r
θ
-1
1
x
-1
Normalised fisheye coordinates
$r = 2\ \mathrm{atan2}(\sqrt{(P_x^2+P_z^2)},P_y)$ / aperture
$\theta = \mathrm{atan2}(P_z,P_x)$
$x = r\cos(\theta)$
$y = r\sin(\theta)$
3D vector to 2D fisheye
2D fisheye to 3D vector
$\phi = r$ aperture / 2
$\theta = \mathrm{atan2}(y,x)$
z
P(x,y,z)
ϕ
y
θ
x
3D vector to longitude/latitude
longitude = $\mathrm{atan2}(P_y,P_x)$
latitude = $\mathrm{atan2}(P_z,\sqrt{(P_x^2+P_y^2)})$
longitude/latitude to 3D vector
P_x = cos(latitude) cos(longitude)
P_y = cos(latitude) sin(longitude)
P_z = sin(latitude)
Normalised Equirectangular coordinates
y
1
(x,y)
-1
1
x
-1
x = longitude / π
y = 2 latitude / π
3D vector to 2D equirectangular
2D equirectangular to 3D vector
longitude = x π
latitude = y π / 2
P_x = cos(latitude) cos(longitude)
P_y = cos(latitude) sin(longitude)
P_z = sin(latitude)
z
P(x,y,z)
latitude
y
longitude
x

Figure 1. Inter-conversion between a fisheye image and an equirectangular image

Projecting onto Equirectangular Plane: The spherical coordinates are then projected onto the equirectangular image plane (Line 9 : Algorithm 1). This projection accounts for the unique geometrical properties of 360-degree images.

Optimizing Alignment: The core of the mapping process is the optimization of the rotation and translation parameters. This is done by minimizing the error between the projected points and their corresponding locations on the equirectangular images (Line 12 and 13 : Algorithm 1). The optimization aims to find the best-fit transformation that aligns the LiDAR data accurately with the camera images.

After optimization, we obtain the final rotation matrix and translation vector. These optimized parameters are the best-fit transformation for aligning the LiDAR point cloud with the equirectangular images.

Algorithm 1 Mapping LiDAR Points to Equirectangular Images

```
1: Input: LiDAR points $P_{\text{LiDAR}} \in \mathbb{R}^{n\times 3}$, Corresponding image points $P_{\text{image}} \in \mathbb{R}^{n\times 2}$, Image width $W$, Image height $H$
2: Output: Optimized rotation matrix $R_{\text{opt}} \in \mathbb{R}^{3\times 3}$, Optimized translation vector $T_{\text{opt}} \in \mathbb{R}^3$
3: Initialize source points $P_{\text{LiDAR}}$ and destination points $P_{\text{image}}$
4: Set image dimensions $W = 3840$, $H = 1920$
5: Define objective function $f(R, T)$ for error calculation:

6: for $i = 1$ to $n$ do
7:     Transform $P_{\text{LiDAR},i}$ to $P'_{\text{LiDAR},i}$ using $R$ and $T$: $P'_{\text{LiDAR},i} = RP_{\text{LiDAR},i} + T$
8:     Convert $P'_{\text{LiDAR},i}$ to spherical coordinates $(\rho, \theta, \phi)$
9:     Project to equirectangular coordinates $(x_{\text{eq}}, y_{\text{eq}})$ using $(\theta, \phi, W, H)$
10:    Calculate error $e_i = \|P_{\text{image},i} - (x_{\text{eq}}, y_{\text{eq}})\|$
11: end for
12: Total error $E = \sum_{i=1}^{n} e_i^2$
13: Use BFGS optimization method to minimize $E$: $(R_{\text{opt}}, T_{\text{opt}}) = \arg\min_{R,T} f(R, T)$
14: Extract optimized rotation matrix $R_{\text{opt}}$ and translation vector $T_{\text{opt}}$
```

Equirectangular Image Challenges: The methodology for mapping LiDAR points onto equirectangular images significantly diverges from traditional stereo setups, commonly used with tools like OpenCV for standard images. In standard stereo imaging, the mapping techniques are straightforward due to the linear nature of the image format. However, the spherical coordinate system inherent in equirectangular images demands a more specialized approach. Our method ensures that our dataset not only maintains the integrity of the spatial data but also aligns accurately with the visual data. This approach is essential in making the dataset a valuable resource for advancing research in 360-degree stereo vision, where conventional mapping techniques fall short.

4. Method

The 360SD-Net framework, building upon the foundation of PSM-Net, presents a novel approach to stereo depth estimation for 360° cameras. It is important to note that while 360SD-Net was originally implemented using disparities, in our case, we only have depth information. This should be acceptable since disparity and depth share a linear relationship. Initially, the framework introduces a specific camera configuration and defines the concept of spherical disparity. In this context, we explore the implementation of the end-to-end trainable model as outlined in Figure 2. The major contributions of 360SD-Net, which set it apart from its predecessor PSM-Net, are as follows:

Camera Setting and Spherical Disparity: 360SD-Net employs a top-bottom camera setting, which is cost-effective and aligns with consumer-level 360° cameras' equirectangular projection. This setting ensures that stereo correspondences lie on the same vertical line in both the camera spheres and the 360° images. The disparity is calculated as the difference in angles between projection points from a 3D point onto the camera spheres of the top and bottom cameras. As stated, we will use depth (in meters) instead of disparities (in pixels).

Incorporation of Polar Angle: To address the distortion in equirectangular images, 360SD-Net adds the polar angle as an input to the model. This addition provides essential geometric information related to the distortion. The model employs a late fusion design where RGB inputs are processed through residual blocks, and the polar angle through Conv2D layers. The features are then concatenated after extraction, enhancing the model's ability to separate geometric information from RGB appearance.

ASPP Module: The Atrous Spatial Pyramid Pooling (ASPP) module is adopted to manage the spatial relationship among pixels in 360° images, which have a larger field-of-view than regular images.

Learnable Cost Volume: 360SD-Net introduces a novel Learnable Cost Volume (LCV) to construct a 3D cost volume. This volume adapts to the distortion of equirectangular images by using a shifting filter, which searches for the optimal step-size in degree units, resulting in a more precise cost volume construction.

3D Encoder-Decoder and Regression Loss: The model uses a stacked hourglass architecture for the 3D encoder-decoder and employs regression to deduce continuous disparity values. This approach is shown to be more robust than classification-based methods, and the model utilizes a smooth L1 loss with the ground truth disparity for loss calculation.

These changes in 360SD-Net represent a comprehensive adaptation of the stereo depth estimation process to the unique demands of 360-degree equirectangular imaging.

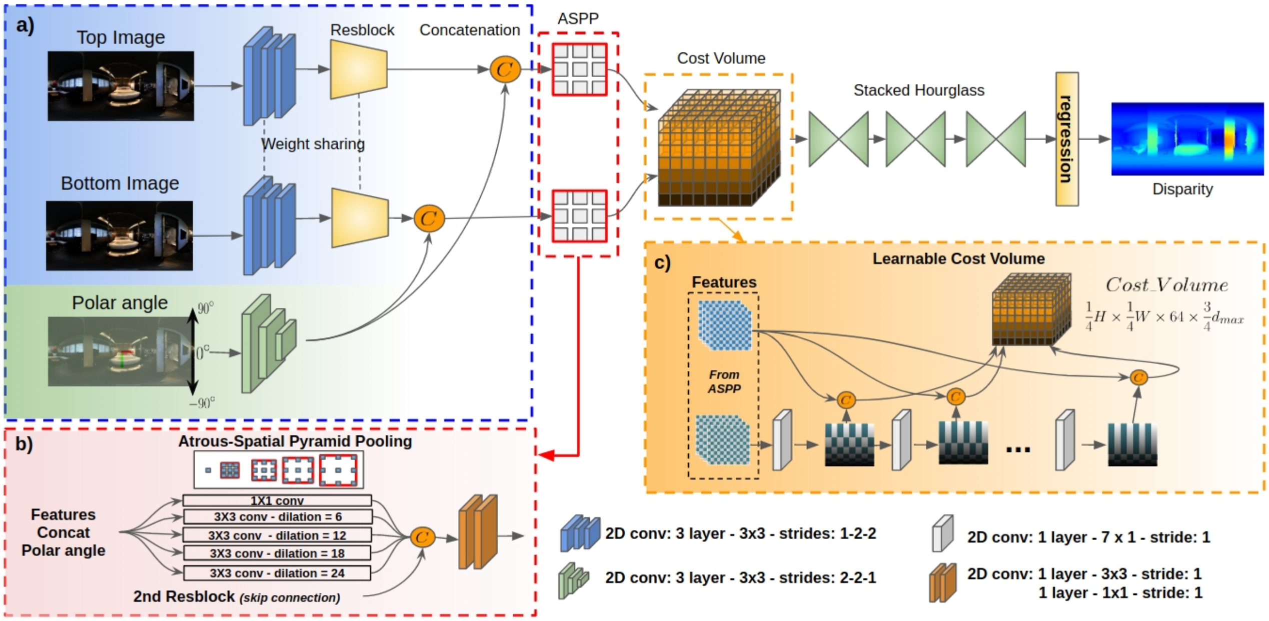
a)
Top Image
Resblock
Concatenation
ASPP
Cost Volume
Stacked Hourglass
regression
Disparity
Weight sharing
Bottom Image
Polar angle
90°
0°
−90°
C
c)
Learnable Cost Volume
Features
From ASPP
Cost_Volume
$\frac{1}{4}H \times \frac{1}{4}W \times 64 \times \frac{3}{4}d_{max}$
...
b)
Atrous-Spatial Pyramid Pooling
Features Concat Polar angle
1X1 conv
3X3 conv - dilation = 6
3X3 conv - dilation = 12
3X3 conv - dilation = 18
3X3 conv - dilation = 24
2nd Resblock (skip connection)
2D conv: 3 layer - 3x3 - strides: 1-2-2
2D conv: 3 layer - 3x3 - strides: 2-2-1
2D conv: 1 layer - 7 x 1 - stride: 1
2D conv: 1 layer - 3x3 - stride: 1
1 layer - 1x1 - stride: 1

Figure 2. The network mainly consists of three parts: a) two-branch feature extractor that concatenates the stereo equirectangular images and the polar angle in a late fusion setting, b) the ASPP module to enlarge the receptive field, and c) the learnable cost volume to account for the nonlinear spherical projection. Finally, it uses the Stacked-Hourglass module to output the final disparity map.

The model's modifications address the inherent challenges of this imaging format, including distortion handling and disparity calculation. For a detailed technical exploration of these contributions, please refer to the 360SD-Net paper (Ning-Hsu Wang, 2020).

5. Experimental results

5.1. Experimental setup

5.1.1. Dataset

Our dataset comprises three unique sequences captured at different outdoor locations on the EPFL campus. These sequences contain diverse conditions and depth ranges.

For our training set, we used 80% of each sequence, merging these portions to form a varied and extensive dataset that exposes the model to a wide range of scenarios during the learning process, while a remaining 10% from each sequence was combined to serve as the testing set, crucial for assessing the model's ability to generalize and perform accurately on unseen data.

5.1.2. Evaluation Metrics

To evaluate the 360SD-Net model on our dataset, we have chosen several metrics, each providing insights into different aspects of the model's depth estimation accuracy:

- **Mean Absolute Error (MAE) and Root Mean Square Error (RMSE):** MAE calculates the average magnitude of errors between predicted and actual depth values for a direct assessment of overall error, while RMSE measures the square root of the average squared differences, giving more weight to larger errors. Both are in meters (m).

- **Mean Absolute Relative Error (MARE):** Given the broad range of depth values in outdoor environments, MARE is particularly relevant. It normalizes the error against the actual depth values, providing a more meaningful measure of accuracy in diverse depth scenarios.

- **Threshold Accuracy:** Denoted as δ_1, δ_2, and δ_3, are crucial for evaluating the model's depth estimation accuracy in our outdoor environments. These metrics assess the percentage of predicted depth values that are within a specific factor of the true depth values. Specifically, δ_i (where $i = 1, 2, 3$) represents the percentage of predictions falling within a factor of 1.25^i of the actual depths.

This metrics, specially Threshold Accuracy metrics δ_1, δ_2, and δ_3, and the Mean Absolute Relative Error (MARE), are especially important in scenarios with wide-ranging depth values. It provides a more detailed understanding of the model's performance across different depth ranges.

For instance, a small absolute error might be negligible for distant objects but crucial for closer ones, emphasizing the importance of these metrics in adapting to the range of depth values.

5.2. Results

Quantitative : The performance of the 360SD-Net model on the Stereo360 dataset, as indicated by the presented results (see Table 1), shows promising capabilities in depth estimation, particularly in threshold accuracies with 80% for d1, 89% for d2, and an impressive 93% for d3. This results are particularly promising, considering that the model was trained from scratch and using depth instead of disparity. This aspect underscores the model's inherent strength and potential, as stereo depth estimation in 360-degree imagery is a complex and challenging task. The reported Mean Absolute Error (MAE) of 2.24 meters and the Root Mean Square Error (RMSE) of 5.59 meters, while indicating areas for improvement, are noteworthy achievements in this context.

	360SD-Net					
Dataset	δ_1	δ_2	δ_3	**mae**	**mare**	**rmse**
Stereo360	80	89	93	2.24	0.180	5,59

Table 1. Experimental results of the proposed method 360SD-Net on our Stereo360 dataset

The fact that the images are captured in outdoor settings, where most objects are at a distance, further adds to the complexity of accurate depth estimation. In such scenarios, even small errors in meters can be significant, especially for closer objects. The moderate level of relative error, as reflected by the Mean Absolute Relative Error (MARE) of 18%, also suggests the model's effectiveness in handling the vast range of depth values typically encountered in outdoor environments. Overall, the model's performance, particularly its threshold accuracies and its ability to achieve this with training from scratch, sets a solid foundation for further advancements in this project.

Qualitative: The qualitative depth maps shown in Figure 3 further reinforces the promising nature of the model depth estimation capabilities. By superposing the predicted depth maps onto their respective images, we gain valuable insights into the model's practical effectiveness. Notably, the model demonstrates a good ability to recognize people and objects. Although the clarity and sharpness of these recognitions are not perfect, the level of detail achieved is noticeable.

However, the model exhibits some challenges in accurately interpreting the sky, occasionally misjudging it as being closer than it is. This issue can be attributed partly to the data preprocessing and training approach. During the training phase, efforts were made to minimize the sky's presence in the dataset, as the LiDAR technology used for capturing ground truth depth values cannot effectively measure the sky's depth (would not make sens either). Consequently, the model has limited exposure and guidance on how to interpret this particular aspect of the images. This limitation in the training data naturally leads to some inaccuracies in the model's depth predictions for sky regions.

Despite this, the overall qualitative results are very encouraging. The model's ability to distinguish and provide depth estimates for various elements within the scenes, with some limitations, is a testament to its potential utility in real-world applications.

5.3. Ablation Study

We undertake two key experiments to understand the impact of dataset diversity on model performance and to assess the model's efficacy at various depth ranges.

5.3.1. Model Training and Evaluation Across Different Sequences:

In this experiment, we train the 360SD-Net model separately on each of the three sequences (seq1, seq2, seq3) and then on the combined dataset of all sequences. This approach enables us to compare the model's performance when trained on data from a single sequence versus a more diverse dataset. We evaluate each of these four models on four different validation sets: the individual validation sets of seq1, seq2, and seq3, and the combined validation set of all sequences.

The results from this experiment given in Table 2 (5.3.1), reveal insightful trends in the model's performance.

- When trained and tested on the *same sequence* (S1 on S1, S2 on S2, S3 on S3), the model demonstrates its highest efficiency, as indicated by the superior δ_1, δ_2, δ_3, and the lowest MAE, MARE, and RMSE values. This trend suggests a strong specialization of the model to the characteristics of the specific data it was trained on.

- A notable decrease in performance is observed when the model is *trained and tested on different sequences*. This decrease in metrics such as δ values and an increase in errors like MAE and RMSE indicates challenges in generalizing the learned features to new, unseen data conditions.

- Training on the *combined dataset* (All) yields consistently robust performance across all test sets. This approach, while not always achieving the peak perfor-

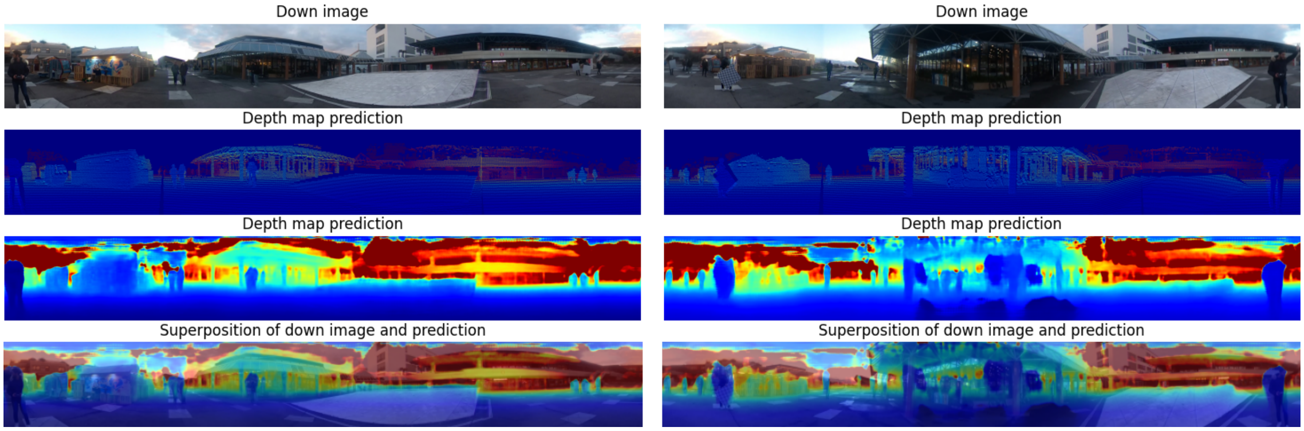

Figure 3. Qualitative examples of 360SD-Net on our Stereo360 (test set)

Test	Train	δ_1	δ_2	δ_3	mae	mare	rmse
S1	S1	**80**	**91**	**95**	**1.94**	**0.164**	**4.44**
	S2	64	79	87	3.19	0.23	6.54
	S3	66	82	90	3.65	0.219	7.52
	All	**80**	**91**	**95**	2.05	0.176	4.46
S2	S1	64	79	87	3.43	0.272	7.48
	S2	**90**	**94**	**96**	**1.47**	**0.125**	**4.44**
	S3	59	74	84	3.56	0.284	7.32
	All	86	92	95	1.64	0.139	5.01
S3	S1	66	82	90	5.31	0.387	9.756
	S2	40	59	71	5.83	0.389	10.56
	S3	**73**	**85**	**90**	**3.16**	**0.236**	**6.98**
	All	71	84	89	3.20	0.240	7.08
All	S1	63	77	86	3.77	0.290	7.66
	S2	69	81	87	3.25	0.233	6.86
	S3	65	79	87	3.45	0.256	7.24
	All	**80**	**89**	**93**	**2.24**	**0.180**	**5,59**

Table 2. Performance Comparison of 360SD-Net Across Different Training and Testing Sequences on the Stereo360 Dataset

mance of sequence-specific training, provides a well-balanced and generalized model capability.

These findings highlight the critical role of diverse training data in enhancing the generalization ability of machine learning models. The ability of the model to adapt and perform accurately in varied settings is significantly improved with exposure to a comprehensive dataset during training.

5.3.2. Depth Range-Based Model Evaluation:

In the second part of our ablation study, we focus on evaluating the 360SD-Net model, trained on sequence 1, across various depth ranges. This experiment is designed to determine if the model's accuracy in depth estimation varies with distance.

Depth Range	δ_1	δ_2	δ_3	mae	mare	rmse
0-5	88	96	97	0.493	0.136	1.91
5-10	76	89	95	1.43	0.200	2.72
10-20	71	89	94	2.62	0.193	4.13
20-50	79	91	95	3.86	0.127	6.52
50+	19	39	54	25.9	0.430	29.0

Table 3. Depth Range-Based Performance Evaluation of 360SD-Net on Sequence 1

The results from this experiment given in Table 3 (3), reveal insightful trends in the model's performance.

- In the **0-5 meters** range, the model achieves high accuracy with δ_1 at 88%, δ_2 at 96%, and δ_3 at 97%. The MAE and RMSE are relatively low at 0.493 and 1.91 respectively, indicating strong performance in close-range depth estimation.
- As the depth range increases (**5-10 meters** and **10-20 meters**), there is a noticeable decrease in δ_1 and a slight increase in MAE and RMSE, suggesting that the model finds medium-range depth estimation more challenging than close-range.
- In the **20-50 meters** range, the model still performs reasonably well, but the errors (MAE and RMSE) continue to increase, reflecting the growing challenge in depth estimation at these distances.
- For distances beyond **50 meters**, the model's performance significantly drops, as seen in the low δ values

and high errors (MAE of 25.9 and RMSE of 29.0), indicating substantial difficulty in accurately estimating depth at long ranges.

These results demonstrate that the 360SD-Net model's accuracy in depth estimation diminishes with increasing distance. The model shows high proficiency in close-range depth estimation but struggles with accuracy as the distance increases, particularly beyond 50 meters. This finding is crucial for applications involving 360-degree stereo vision, as it highlights the need for enhanced techniques or additional data to improve depth estimation at longer ranges. Enriching the dataset with far-range depth information might also significantly improve the model's ability to generalize and perform accurately for distant depth estimation.

6. Conclusion

This report presented a comprehensive study on stereo depth estimation using 360-degree imagery, encapsulating the creation of a novel Stereo360 dataset and the implementation of the 360SD-Net model. The dataset incorporated LiDAR-measured point clouds mapped onto equirectangular images, a process tailored to address the unique challenges of 360-degree imaging.

The 360SD-Net model, trained from scratch, demonstrated promising results, particularly in threshold accuracies, showcasing its potential in handling the complexities of 360-degree stereo imagery. Despite its notable proficiency in recognizing people and objects, the model revealed limitations in depth estimations for distant objects and struggled with interpreting the sky, partly due to the nature of our dataset and the LiDAR's constraints in capturing depth for such regions.

Our ablation study further illuminated the model's behavior under varying conditions. Training and testing on the same sequence yielded the best performance, underscoring the model's specialization capabilities. However, training on a combined dataset of all sequences enhanced its generalizability, a crucial factor for real-world applications. Additionally, the depth range-based evaluation revealed a decline in accuracy with increasing distance, highlighting the need for more data representing distant objects to improve long-range depth estimations.

7. Further work

Looking ahead, several promising avenues can be pursued. A primary objective would be to construct a more extensive and diverse Stereo360 dataset. This expansion would involve gathering more data across various sequences and settings, including both indoor and outdoor environments. Such diversity in data would better equip the model to handle different scenarios and improve its generalizability.

Another key area of future work is the optimization of the model training process. In this study, extensive parameter tuning was not the primary focus; thus, there is substantial potential for enhancing model performance through the careful selection and optimization of training parameters. Additionally, it is important to consider using disparities instead of depth in future implementations, as stereo depth estimation models are generally designed around disparity information.

Furthermore, a comparative analysis with other models, such as PSM-Net, would provide valuable insights. This comparison would be instrumental in evaluating the specific contributions of the 360SD-Net model, particularly in its ability to process 360-degree images effectively. Such an analysis would help in substantiating the value added by 360SD-Net and could guide further enhancements to the model.

References

A. Chang, A. Dai, T. F.-M. H. M. N. M. S. S. S. A. Z. Y. Z. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision*, 2017.

Guorun Yang, Xiao Song, C. H.-Z. D. J. S. B. Z. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. *CVPR*, 2019.

Jia-Ren Chang, Y.-S. C. Pyramid stereo matching network. *IEEE*, 2018.

Ning-Hsu Wang, Bolivar Solarte, Y.-H. T. W.-C. C. M. S. 360sd-net: 360° stereo depth estimation with learnable cost volume. *ICRA*, 2020.

Yiyi Liao, Jun Xie, A. G. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE*, 2021.