Current SOTA Stereo Depth Estimation Techniques

TPA 2

AE21B108 - Reva Dhillon
AE21B002 - Abhigyan Roy

COMPUTER VISION CS6350

Table of Contents

Problem Statement

1. Depth estimation from stereo images is a critical problem in computer vision with widespread applications, including autonomous driving, robotics, and augmented reality.

2. Accurate depth estimation is vital for the functionality and safety of these applications.

3. The methods for stereo depth estimation can be broadly divided into shallow learning-based methods, which rely on traditional computer vision techniques, and deep learning-based methods, which utilize neural networks to learn depth information from large datasets.

Stereo Depth Estimation

What is Depth Estimation?

- The process of determining the distance of objects from the camera.
- Achieved by comparing two or more images of the same scene from slightly different viewpoints.

Core Idea: Stereo Vision

- Two cameras (Left & Right) capture the same scene.
- By finding corresponding pixels between the images, we compute disparity (pixel shift).
- The depth Z is calculated using the triangulation formula:

$$Z = \frac{f \cdot B}{d}$$

where:

- f = focal length of camera
- B = baseline distance between cameras
- d = disparity (horizontal shift between corresponding pixels)

Depth Map

- A 2D image where each pixel's intensity represents distance from the camera.
- Brighter = Closer, Darker = Farther.

Why is it Challenging?

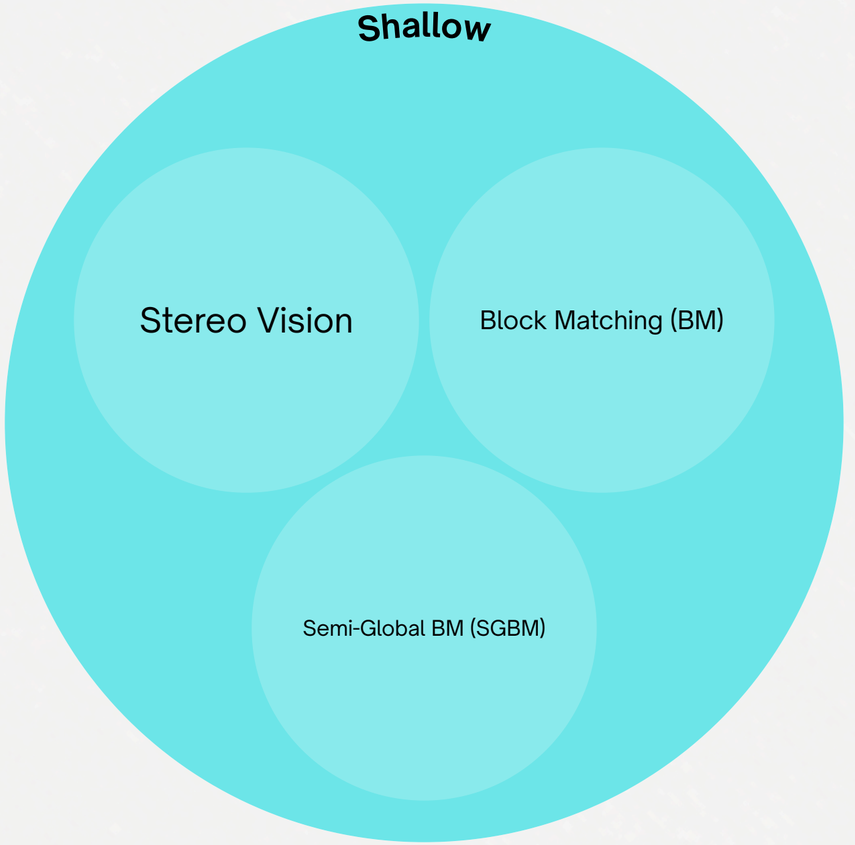
Low-texture or reflective regions make correspondence hard.

Illumination and environmental changes degrade performance.

Deep methods improve accuracy but require large datasets and compute power.

Shallow Methods

Method	Core Idea	Pros	Cons	Type
Stereo Vision	Geometry-based depth from epipolar constraint	Simple	Sensitive to noise	Shallow
Block Matching (BM)	Match small pixel blocks	Fast, easy	Grainy edges, errors in flat regions	Shallow
Semi-Global BM (SGBM)	Adds smoothness + global optimization	Accurate, smooth	Slower	Shallow



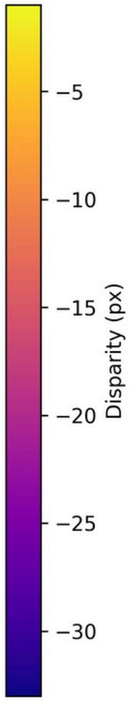
KITTI Sample: 000130_10

BLOCK MATCHING

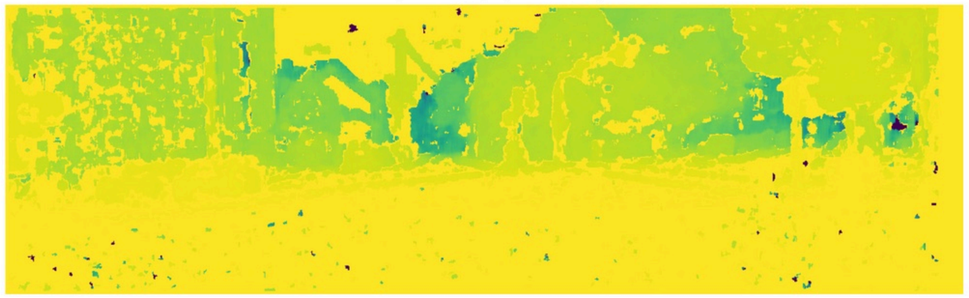
Left Image

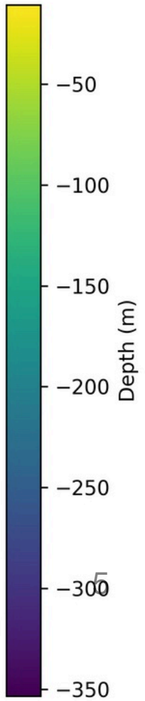
Right Image

Disparity Map (pixels)



Depth Map (meters)





Modifications to Shallow Methods

Edge-Overlayed

Intensity-Gradient

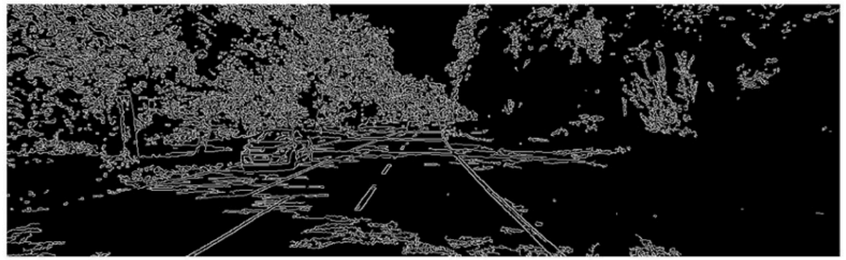
Census Transform

Coarse-to-Fine Strategy

EDGE-OVERLAYED BM & SGBM

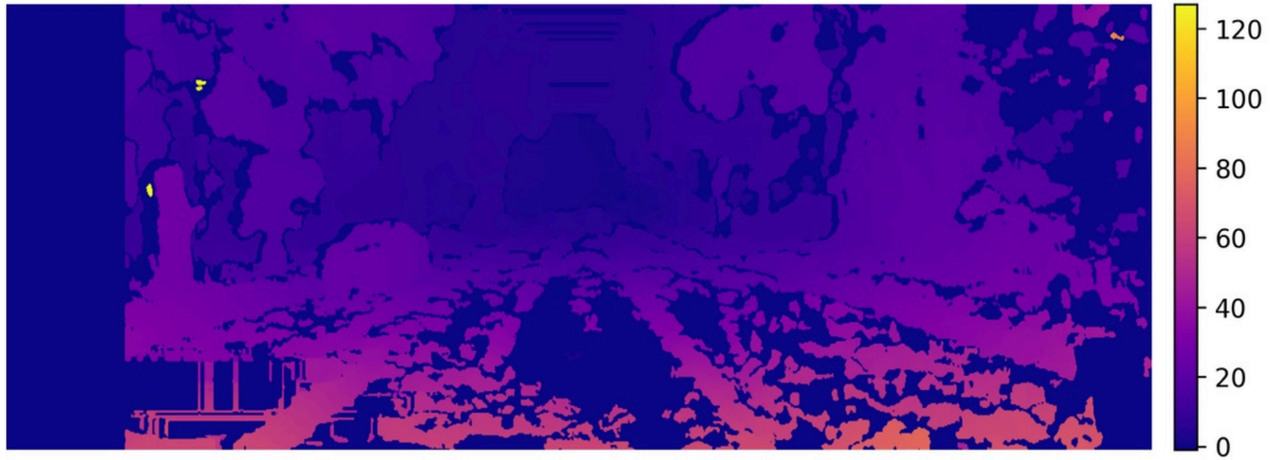
Original Left Image

Canny Edges

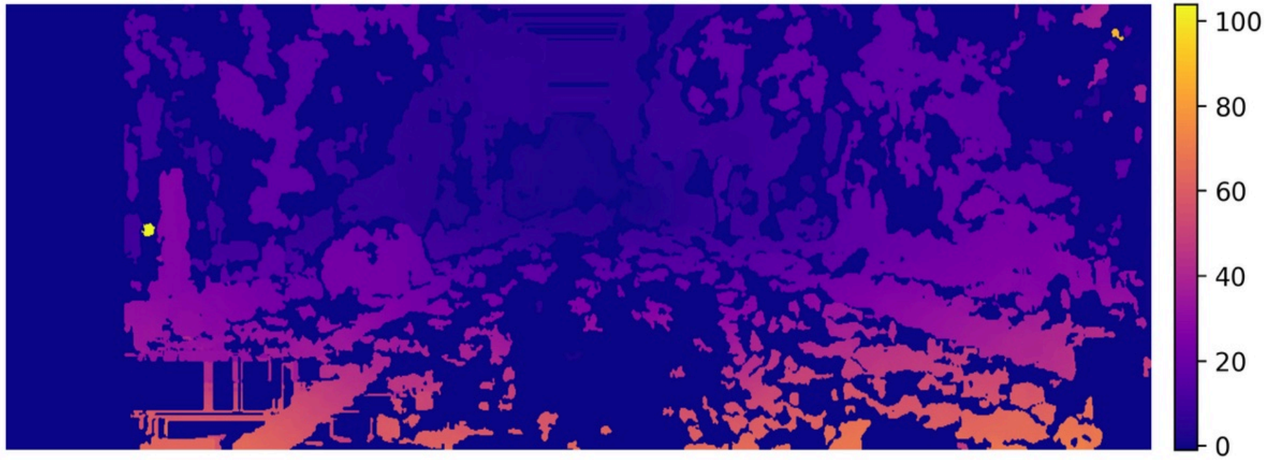


Edge Overlay (Black Edges)

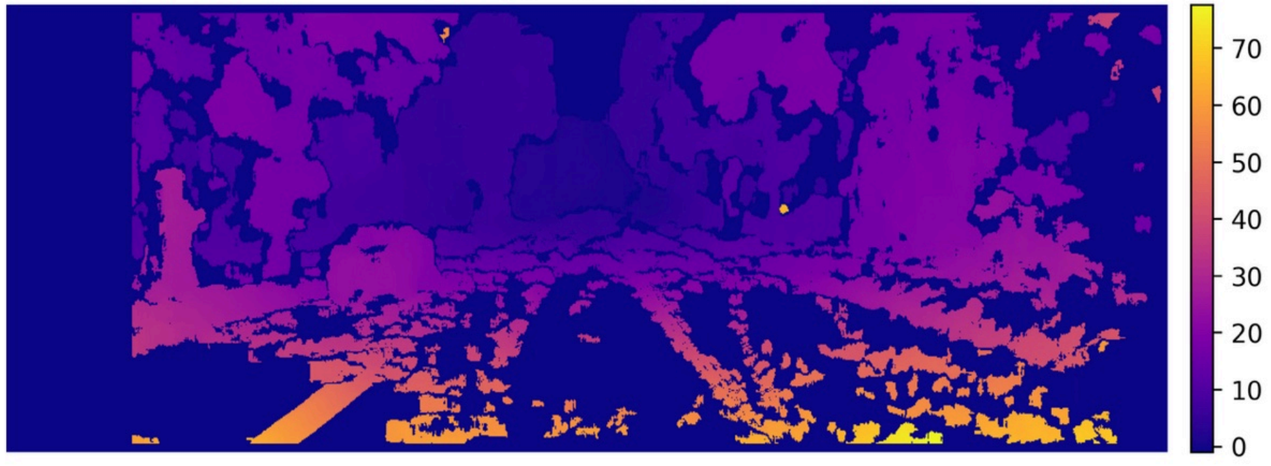
SGBM Disparity (Original) | MAE=1.00



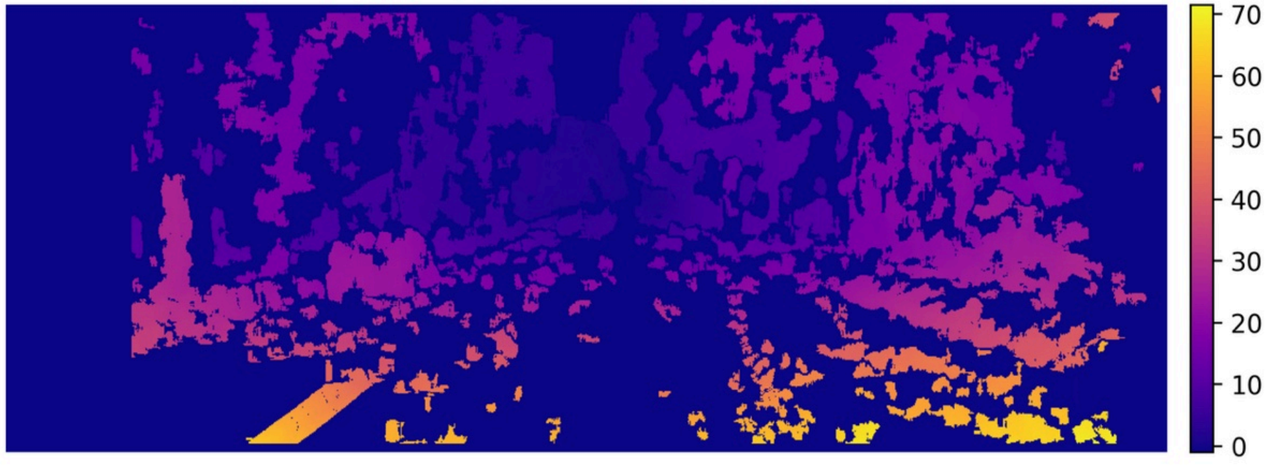
SGBM Disparity (Edge Overlay) | MAE=1.10

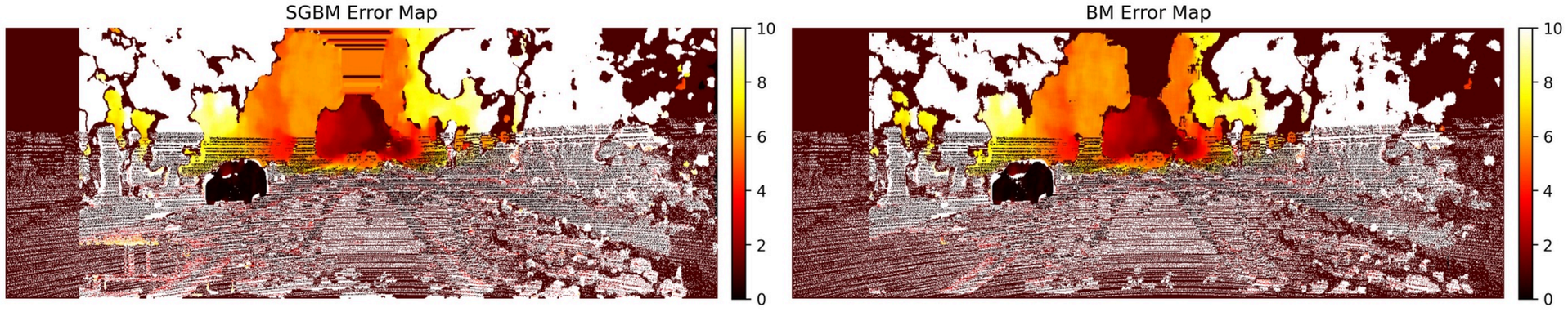


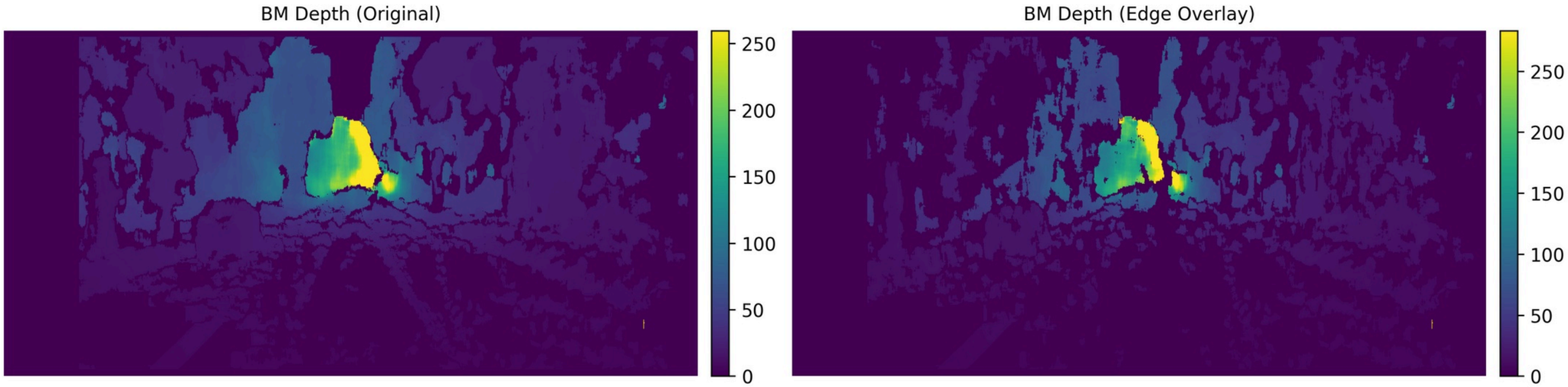
BM Disparity (Original) | MAE=0.83

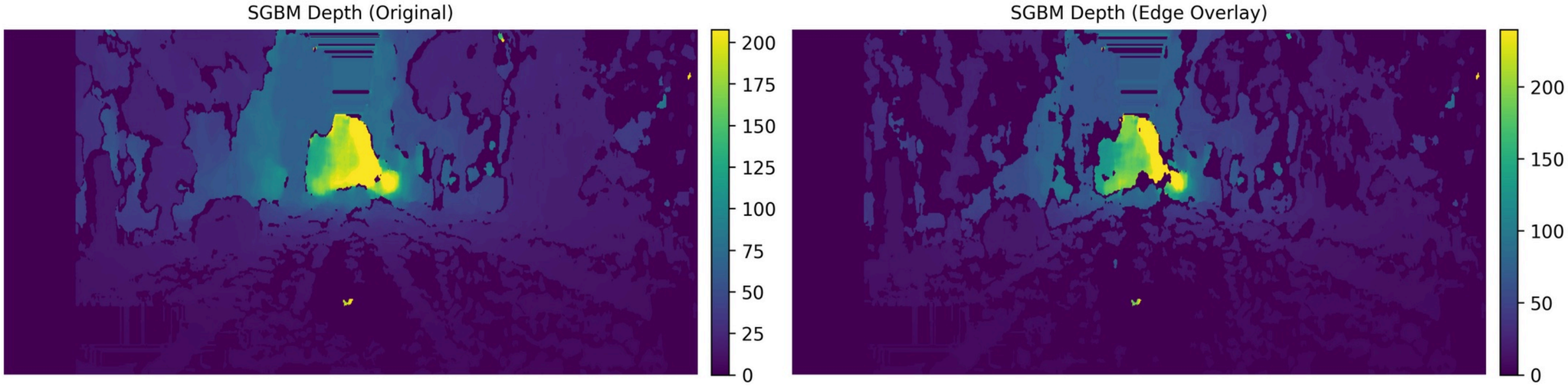


BM Disparity (Edge Overlay) | MAE=0.84

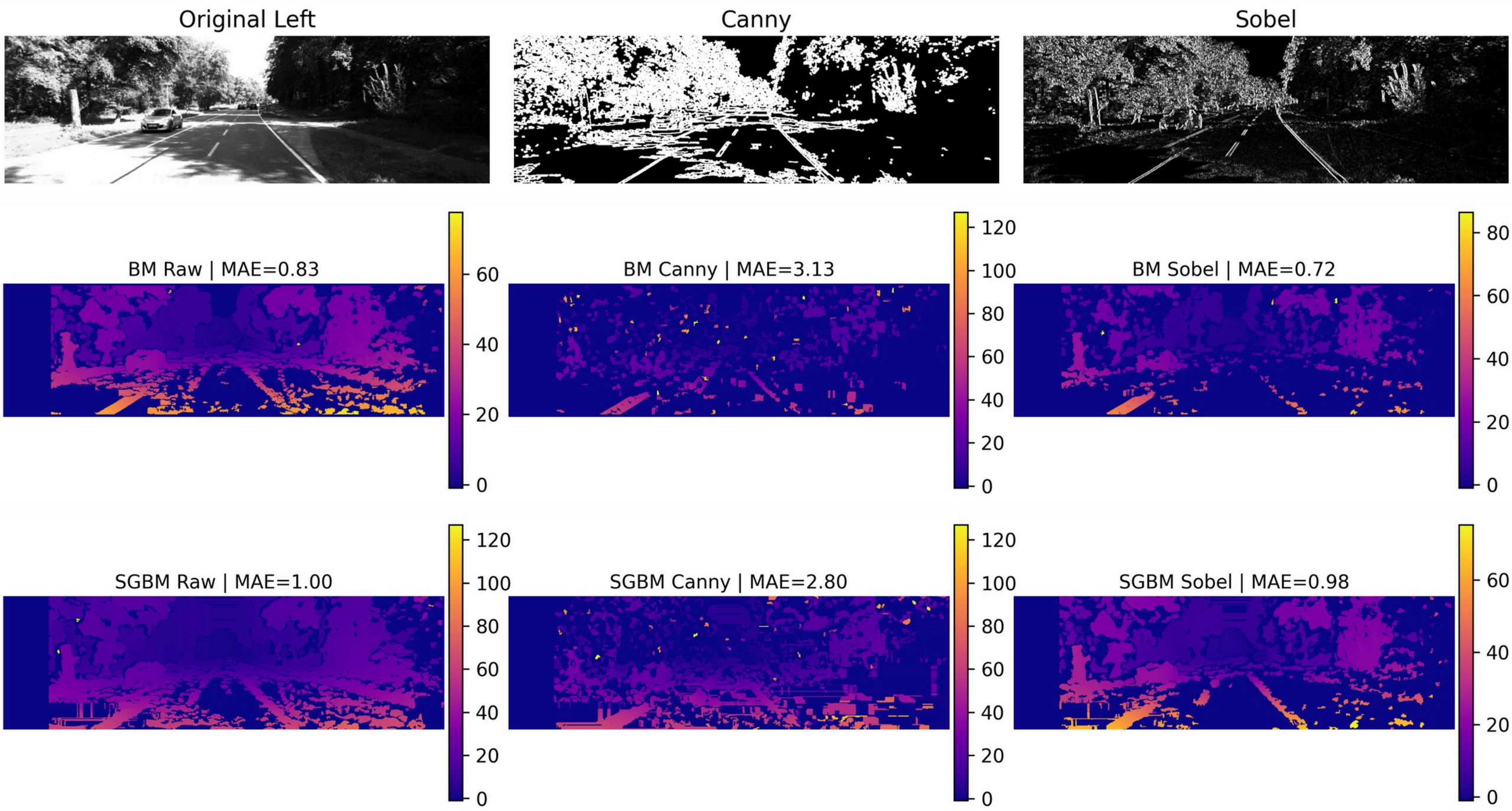


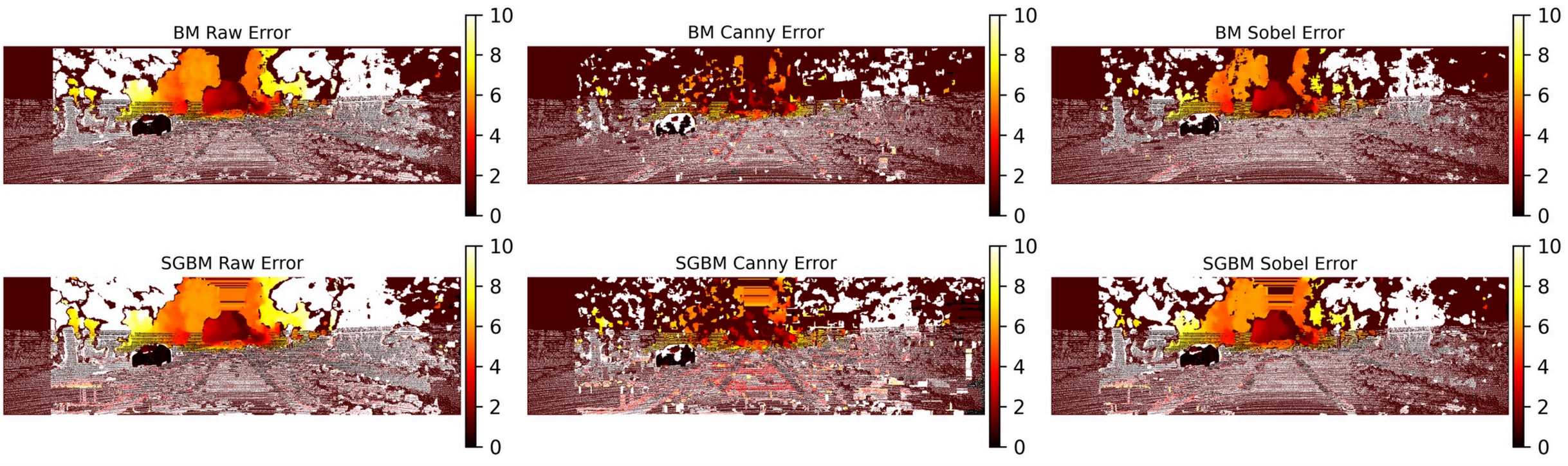

SGBM Error Map
BM Error Map
0
2
4
6
8
10

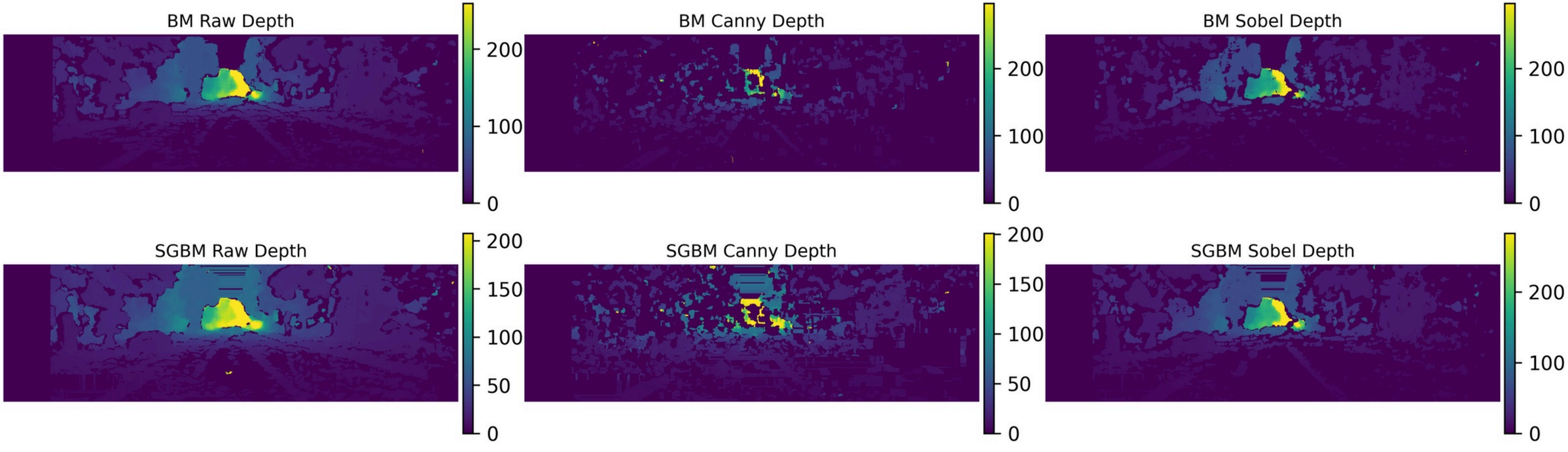

BM Depth (Original)
BM Depth (Edge Overlay)
0
50
100
150
200
250


SGBM Depth (Original)
SGBM Depth (Edge Overlay)
0
25
50
75
100
125
150
175
200

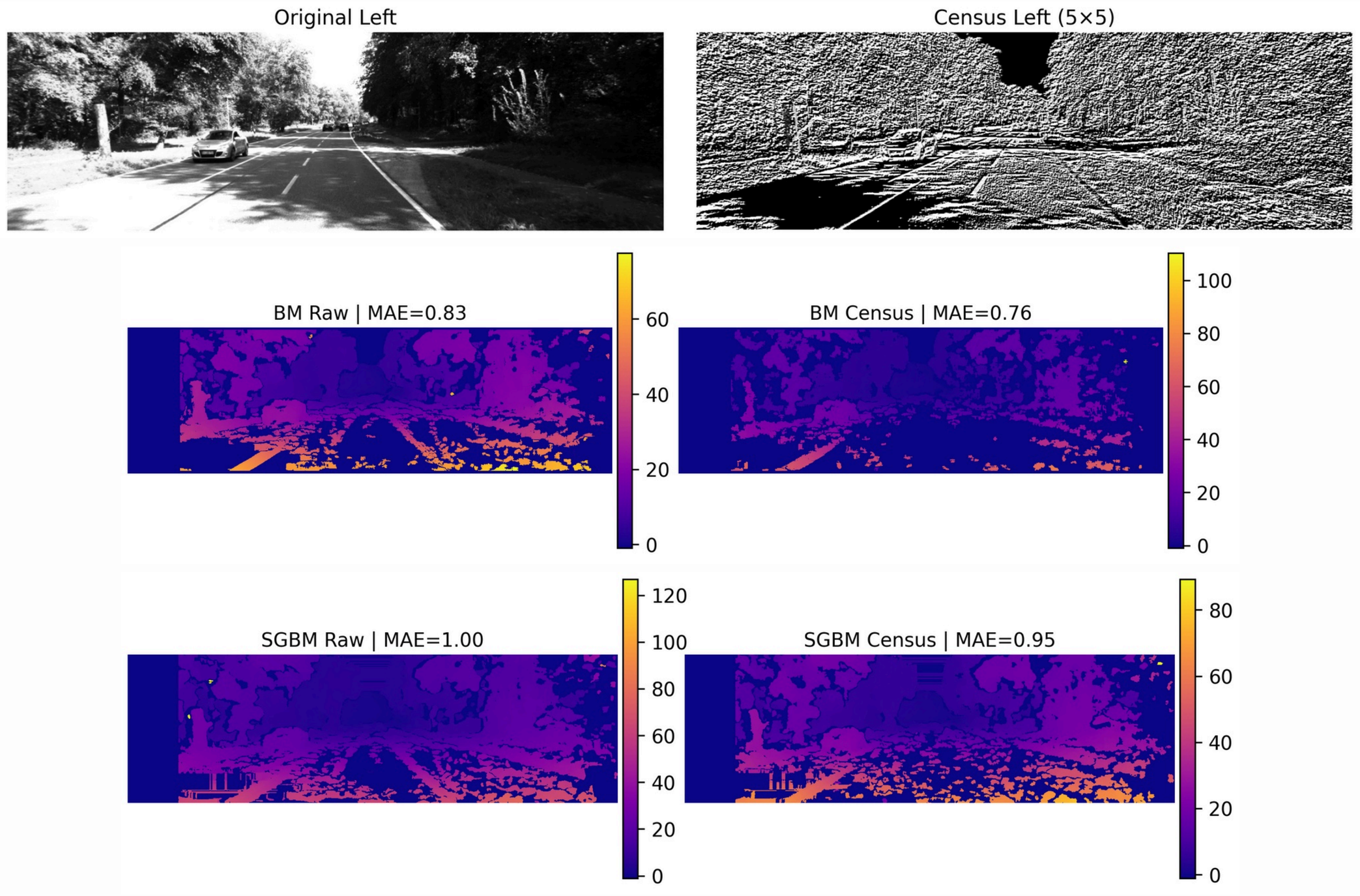
INTENSITY-GRADIENT BM & SGBM

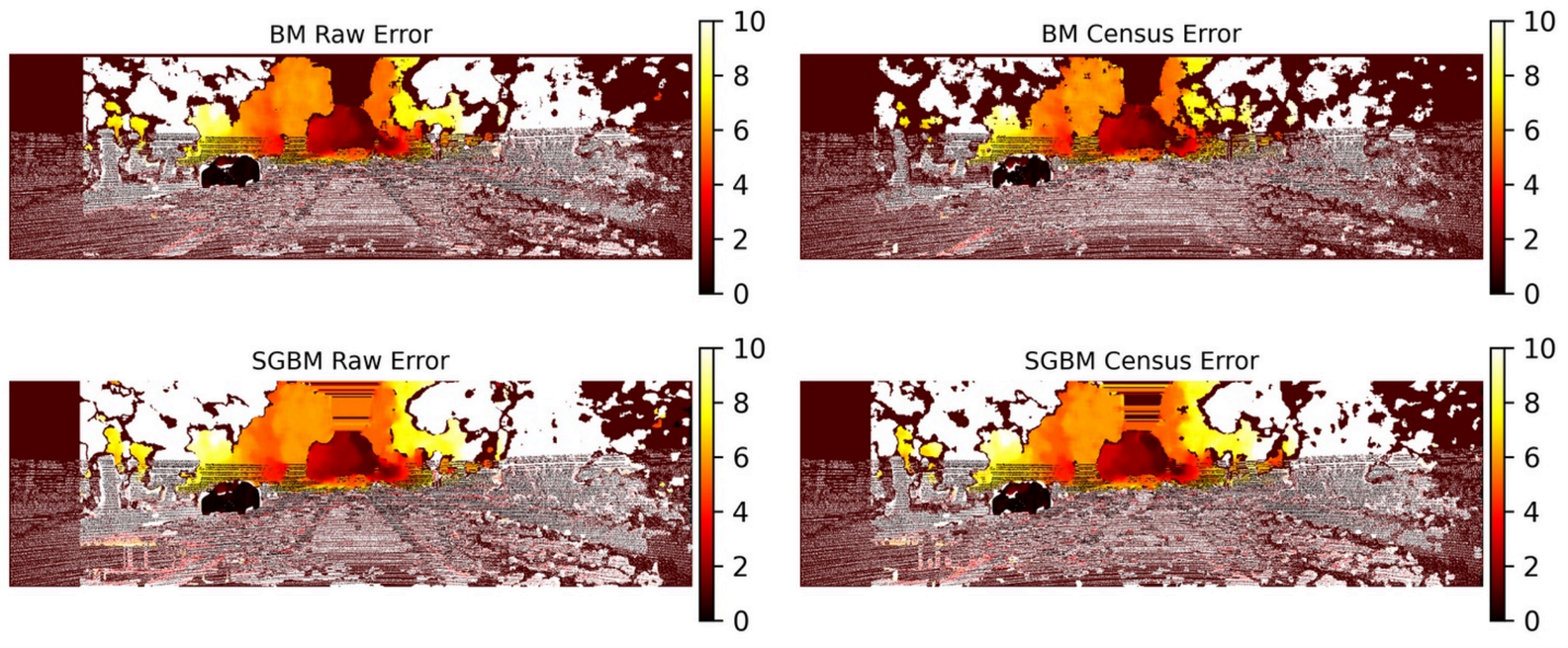


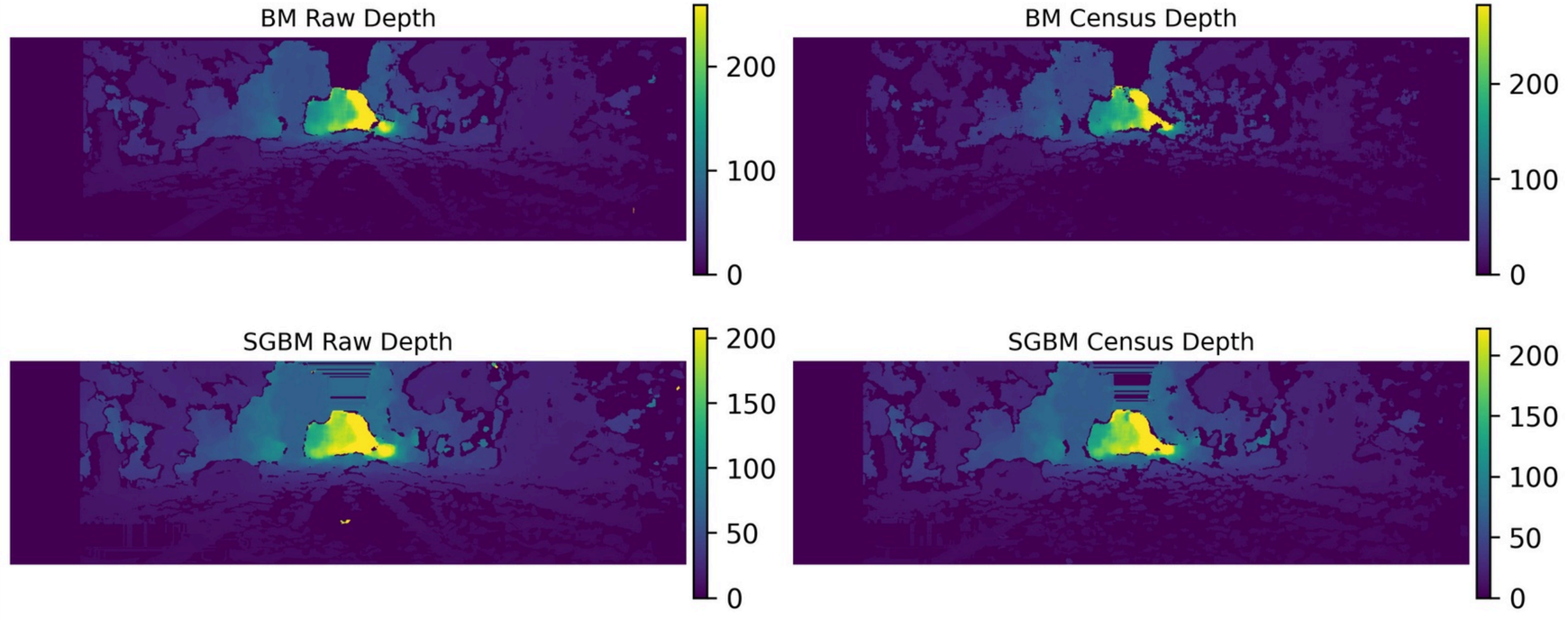

BM Raw Error
BM Canny Error
BM Sobel Error
SGBM Raw Error
SGBM Canny Error
SGBM Sobel Error
0
2
4
6
8
10


BM Raw Depth
BM Canny Depth
BM Sobel Depth
SGBM Raw Depth
SGBM Canny Depth
SGBM Sobel Depth
0
50
100
150
200

CENSUS TRANSFORMED BM & SGBM

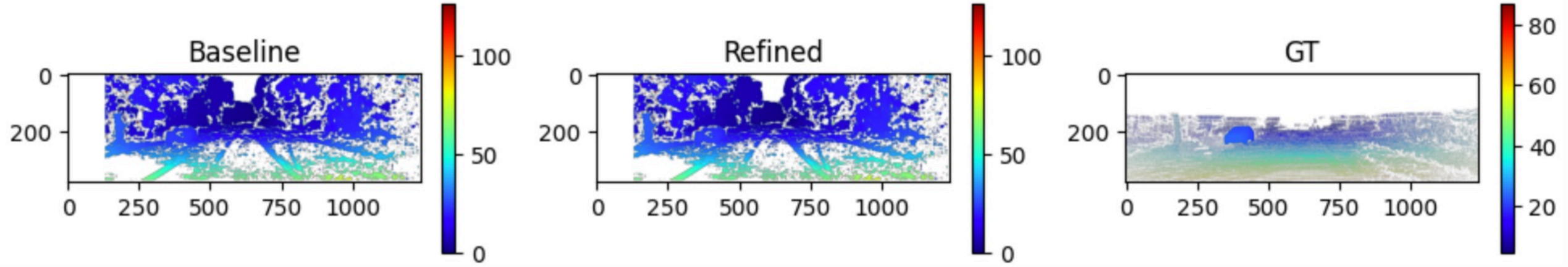



BM Raw Error
BM Census Error
SGBM Raw Error
SGBM Census Error
0
2
4
6
8
10

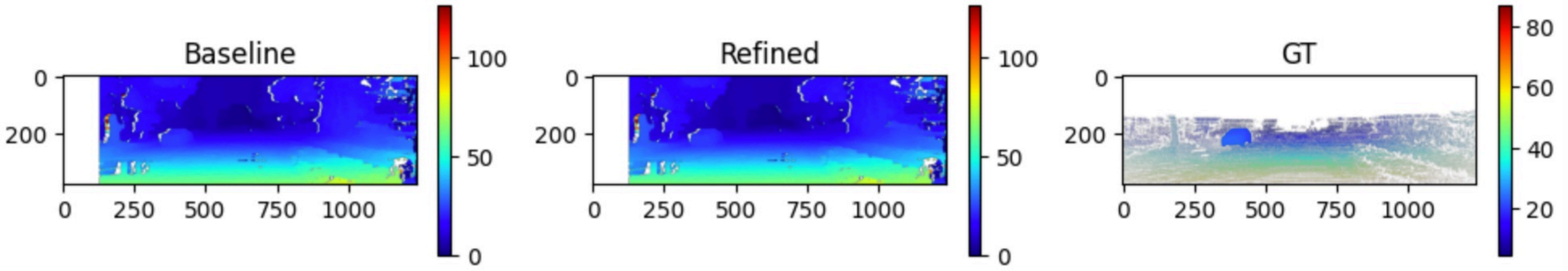

BM Raw Depth
BM Census Depth
SGBM Raw Depth
SGBM Census Depth
0
50
100
150
200

COARSE-TO-FINE BM & SGBM

SGBM



BM



Deep Learning Methods

Dusk till Dawn

ACVNet
Fast-ACVNet

Geometry Informed Neural Operator (FNO)

Dusk till Dawn

Paper: Dusk Till Dawn: Self-Supervised Nighttime Stereo Depth Estimation using Visual Foundation Models (2024)

Goal:

- Robust depth estimation in low-light/nighttime scenes

Key Ideas:

- Uses Visual Foundation Models (VFMs) for lighting-invariant features
- Self-supervised → no ground-truth depth
- Combines photometric + smoothness + consistency losses

Architecture:

Left/Right images → VFM Encoder → Cost Volume → Depth Decoder

Dataset:

Oxford RobotCar

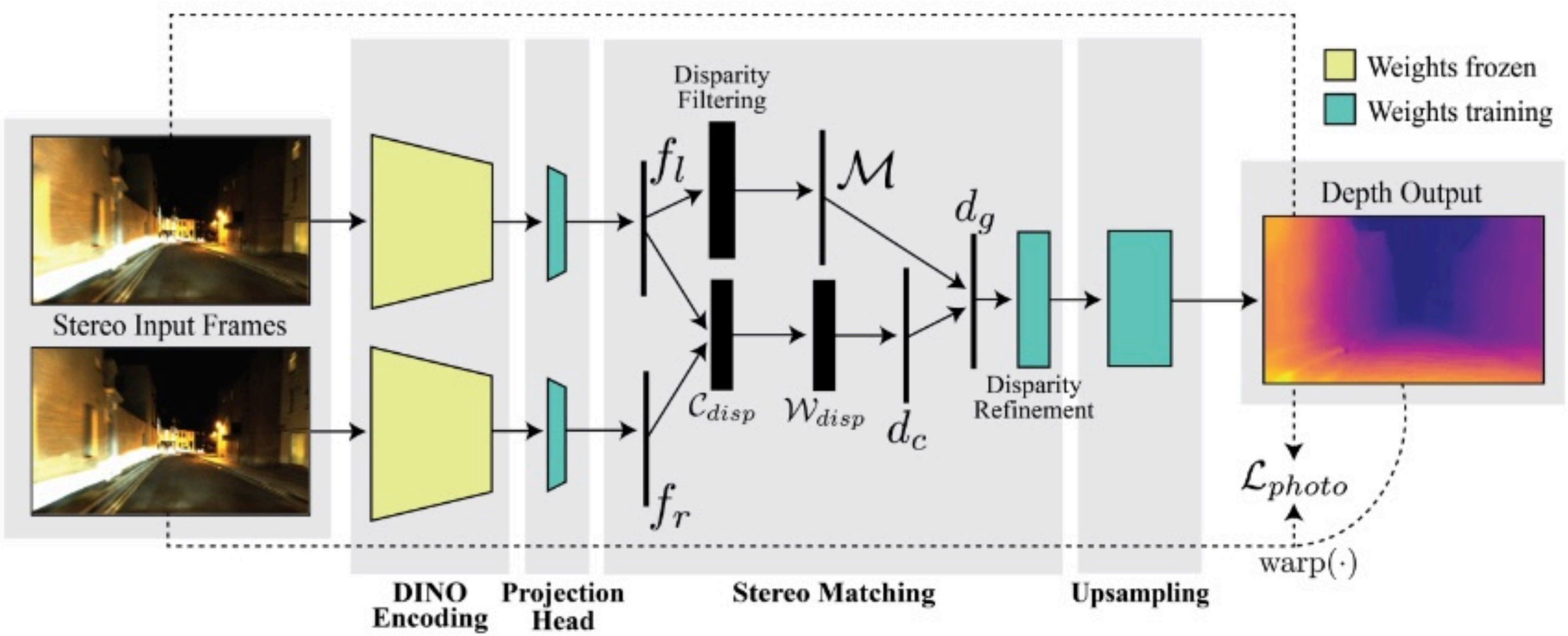
Results:

Better nighttime accuracy than supervised baselines

Highlights:

- Illumination-robust
- Label-free training
- First day–night stereo learning framework

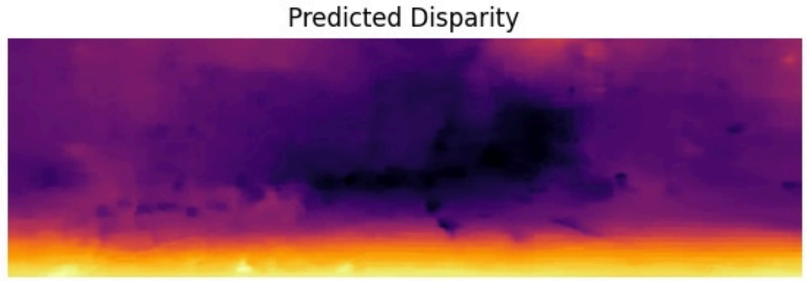
Architecture:

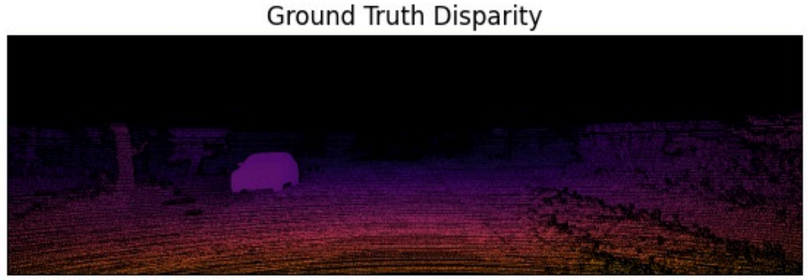


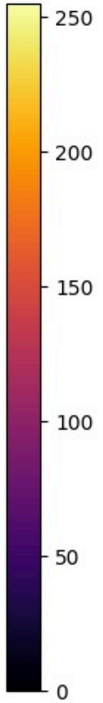
Our approach consists of four main elements. Features are encoded independently for each input using DINO [13], a learnable projection head adapts these features and reduces their dimension, giving fl and fr . Stereo matching of the features then takes place, with disparity filtering yielding the mask M, and the combination of fl and fr providing the correspondence volume Cdisp. Wdisp is found by using softmax on Cdisp, which is used to find coarse disparity dc. Coarse disparity and the mask combine to give global disparity dg, which is refined and upsampled to give final depth.

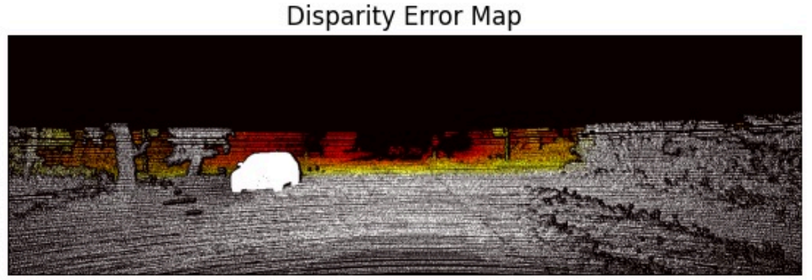
Madhu Vankadari et al., "Dusk Till Dawn: Self-supervised Nighttime Stereo Depth Estimation using Visual Foundation Models," 2024.

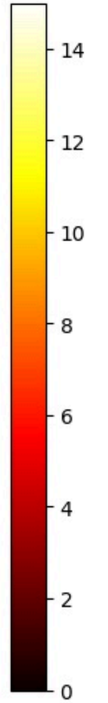
Predictions from DTD:











ACVNet and Fast-ACVNet

Paper: Accurate and Efficient Stereo Matching via Attention Concatenation Volume (2022)

Goal:

- SOTA algorithm for real-time depth estimation from stereo images.

Key Ideas:

- A novel cost volume construction method (attention concatenation volume - ACV), which generates **attention weights** from **correlation clues** to suppress redundant information and **enhance matching-related information** in the concatenation volume.
- Propose a **Volume Attention Propagation module** and a **Fine-to-Important sampling** strategy, which are key success factors of Fast-ACV.

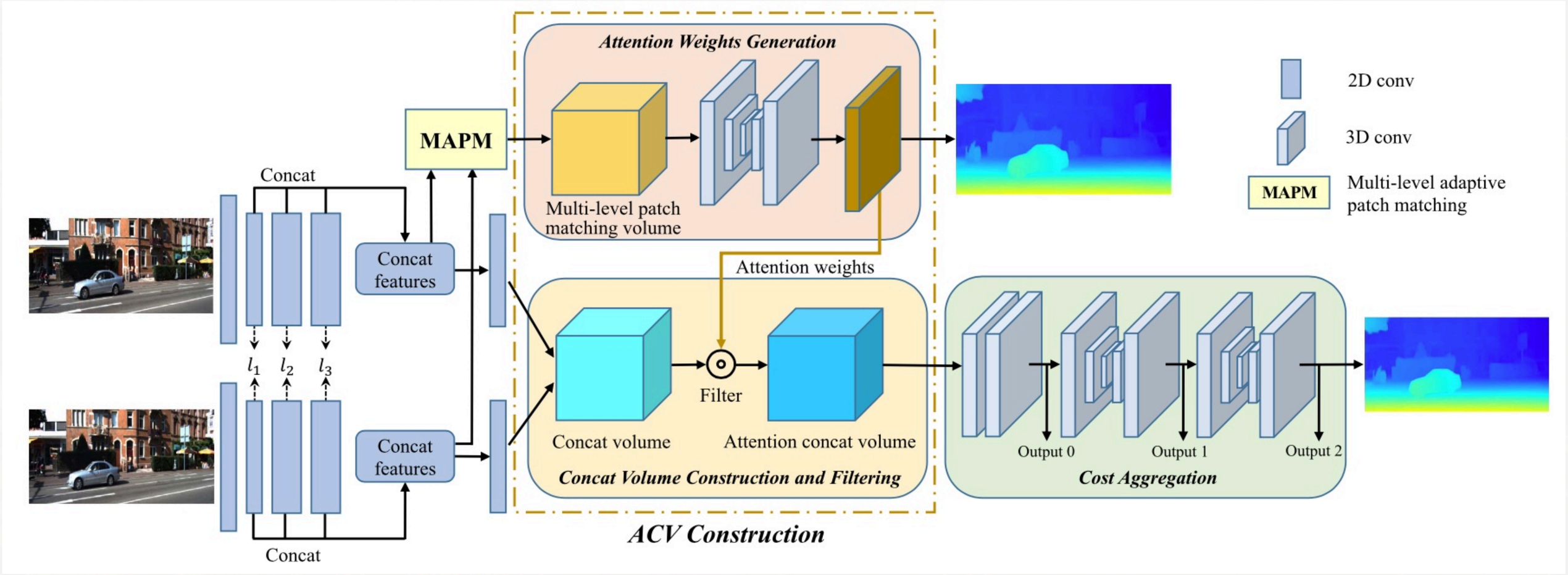
Results:

Outperforms other SOTA methods

Dataset:

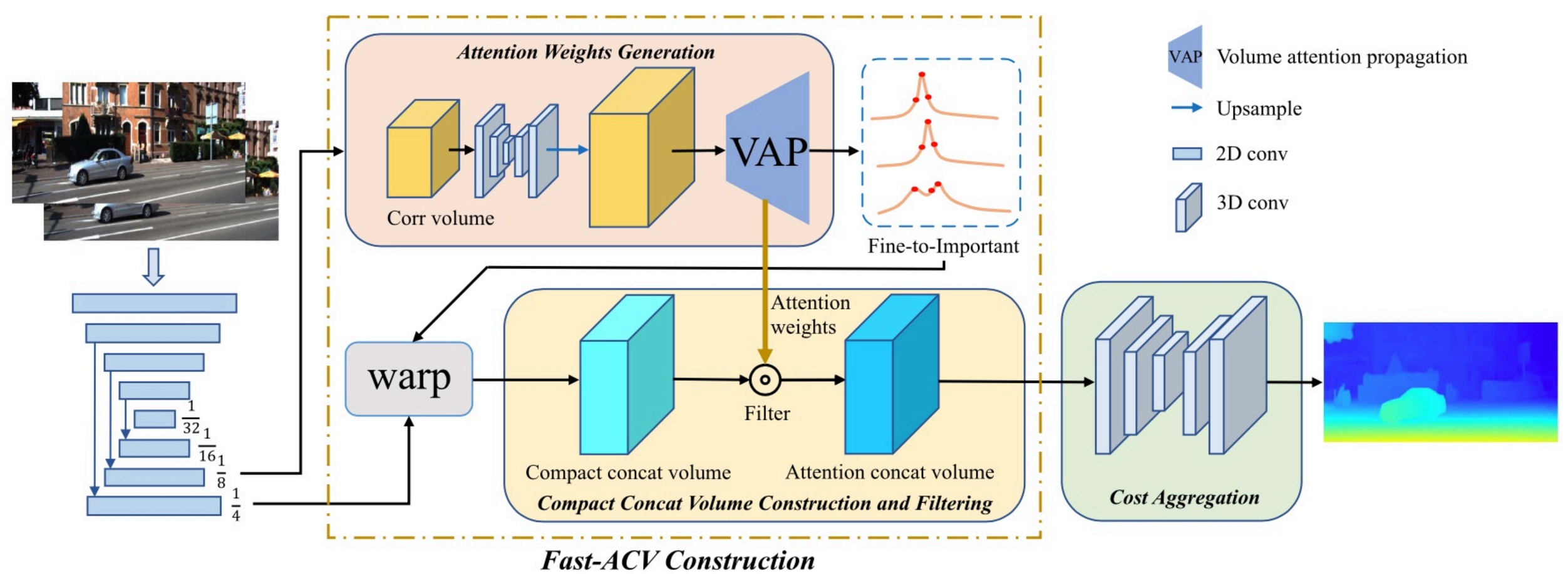
Scene Flow, ETH3D, KITTI

ACVNet Architecture



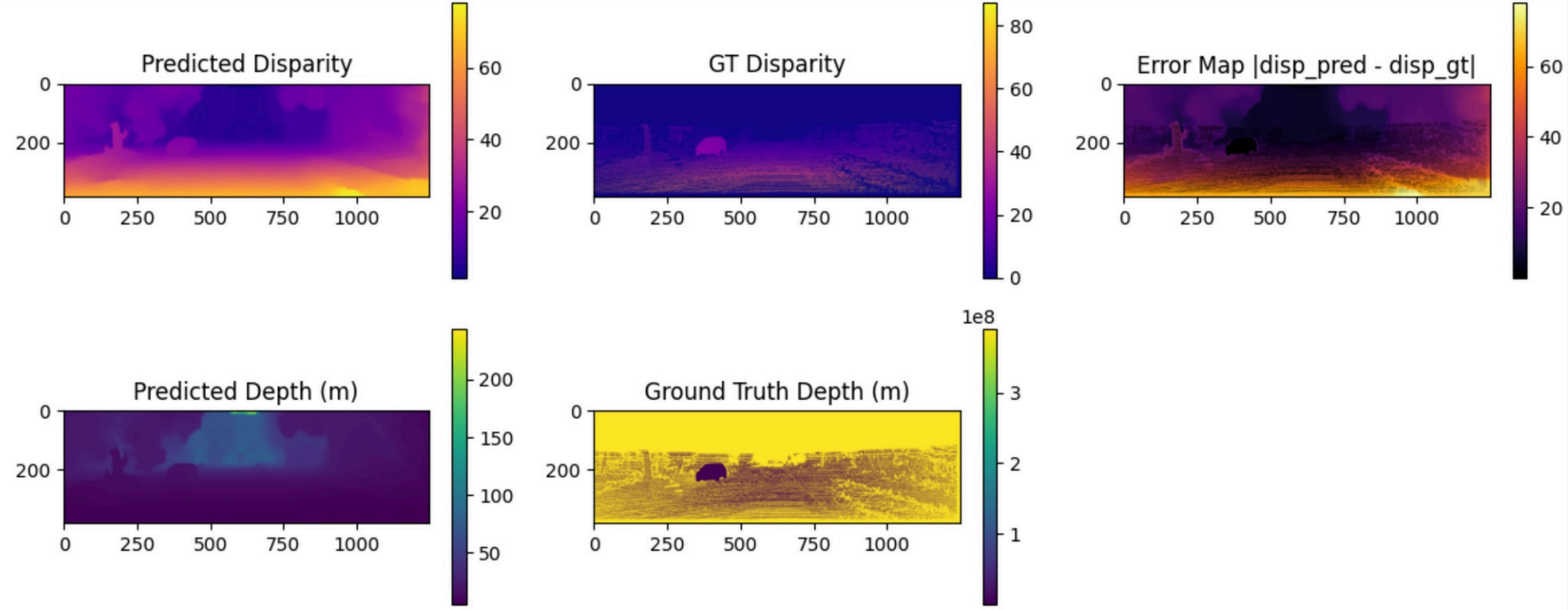
The construction process of ACV consists of three steps: Attention Weights Generation, Initial Concatenation Volume Construction and Attention Filtering. First obtain feature maps at three different levels l1, l2 and l3 from the feature extraction module, and the number of channels for l1, l2 and l3 is 64, 128 and 128 respectively. l1, l2 and l3 are concatenated to form 320-channel concat features for the generation of attention weights. Then two convolutions are applied to compress the 320-channel concat features to 32-channel features for construction of the initial concatenation volume.

Fast-ACVNet Architecture



First exploit a correlation volume to generate disparity hypotheses with high likelihood and the corresponding attention weights. Then we use the attention weights to filter the compact concatenation volume constructed based on disparity hypotheses, deriving our Fast-ACV.

Results


Predicted Disparity
GT Disparity
Error Map |disp_pred - disp_gt|
Predicted Depth (m)
Ground Truth Depth (m)
1e8

Geometry-Informed Neural Operator (FNO)

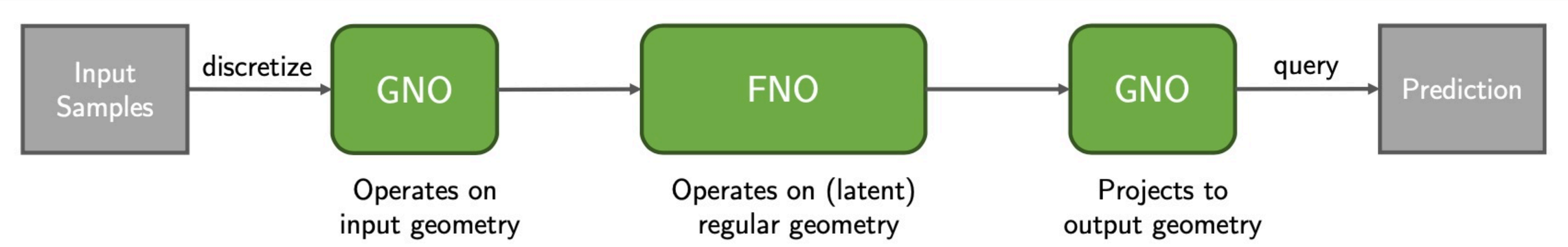
Paper: Geometry-Informed Neural Operator for Large-Scale 3D PDEs (2023)

Goal: • To learn the solution operator of large-scale partial differential equations with varying geometries.

Key Ideas:

- Uses a signed distance function (SDF) and point-cloud representations of the input shape and neural operators based on graph and Fourier architectures to learn the solution operator.
- The graph neural operator handles irregular grids and transforms them into and from regular latent grids on which Fourier neural operator can be efficiently applied. GINO is discretization-convergent.
- Dataset: Industry-standard aerodynamics dataset of 3D vehicle geometries with Reynolds numbers as high as five million. GINO successfully trained to predict the pressure on car surfaces using only five hundred data points.

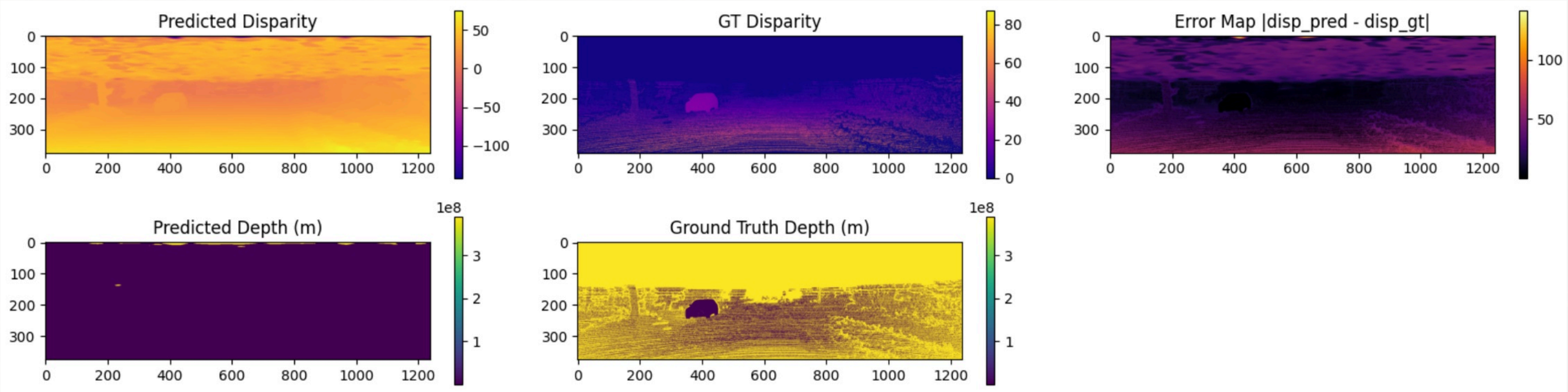
GINO Architecture



Use this architecture for stereo depth estimation

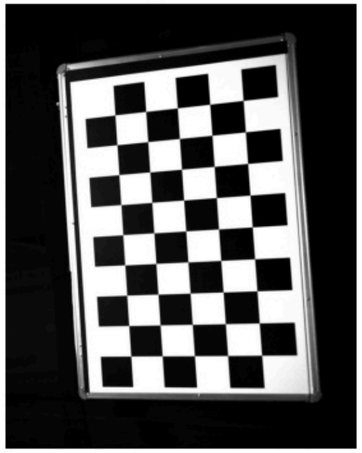
- The GNO block will allow us to interpolate which will introduce a smoothing effect and generalize to different image sizes.
- The FNO layers project the data to fourier space (CONVOLUTIONS TO MULTIPLICATION). Since it is natural to represent image data in fourier space we might obtain highly accurate results from this.
- The memory requirements for the entire model are extremely high. Hence, we have just trained a single FNO unit.

Results

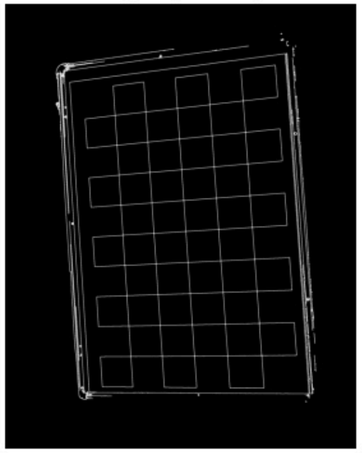

Predicted Disparity
GT Disparity
Error Map |disp_pred - disp_gt|
Predicted Depth (m)
Ground Truth Depth (m)
1e8
1e8

Edge OverLay BM + SGBM

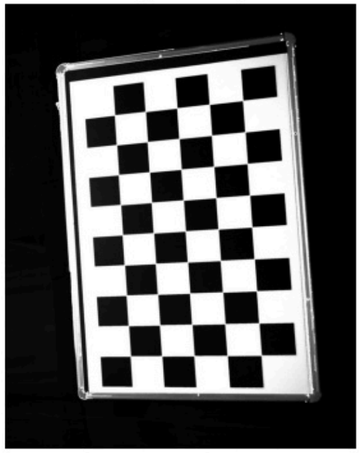
Original Left



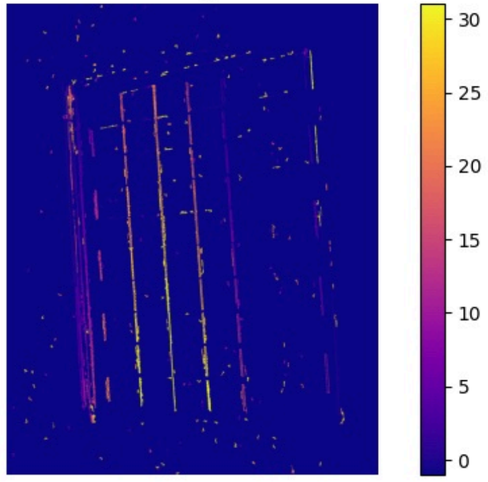
Edges (dilated)



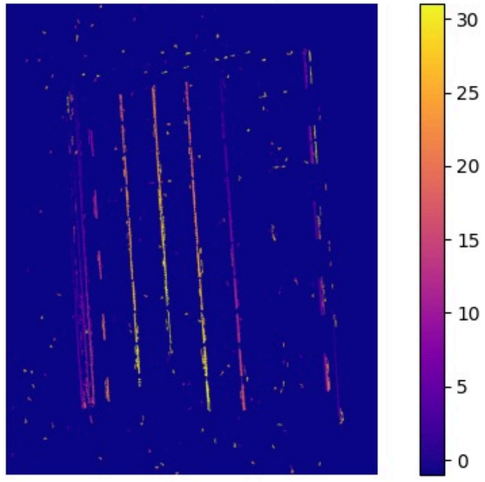
Edge Overlay



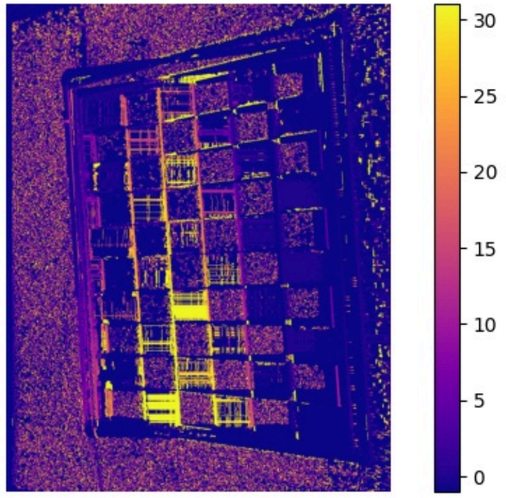
BM Original



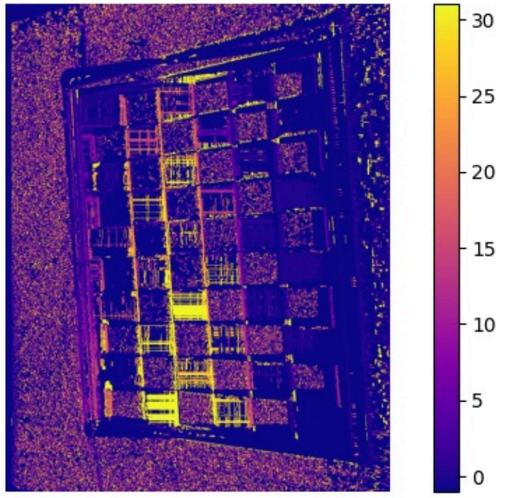
BM Edge Overlay



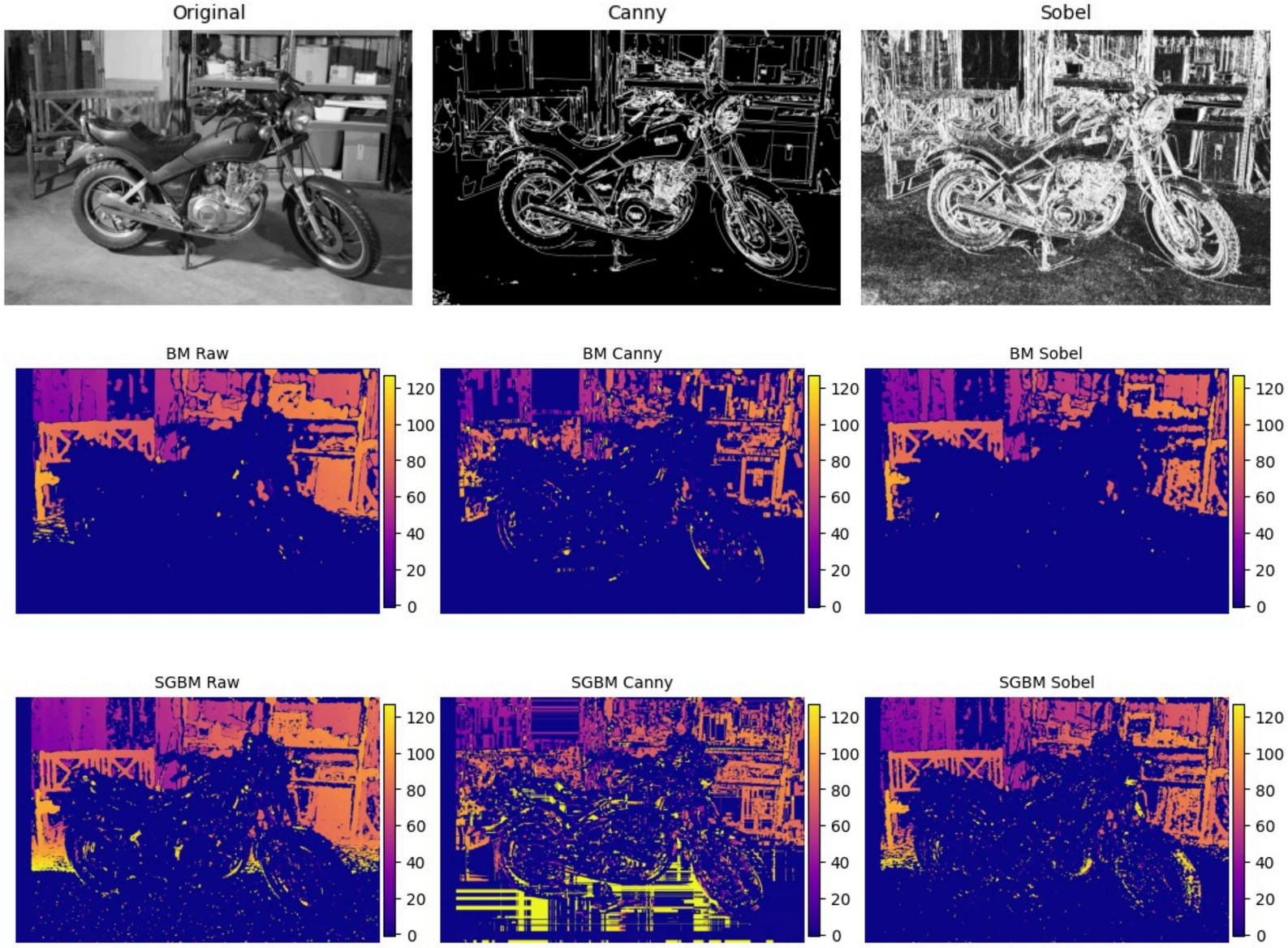
SGBM Original



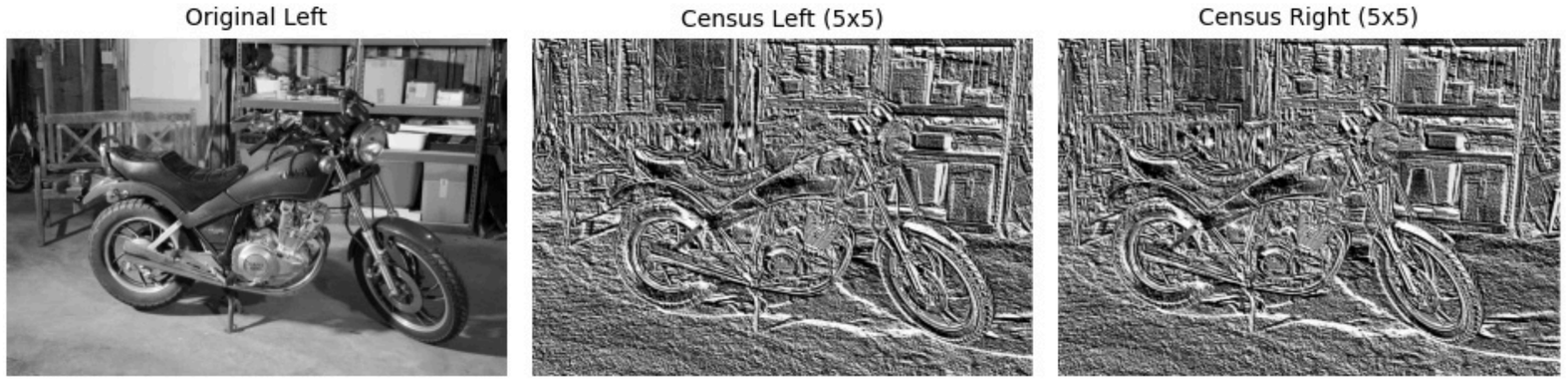
SGBM Edge Overlay

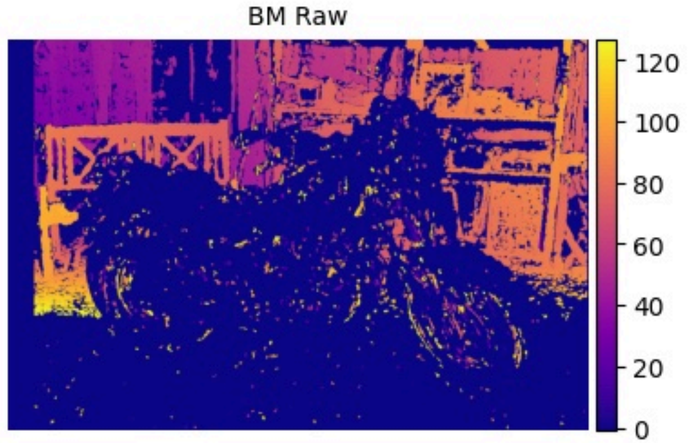


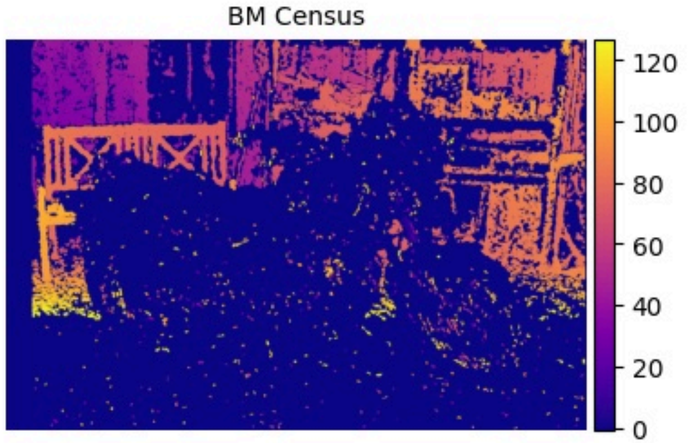
Intensity Gradient BM + SGBM

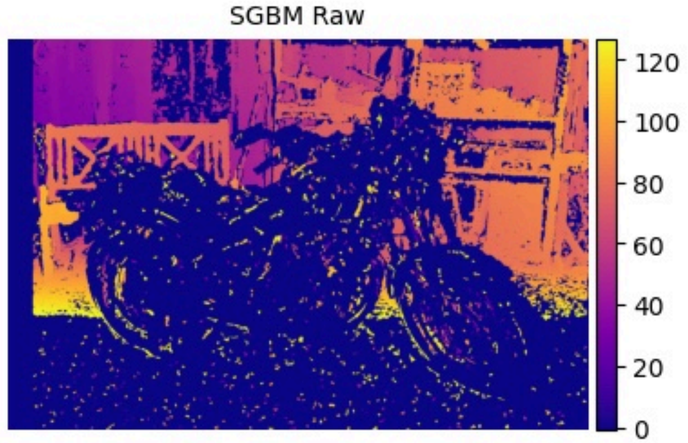


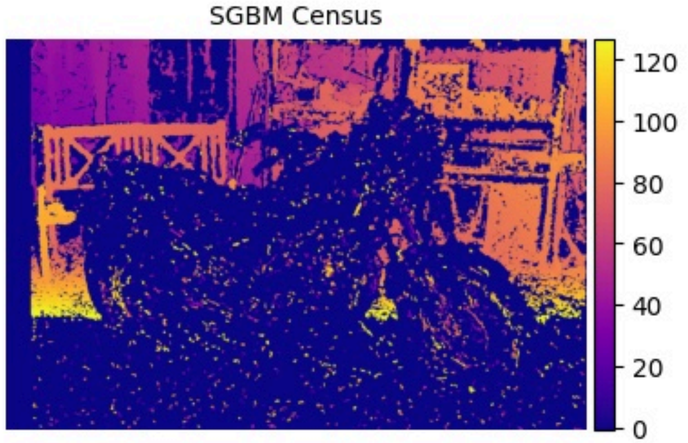
Census Transform BM SGBM



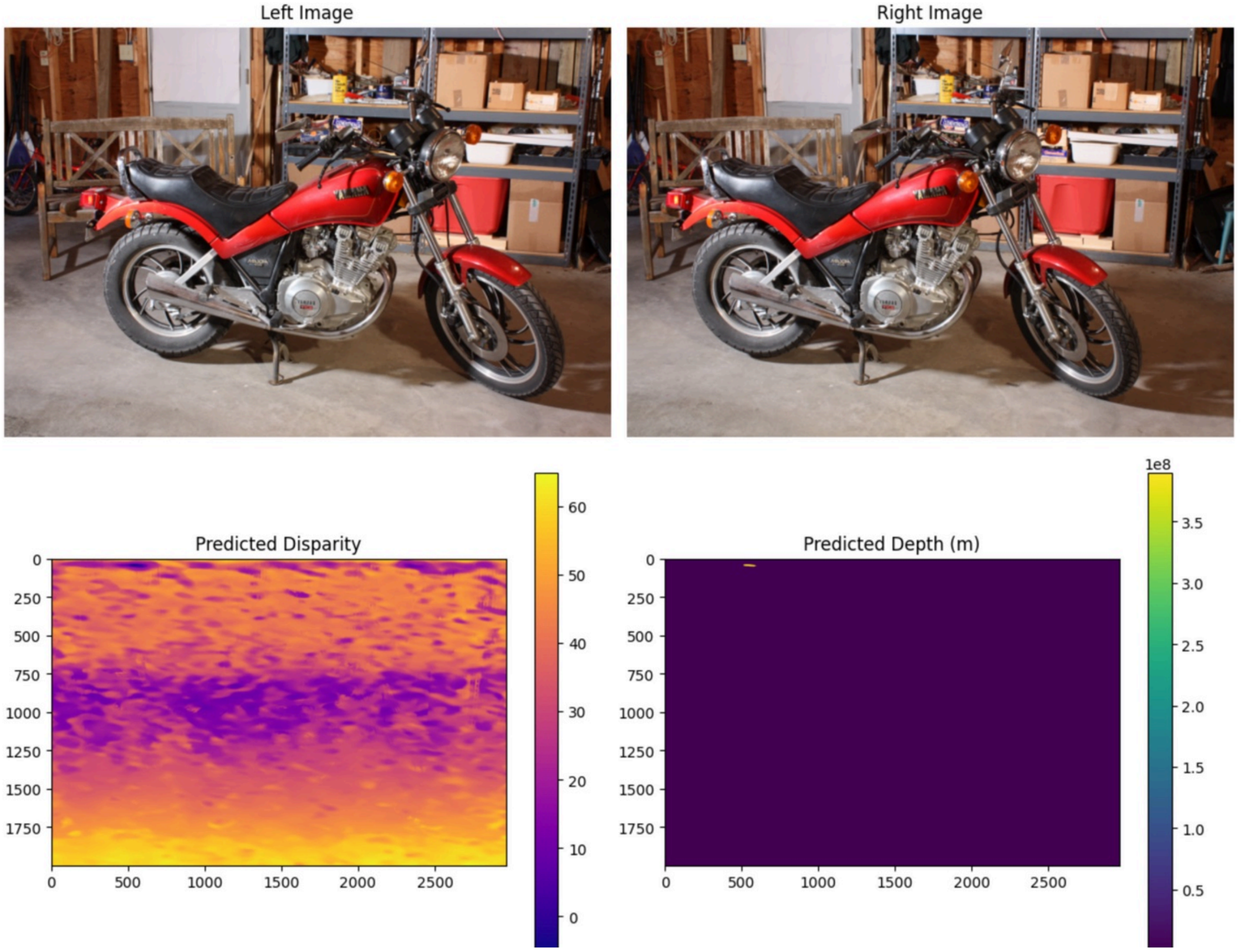




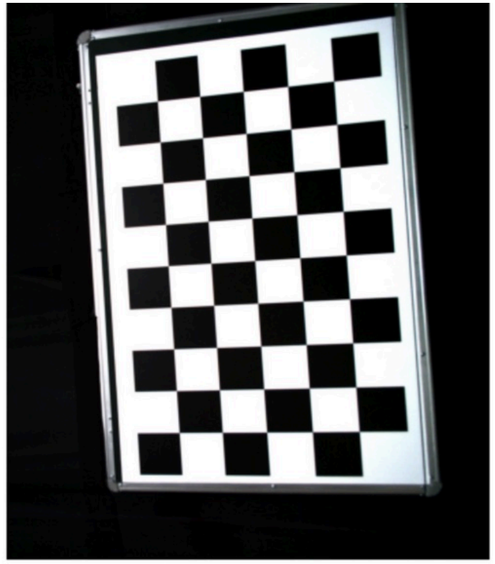


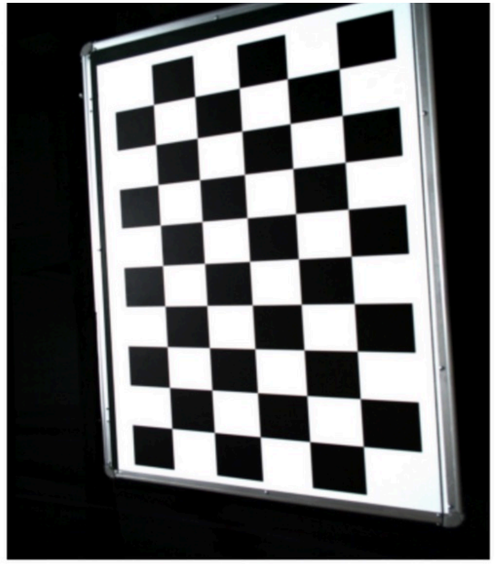


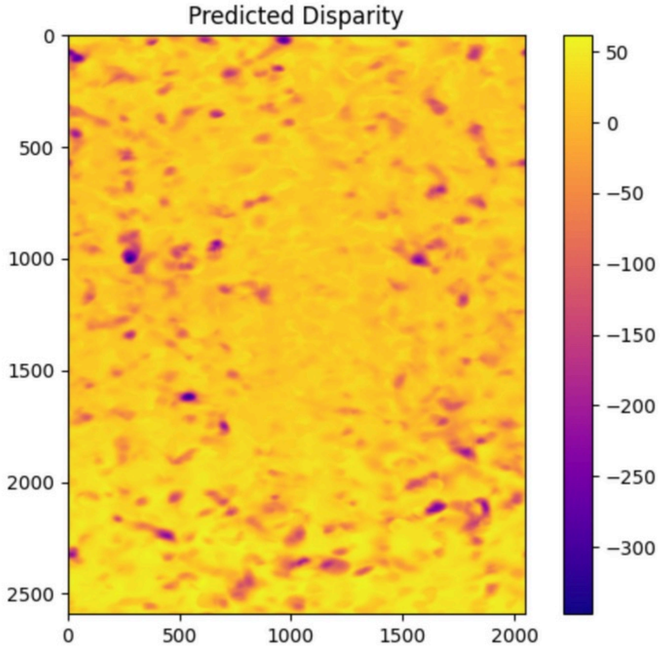
FNO

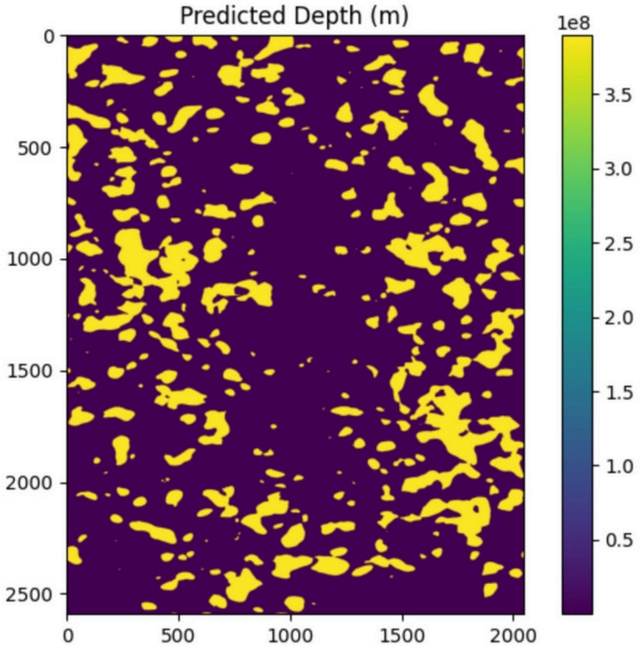

Left Image
Right Image
Predicted Disparity
0
250
500
750
1000
1250
1500
1750
0
500
1000
1500
2000
2500
60
50
40
30
20
10
0
Predicted Depth (m)
1e8
3.5
3.0
2.5
2.0
1.5
1.0
0.5

FNO

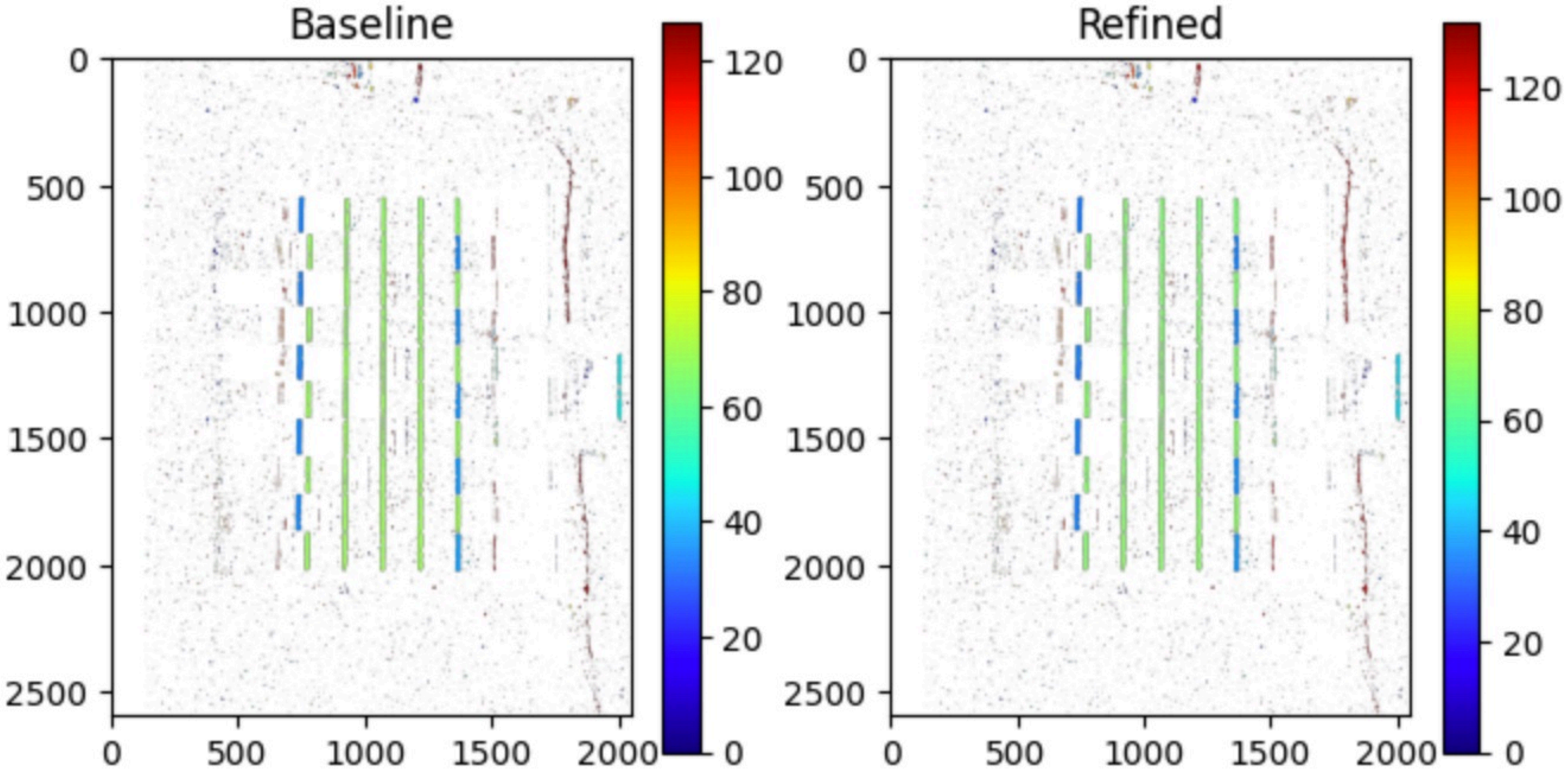




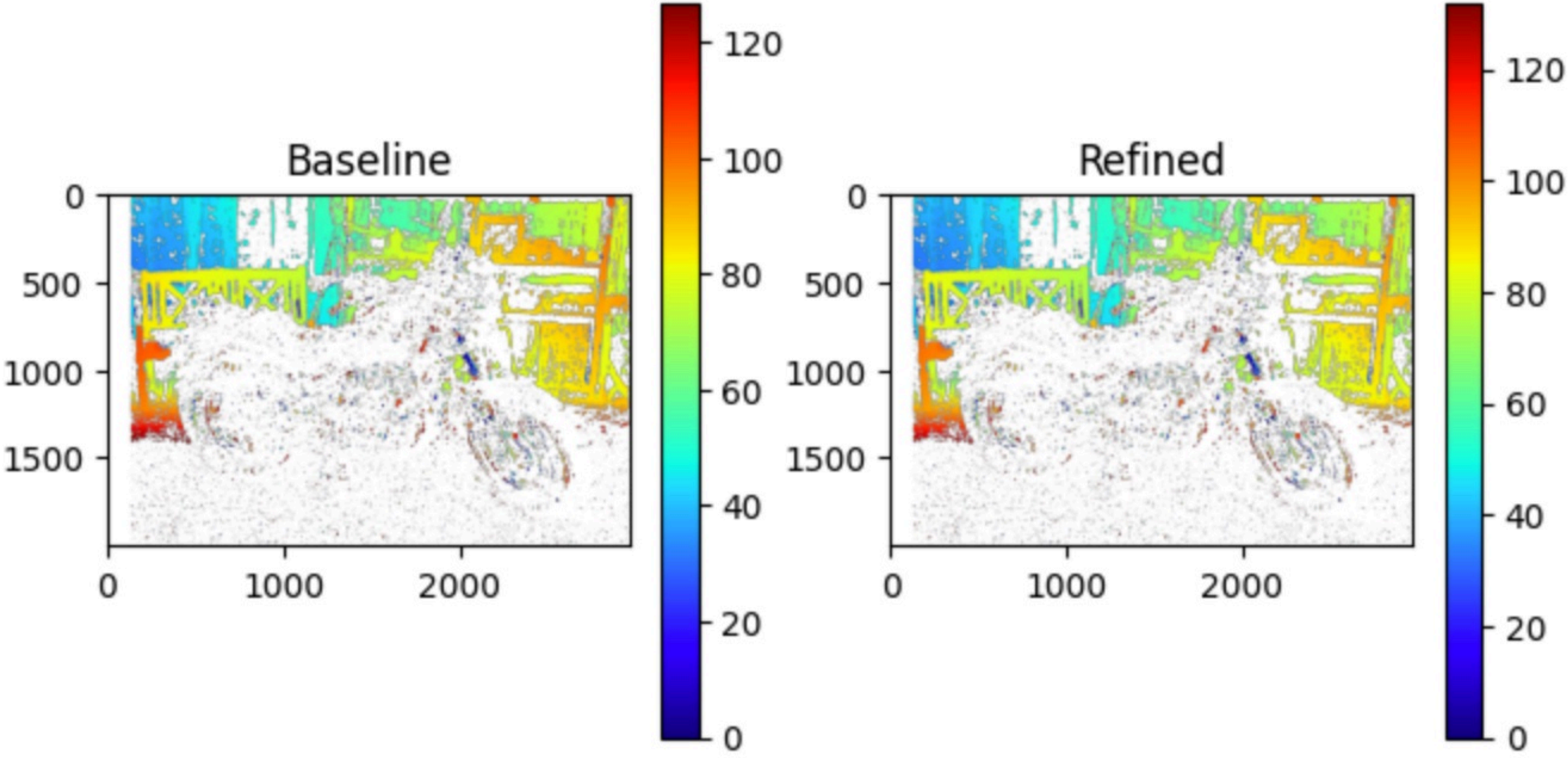

Predicted Disparity
0
500
1000
1500
2000
2500
0
500
1000
1500
2000
50
0
−50
−100
−150
−200
−250
−300


Predicted Depth (m)
1e8
0
500
1000
1500
2000
2500
0
500
1000
1500
2000
3.5
3.0
2.5
2.0
1.5
1.0
0.5

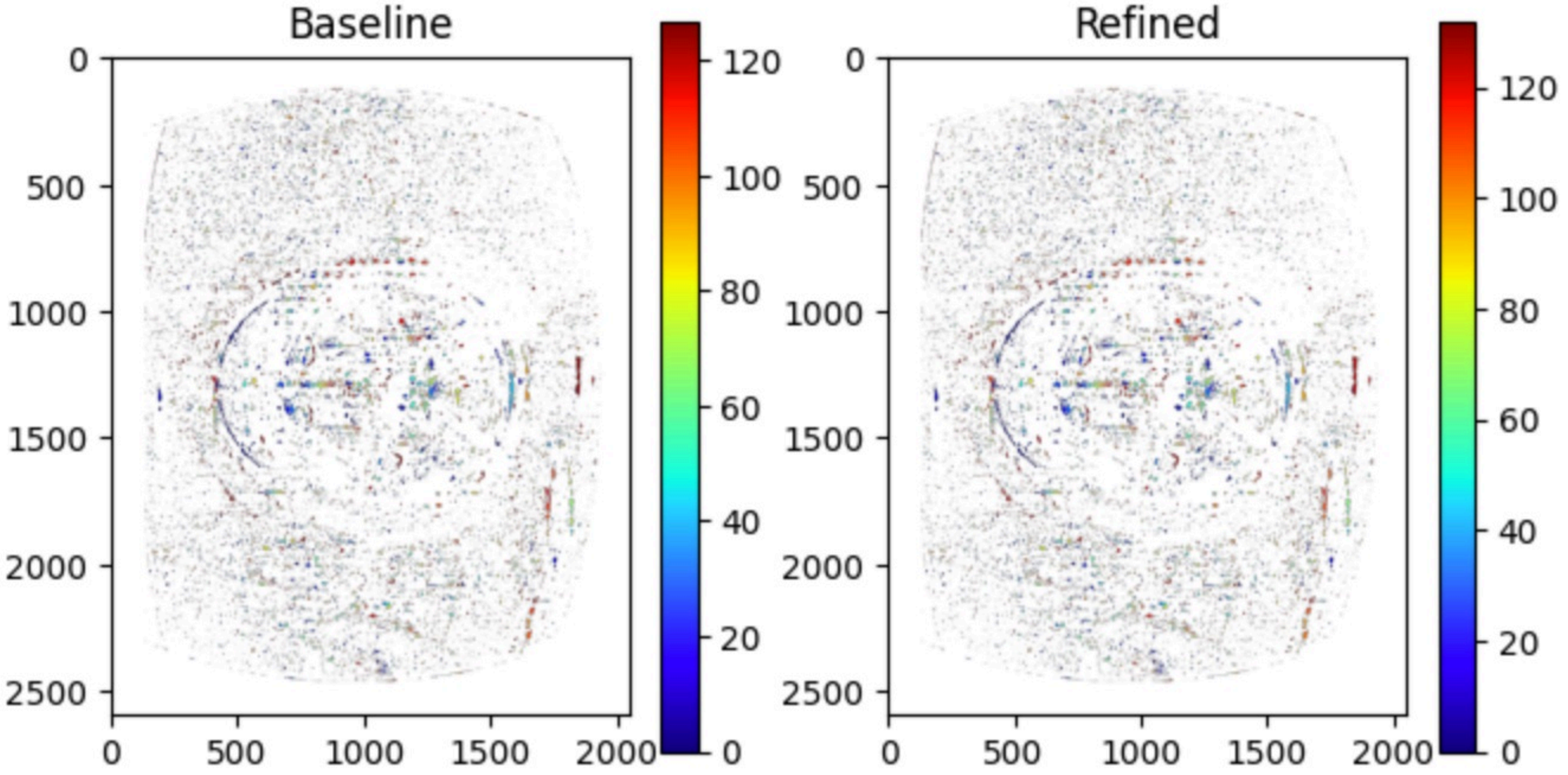
Coarse-to-Fine



Coarse-to-Fine



Coarse-to-Fine



The End

THANK YOU