

# Towards Understanding 3D Vision: the Role of Gaussian Curvature

Sherlon Almeida da Silva<sup>1,2</sup> Davi Geiger<sup>2</sup> Luiz Velho<sup>3</sup> Moacir Antonelli Ponti<sup>1</sup>

<sup>1</sup>Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP)

<sup>2</sup>Courant Institute of Mathematical Sciences, New York University (NYU)

<sup>3</sup>Instituto de Matemática Pura e Aplicada (IMPA)

{sherlon.a, dg1}@nyu.edu, lvelho@impa.br, moacir@icmc.usp.br

## Abstract

Recent advances in computer vision have predominantly relied on data-driven approaches that leverage deep learning and large-scale datasets. Deep neural networks have achieved remarkable success in tasks such as stereo matching and monocular depth reconstruction. However, these methods lack explicit models of 3D geometry that can be directly analyzed, transferred across modalities, or systematically modified for controlled experimentation. We investigate the role of Gaussian curvature in 3D surface modeling. Besides Gaussian curvature being an invariant quantity under change of observers or coordinate systems, we demonstrate using the Middlebury stereo dataset that it offers a sparse and compact description of 3D surfaces. Furthermore, we show a strong correlation between the performance rank of top state-of-the-art stereo and monocular methods and the low total absolute Gaussian curvature. We propose that this property can serve as a geometric prior to improve future 3D reconstruction algorithms.

## 1. Introduction

Vision is a fundamental sensory modality that allows us to perceive and interpret the structure of our surroundings, playing a critical role in numerous robotics applications. However, purely data-driven deep learning approaches – despite their success in tasks such as depth estimation – often lack explicit and transferable geometric representations.

We emphasize that *understanding* or *explaining* the visual world is not the same scientific goal as developing algorithms purely for *prediction*. For example, a predictive model might use visual video data and machine learning to accurately forecast the trajectory of a falling object. Although such a system may yield precise results, it does not necessarily uncover or convey the underlying physical law: namely,  $F = ma$ , where  $F$  is the gravitational force and  $a = 9.8 \text{ m/s}^2$  is the acceleration due to gravity.

More generally, explanatory models aim to reveal the underlying structure of the world. It is commonly believed that such models offer greater **simplicity** and **generalization power**, as exemplified in the classical laws of physics where  $F = ma$  is applied to other physical scenarios (not just objects falling under gravity).

The world around us is shaped both by natural processes and by human-made structures, which are commonly represented in stereo-vision by depth and disparity measurements. These metrics can accurately describe 3D geometry but depend heavily on the observer’s position and viewpoint, limiting their robustness. In contrast, Gaussian curvature is a purely local geometric property of a surface that remains invariant under changes in viewpoint, making it ideal for robust 3D surface analysis.

Our main contribution is on *Foundational Vision Understanding* by deepening our understanding of 3D scene geometry, paving the way for more **interpretable**, **generalizable**, and **reliable** vision systems. This is an analytical work grounded on data from state-of-the-art (SOTA) Deep Learning (DL) stereo algorithms. We investigate their benchmarking performance and how this insight about the importance of Gaussian curvature description of the 3D world is implicitly captured by stereo algorithms with indoor scenes on Middlebury dataset. We show that (i) Gaussian curvature is sparsely distributed across natural 3D surfaces; (ii) Low Gaussian curvature magnitude can serve as a prior for regularizing or modeling 3D surface geometry; (iii) This prior may be implicitly captured by current SOTA algorithms, but it is not explicit as an independent module and so we cannot use it elsewhere; (iv) Low Gaussian curvature magnitude enables the definition of a novel unsupervised evaluation metric for assessing the quality of 3D surface reconstruction algorithms.

The contributions fall under the following two umbrella concepts:

- **Scene Understanding:** The identification of quantities that are zero or nearly zero across most of a scene – thus

encoding data with a minimal set of active components – forms the foundation of sparse representation. We show in Section 4 that the Gaussian curvature magnitude fulfills this role for 3D data representations of indoor environments.

**– Explainable AI:** In Section 5 we empirically verify that the SOTA stereo algorithms seem to apply low Gaussian curvature magnitude priors. However, it is unclear whether or where these priors are explicitly used, making it impossible to isolate reusable algorithmic modules, in analogy to modules that perform feature extraction, from current depth reconstruction methods.

A potential immediate application of this explainability study is to use low Gaussian curvature magnitude as an unsupervised metric to regularize models and/or evaluate 3D reconstruction algorithms.

## 2. Previous Work

We begin by highlighting the influential book by Koenderink [15], in which a differential geometry framework is used to provide a rich theoretical foundation for understanding surface geometry in vision. However, it does not address the construction of a sparse representation nor explore Gaussian curvature as a key feature for computer vision. The field of sparse image representation gained prominence in image analysis through the work of Field and Olshausen [5, 21], who highlighted the statistical regularities present in natural images and proposed sparse coding models inspired by the processing mechanisms of the visual cortex. Since then, these ideas have been extended to various visual domains, including, but not limited to, texture modeling [33], illumination and shape analysis [1, 8], template matching [10], and image restoration [18].

Some work worth mentioning as they address some aspects of the topics presented here. In studying visual illusions, Ishikawa and Geiger [13] noted that humans may have low Gaussian curvature priors when reconstructing surfaces. A recent application of imposing the zero Gaussian curvature on man-made CAD reconstruction indicates the value of such extreme case (zero Gaussian curvature) for designing industrial applications [7]. Also, Guo [11] introduced the Gaussian Curvature Co-occurrence Matrix, combining curvature with co-occurrence statistics for 3D shape representation. Zhong and Qin [30] proposed a sparse approximation for 3D shapes using spectral graph wavelets to capture local geometric details. Ververas et al. [24] apply a curvature-based densification step to populate an under-represented area. These studies highlight the potential of integrating curvature and sparse representations in 3D geometry analysis.

In contrast to previous approaches to 3D surface understanding, which often focus on quantities defined within specific coordinate frames or based on appearance [25, 32], our work seeks a geometric quantity that is invariant to the

choice of observer or coordinate system, a property that is especially desirable for 3D surface analysis. To the best of our knowledge, there have been no prior studies applying sparse representation principles to Gaussian curvature in 3D geometry.

## 3. Gaussian Curvature (GC)

We begin by briefly reviewing the concept of GC and explaining its relevance to our study. Next, we discuss existing methods for estimating GC. Finally, we describe the synthetic 3D scenes we generated to enable controlled curvature analysis and benchmarking.

### 3.1. Brief Review

GC of a smooth surface at a given point is given by [6]:

$$K = \kappa_1 \kappa_2, \quad (1)$$

where  $\kappa_1$  and  $\kappa_2$  are the principal curvatures of a surface, i.e., the maximum and minimum normal curvatures. The sign of the GC indicates the local shape of the surface: (i) Positive GC ( $K > 0$ ) and the surface is locally convex. Both principal curvatures have the same sign, e.g. sphere of radius  $r$  has constant curvature  $K = \frac{1}{r^2}$ ; (ii) Zero GC ( $K = 0$ ) and one of the principal curvatures is zero, e.g. plane or cylinder; (iii) Negative GC ( $K < 0$ ) and the principal curvatures have opposite signs, e.g. a hyperboloid of one sheet or the inner region of a torus. Note that a GC has dimensions inverse to the square of distance. Gauss’s Theorema Egregium states that **GC is an intrinsic property** of a surface.

This invariance makes GC particularly valuable for computer vision and geometric modeling, as it provides a consistent descriptor regardless of viewpoint or deformation. The Gauss–Bonnet theorem further reinforces its significance by relating the integral of GC over a surface to its Euler characteristic, thereby bridging local geometric information with global topological structure. Moreover, previous studies have shown that low GC plays a role in human shape understanding (see [13]).

### 3.2. Methods to estimate GC from data

In order to estimate a GC we use the formula [6]:

$$K = \frac{\det(\text{II})}{\det(\text{I})}. \quad (2)$$

where the fundamental forms I and II can be obtained from a parametric surface  $\mathbf{Z}(u, v)$ , its derivatives  $\mathbf{Z}_u(u, v), \mathbf{Z}_v(u, v)$  and second derivatives  $\mathbf{Z}_{uu}, \mathbf{Z}_{uv}, \mathbf{Z}_{vv}$ . More details of the formula are in the supplementary material.

We construct the parametrized surface from the depth data. More precisely, given a set of  $n$  depth measurements

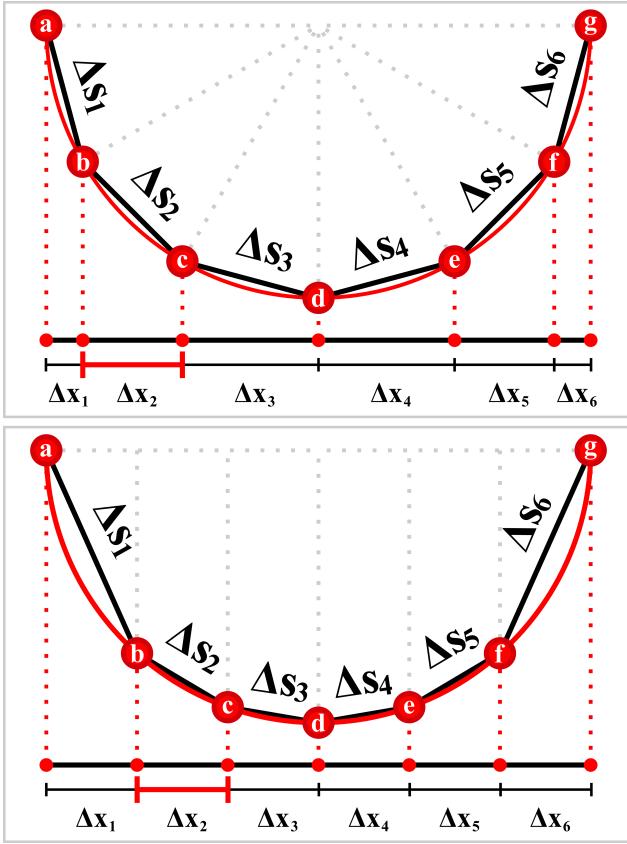


Figure 1. **a,b,...,g** are sample points of a surface  $z(x, y)$  along a slice  $z(x, y_0)$ . GC require calculations with surface distances between points, here denoted by  $\{\Delta s_i; i = 1 \dots 6\}$  and shown on a curve, a slice of a surface. a. Top: The distances are constant, i.e.,  $\Delta s_1 = \Delta s_2 = \dots = \Delta s_6$ . Resulting in non-uniform projected distances  $\Delta x_i; i = 1 \dots 6$ . b. Bottom: The projected distances  $\Delta x_i; i = 1 \dots 6$  are constant, but then the distances  $\Delta s_i; i = 1 \dots 6$  between surface points are non-uniform.

$\{d_i(x_i, y_i)\}_{i=1}^n$ , we represent the surface as a 3D point cloud  $\{(X_i, Y_i, d_i(x_i, y_i))_{i=1}^n\}$  where  $X_i, Y_i$  are obtained by inverting the projection, that is,  $X_i = \frac{(x_i - c_x)}{f_x} d_i(x_i, y_i)$  and  $Y_i = \frac{(y_i - c_y)}{f_y} d_i(x_i, y_i)$ , before assigning a parametrized surface  $Z(u, v)$ . Since numerical differentiation tends to amplify noise, especially when applied to noisy depth data, we apply smoothing when necessary. Importantly, we smoothed the 3D point cloud with Gaussian smoothing over each 3D component,  $X, Y, Z$ , independently. Note that one should not smooth the raw depth map, as the depth values are not uniformly sampled with respect to surface distances, and the smoothing in the image space does not adequately account for the true geometry (see Figures 1 and 2).

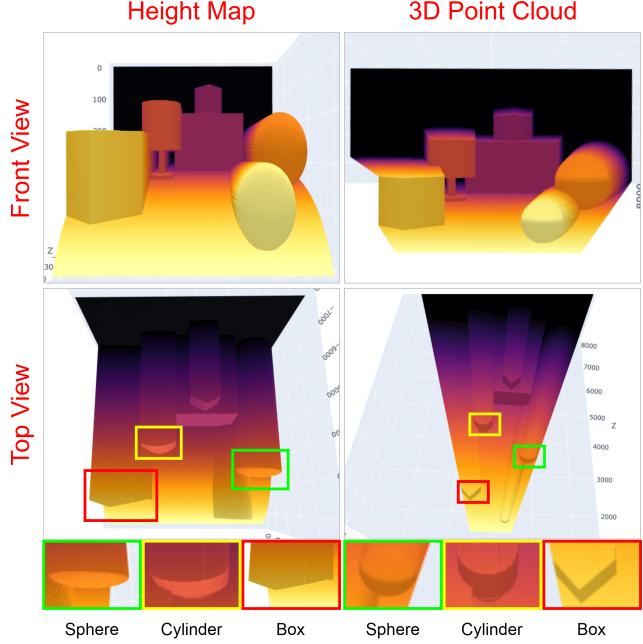


Figure 2. This figure visually represents the  $\Delta x$  uniform distances on the Height Map (left) and the  $\Delta s$  uniform distances on the 3D Point Cloud. Observe that not dealing with the correct GC formula may lead to a wrong curvature analysis.

### 3.3. 3D Scenes for Curvature Analysis

In order to evaluate algorithms that estimate GC on shapes we created five 3D synthetic scenes<sup>1</sup> composed of developable surfaces (zero GC), such as cylinders, boxes, and planes, and spheres representing a convex surface with constant positive curvature. Despite the simplicity of the scenes, it provides complete control over stereo image acquisition and precise ground truth for GC analysis.

The simulated environment was created using Unity 3D. We created two physical cameras inspired on Middlebury settings, with **Sensor Size (H,W)** = (14.8, 22.2)mm, **Image Size (H,W)** = (2000, 3000)px, **Focal Length (mm, px)** = (35, 4729.73) and **Pixel Size** = 0.0074mm. The cameras were YZ-aligned, and X-shifted by 200 millimeters to guarantee the images are rectified. Thus, **Baseline** = 200mm.

Figure 3 illustrates the five synthetic scenes created. As part of our ongoing work, we plan to expand the number of 3D synthetic scenes by including additional objects with known GC. For a more detailed analysis of GC, particularly in the MainScene, please refer to Figure 4.

<sup>1</sup>3D Synthetic Scenes Page: <https://github.com/SherlonAlmeida/Stereo-Cameras-Simulation>

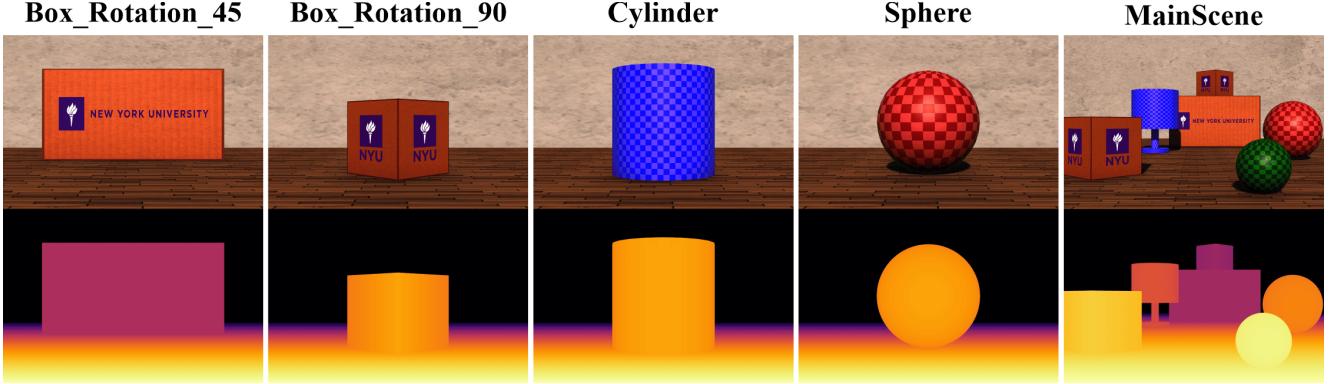


Figure 3. **3D Synthetic Scenes:** Five scenes for depth estimation and curvature analysis, where: Box.Rotation.45, Box.Rotation.90, and Cylinder have  $K = 0$ ; Sphere  $K = \frac{1}{r^2}$ ; and MainScene mix all these objects in a scene. We provide the depth in meters (in this paper), left and right images with (2000, 3000)px resolution. The two spheres in the MainScene have the radius of  $r = 0.25\text{m}$  and  $r = 0.125\text{m}$ .

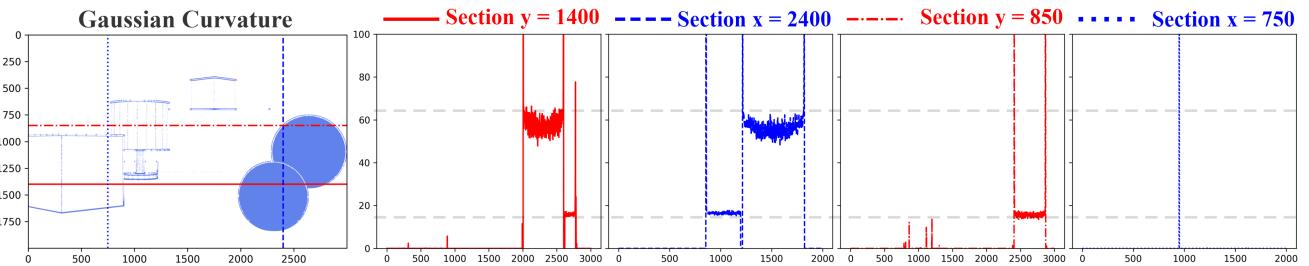


Figure 4. GC Cross-sections on the MainScene data. Since second-order derivatives are sensitive to small changes, we smooth the slices with a Gaussian filter of  $\sigma = 2\text{ m}$ . The Y-section is plotted from left to right, and each X-section from top to bottom. Note (1) high curvature at the edges of 3D objects in the scene; (2) Approximate constant positive curvature inside the spheres, specifically  $16\text{ m}^{-2}$  for the red sphere, and  $64\text{ m}^{-2}$  for the green sphere, which are the correct GC values. (3) Zero curvature in the remaining areas.

## 4. A Sparse Representation of 3D data

We investigate a sparse representation associated with depth data by analyzing the Middlebury stereo dataset [22], which provides accurate ground truth (GT) depth maps.

The Middlebury dataset includes a training partition, where GT data is publicly available. The training dataset contains 15 images, each of size approximately  $2,000 \times 3,000$  pixels, yielding on the order of  $9 \times 10^7 = 15 \times 2,000 \times 3,000$  ground truth depth values.

Middlebury results are provided as disparity maps, so we evaluated the GC of the 15 training images as follows: we computed  $Depth = \frac{f \cdot b}{d + doffs}$ , where  $f$  is the focal length in pixels,  $b$  is the baseline in meters, and  $d + doffs$  is the disparity value in pixels corrected by the center of projection offset. For each pixel, we compute the GC (throwing away boundary data) and aggregate the results and normalize them into a histogram, shown in Figure 5.

Let us denote by  $h(K)$  the *normalized histogram* of GC values computed empirically over the Middlebury dataset. This histogram can be interpreted as an *empirical prior distribution* over GC. Based on this, we define a prior model over 3D surface geometry as

$$P(K) = e^{-\mathcal{L}(K)}, \quad (3)$$

where the loss function  $\mathcal{L}(K)$  is defined by

$$\mathcal{L}(K) = -\ln h(K). \quad (4)$$

Our empirical findings suggest that GC is sparsely distributed in real-world 3D scenes, supporting our hypothesis that it captures structural regularities in surface geometry.

A practical and interpretable approximation to this loss is a regularizer:

$$\mathcal{L}(K) = \alpha |K|^{\frac{1}{2}} = \alpha \sqrt{|\kappa_1 \kappa_2|}, \quad (5)$$

where  $\alpha > 0$  is a weighting parameter in units of distance that helps match the empirical distribution. This loss function is the sparse  $L^0$  norm of quantities  $|\kappa_1|$  and  $|\kappa_2|$ , i.e.,

$$\sqrt{|\kappa_1 \kappa_2|} = \lim_{p \rightarrow 0} \left( \frac{1}{2} (|\kappa_1|^p + |\kappa_2|^p) \right)^{\frac{1}{p}}, \quad (6)$$

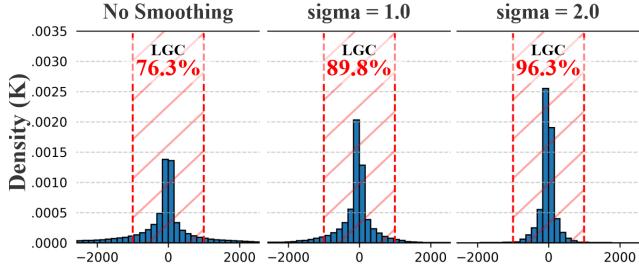


Figure 5. Left: Normalized histogram of the GC, in units of inverse of square meters, across the 15 training images from the Middlebury dataset. Middle and Right: smoothing by  $\sigma = 1.0\text{ m}, 2.0\text{ m}$  respectively. We discarded the highest 20% of  $|K|$  values, so we discard depth boundary data to focus on items, and the remaining  $K$  values range between  $[-32, 370.4; 14, 038.9]$ . For visualization we plot  $K$  values within  $[-2, 500; 2, 500]\text{ m}^{-2}$  in 30 bins uniformly distributed. Note that increased smoothing results in a higher LGC measure.

(see [12]. For completion we placed a proof of it in the supplementary material). Thus equation (5) is indeed the sparse loss function that aim to have as few components as possible of the principal curvatures.

A natural metric to evaluate the sparsity of the GC distribution in an image or dataset is the Shannon entropy of  $h(K)$ . Alternatively, we propose a simpler to compute and interpretable metric, which we call *Low Gaussian Curvature (LGC)* metric, where a low entropy should map to a high LGC. This metric is defined as the percentage of all computed GC values that lie within a specified range,  $[-W, W]$ . To avoid the impact of high values near surface boundaries (depth discontinuities), we remove the top 20% of the highest  $|K|$  values from the entire analysis. As shown in Figure 5, 76.3% of the GC values in the Middlebury dataset are concentrated in the range of  $W = 1,000\text{ m}^{-2}$ . After applying smoothing with  $\sigma = 2\text{ m}$  to the data, the LGC increases to 96.3% of values that fall within  $[-1,000; 1,000]\text{ m}^{-2}$ , while the maximum absolute curvature reaches  $|K|_{\max} = 32,370.4\text{ m}^{-2}$ .

## 5. An analysis on depth reconstruction

We design experiments to evaluate to what extent depth-reconstruction algorithms incorporate a sparse representation for the magnitude of the GC. These experiments are carried out on indoors scenes from both real-world data, using the Middlebury dataset and controlled synthetic environments from our created 3D synthetic scenes.

### 5.1. Middlebury Dataset

The dataset includes a training partition, where GT data is publicly available, and a test partition, where only some cal-

Metric	Definition
Average Absolute Error (AvgError)	$\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N  d_{ij} - \hat{d}_{ij} $
Root Mean Squared Error (RMS)	$\sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (d_{ij} - \hat{d}_{ij})^2}$
Bad-N (% of pixels with error > N)	$100 \times \frac{ \{(i, j) \mid  d_{ij} - \hat{d}_{ij}  > N\} }{ \{(i, j)\text{ valid pixels}\} }$
Average Absolute Gaussian Curvature (Avg K )	$\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N  K_{ij} $
Normal Error (NormalsErr)	$\frac{1}{ \Omega } \sum_{(i,j) \in \Omega} \left( 1 - \frac{\mathbf{n}_{ij}^{\text{gt}} \cdot \mathbf{n}_{ij}^{\text{pred}}}{\ \mathbf{n}_{ij}^{\text{gt}}\  \ \mathbf{n}_{ij}^{\text{pred}}\ } \right)$ where $\Omega$ is the set of valid pixels.
Low Gaussian Curvature (LGC)	$100 \times \frac{ \{(i, j) \mid -W \leq K_{ij} \leq W\} }{ \{(i, j)\text{ valid pixels}\} }$

Table 1. Summary of evaluation metrics adopted in this work, covering disparity and depth errors, curvature measurements, and surface-normal consistency.

ibration information and the left-right image pairs are provided. In both partitions, algorithm rankings are available; however, only in the training set can the disparity map results of submitted methods be downloaded for further analysis. We obtained the full-resolution results of SOTA techniques<sup>2</sup>, organized into two groups: **Group A**) FoundationStereo [27], LG-Stereo<sup>3</sup>, DEFOM-Stereo [14], MonoStereo [4], MonSter++ [3], and AIO-Stereo [31] which represent the new generation of stereo approaches that combine stereo matching with monocular features; and **Group B**) RAFT-Stereo [17], CREStereo [16], DLNR [29], Selective-IGEV [26], and S2M2 [20], corresponding to previous SOTA techniques. We also evaluated the results of BLMT-Stereo<sup>3</sup> and DepthFocus<sup>3</sup>, two new approaches currently under review for CVPR 2026. Although the papers and source codes were not available at the time of our submission, we were able to analyze their results using the GC evaluation.

The experiment on Figure 6 compares the disparity AvgError performance and the median  $|K|$  values of GC for all techniques. Note that the new SOTA methods (Group A) are the best-performing approaches in the Middlebury ranking and tend to estimate lower curvature magnitudes compared to the Group B methods. Table 1 presents a summary of the evaluation metrics.

A more detailed experiment can be seen on Figure 7 and Table 2, where we analyze the GC distribution for each

<sup>2</sup>There are other relevant methods reported in the literature; however, our selection was based on top-ranked methods on the Middlebury, KITTI, and ETH3D benchmarks and the availability of source code.

<sup>3</sup>Papers and code for LG-Stereo, BLMT-Stereo, and DepthFocus are not available yet.

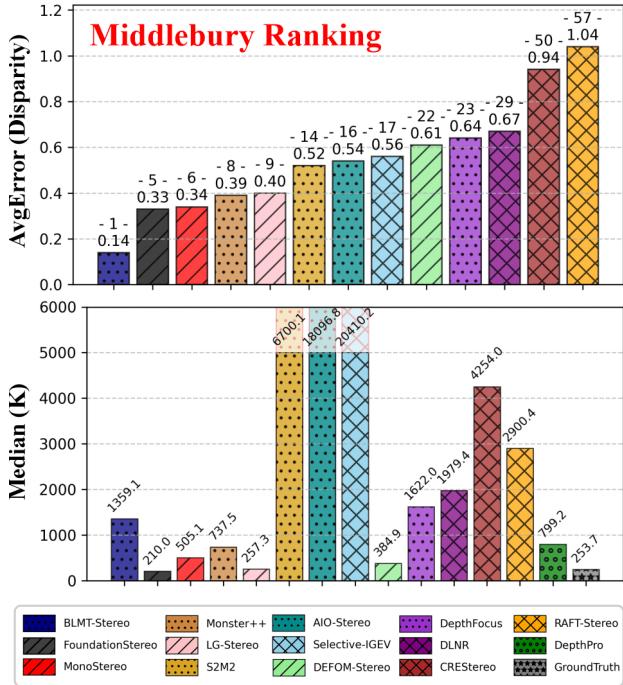


Figure 6. **Middlebury Dataset - Curvature x AvgError Analysis:** Top: shows the overall position of each approach in the Middlebury ranking, along with its average disparity error (AvgError). Bottom: presents the median of absolute GC for all techniques and the GT.

technique. Note that FoundationStereo has 81.3% of its GC between  $[-1,000, 1,000]m^{-2}$   $K$  values. Additionally, the same trend of minimizing GC across the 15 Middlebury training set was observed in the Top 5 best techniques, all of them from Group A<sup>4</sup>. Despite achieving low AvgError, techniques from Group B exhibit also lower LGC, with a tendency toward higher  $|K|$  values compared to those observed in the Top 5 Group A techniques.

Considering that humans easily perceive the shape of flat objects (*e.g.* walls, ground, doors, etc), in the following experiment we present a visual analysis of curvature, using the “Piano” data from Middlebury dataset. We compared the  $K$  values of GC from the Middlebury’s provided GT and the FoundationStereo reconstruction, which is the best (lowest AvgError) on the Middlebury ranking for this data with public code. We show the left image, and depth for GT and FoundationStereo in Figure 8, emphasizing the piano bench and sheet music, that we know are developable surfaces with  $K = 0$ . Despite good depth estimation, Middlebury’s GT provides a binary mask with measurement problems for each disparity map. Besides filtering the highest  $|K|$  values

<sup>4</sup>We invite the reader to refer to the Supplementary Material for a more comprehensive study.

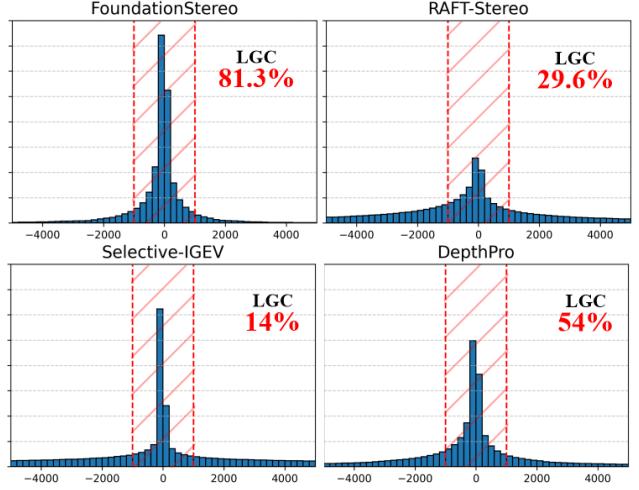


Figure 7. **Curvature Distribution:** each plot presents a normalized histogram of the GC distribution for all 15 training images from the Middlebury dataset. We discarded the highest 20% of  $|K|$  values to avoid depth discontinuities, and plotted the remaining  $K$  values within  $[-4,000, 4,000]m^{-2}$  in 50 bins uniformly distributed.

of the GT, previously we removed these NaN values from the GC analysis for GT. An example of these problematic data can be seen as black positions on the GT depth in Figure 8.

In Figure 9 the GT, FoundationStereo, and Selective-IGEV present a noisy representation of GC (first row), then we also present  $K$  values smoothed ( $\sigma = 2.0 m$ ). Also, the non-black coordinates represent  $|K|$  values lower than  $1,000m^{-2}$ .

Observe that edges throughout the image exhibit abrupt changes in curvature, while flat surfaces do not. Another interesting observation is that the non-smoothed reconstruction from FoundationStereo produces  $|K| > 1,000m^{-2}$  in some areas of the lampshade — a cylindrical surface that should have a constant  $K = 0$  — as well as higher  $K$  values on the music notes printed on the sheet music, which, as a flat piece of paper, should also have  $K = 0$ . On the other hand, the bottom part of the lampshade shows an approximation of positive (red)  $K$  values, which is expected given the spherical shape’s positive principal curvatures. Looking at the  $K$  values estimated by Selective-IGEV, it is evident that the output is noisier than those of other approaches. While FoundationStereo exhibited low GC on flat surfaces even in the original output, Selective-IGEV only revealed clearer structures after smoothing.

In Table 3, we extend our analysis to the normal error between the reconstructed point clouds from each technique and the GT across all 15 Middlebury training images. Notably, the top-performing (Top 5) methods in terms of Nor-

Technique	LGC	AvgError ↓	RMS ↓	Bad 2.0 ↓	Bad 4.0 ↓
FoundationStereo [27]	81.3	0.33 <sup>2</sup>	2.86 <sup>4</sup>	0.79 <sup>2</sup>	0.40 <sup>3</sup>
LG-Stereo	78.1	0.40 <sup>5</sup>	2.94 <sup>5</sup>	1.04 <sup>5</sup>	0.44 <sup>4</sup>
DEFOM-Stereo [14]	69.2	0.61 <sup>9</sup>	3.66 <sup>8</sup>	2.26 <sup>10</sup>	1.19 <sup>10</sup>
MonoStereo [4]	64.6	0.34 <sup>3</sup>	2.46 <sup>2</sup>	0.94 <sup>3</sup>	0.35 <sup>2</sup>
Monster++ [3]	56.5	0.39 <sup>4</sup>	2.54 <sup>3</sup>	1.17 <sup>6</sup>	0.47 <sup>5</sup>
BLMT-Stereo	43.1	0.14 <sup>1</sup>	1.33 <sup>1</sup>	0.17 <sup>1</sup>	0.09 <sup>1</sup>
DepthFocus	39.9	0.64 <sup>10</sup>	4.27 <sup>11</sup>	1.84 <sup>7</sup>	1.08 <sup>9</sup>
DLNR [29]	36.3	0.67 <sup>11</sup>	3.90 <sup>10</sup>	2.92 <sup>11</sup>	1.41 <sup>11</sup>
RAFT-Stereo [17]	29.6	1.04 <sup>13</sup>	5.25 <sup>13</sup>	5.25 <sup>13</sup>	2.89 <sup>13</sup>
CREStereo [16]	25.8	0.94 <sup>12</sup>	5.21 <sup>12</sup>	4.01 <sup>12</sup>	2.04 <sup>12</sup>
S2M2 [20]	17.5	0.52 <sup>6</sup>	3.73 <sup>9</sup>	1.00 <sup>4</sup>	0.59 <sup>6</sup>
AIO-Stereo [31]	15.3	0.54 <sup>7</sup>	3.50 <sup>6</sup>	1.97 <sup>8</sup>	0.82 <sup>7</sup>
Selective-IGEV [26]	14.0	0.56 <sup>8</sup>	3.57 <sup>7</sup>	2.18 <sup>9</sup>	0.92 <sup>8</sup>
Ground Truth	76.3	—	—	—	—

Table 2. This table presents the Middlebury benchmark ranking for the 15 training images, with techniques listed in descending LGC order. Superscripts indicate each method’s rank among the compared techniques for the metrics AvgError, RMS, Bad 2.0, and Bad 4.0. Darker cell shading highlights better performance, indicating the technique is among the Top 1, Top 3, or Top 5 best approaches. Notably, top-performing methods (Group A) generally exhibit higher LGC (i.e., lower GC).

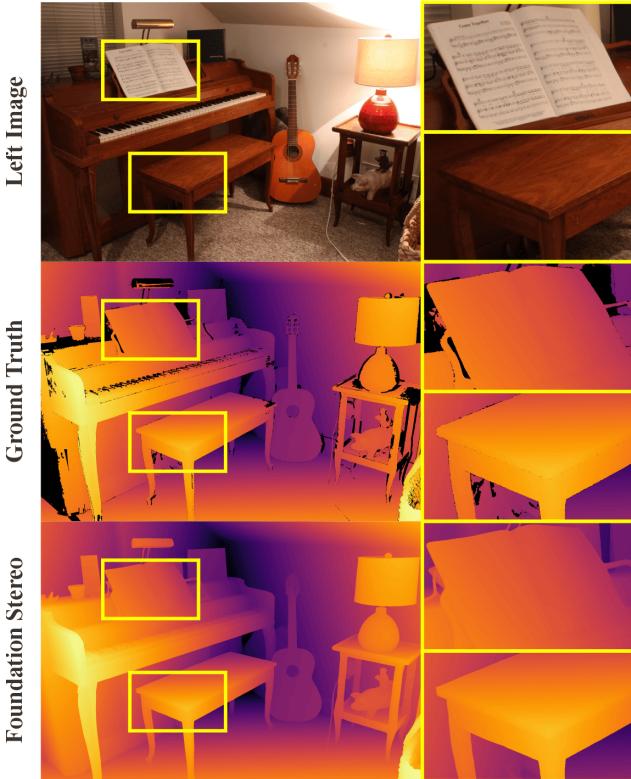


Figure 8. RGB left image x Depth.

mal Average Error (NormAvg) belong to Group A. Moreover, Selective-IGEV showed poor alignment of surface normals with the expected GT across all 15 images.

In Figures 10 and 11, we qualitatively present the normal reconstructions for the GT, BLMT-Stereo, Foundation-

Stereo, and Selective-IGEV on the Piano and Playable data, respectively. BLMT-Stereo and Foundation-Stereo produce depth maps with more coherent surface normals compared to Selective-IGEV. While BLMT-Stereo aligns normals more accurately in thin structures and near depth discontinuities, Foundation-Stereo achieves better consistency in regions with low absolute Gaussian curvature. In both cases, Selective-IGEV struggles, producing surfaces with high absolute GC and poorer normal alignment with the GT.

In Figure 12, we also present additional qualitative and quantitative results for the Adirondack data. Although both Foundation-Stereo and Selective-IGEV achieve accurate depth reconstruction (low AvgError), their Gaussian curvature and normal consistency differ substantially. Foundation-Stereo exhibits lower Avg|K| and lower NormalsErr, which may account for its overall better performance.

Our experiments demonstrate that the new SOTA approaches (Group A) are able to estimate more concise structures during disparity/depth reconstruction, which means the global structure of a surface is kept. Unlike new SOTA methods, Group B techniques do not preserve global surface structures, despite high accuracy in AvgError, which suggests that unifying stereo and monocular features is crucial to preserve 3D data relationships.

## 5.2. 3D Synthetic Scenes

In our 3D synthetic scenes, the data were designed to make the GC analysis easier by using simple objects with well-known curvatures. We ensured that planar surfaces – such as boxes, the ground, and walls – have  $K = 0$  in all scenes by generating the GT depth map through orthogonal projection of each surface point onto the left camera’s image

Technique	NormAvg ↓	Adirondack	ArtL	Jadeplant	Motorcycle	MotorcycleE	Piano	PianoL	Pipes	Playroom	Playable	PlayableP	Recycle	Shelves	Teddy	Vintage
BLMT-Stereo	6.44	2.09 <sup>1</sup>	18.08 <sup>4</sup>	4.90 <sup>3</sup>	6.31 <sup>1</sup>	6.30 <sup>1</sup>	2.93 <sup>1</sup>	3.01 <sup>1</sup>	8.36 <sup>1</sup>	3.81 <sup>1</sup>	2.60 <sup>1</sup>	2.98 <sup>1</sup>	1.86 <sup>1</sup>	4.40 <sup>1</sup>	25.92 <sup>2</sup>	3.08 <sup>1</sup>
FoundationStereo	6.75	2.24 <sup>2</sup>	15.92 <sup>1</sup>	4.84 <sup>2</sup>	7.88 <sup>4</sup>	7.84 <sup>3</sup>	3.70 <sup>3</sup>	3.97 <sup>3</sup>	9.04 <sup>3</sup>	4.91 <sup>3</sup>	3.09 <sup>3</sup>	3.52 <sup>4</sup>	2.74 <sup>5</sup>	5.03 <sup>3</sup>	23.23 <sup>1</sup>	3.37 <sup>3</sup>
LG-Stereo	6.78	2.09 <sup>1</sup>	17.97 <sup>3</sup>	4.91 <sup>4</sup>	7.00 <sup>2</sup>	7.05 <sup>2</sup>	3.36 <sup>2</sup>	3.68 <sup>2</sup>	8.46 <sup>2</sup>	4.81 <sup>2</sup>	2.92 <sup>2</sup>	3.24 <sup>2</sup>	2.05 <sup>2</sup>	4.90 <sup>2</sup>	25.98 <sup>3</sup>	3.24 <sup>2</sup>
MonoStereo	7.29	2.32 <sup>3</sup>	17.75 <sup>2</sup>	4.60 <sup>1</sup>	7.86 <sup>3</sup>	8.32 <sup>4</sup>	3.88 <sup>4</sup>	4.63 <sup>4</sup>	9.92 <sup>4</sup>	5.24 <sup>4</sup>	3.14 <sup>4</sup>	3.36 <sup>3</sup>	2.36 <sup>3</sup>	6.33 <sup>4</sup>	26.02 <sup>4</sup>	3.64 <sup>4</sup>
Monster++	8.13	2.70 <sup>4</sup>	18.26 <sup>5</sup>	5.05 <sup>5</sup>	8.62	8.38 <sup>5</sup>	4.81	5.86 <sup>5</sup>	11.43	6.04 <sup>5</sup>	4.05	4.00 <sup>5</sup>	2.76	9.57	26.26 <sup>5</sup>	4.19
DEFOM-Stereo	8.18	2.77 <sup>5</sup>	19.18	6.82	8.56 <sup>5</sup>	8.70	4.71	6.18	11.31 <sup>5</sup>	6.53	3.90 <sup>5</sup>	4.15	2.68 <sup>4</sup>	7.16 <sup>5</sup>	26.43	3.66 <sup>5</sup>
DepthFocus	8.93	4.24	19.38	10.49	8.79	8.83	4.61 <sup>5</sup>	7.85	12.05	6.97	4.01	4.17	3.43	7.35	27.34	4.39
DLNR	9.05	3.50	19.76	7.20	9.40	9.50	6.22	8.88	11.60	6.92	4.20	4.47	3.36	8.68	26.65	5.42
S2M2	9.82	5.51	19.50	6.40	10.63	10.59	5.99	6.07	15.46	8.02	5.82	5.94	4.58	8.54	28.00	6.20
RAFT-Stereo	11.11	4.94	21.57	9.57	11.44	11.55	7.77	11.84	14.67	8.45	5.90	6.16	4.38	12.06	27.08	9.33
CREStereo	11.61	5.69	20.46	10.30	12.55	12.52	9.73	12.94	15.87	9.78	7.80	7.84	5.19	10.17	26.95	6.43
AIO-Stereo	16.65	10.03	22.05	12.28	15.49	15.55	14.60	16.45	23.45	13.21	14.20	14.81	10.46	17.65	28.48	20.99
Selective-IGEV	17.90	10.80	22.62	13.61	17.23	17.28	16.88	18.70	24.09	14.31	15.87	16.44	11.57	18.15	28.98	22.00
DepthPro	19.86	10.74	42.77	31.17	23.37	23.37	11.89	11.89	31.91	18.42	8.87	9.17	9.62	14.88	42.35	7.46

Table 3. This table presents techniques listed in ascending Normal Average Error (NormAvg) order for the 15 Middlebury training images. Superscripts indicate the lowest 1-5th Normal Error (NormalsErr) per technique for each one of the 15 Middlebury images. Darker cell shading highlights lower NormalsErr, indicating the technique is among the Top 1, Top 3, or Top 5 approaches with the lowest NormalsErr. Notably, the top-performing methods are from the Group A, and predict surface normals that are closely aligned with the ground truth.

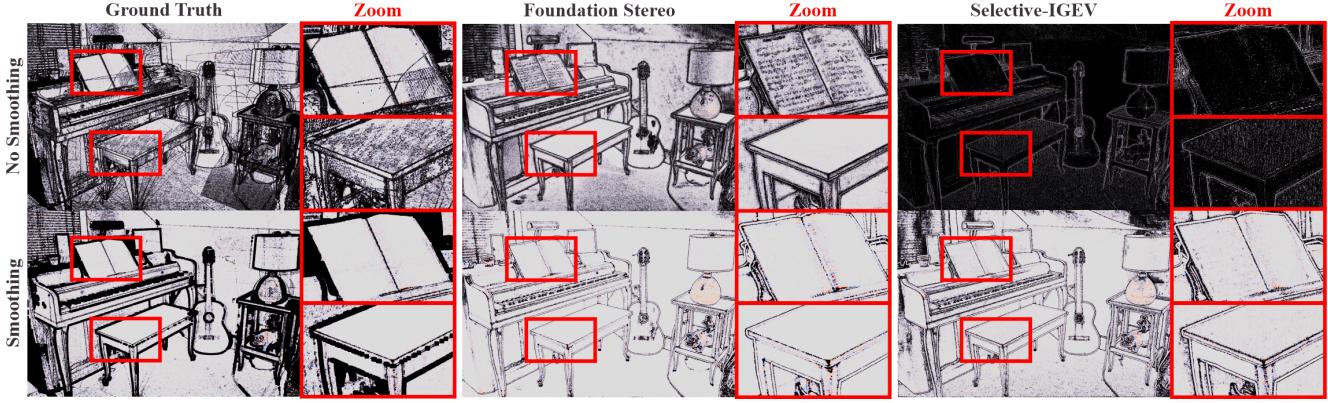


Figure 9. **GT x SOTA approaches:** a point-wise analysis of curvature for "Piano" image. Black coordinates represent values of  $|K| > 1,000m^{-2}$ . For the GT, black coordinates also represent NaN values, which are measurement inconsistencies during Middlebury disparity estimation. In the second row, we applied smoothing with  $\sigma = 2 m$  in the 3D point cloud before computing the GC.

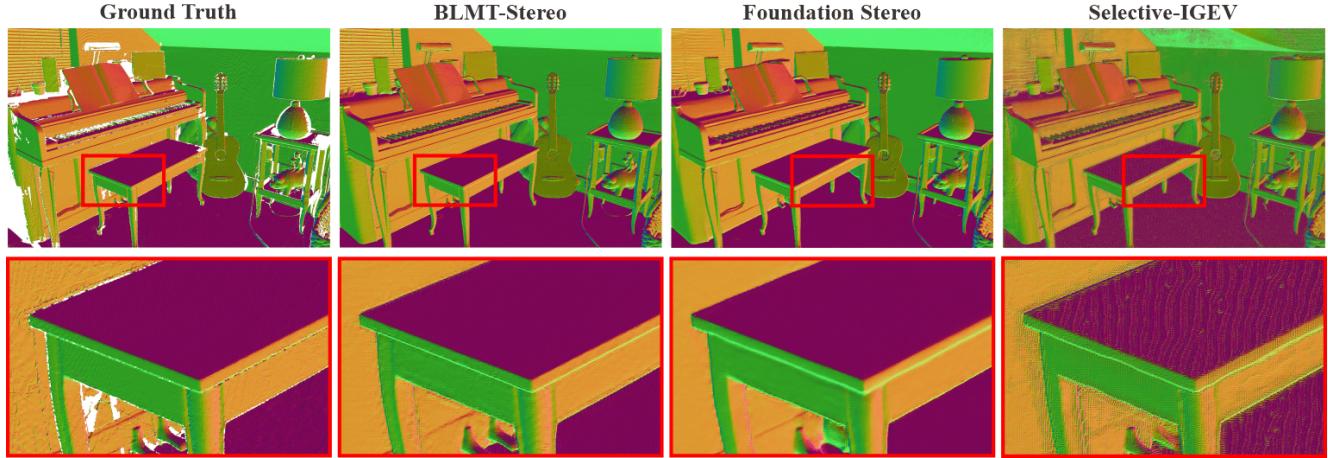


Figure 10. **Normals Analysis:** Qualitative comparison on normals reconstruction for Piano data.

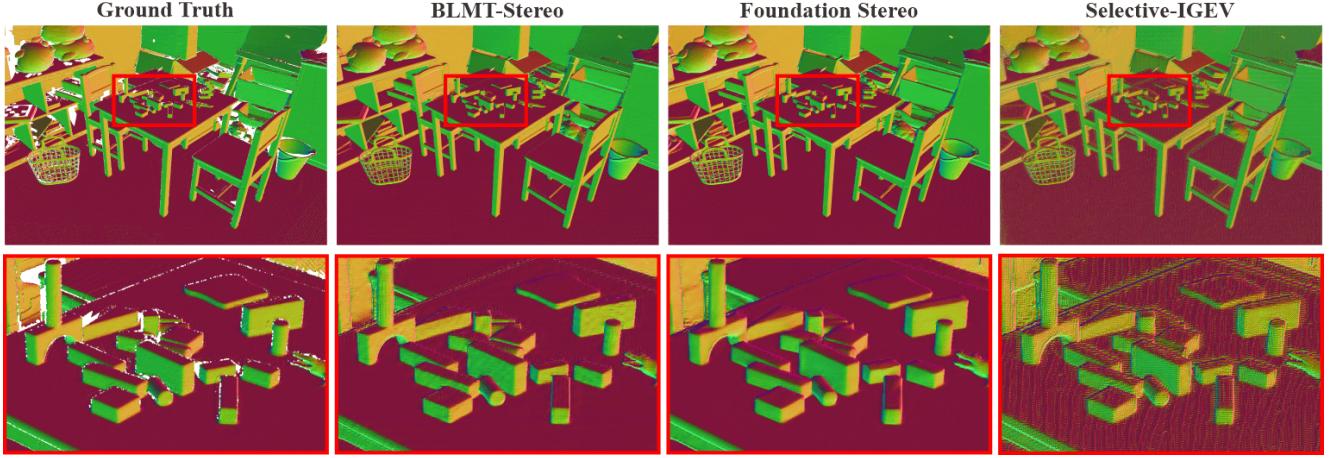


Figure 11. **Normals Analysis:** Qualitative comparison on normals reconstruction for Playable data.

plane, with a (2000, 3000)px resolution.

For the following experiments, we obtained the source code for: FoundationStereo, RAFT-Stereo, and Selective-IGEV, representing the aforementioned stereo approaches; and for DepthAnythingV2 [28] and DepthPro [2], representing the SOTA in Monocular Depth Estimation (MDE).

[Figure 13](#) presents a quantitative analysis of GC and average depth error across the five scenes in the 3D synthetic scenes. The left-most bar in each graph represents the curvature of the GT, while the other bars show the curvature estimated by stereo and monocular methods. The right-most plot for each scene shows the average depth error (AvgError) in centimeters. Notably, FoundationStereo, which achieves the lowest average error – less than one centimeter in all scenes – also exhibits the lowest GC. See the depth and curvature results for each one of them in [Figure 14](#).

We do not report average error for DepthAnythingV2 and DepthPro, as these methods provide only relative depth. Although DepthPro estimates depth in meters and predicts the focal length from an image, our investigation revealed that its depth predictions face real-scale limitations, which is expected for monocular methods. However, to the best of our knowledge, DepthPro provides the most detailed surface estimation among current MDE approaches. While DepthPro tends to minimize GC on the Middlebury dataset (see Figures 6 and 7), both DepthPro and DepthAnythingV2 reconstruct scenes with higher curvatures in our 3D synthetic scenes (see [Figure 13](#)). To visualize the results, [Figure 15](#) shows the 3D surface and GC estimated from SOTA MDE approaches, like DepthPro and DepthAnythingV2.

## 6. Limitations and Future Directions

This paper highlights Gaussian curvature as a powerful measure for analyzing 3D scenes, given its invariance to viewpoint changes and sparsity in man-made environments. However, our analysis is limited to the 15 training images from the Middlebury dataset, as it is the only benchmark that publicly provides the techniques’ results. Although we did not directly evaluate other benchmarks, the strong performance of BLMT-Stereo as the top method on Middlebury [22] and ETH3D [23], and of MonSter++ as the leading method on KITTI 2012 [9] and KITTI 2015 [19], suggests that our findings may generalize to outdoor scenes. As future work, we plan to extend our investigation to additional datasets.

In this paper, we focused on *analyzing and understanding* the role of Gaussian curvature in 3D reconstruction using SOTA stereo-vision techniques. We did not aim to develop a new algorithm at this stage. As future work, we plan to incorporate low absolute Gaussian curvature as an unsupervised metric for stereo 3D reconstruction, which also includes treating depth discontinuities, occlusions, repetitive patterns, and textureless regions.

## 7. Conclusion

Gaussian curvature (GC) plays a significant role in the reconstruction of 3D surfaces. We showed that the histogram of GC reveals a sparse distribution, suggesting that GC is a compact descriptor of surface geometry. Furthermore, we proposed that the square root of the GC magnitude can serve as a sparse loss function consistent with the observed normalized histogram distribution. In addition, we introduced a simple and efficient metric, termed Low Gaussian Curvature (LGC), which can be used as a proxy for the inverse of the Shannon entropy of the normalized histogram distribu-

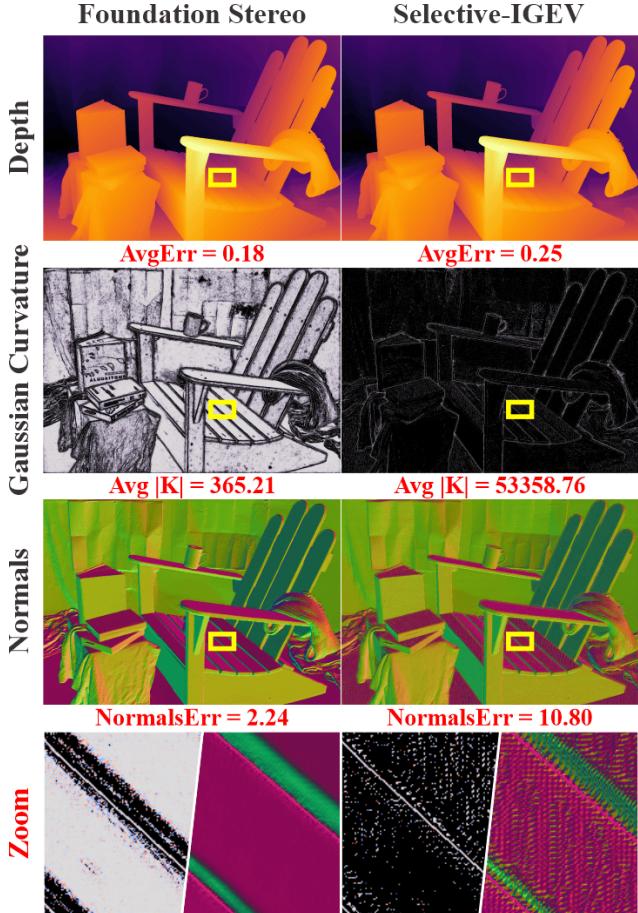


Figure 12. **Quantitative vs. Qualitative comparison:** We present the average disparity error (AvgError), average absolute Gaussian curvature ( $\text{Avg}|K|$ ), and normal error (NormalsErr) for the Foundation-Stereo and Selective-IGEV techniques on the Adirondack data. We also include a zoomed-in region to illustrate the pointwise consistency of Gaussian curvature and surface normals. Note that despite its low AvgError, the reconstruction produced by Selective-IGEV lacks smoothness.

tion, as higher LGC imply low entropy for the histogram.

Our empirical evaluation of state-of-the-art methods on the Middlebury benchmark revealed that the GC normalized histograms generated by these methods approximate the high LGC values of the ground truth GC distribution both on the Middlebury dataset and on our 3D synthetic scenes. This observation suggests that very likely state-of-the-art methods do incorporate a prior that minimizes GC values in regions where data matching is not sufficient to infer a solution. However, one does not know where in the algorithm such a prior occurs nor, in analogy to modules that perform feature extraction, how to extract a module with such a prior to be used in other applications.

In summary, our work enhances the *interpretability* and

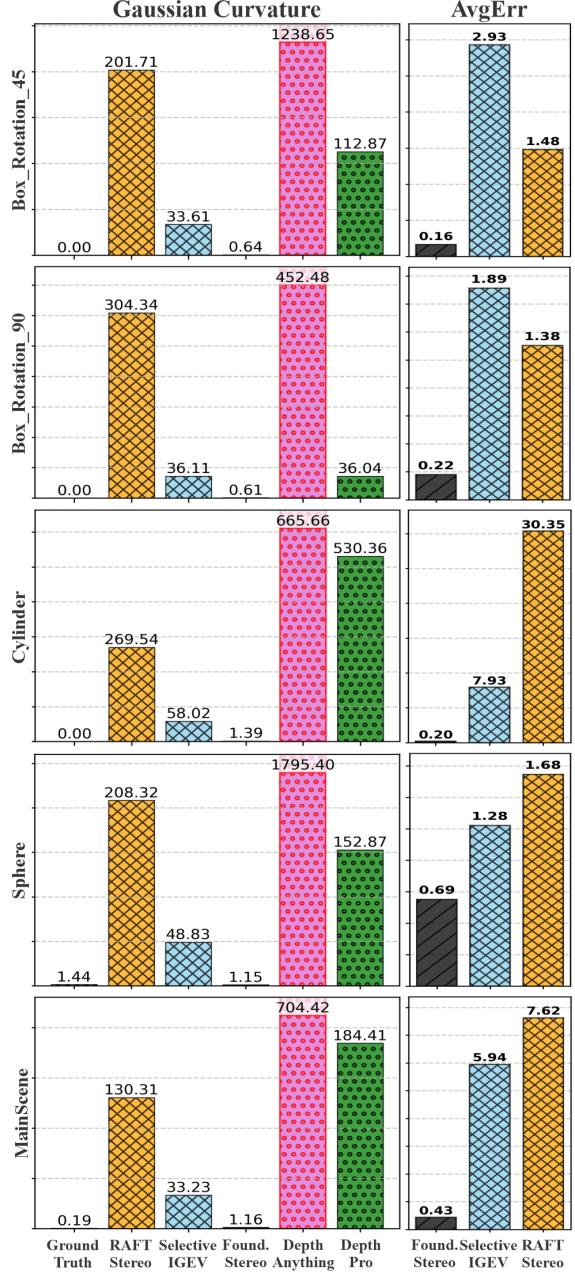


Figure 13. This figure presents an quantitative analysis on the average  $|K|$  values of GC (left) and the depth (cm) AvgError performance (right). Notice that stereo approaches with the lowest GC values achieve the lowest AvgError. We do not measure AvgError for monocular approaches since they usually provide relative depth.

*understanding* of 3D vision by highlighting Gaussian Curvature as an intrinsic geometric prior for indoor 3D surfaces. Grounded in modern deep learning data, our approach underscores the importance of 3D geometric modeling in cap-

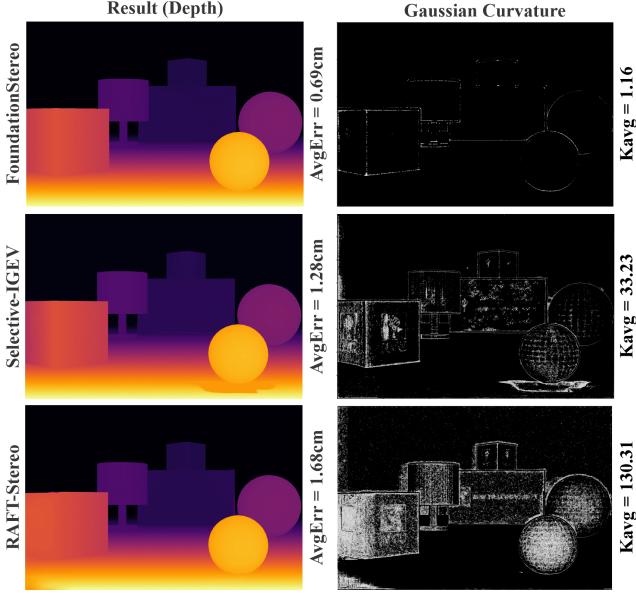


Figure 14. This figure visually demonstrates the superiority of FoundationStereo in reconstructing the 3D scene with depth error inferior to 1cm in average, and low GC, preserving 3D geometry more consistently than other stereo techniques.

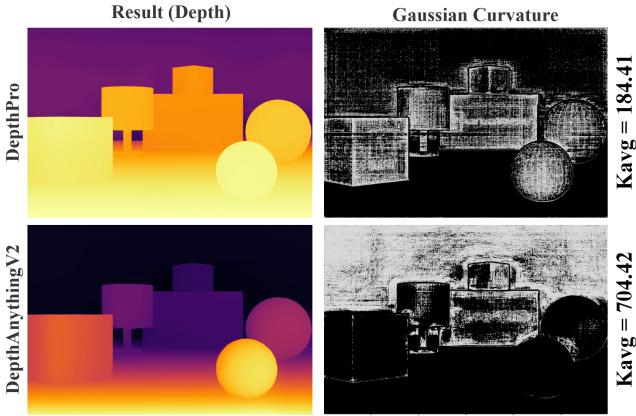


Figure 15. This figure shows the 3D surfaces obtained from the SOTA MDE approaches, DepthPro and DepthAnythingV2, for the left view of the Main Scene. Observe that meaningful 3D patterns are extracted, however depth are relative and curvature presents minor inconsistencies.

turing critical visual information and can guide the development of next-generation vision systems.

As a possible immediate consequence of this study, LGC could be used as a quality measure in multiple 3D reconstruction modalities, including stereo-vision, monocular depth estimation, and, by inference, structure from motion.

## Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

## References

- [1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. In *In Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 383–390 vol.2, 2001. 2
- [2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 9
- [3] Junda Cheng, Wenjing Liao, Zhipeng Cai, Longliang Liu, Gangwei Xu, Xianqi Wang, Yuzhou Wang, Zikang Yuan, Yong Deng, Jinliang Zang, Yangyang Shi, Jinhui Tang, and Xin Yang. Monster++: Unified stereo matching, multi-view stereo, and real-time stereo with monodepth priors, 2025. 5, 7
- [4] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 5, 7
- [5] Field D. What is the goal of sensory coding. *Neural Comp.*, 6:559–601, 1994. 2
- [6] M.P. do Carmo. *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Dover Publications, 2016. 2, 1
- [7] Qiujie Dong, Rui Xu, Pengfei Wang, Shuangmin Chen, Shiqing Xin, Xiaohong Jia, Wenping Wang, and Changhe Tu. Neurcadrecon: Neural representation for reconstructing cad surfaces by enforcing zero gaussian curvature. *ACM Trans. Graph.*, 43(4), 2024. 2
- [8] R. Epstein, P.W. Hallinan, and A.L. Yuille. 5 plus or minus 2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. In *In Proceedings of IEEE Workshop on Physics-Based Modeling in Computer Vision*, pages 108–116, 1995. 2
- [9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 9
- [10] D. Geiger, B. Ladendorf, and A. L. Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14(3):211–226, 1995. 2
- [11] Kehua Guo. 3d shape representation using gaussian curvature co-occurrence matrix. In *Artificial Intelligence and Computational Intelligence: International Conference, AICI 2010, Sanya, China, October 23–24, 2010, Proceedings, Part I 2*, pages 373–380. Springer, 2010. 2

- [12] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, Cambridge, UK, 1952. 5, 1
- [13] Hiroshi Ishikawa and Davi Geiger. Illusory volumes in human stereo perception. *Vision research*, 46(1-2):171–178, 2006. 2
- [14] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 5, 7
- [15] Jan J. Koenderink. *Solid shape*. MIT Press, Cambridge, MA, USA, 1990. 2
- [16] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 5, 7
- [17] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 5, 7
- [18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *In Proceedings IEEE International Conference on Computer Vision. ICCV 2009*, 2009. 2
- [19] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 9
- [20] Junhong Min, Youngpil Jeon, Jimin Kim, and Minyong Choi. S2m2: Scalable stereo matching model for reliable depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26729–26739, 2025. 5, 7
- [21] B. A. Olshausen and D. J. Field. Sparse coding with an over-complete basis set:a strategy employed by v1? *Vis. Res.*, 37: 3311–3325, 1997. 2
- [22] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 4, 9
- [23] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 9
- [24] Evangelos Ververas, Rolando Alexandros Potamias, Jifei Song, Jiankang Deng, and Stefanos Zafeiriou. Sags: structure-aware 3d gaussian splatting. In *European Conference on Computer Vision*, pages 221–238. Springer, 2024. 2
- [25] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2
- [26] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 5, 7
- [27] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 5, 7
- [28] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 9
- [29] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1327–1336, 2023. 5, 7
- [30] Ming Zhong and Hong Qin. Sparse approximation of 3d shapes via spectral graph wavelets. *The Visual Computer*, 30:751–761, 2014. 2
- [31] Jingyi Zhou, Haoyu Zhang, Jiakang Yuan, Peng Ye, Tao Chen, Hao Jiang, Meiya Chen, and Yangyang Zhang. All-in-one: Transferring vision foundation models into stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10797–10805, 2025. 5, 7
- [32] Linglong Zhou, Guoxin Wu, Yunbo Zuo, Xuanyu Chen, and Hongle Hu. A comprehensive review of vision-based 3d reconstruction methods. *Sensors*, 24(7):2314, 2024. 2
- [33] S-C. Zhu, W-Y. Nian, and D. Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27:107–126, 1998. 2

# Towards Understanding 3D Vision: the Role of Gaussian Curvature

## Supplementary Material

### 8. Introduction

This supplementary material provides a deeper exploration of the concepts, methods, and results introduced in the main paper. While the main text presents a concise overview of our findings, certain theoretical insights and experimental details require further elaboration to fully support our claims and offer transparency in our methodology. This document is intended to complement the main paper by offering readers a more comprehensive understanding of the mathematical foundations and empirical behavior of state-of-the-art (SOTA) techniques with respect to Gaussian Curvature analysis in depth estimation.

In section 9, we describe the methods used to estimate Gaussian Curvature from discrete depth data and provide a formal derivation linking curvature minimization to the  $L^0$  loss. This section is essential for grounding our curvature analysis in a solid mathematical framework. Then, in section 10, we expand our evaluation of SOTA techniques by presenting additional experimental results on both the Middlebury and our 3D synthetic scenes. These extended results not only reinforce the findings of the main paper but also reveal deeper patterns and behaviors that are critical for interpreting the performance of modern depth estimation methods.

### 9. Geometrical Supplemental Material

The material below are known geometrical properties (see [6, 12]) that we present here to "refresh" the interested reviewer, just in case. First we describe the two fundamental forms used to compute the Gaussian curvature (GC) and then we prove the sparsity result associated with the GC measure.

#### 9.1. Methods to estimate Gaussian curvature from data

For completion, we expand here the Gaussian curvature formula:

$$K = \frac{\det(\mathbb{I})}{\det(\mathbb{I})} = \frac{LN - M^2}{EG - F^2}. \quad (7)$$

where the fundamental forms I and II can be obtained as follows

**The first fundamental form I:** Let  $\mathbf{Z}(u, v)$  be a parametric surface and let  $\mathbf{Z}_u(u, v), \mathbf{Z}_v(u, v)$  denote the partial

derivatives that are independent tangent vectors to the surface. Then the inner product of two tangent vectors is

$$\mathbb{I}(a\mathbf{Z}_u + b\mathbf{Z}_v, c\mathbf{Z}_u + d\mathbf{Z}_v) \quad (8)$$

$$= ac\langle \mathbf{Z}_u, \mathbf{Z}_u \rangle + (ad + bc)\langle \mathbf{Z}_u, \mathbf{Z}_v \rangle + bd\langle \mathbf{Z}_v, \mathbf{Z}_v \rangle \\ = (a \ b) \begin{pmatrix} E & F \\ F & G \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix}, \quad (9)$$

where E, F, and G are the coefficients of the first fundamental form.

**The second fundamental form II:** The vector normal to the surface is given by

$$\mathbf{n} = \frac{\mathbf{Z}_u(u, v) \times \mathbf{Z}_v(u, v)}{|\mathbf{Z}_u(u, v) \times \mathbf{Z}_v(u, v)|}. \quad (10)$$

The second fundamental form is then written as

$$\mathbb{II} = (du \ dv) \begin{pmatrix} L & M \\ M & N \end{pmatrix} \begin{pmatrix} du \\ dv \end{pmatrix}, \quad (11)$$

where

$$L = \mathbf{Z}_{uu} \cdot \mathbf{n}, \quad M = \mathbf{Z}_{uv} \cdot \mathbf{n}, \quad N = \mathbf{Z}_{vv} \cdot \mathbf{n}. \quad (12)$$

#### 9.2. Proof of GC minimization association with $L^0$ Loss

The result of Equation 6 is known and quite important and so we repeat here in a form of a known theorem  $\sqrt{|\kappa_1 \kappa_2|} = \lim_{p \rightarrow 0} \left( \frac{1}{2}(|\kappa_1|^p + |\kappa_2|^p) \right)^{\frac{1}{p}}$ . We prove this theorem by breaking it into three lemmas before the final proof.

**Lemma 1.** For  $r > 1$ ,  $\left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^r \leq \frac{|\kappa_1|^{pr} + |\kappa_2|^{pr}}{2}$ ,

*Proof.* Define  $a_1 = |\kappa_1|^p \geq 0$  and  $a_2 = |\kappa_2|^p \geq 0$ , and since  $r \geq 1$ , then  $(a_1 + a_2)^r$  is convex. From Jensen inequality we then have  $\left( \frac{a_1 + a_2}{2} \right)^r \leq \frac{1}{2}(a_1^r + a_2^r)$ . Replacing back  $a_1 = |\kappa_1|^p$  and  $a_2 = |\kappa_2|^p$  completes the proof.  $\square$

**Lemma 2.** For  $1 \geq q \geq p \geq 0$ ,  $\left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}} \leq \left( \frac{|\kappa_1|^q + |\kappa_2|^q}{2} \right)^{\frac{1}{q}}$ ,

*Proof.* Replacing  $r$  by  $\frac{q}{p} \geq 1$  in lemma 1  $\left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{q}{p}} \leq \frac{|\kappa_1|^q + |\kappa_2|^q}{2}$ . Then taking the  $q$  root on both sides does not change the order of the inequality and completes the proof.  $\square$

**Lemma 3.** For  $1 \geq p \geq 0$ ,

$$\frac{\log |\kappa_1| + \log |\kappa_2|}{2} \leq \log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}}, \quad (13)$$

and thus  $\left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}}$  is bounded from below.

*Proof.* The log function is a concave function. Thus,

$$\frac{\log |\kappa_1|^p + \log |\kappa_2|^p}{2} \leq \log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right) \quad (14)$$

follows from Jensen's inequality. Since  $\log |\kappa_1|^p + \log |\kappa_2|^p = p(\log |\kappa_1| + \log |\kappa_2|)$ , then back to the Jensen's inequality  $p \left( \frac{\log |\kappa_1| + \log |\kappa_2|}{2} \right) \leq \log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)$  and so  $\left( \frac{\log |\kappa_1| + \log |\kappa_2|}{2} \right) \leq \frac{1}{p} \log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)$  completes the proof.  $\square$

From lemma (2) and lemma (3), it follows that  $\left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}}$  decreases as  $p$  decreases and it is bounded from below. Therefore, it converges as  $p \rightarrow 0$ .

**Theorem 1.**

$$\lim_{p \rightarrow 0} \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}} = e^{\frac{1}{2} \log(|\kappa_1| |\kappa_2|)}, \quad (15)$$

*Proof.* We use a known inequality that for  $0 \leq p \leq 1$ ,  $\log x \leq \frac{1}{p}(x^p - 1)$ . Thus, for  $x = \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}}$  we have

$$\log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}} \leq \frac{1}{p} \left( \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right) - 1 \right) \quad (16)$$

$$= \frac{1}{2} \left( \frac{1}{p} (|\kappa_1|^p - 1) + \frac{1}{p} (|\kappa_2|^p - 1) \right) \quad (17)$$

Taking the limit  $p \rightarrow 0$  and using the equality  $\log x = \lim_{p \rightarrow 0} \frac{1}{p}(x^p - 1)$  (use L'Hôpital rule to check) we finally obtain

$$\lim_{p \rightarrow 0} \log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}} \leq \frac{\log |\kappa_1| + \log |\kappa_2|}{2} \quad (18)$$

Thus the  $\lim_{p \rightarrow 0} \log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}}$  is bounded by  $\frac{\log |\kappa_1| + \log |\kappa_2|}{2}$  from above and, by lemma 3, bounded from below by the same quantity and so

$$\lim_{p \rightarrow 0} \log \left( \frac{|\kappa_1|^p + |\kappa_2|^p}{2} \right)^{\frac{1}{p}} = \frac{\log |\kappa_1| + \log |\kappa_2|}{2}. \quad (19)$$

Taking the exponential from both sides completes the proof.  $\square$

## 10. In Depth Analysis on the SOTA approaches

In this section we present in details more experiments we have conducted for the GC analysis and understanding. In the following subsections we discuss the results for Middlebury Dataset, and for our 3D synthetic scenes, respectively.

### 10.1. Middlebury Dataset

In the main paper, we presented a normalized histogram distribution in Figure 7, which shows the LGC metric for FoundationStereo, DepthPro, RAFT-Stereo, Selective-IGEV, and in Figure 5 for the ground truth (GT). We also present in Table 2 a ranking that includes several benchmarking metrics alongside LGC. We selected these techniques for inclusion in the main text because we had already tested their code on our 3D synthetic scenes. Additionally, they represent key categories: FoundationStereo belongs to Group A (new SOTA), RAFT-Stereo and Selective-IGEV to Group B (previous SOTA), and DepthPro is the best-performing monocular depth estimation (MDE) method in terms of depth reconstruction.

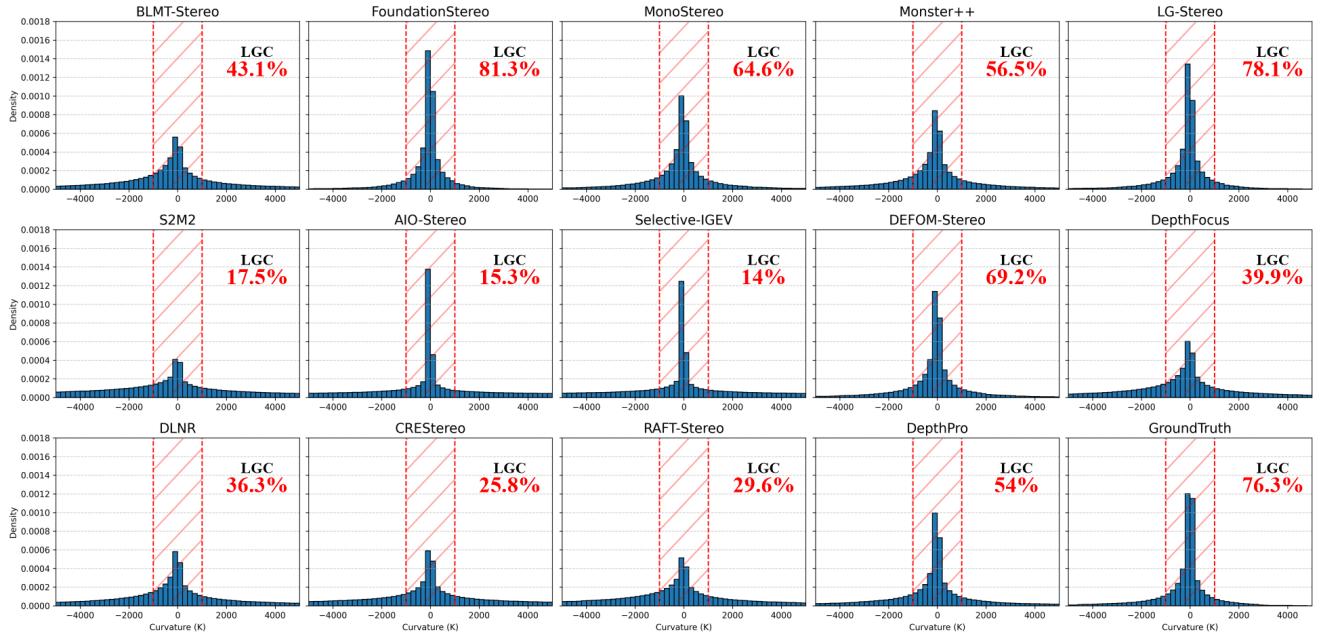
Here, in the supplementary material (see Figure 16), we take advantage of the additional space to extend our analysis to all evaluated techniques. We emphasize our earlier observation: Group A approaches not only achieve the lowest average disparity errors in the Middlebury ranking but also tend to minimize Gaussian Curvature—thus maximizing the LGC metric. Notably, FoundationStereo, MonoStereo, LG-Stereo, and DEFOM-Stereo exhibit LGC values above 65%, while Selective-IGEV, RAFT-Stereo, DLNR, and CREStereo show LGC values below 36%.

We also present a detailed analysis of the Middlebury ranking and Gaussian Curvature for each of the 15 training images. While the main paper shows the average disparity error aggregated across all 15 images, Figure 17 provides a per-image breakdown, showing each method's ranking position alongside its average disparity error for each individual image. In Figure 18, we show the average Gaussian Curvature for each of the 15 training images.

It can be observed in Figure 17 that the Group A approaches consistently achieve the lowest average disparity error across all images. Furthermore, Figure 18 shows that the Gaussian Curvature values produced by Group A approaches tend to have lower absolute magnitudes ( $|K|$ ) compared to those of Group B approaches.

### 10.2. 3D Synthetic Scenes

In the following experiments, we analyze quantitatively and qualitatively the depth and curvature results for each of the 5 scenes: Box\_Rotation\_45, Box\_Rotation\_90, Cylinder, Sphere, and MainScene. As already mentioned, we obtained the source code of: Group A) FoundationStereo;



**Figure 16. Curvature Distribution:** each plot presents a normalized histogram of the GC distribution for all 15 training images from the Middlebury dataset. We discarded the highest 20% of  $|K|$  values, and plotted the remaining  $K$  values within  $[-5,000, 5,000] \text{m}^{-2}$  in 50 bins uniformly distributed.

Group B) Selective-IGEV, and RAFT-Stereo; and MDE) DepthPro, and DepthAnythingV2.

The results of each aforementioned approach are shown in Figures 19, 20, 21, 22, and 23, respectively. For the Group A and Group B approaches, which estimate the disparity of a scene, we computed depth using the equation  $\text{Depth} = \frac{f \cdot b}{d}$ , where  $f$  is the focal length in pixels,  $b$  is the baseline in meters, and  $d$  is the disparity value in pixels. Note that no pixel offset correction is needed ( $\text{dof}_{fs} = 0$ ), as our simulated environment ensures that the left and right images are rectified. The resulting depth for each technique is shown in the first column of its respective figure.

For the Group A and Group B approaches, we also computed the difference between each method’s estimated depth, denoted as  $D_{\text{technique}}$ , and the expected ground truth ( $GT$ ). The second column in each technique’s figure presents a visual comparison of the difference  $GT - D_{\text{technique}}$  for each scene in the 3D synthetic scenes. Since depth values are strictly positive, this difference can yield both positive (red) and negative (blue) values. A positive value ( $GT - D_{\text{technique}} > 0$ , shown in red) indicates that the method predicted a depth further ahead than it should have. Conversely, a negative value ( $GT - D_{\text{technique}} < 0$ , shown in blue) indicates that the method predicted a depth further back than the correct position.

For all techniques, we computed the Gaussian Curvature, which is shown in the third column of each figure.

To facilitate the analysis, we masked regions where  $|K| <$  threshold (with a threshold of  $1000 \text{ m}^{-2}$ ), displaying them in black. This masking step allows us to focus on regions with higher curvatures and better understand the behavior of each method.

All relevant insights and observations are described in the figure’s captions to support and streamline the reading process. We also omitted some parts of the figures for Blind Review.

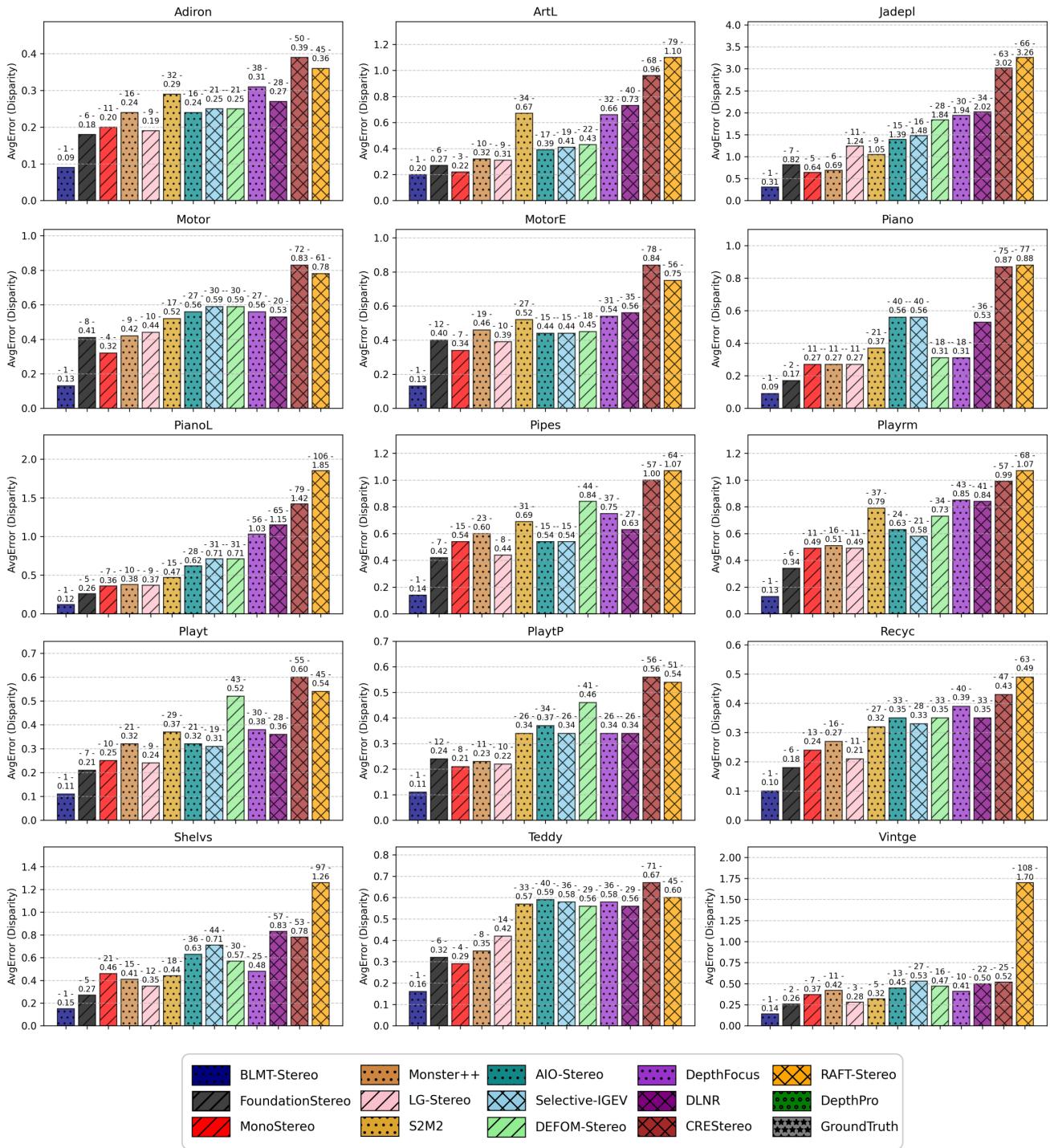


Figure 17. **Middlebury Ranking:** Average disparity error (AvgError) per image (see subsection 10.1 for more information).

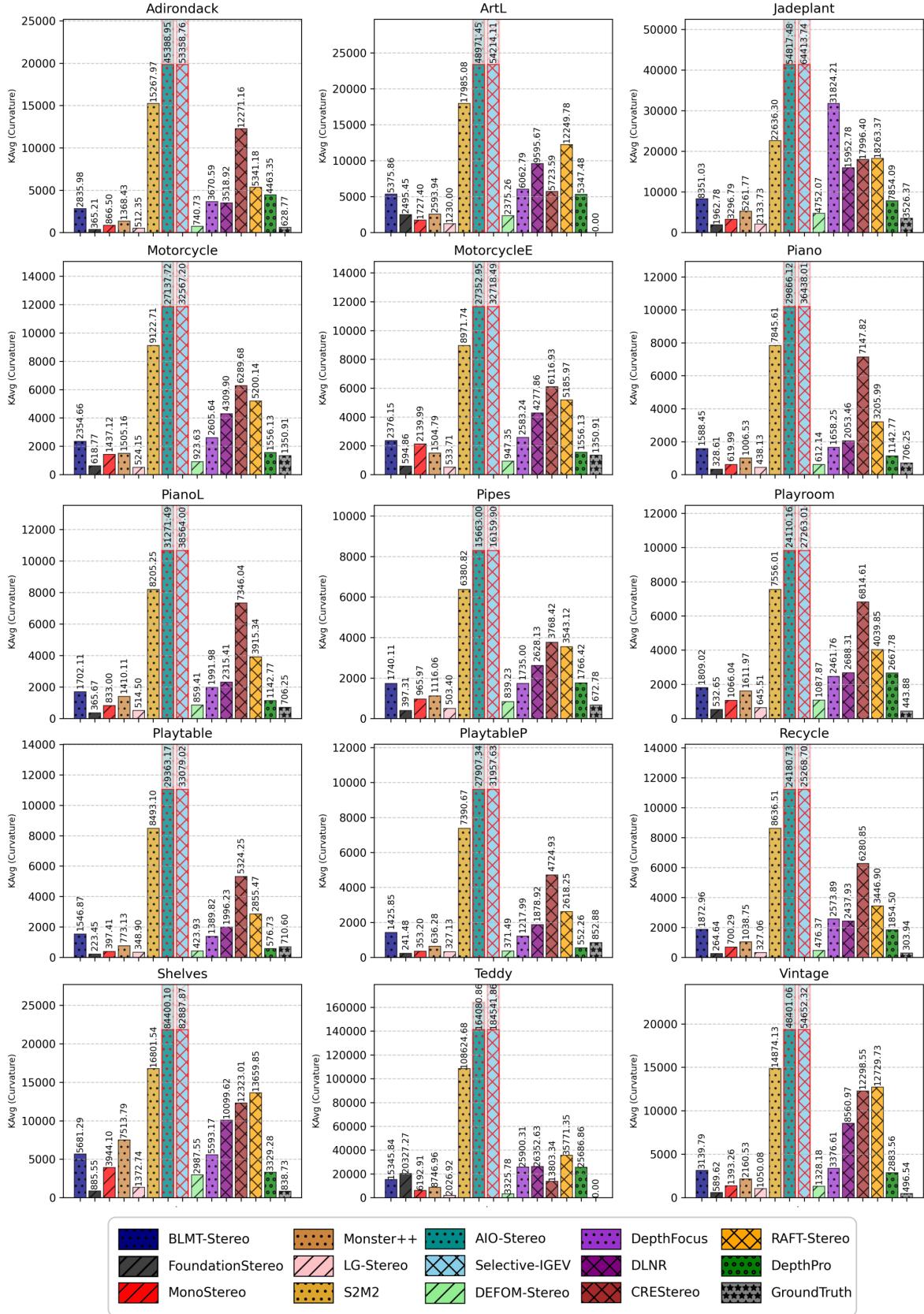


Figure 18. **Middlebury Curvature:** Average Absolute Gaussian Curvature (Avg| $K$ |) per image (see subsection 10.1 for more information).

## FoundationStereo

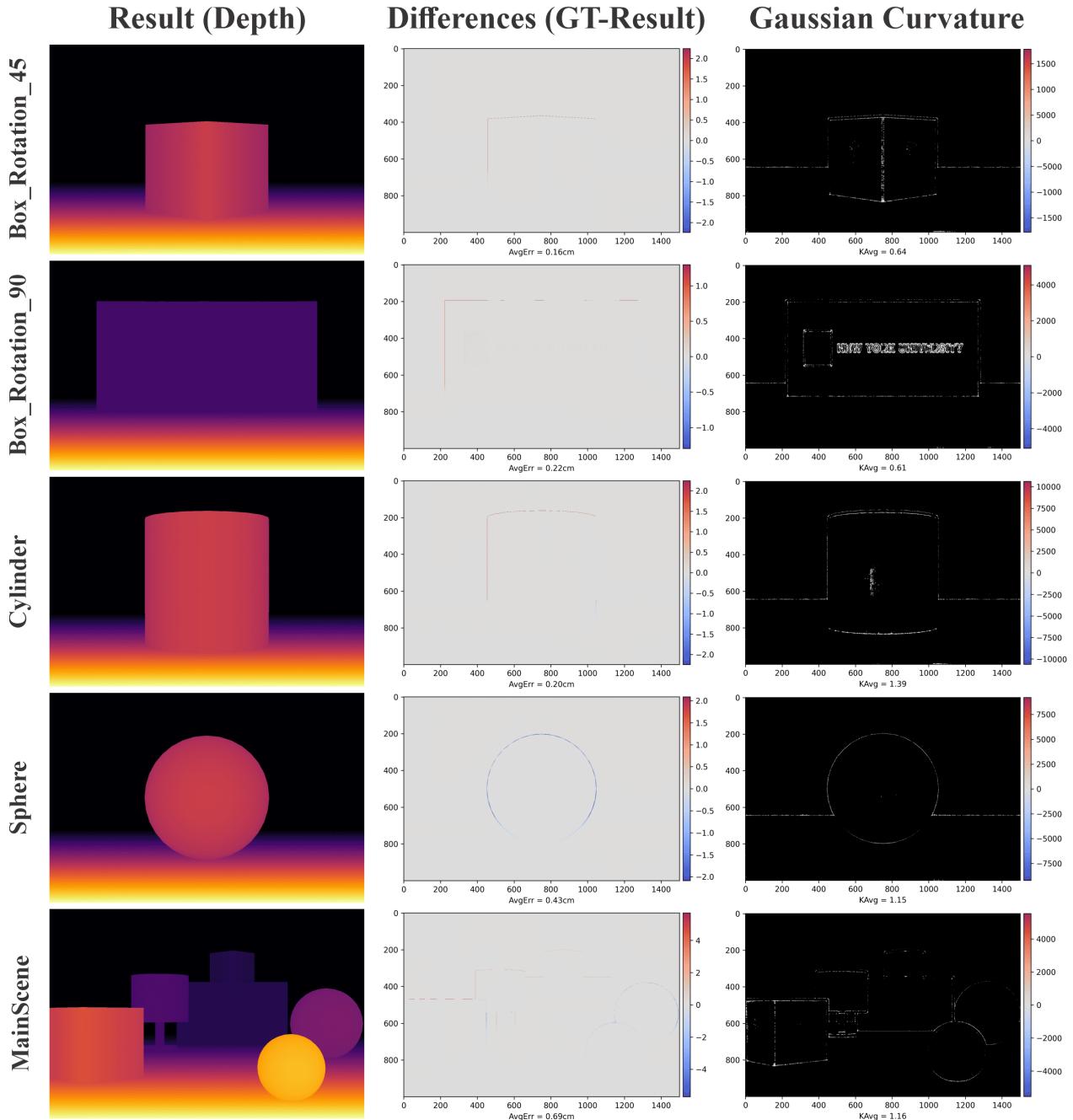


Figure 19. Results of **FoundationStereo** on our 3D synthetic scenes. FoundationStereo achieved a depth error of less than one centimeter across all five scenes. Additionally, it estimated the lowest Gaussian Curvature among all evaluated approaches. Notably, regions with  $|K| > 1000 \text{ m}^{-2}$  appear only near edges. Once again, FoundationStereo demonstrates its position as the best-performing method in both the Middlebury dataset and our 3D synthetic scenes.

## Selective-IGEV

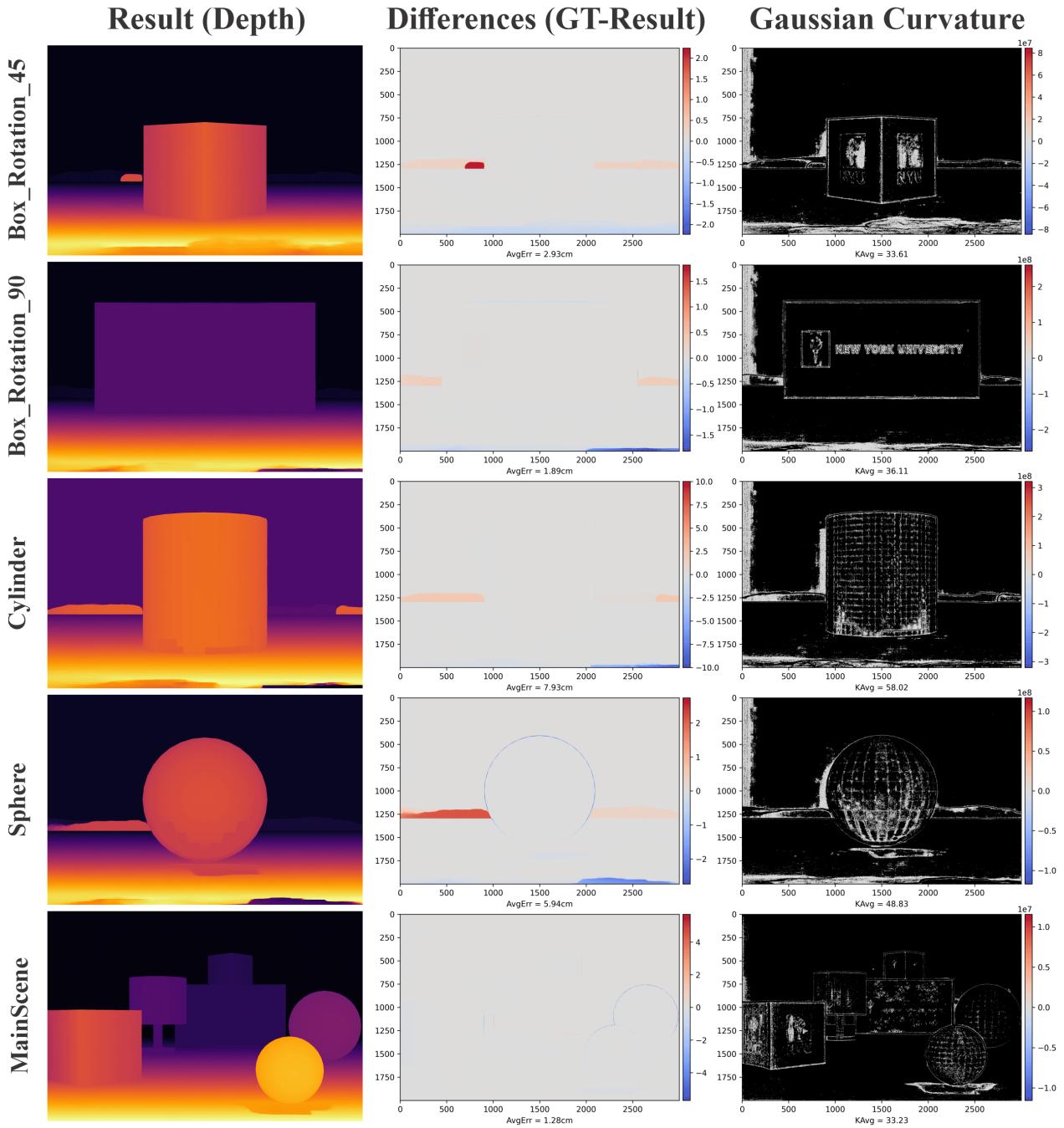


Figure 20. Results of **Selective-IGEV** on our 3D synthetic scenes. Selective-IGEV showed some inaccuracies in depth estimation and produced higher  $|K|$  values, as observed in the curvature plots. While it exhibited the highest Gaussian Curvature among all methods in the Middlebury dataset, in our 3D synthetic scenes it estimated lower curvature values than RAFT-Stereo (see [Figure 13](#) in the main paper).

## RAFT-Stereo

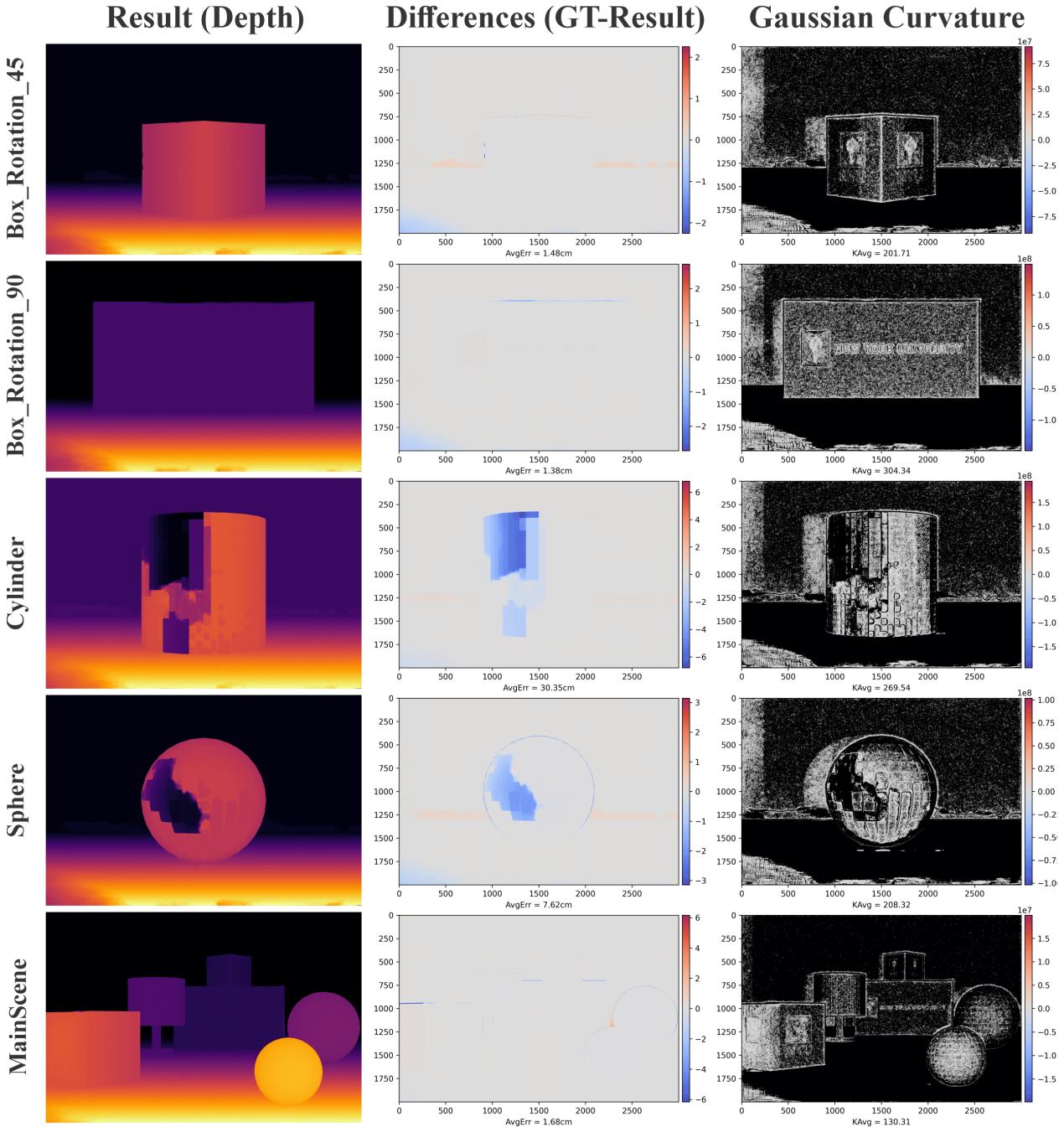


Figure 21. Results of **RAFT-Stereo** on our 3D synthetic scenes. RAFT-Stereo exhibited several inconsistencies in depth estimation. Notably, the Cylinder and Sphere scenes showed average errors exceeding 30 cm and 7 cm, respectively. Despite these limitations, RAFT-Stereo performed reasonably well in the remaining scenes. However, it produced the highest Gaussian Curvature among all evaluated approaches in our 3D synthetic scenes, including both Group A and Group B methods.

## DepthPro

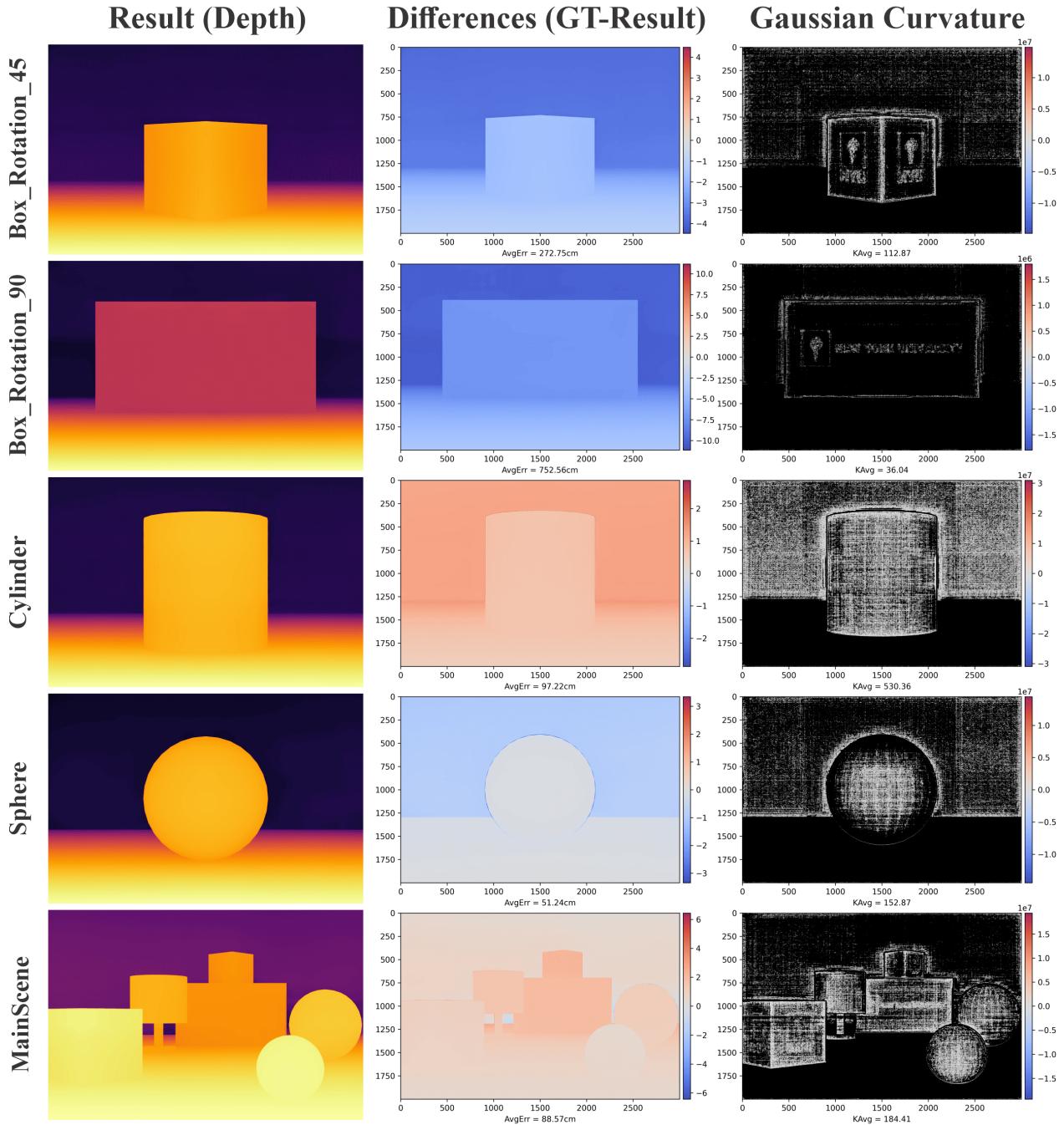


Figure 22. Results of **DepthPro** on our 3D synthetic scenes. DepthPro estimates depth in meters and also predicts focal length from a single image. In the main paper, we do not report DepthPro's average depth error due to observed limitations related to real-scale accuracy. As shown in the depth difference maps, DepthPro's minimum and maximum errors range from 51 cm to 752 cm. On the other hand, DepthPro produces reasonably accurate Gaussian Curvature estimates for flat surfaces such as the ground, boxes, and walls. As discussed in the paper, DepthPro achieves an LGC of approximately 54% in Middlebury Dataset, positioning it between the Group A methods, which tend to produce higher LGC values, and Group B methods, which generally show lower LGC values.

## DepthAnythingV2

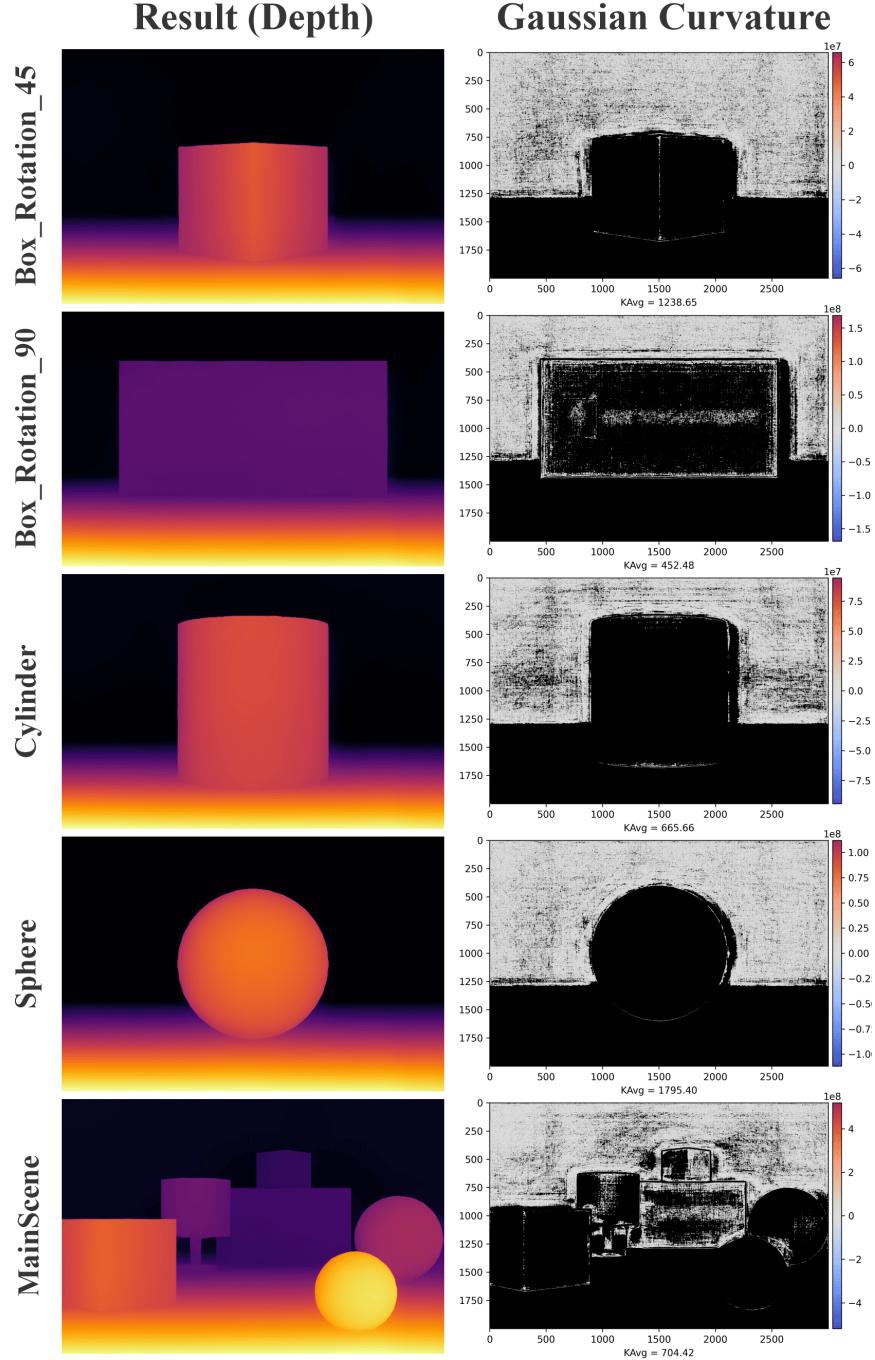


Figure 23. Results of **DepthAnythingV2** on our 3D synthetic scenes. DepthAnythingV2 provides relative depth; therefore, we do not report its average depth error. Qualitatively, the depth reconstruction appears consistent, correctly preserving the relative positioning of objects in the scene. However, DepthAnythingV2 exhibits the highest average Gaussian Curvature in our 3D synthetic scenes, indicating limitations in capturing intrinsic geometric relationships. As seen in the curvature plots, high  $|K|$  values are distributed across the entire scene—especially on the wall—despite the expected curvature being nonzero only near edges and on spherical surfaces.