Supplemental Material
The HCI Stereo Metrics:
Geometry-Aware Performance Evaluation of Stereo Algorithms

Katrin Honauer[1] Lena Maier-Hein[2] Daniel Kondermann[1]

[1]HCI, Heidelberg University
firstname.lastname@iwr.uni-heidelberg.de

[2]German Cancer Research Center (DKFZ)
l.maier-hein@dkfz-heidelberg.de

1. Overview

On the following pages, we first provide further details on the extraction of pixel sets at depth discontinuities, planar surfaces and fine structures. Secondly, we depict heuristics for the metric ranges and present further results on metric orthogonality. Finally, we provide additional examples of stereo evaluation based on the novel metrics. The figures are best viewed in color and on screen.

For the sake of completeness and reproducibility, Figure 1 indicates which regions of the Middlebury disparity maps were used for the experiments and illustrations in the main paper.

2. Extraction of Pixel Sets

In this section, we provide further detail on how we extract pixel sets for the geometry-aware evaluation of stereo results. The explanations build upon the pixel sets and evaluation principles introduced in Section 3 of the main paper.

2.1. Extraction of Depth Discontinuities

To assess stereo results near depth discontinuities, we use the pixel sets $\mathcal{M}_d$, $\mathcal{M}_f$, $\mathcal{M}_b$ as displayed in Figure 2 and described in Section 3.1 of the main paper. With the GT disparities D_{gt}, we define the discontinuity set $\mathcal{M}_d$ as:

$$\mathcal{M}_d = \{\vec{x} \in D_{gt} \colon |\nabla D_{gt}(\vec{x})| > c_d\}$$

For our experiments on the Middlebury dataset, we found $c_d = 8$ to be a good value for the major depth discontinuities. We further extract $\mathcal{M}_f$ and $\mathcal{M}_b$ by linearly following local gradient directions on both sides of the discontinuity and applying a median filter to fill gaps (see Figure 2).

We apply a similar procedure to compute D_f and D_b (second row of Figure 2). Instead of just identifying the affected pixels, we further set their values to the nearest disparity value on the other side of the discontinuity.

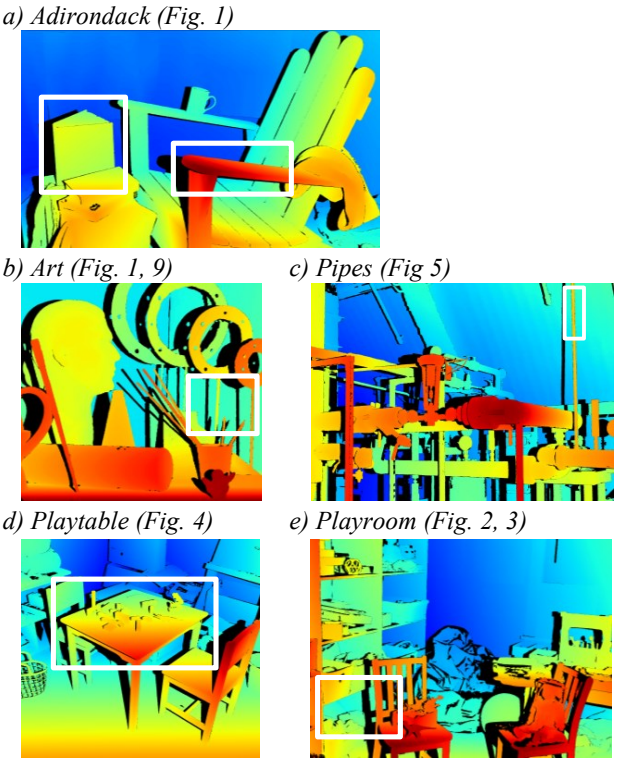

Figure 1: The white boxes illustrate which image regions were used in the main paper. For instance, the book and the arm rest of the Adirondack image were used in Figure 1.

2.2. Extraction of Planar Surfaces

To assess disparity maps at planar surfaces, we extract pixel sets of planar surfaces from D_{gt} and fit planes to those sets. As shown in Figure 3.a, we first apply Gaussian smoothing to the gradient directions of D_{gt}:

$$D_g = G_{\sigma=6}\left(atan2(\frac{\partial D_{gt}}{\partial y}, \frac{\partial D_{gt}}{\partial x})\right)$$

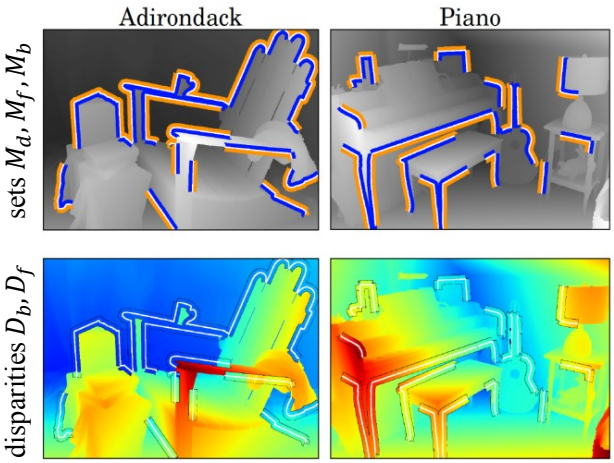

Figure 2: *Top:* The pixel set $\mathcal{M}_d$ (white) represents regions of depth discontinuities. The sets $\mathcal{M}_f$ (blue) and $\mathcal{M}_b$ (orange) represent foreground and background regions on either side of the discontinuities.
Bottom: To identify foreground thinning, the background disparities $\mathcal{M}_b$ which are close to the discontinuity are propagated into $\mathcal{M}_f$. The resulting disparity map D_b is depicted for the Adirondack image. To identify foreground fattening, the disparities $\mathcal{M}_f$ are propagated into $\mathcal{M}_b$, as depicted in D_f for the Piano image.

We then obtain an initial guess for planar surfaces (compare Figure 3.b) by computing the change in gradient direction. For the Middlebury dataset, we keep those pixels for which two conditions apply: first, the local change in direction is below 0.5 and second, the pixels are part of connected components whose size is at least 1% of the total pixel count.

For each of these components, we use RANSAC to robustely fit planes to the respective pixel sets in D_{gt}. As shown in Figure 3.c, we remove outliers which do not accurately belong to the fitted planes. Figure 3.d displays the normal directions of the final planar surfaces.

2.3. Extraction of Fine Structures

To obtain $\mathcal{M}_s$, the set of fine structure pixels in D_{gt}, we first shift negative disparity gradients to the left and positive gradients to the right, to compute the degree of overlap as shown in Figure 4.b and 4.c. Regions of high overlap potentially belong to thin structures.

We remove fragments which are smaller than 0.05% of the total pixel count to obtain the final set $\mathcal{M}_s$ as depicted in Figure 4.d.

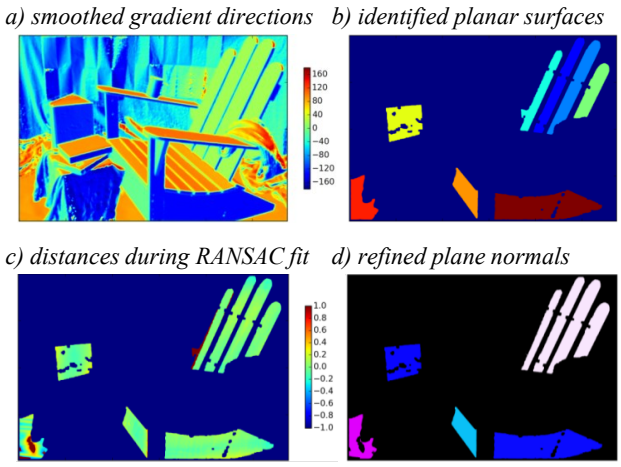

Figure 3: To obtain the set of planar surface pixels $\mathcal{M}_p$ and the set of fitted planes $\mathcal{P}$, we first compute a smoothed version of the local gradient directions of D_{gt} (a). We then identify connected regions of homogenous gradient directions (b) and iteratively fit planes to each of these components (c). We finally keep the inliers of the plane fits in $\mathcal{M}_p$ and compute the surface normals (d) for each plane $p_i = (\vec{n}_i, P_i)$ in $\mathcal{P} = \{p_0, ..., p_m\}$.

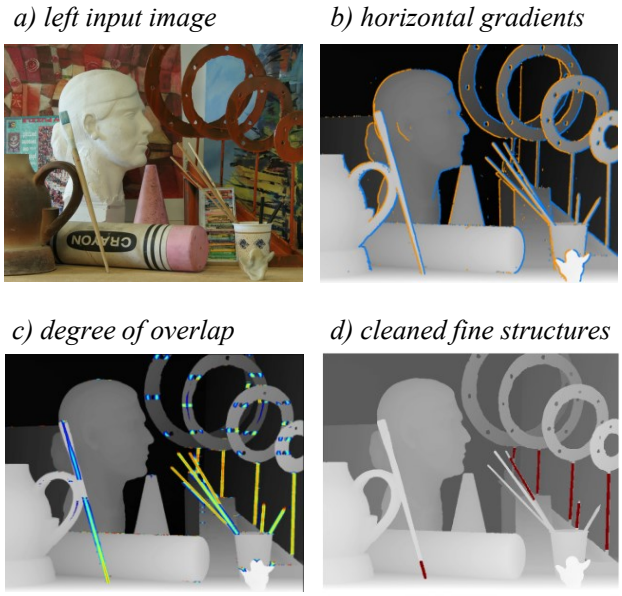

Figure 4: To obtain the fine structure set $\mathcal{M}_s$, we first compute the overlap resulting from horizontally shifting negative and positive gradients towards each other (b, c). After thresholding and removing small fragments, we obtain the final set $\mathcal{M}_s$ (d) which is depicted in red, overlaid onto the original disparity map.

3. Metric Evaluation

In this Section, we present heuristics on the value distributions of the metrics. We further provide additional data and figures for the evaluation of metric orthogonality.

3.1. Metric Ranges

Just like for the RMS, example scores are necessary for the novel metrics to get a good grasp of the scale of different metric scores. For instance, a score difference of 0.1 between two algorithms may be negligible for a metric in range $[0, 100]$ but most likely not for a metric in range $[0, 1]$.

We therefore provide an overview of the metric ranges in Figure 5. The table indicates theoretical and heuristic maxima for each metric. The latter are based on 13 algorithms and 12 images of the Middlebury benchmark as explained in the Experimental Setup of Section 4.1 in the main paper. Together with the metric histogram in the main paper, these metric heuristics help to get a better grasp on the relative differences and to get used to the meaning of particular metric scores, just as for the RMS.

	RMS	***Bad1.0***	***Bad4.0***	D_{fat}	D_{thin}	D_{fuz}
min	*0*	*0*	*0*	*0*	*0*	*0*
heuristic max	*70*	*80*	*45*	*0.6*	*0.3*	*3.4*
theoretical max	*∞*	*100*	*100*	*1*	*1*	*∞*

	P_{bump}	P_{dist}	P_{mis}	F_{por}	F_{frag}	F_{fat}
min	*0*	*0*	*0*	*0*	*0*	*0*
heuristic max	*4.5*	*25*	*60*	*1.2*	*0.5*	*0.7*
theoretical max	*∞*	*∞*	*90*	*∞*	*1*	*1*

Figure 5: Similar to the RMS metric, example scores for the novel metrics are necessary to get a feel for which scores are good or bad. We therefore denote the range of possible values as well as the heuristic maxima which were obtained based on the Middlebury dataset.

3.2. Orthogonality of Metrics

The orthogonality between existing and novel measures was evaluated in Section 4.4 of the main paper. For reasons of clarity, we depicted plots for the *Jadeplant* image only. For the sake of completeness and more comprehensive evaluations, we depict scatter plots for all algorithms (labeled by shape) and all images (labeled by color) in Figure 6 .

In analogy to the first row of Figure 10 in the paper, the first column in Figure 6 plots the established RMS measure against two BadPix metrics and three of our metrics. One can see, that there are correlations between the RMS and the BadPix metrics and almost no correlations between the RMS and the edge thinning and fragmentation metrics. There are moderate correlations between the RMS and the plane orientation. This makes sense since first, accurate plane orientation is indeed linked to accurate depth estimates and second, planar surfaces comprise relatively big pixel sets of the image and have thus a stronger influence on the RMS average. On the fourth subplot of the first column, one can further see that algorithms with very similar RMS measures yield very different scores on the fragmentation metric (compare Figure 6, esp. ArtL in light blue, Pipes in red and Shelves in olive).

The third column plots the novel edge fattening metric against the same metrics as for the RMS column. Correlations are much weaker between D_{fat} and the other metrics. We further plot the RMS values computed on the pixel subset for D_{fat} against the same metrics (second column in Figure 6). As a general observation over all rows of this column, the principal distribution of scores remains similar but not identical and it has generally higher RMS values. Hence, it does not suffice to limit the geometry-aware evaluation to computing RMS values on geometric pixel subsets.

4. Geometry-Aware Stereo Evaluation

In this Section, we provide exemplary metric scores and rankings for multiple algorithms and each metric category.

4.1. Evaluation at Depth Discontinuities

Figure 7 and Figure 8 depict sorted algorithm results according to their score on D_{fat} and D_{thin}. We refer to the captions of the Figures for further evaluation details.

4.2. Evaluation at Planar Surfaces

Our pixel sets for geometry aware stereo evaluation allow further automated quantifications of stereo performance. Furthermore, Figures 9 and 10 illustrate how algorithm performance depends on both, texture and surface orientation.

4.3. Evaluation at Fine Structures

Performance evaluation at fine structures particularly requires specific metrics as their relative weight is very low on averaging metrics such as the RMS. Figure 11 and Figure 12 depict sorted algorithm results according to their score on F_{frag} and F_{sampl}. For the fine structure metrics, changes in rank order are particularly large. For further evaluation details, we refer to the captions of the mentioned figures.

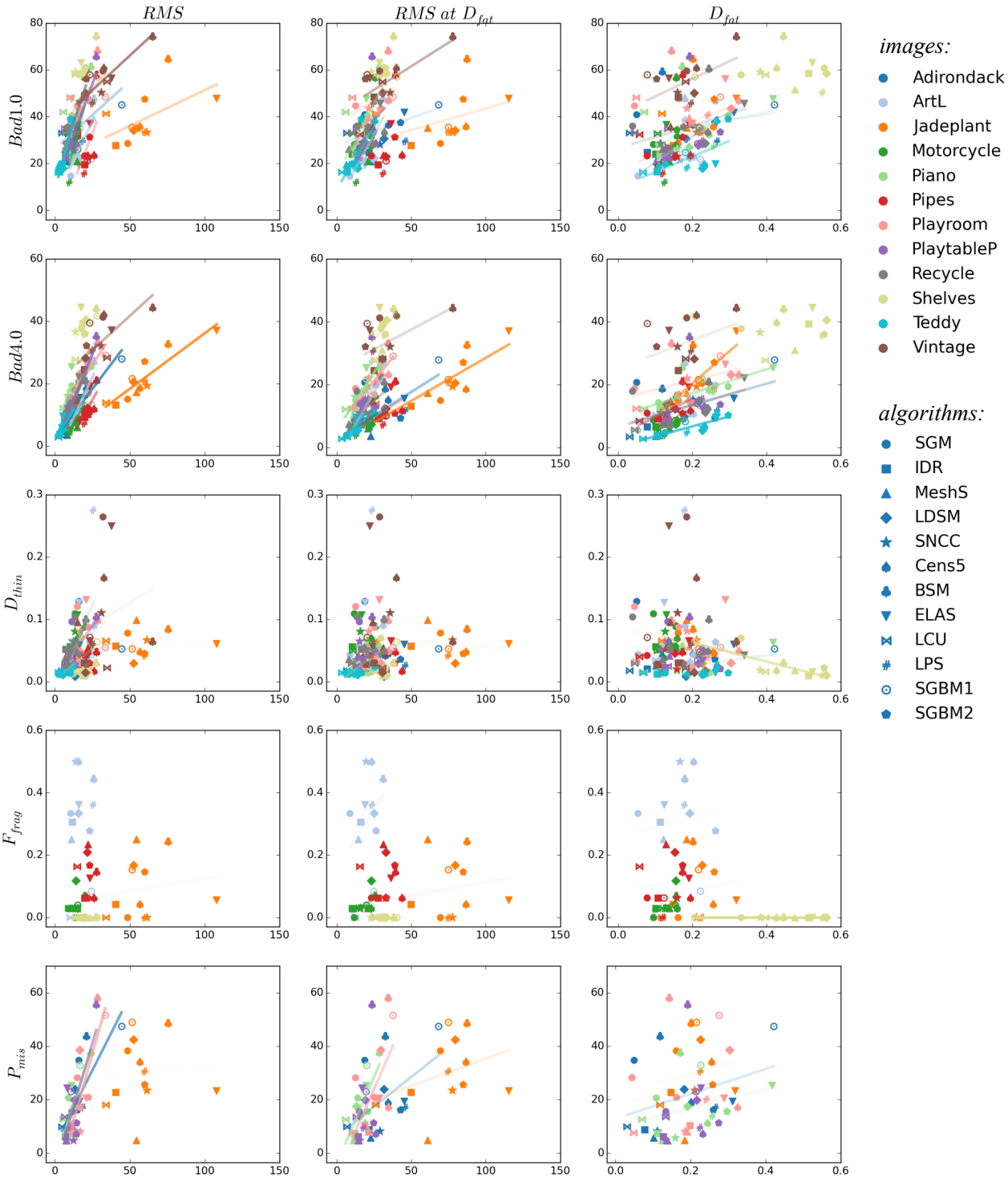

Figure 6: Each plot compares algorithm results on multiple images measured with two metrics. Colors identify images and shapes identify algorithms. For each image, the resulting line of a linear least squares fit is depicted in the color of the image. Its transparency is set to r^2, the coefficient of determination, i.e. lines are more opaque for better linear fits. We plot RMS and D_{fat} metrics against other metrics (first and third column). In the second column we further plot RMS values measured on the pixel subset for D_{fat}. One can see that RMS and $Bad4.0$ are the most correlated metrics and that the novel metrics yield lower correlations.

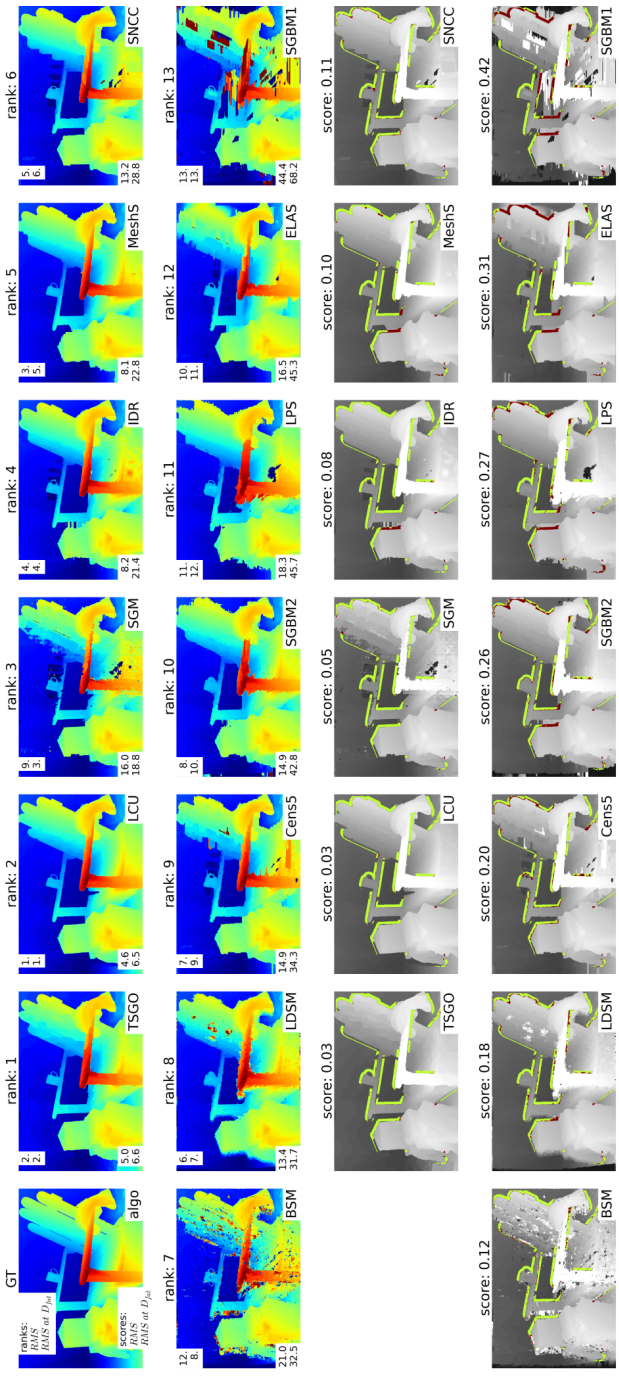

Figure 7: The stereo results are sorted by edge fattening performance. Numbers on top of the disparity maps and the metric visualizations denote the rank and the metric scores for D_{fat}. The tiny numbers on the upper left image corners denote the ranks and scores of the RMS computed on the entire image (upper number) or on the same pixel set as D_{fat} (lower number). Red areas in the green visualization of the subset $\mathcal{M}_b$ denote edge fattening. Note how the metric scores reflect the increasing amount of fattening present in the different disparity maps.

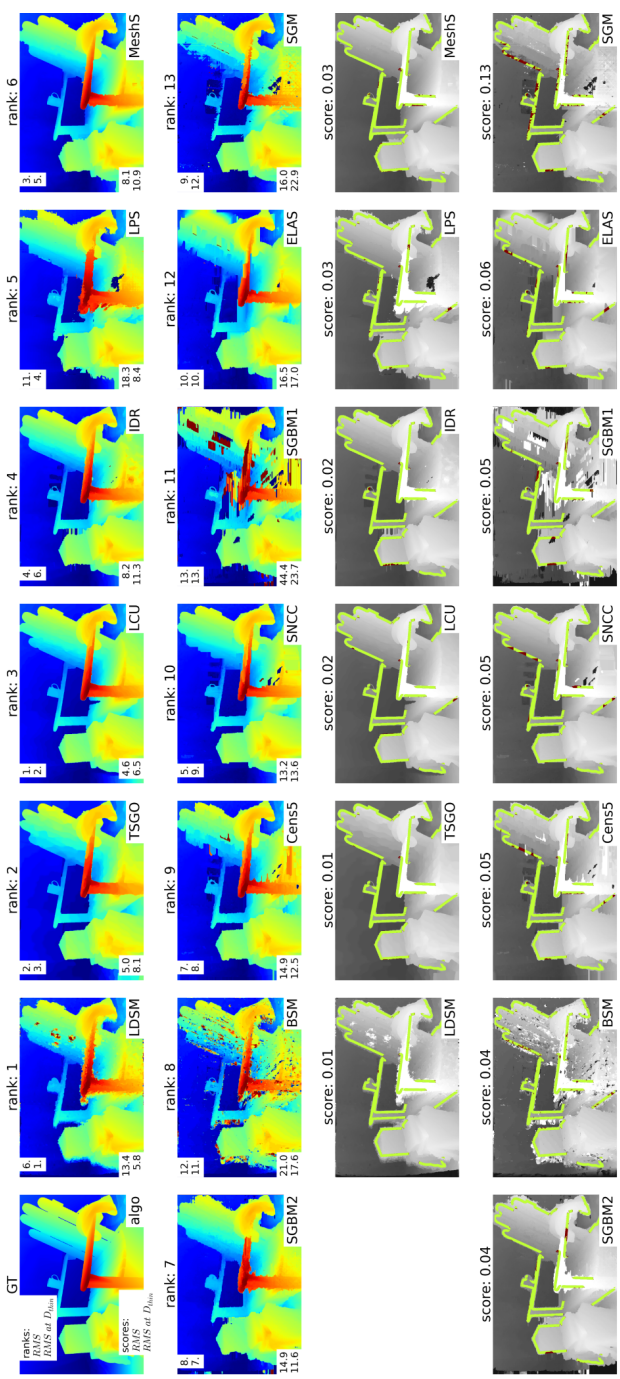

Figure 8: The stereo results on the Adirondack image are ranked by edge thinning performance and depicted analogously to Figure 7. Note how much the relative rankings of LPS and BSM differ when sorting by D_{thin} or RMS. Both algorithms rank much better for edge thinning as compared to their overall RMS rank. By contrast, the MeshS algorithm is the third best algorithm by RMS score but it is only the fifth and sixth best algorithm in terms of the edge fattening and edge thinning scores.

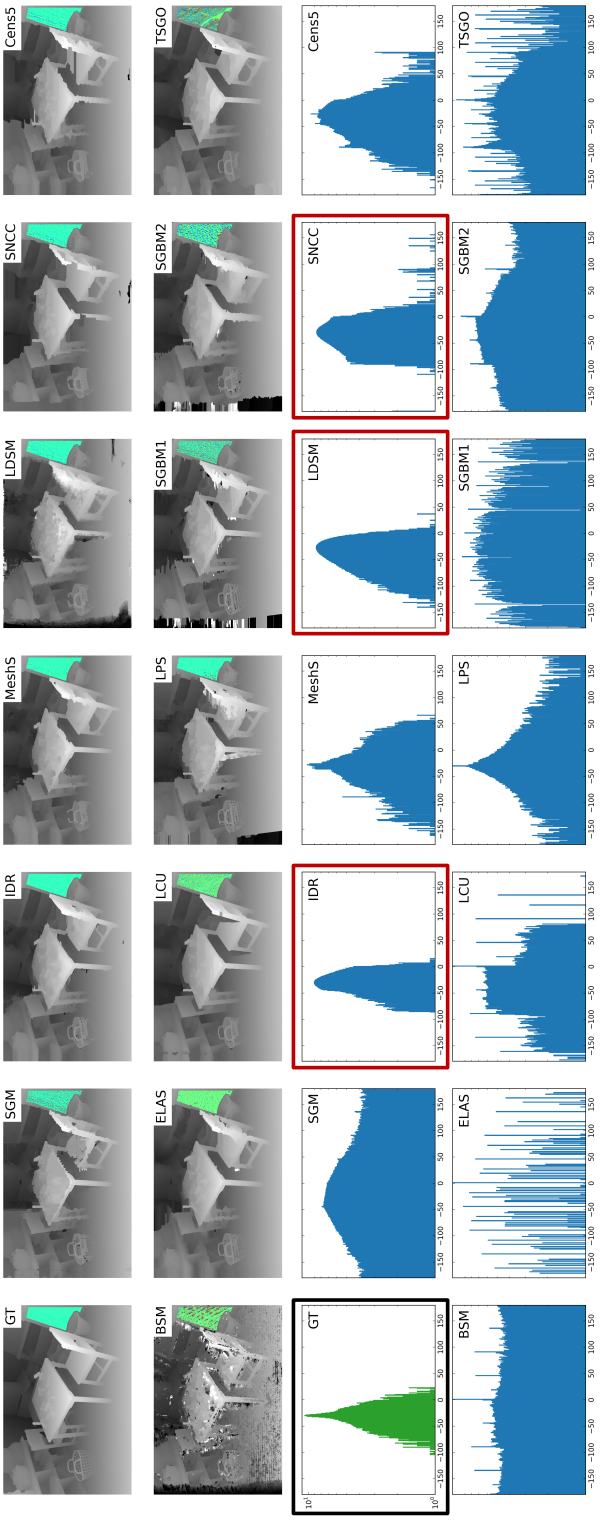

Figure 9: The top rows depict algorithm results where the colors of the cupboard surface indicate local gradient directions. The bottom rows depict logarithm histograms of the direction counts for each angle in $[-180, 180]$. Algorithms like IDR, LDSM and SNCC have very similar distributions as the GT whereas others yield much more diverse directions.

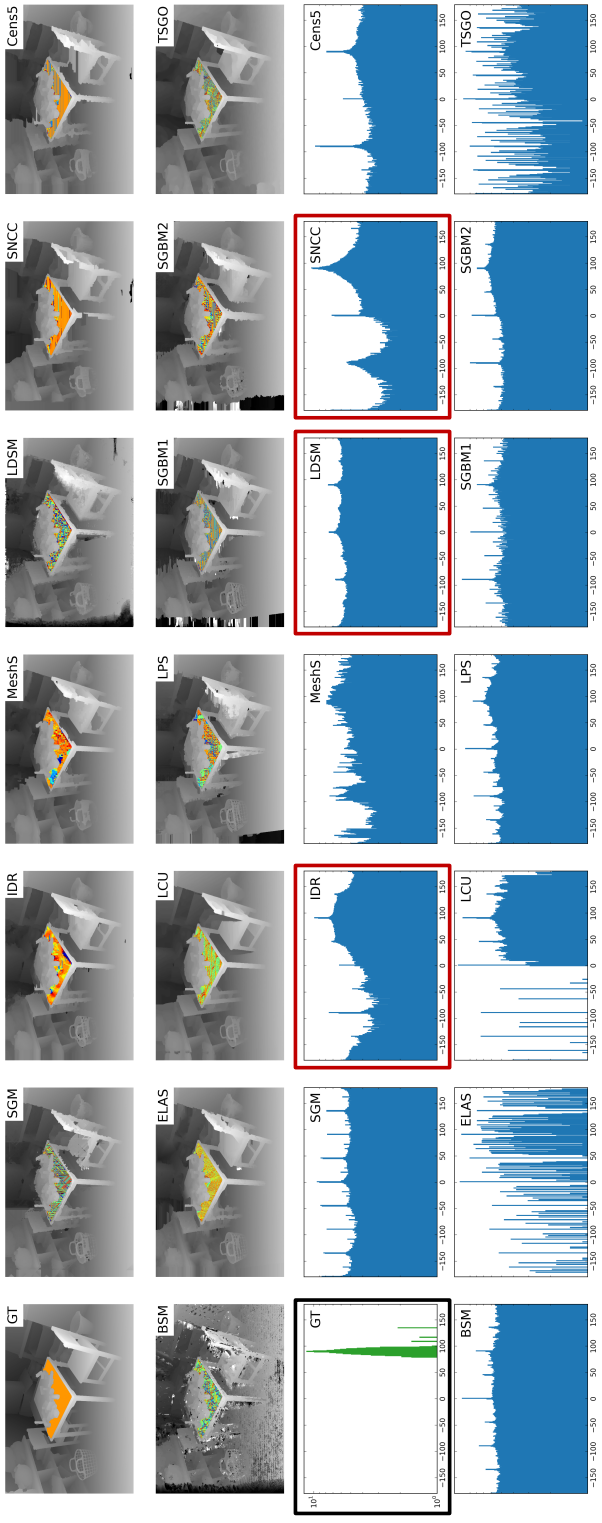

Figure 10: For the cupboard in Figure 9, the direction histograms of IDR, LDSM and SNCC are much more similar to the respective GT histogram as compared to the analogous histograms of IDR, LDSM and SNCC for the table surface. The cupboard plane is much better reconstructed as it features more texture and its orientation is more similar to *fronto-parallel* surfaces.

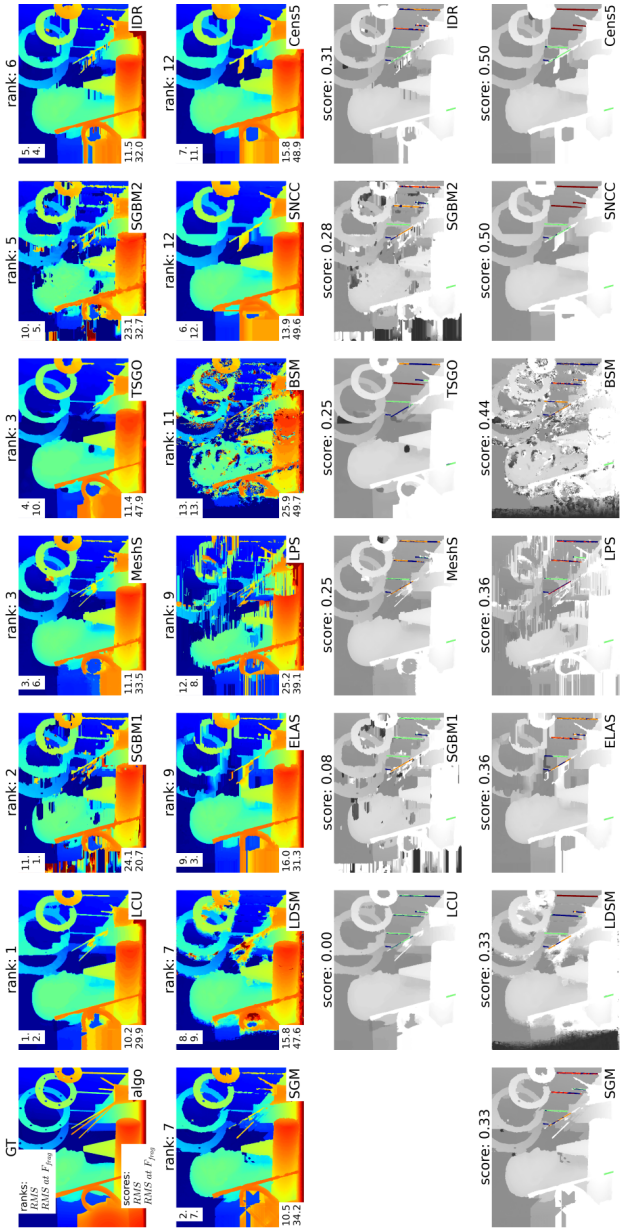

Figure 11: The stereo results are ranked by fragmentation performance at fine structures and depicted analogously to Figure 7. Note how much the relative rankings of F_{frag}, the RMS and the RMS at $\mathcal{M}_s$ differ from each other. Many local or moderately regularizing algorithms perform better at fine structures and worse on RMS scores. The results are best assessed when zooming to the thin bars on the lower right corner of the images. Green indicates no fragmentation, yellow and orange indicate moderate fragmentation and read indicates completely missing structures.

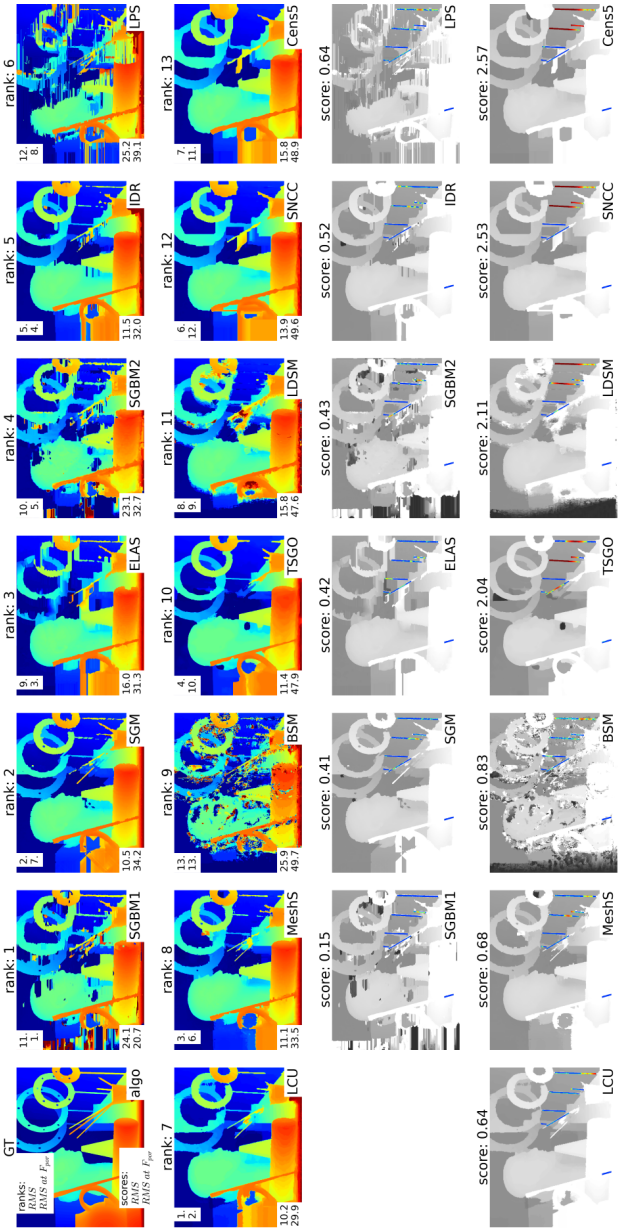

Figure 12: The top rows depict algorithm results ranked by porosity at fine structures. Note how the metric scores at the bottom rows reflect the degree of missing structures. The almost perfect SGBM1 result has a very good F_{por} score. The metric values for SGM, ELAS and SGBM2 reflect that their performance is lower yet very similar among each other. The values for TSGO, LDSM and SNCC accurately indicate that a lot of structure detail is missing. Similar to F_{frag} in Figure 11, many of the top performing algorithms in terms of RMS rank relatively low at F_{por}.