DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

ORNL/SPR-2024/3443

Dense Image Matching Uncertainty Estimation and Confidence Metrics

Jim Massaro
Orrin Thomas
Tyler Frazier
Doug Lepro

Approved for public release.
Distribution is unlimited.

May 13, 2025

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/SPR-2024/3443

Geospatial Science and Human Security Division

DENSE IMAGE MATCHING UNCERTAINTY ESTIMATION AND CONFIDENCE METRICS

Jim Massaro
Orrin Thomas
Tyler Frazier
Doug Lepro

May 13, 2025

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

APKR	Average Peak Ratio
AUC	Area Under the Curve
BNN	Bayesian Neural Network
CNN	Convolutional Neural Network
DLB	Distance from Left Border
DSM	Distinctive Similarity Measure
ENS7	Ensemble Learning (7 features)
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago
KL	Kullback-Leibler
LAFNet	Locally Adaptive Fusion Network
LEV	Leveraging Stereo Confidence
LiDAR	Light Detection and Ranging
LRC	Left-Right Consistency
MLM	Maximum Likelihood Measure
MM	Maximum Margin
MPN	Matching Probability Network
NLM	Non-linear Margin
PKR	Peak Ratio
RBM	Ratio Bad Matches
SEDNet	Stereo Error Distribution Network
SGM	Semi-Global Matching
SGMF	SGM Forest
UCN	Unified Confidence Network
VAR	Variance of Disparity

ABSTRACT

Dense stereo matching takes overlapping image pairs as input and outputs a disparity map which encodes pixel-by-pixel matches between the images. Recently, there has been an interest in ranking the quality, or even quantifying the accuracy, of disparity estimates. The proposed methods can be described as either uncertainty estimators or confidence metrics. Uncertainty estimators are a small minority of the research. However, they have the potential to be the most useful because they estimate disparity accuracy (in pixel units) that can be used to threshold matches or carried forward using error propagation. The majority of the research deals with confidence metrics which give an ordinal (or binary) ranking of a match's quality relative to other matches. Confidence metrics do not have units and thus are useful primarily for thresholding matches from mismatches. The methods could also be described as handcrafted or deep-learning based. The majority of the research focused on outdoor driving scenes. Hence, our interest–application to a satellite semi-global matching pipeline–is a domain shift that may challenge deep-learning based methods. We conclude by recommending five handcrafted and two deep-learning based methods for evaluation in our pipeline.

1. INTRODUCTION

This work aims to identify promising algorithms from the literature to add mismatch identification and uncertainty estimation to an operational dense matching pipeline. The pipeline implements semi-global matching [8] using the stable descriptor and is being used to produce a global 2m digital surface model from electro-optical stereo image pairs. Years of optimization and stabilization have been invested in the current pipeline. The system throughput capacity is approximately 2 million square kilometers a day, and our failure rate is approximately 1:100000 stereo pairs. Future work may relax these guidelines, but for now the pipeline needs to remain in production and the development work should follow principles of minimum risk while the error estimation is added. Hence, error estimators will be discussed in terms of suitability for implementation in the pipeline. Poorly suited estimators may serve as benchmark comparisons and guide work on the next generation pipeline.

Dense stereo matching research in the last two decades has garnered a lot of interest due to autonomous vehicle navigation, specifically for automobiles. The self-driving vehicle community needs cheaper and faster processing of dense 3D data. Depth information derived from stereo cameras is beginning to complement Light Detection and Ranging (LiDAR) [39], and research continues to improve the quality and accuracy of the data. Our interest is in the dense matching of satellite images, but the autonomous vehicle research may be applicable because the objectives of the error estimation are the same: identifying occlusions and mismatches, and quantifying the uncertainty of each match. However, the domain shift may be an issue for deep-learning based methods.

Dense matching quality assessment research can be described as either confidence estimation or uncertainty estimation [2, 23]. Confidence estimation results in an ordinal [31, 6, 16] or binary [35, 32] quality metric. These unitless metrics are designed to be related to the probability that a match is accurate, and thus can be used as rejection criteria. In contrast, uncertainty estimation gives an estimate of each match's standard error in pixels [2, 23]. Confidence metrics research has a rich history included review papers [12, 9, 32] and published source code [21]. The most comprehensive and widely referenced of the confidence metric reviews was Poggi *et al.* in 2021 [32].

Confidence metrics are categorized as learned (machine learning based) or handcrafted. The learned metrics can further be categorized as 2D or 3D convolutions [1] by their training data. Most of the methods start by training on synthetic data and then either branch to real data or focus on domain independence (not requiring real data) [15] [36] [22]. Large images were a challenge for learned methods [1].

1.1 DENSE MATCHING CONCEPTS

Dense matching for stereo 3D reconstruction is the process of matching images pixel-by-pixel. Matches are determined by analyzing matching cost function over the range of possible matches (called the disparity range). In the simplest case, the minimum value of the cost function curve indicates the matching disparity at that pixel. To limit noise and improve accuracy, disparity solutions are actually solved semi-globally by taking continuity with other matches into account [8]. An example of a cost curve can be seen in Figure 1. A collection of cost curves over an entire image is a cost volume. An example of a cost volume can be seen in Figure 2. Learned metrics that use cost volumes for training are in the 3D convolution category.

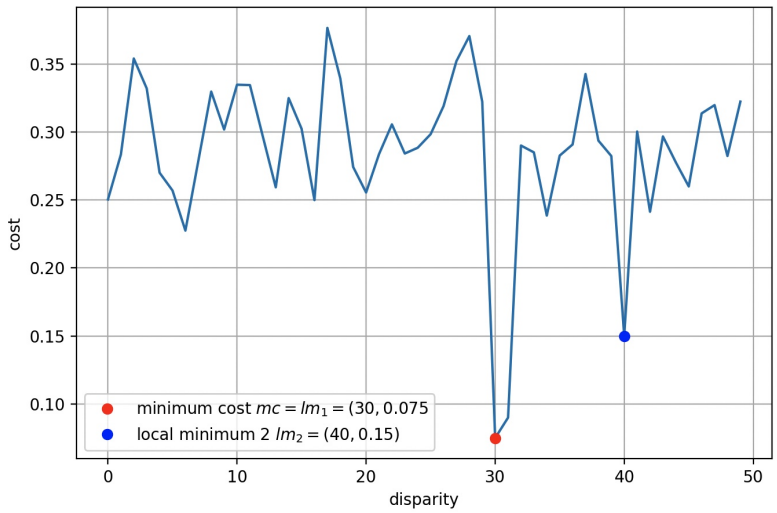

Figure 1. Cost curve with annotations

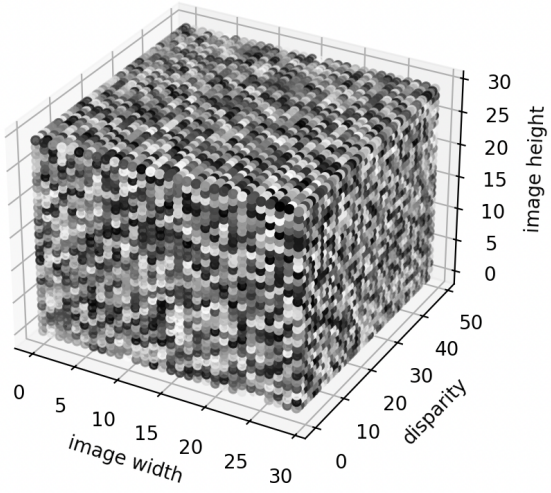

Figure 2. Example of a cost volume, where the grayscale color indicates the cost value, and parallel lines along the disparity axis are cost curves.

For SGM the cost function is minimized pixel by pixel, but considers continuity over a local area, N_p, making it semi-global. The algorithm runs on epipolar images, which simplifies localizing pixel matches to a linear search within a scan line that is bounded by disparity range. The cost function we used in our application

comes from [8]

$$E(D) = \sum_{p}(C(p, D_p) + \sum_{q \in N(p)} P_1 T[|D_p - D_q| < 1] + \sum_{q \in N(p)} P_2 T[|D_p - D_q| > 1])$$

where D is the disparity image, q is a pixel in the left image over the area N_p in the right image. The $C()$ operator indicates the image similarity function (the stable descriptor in our case). The $T[]$ operator is 1 if the condition is true and 0 otherwise. P_1 and P_2 are predefined penalty values also known as smoothing operators where $P_2 > P_1$. Large P_2 values enforce a penalty for larger disparity changes. Using a lower P_2 value can help with resolving edges, slanted, and curved surfaces.

2. UNCERTAINTY ESTIMATION

Uncertainty estimation work is limited to a single paper from Computer Science Department of Stevens Institute of Technology [2], and a series from the Institute of Photogrammetry and GeoInformation, Leibniz University in Hanover, Germany [24, 25, 40, 23]. Only Chen *et al.* [2] published uncertainty estimation source code [27].

The uncertainty estimation work from Leibniz University introduced separately modeling aleatoric and epistemic uncertainty [23]. Aleatoric error is non-deterministic and inherent to the data. Epistemic error is described as inaccuracies in the models, and failure to include it can lead to overconfident error estimation [5, 14]. Mehltretter *et al.* [23] models the epistemic uncertainty by converting their Convolutional Neural Network (CNN) to Bayesian Neural Network (BNN) with two probabilistic layers (the feature extraction convolution filter and the multi-scale feature matching). The probabilistic layers allowed them to run perturbations of the matching to sample the epistemic error.

The paper from the Stevens Institute of Technology is also a deep-learning based estimation of uncertainty. They named their network Stereo Error Distribution Network (SEDNet) [2, 27]. Both SEDNet and [23] used Kullback-Leibler (KL) divergence to derive the Bayesian loss function for the machine learning algorithm. KL divergence is a logarithmic statistical distance that is a measure of dissimilarity between two probability distributions.

Neither Mehltretter or Chen use Census-SGM as their baseline for computing disparity. Mehltretter uses his own CVANet [25] and GCNet [14] as disparity map estimators and adds another network for uncertainty estimation. Mehltretter *et al.* do not compare results to any other methods for uncertainty calculation. SEDNet uses GwCNet [4] as a basis to compute disparity and compared to a single confidence metric (Locally Adaptive Fusion Network (LAFNet)).

2.1 EVALUATING CONFIDENCE

Poggi *et al.* [32] used four benchmark data sets to evaluated the confidence metrics. Middlebury 2014 is a high-resolution oblique indoor stereo data set taken at close range with highly accurate ground truth data acquired using a structured light system [33]. Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago (KITTI) data sets are outdoor driving data sets acquired with a stereo camera and mounted on an automobile with laser scanner ground truth [3]. KITTI 2015 contains more dynamic outdoor driving scenes with motion within the images. The ground truth was prepared in a semi-automatic process [26]. ETH3D, created by the Federal Institute of Technology Zurich, is a high-resolution multi-view stereo data set with high-precision laser scanned ground truth [34].

The metric commonly used to evaluate and compare the performance of stereo confidence algorithms is Area Under the Curve (AUC). It was adopted by [32] from [10] (see Table 1). Both AUC chart axes are ratios and hence range from 0.0 to 1.0. The AUC horizontal axis (x-axis) indicates the proportion of matches sorted from highest confidence to lowest confidence. The vertical axis (y-axis) plots the Ratio Bad Matches (RBM) within the top x proportion of confidence metrics. For example the hypothetical point (0.25, 0.03) indicates that in the 25% of matches with the best confidence scores 3% are bad. Thus, each point on the plot is an estimate of the probability that a match is bad given that it is in the best proportion of the highest confidence matches. 'Bad matches' are defined by a error threshold, τ, from the truth disparity.

The error threshold, τ, ranges from 1-3 pixels in the literature, depending on the evaluation data set. The optimum value $AUC_{min} = ln(1-\epsilon) * (1-\epsilon) + \epsilon$ depends on ϵ, the ratio of bad matches. Results are only strictly comparable if they have the same ϵ. Hence, AUC values for different stereo pairs are not generally

comparable. Nor can they strictly be summarized statistically using averages or medians. The pattern in practice, for ease, has been to assume that ϵ is approximately equal within individual bench mark data sets and report average AUC for each bench mark set [32].

The confidence metric comparison work done by [32] is reorganized in Table 1. The data is from Semi-Global Matching (SGM) tests run on the four bench mark data sets described above. The results are sorted by the confidence metric's overall performance relative to all the others. The results are also color coded into three categories of learned metrics and six categories of handcrafted metrics. These categories will be explained in the following sections of the paper, and each algorithm's performance will be important context for discussion. The results were reorganized for three reasons. First, to ease comparison of handcrafted and deep-learning-based methods. Second, to average the deep learning based performance across the two different training data sets. Third, to remove the performance metrics for synthetic training data. The AUC values are multiplied by 100 for consistency with [32]. As noted above, AUC averaging is not strictly valid (it was done for ease of summary and consistency with past work). Hence, the ranking order in Table 1 should be considered approximate.

Metric	KITTI2012	KITTI2015	Midd.	ETH	average	rank
MPN	1.75	1.95	10.935	6.275	5.2275	1
UCN	1.96	2.21	10.925	6.04	5.28375	2
VAR19	1.97	1.92	12.02	5.37	5.32	3
LAF	1.84	2.07	10.96	6.825	5.42375	4
MM	2.82	2.83	10.68	5.39	5.43	5
NLM	2.82	2.83	10.68	5.39	5.43	6
APKR5	2.64	2.81	11.06	5.48	5.4975	7
WPKR5	2.7	2.86	11.03	5.49	5.52	8
ACN	2.04	2.02	11.045	7.005	5.5275	9
PKR	2.81	2.92	11.15	5.6	5.62	10
CVA	2.52	2.61	11.825	5.815	5.6925	11
MLM	2.74	2.7	11.29	6.13	5.715	12
LGC	2.025	2.14	12.33	6.415	5.7275	13
SGMF	2.66	2.785	11.815	5.925	5.79625	14
WMN	2.89	3.04	11.5	5.95	5.845	15
O2	2.15	2.28	11.795	7.185	5.8525	16
LRD	3.08	3.22	11.28	5.88	5.865	17
PBCPr	2.415	2.38	12.05	6.655	5.875	18
LEV50	1.88	1.985	12.7	7.215	5.945	19
PER	2.91	2.82	11.83	6.42	5.995	20
O1	2.385	2.485	12.425	7.02	6.07875	21
ALM	2.97	2.89	12.05	6.52	6.1075	22
WMNN	3.61	3.47	11.44	5.95	6.1175	23
MMC	2.34	2.33	12.7	7.215	6.14625	24
DS31	3.05	3.47	12.8	5.38	6.175	25
ConfNet	2.375	2.69	13.365	6.5	6.2325	26
PWCFA	3.29	3.22	12.03	6.48	6.255	27
PKRN	4.03	3.88	11.74	6.01	6.415	28
APKRN5	4.13	4.01	12.04	5.95	6.5325	29
CCNN	2.315	2.475	13.825	7.645	6.565	30
SCS	3.55	3.79	13.08	6.54	6.74	31
DA31	3.92	4.23	14.51	4.39	6.7625	32
GCP	2.185	2.495	14.82	7.645	6.78625	33
LEV22	1.815	1.985	14.34	9.06	6.8	34
UCC	3.48	3.48	13.04	7.41	6.8525	35
MND21	2.99	3	14.23	7.27	6.8725	36
SGE	3.38	3.3	13.11	7.83	6.905	37
ENS23	2.43	2.695	14.225	8.44	6.9475	38
MSM	3.55	3.46	13.04	7.82	6.9675	39
WPKRN5	4.64	4.5	12.52	6.21	6.9675	40
FA	2.935	2.93	13.65	8.48	6.99875	41
DSM	4.5	2.78	13.66	7.8	7.185	42
PS	5.37	4.79	11.98	7.24	7.345	43
LFN	3.165	3.375	14.505	8.535	7.395	44
CUR	5.51	4.77	13.05	6.79	7.53	45
LC	5.47	4.91	12.96	6.95	7.5725	46
PBCPd	2.505	2.985	14.415	10.45	7.58875	47
MMN	5.36	5.05	13.4	6.94	7.6875	48
NLMN	5.36	5.05	13.4	6.94	7.6875	49
ENS7	3.505	3.745	15.69	9.1	8.01	50
CRNN	3.425	3.075	17.795	8.84	8.28375	51
SKEW21	4.03	4.12	15.91	9.21	8.3175	52
DTD	3.61	3.8	17.67	8.68	8.44	53
MDD21	3.7	3.64	18.97	10.35	9.165	54
RCN	3.56	3.165	20.835	10.595	9.53875	55
DMV	4.77	4.67	18.77	11.1	9.8275	56
ACC	6.08	5.59	18.43	11.09	10.2975	57
UC	6.28	5.83	18.79	11.37	10.5675	58
LRC	6.16	5.57	19.65	11.38	10.69	59
EFN	6.17	5.455	19.995	11.305	10.73125	60
SAMM	11.89	3.71	19.17	8.97	10.935	61
LMN	7.43	5.9	20.92	11.24	11.3725	62
UCO	7.44	6.41	19.95	11.88	11.42	63
DLB	6.79	6.93	23.26	13.01	12.4975	64
ZSAD	9.49	8.14	19.86	12.59	12.52	65
DAM	8.67	8.21	22.67	13.28	13.2075	66
DB	8.45	8.75	23.84	11.87	13.2275	67
IVAR5	8.64	8.51	25.41	12.47	13.7575	68
HGM	9.27	8.19	25.86	13.99	14.3275	69
DTE	9.36	8.92	26.45	13.55	14.57	70
NEM	10.33	9.04	28.9	14.59	15.715	71
DTS	22.53	6.39	21.17	15.03	16.28	72
NOI	14.77	12.09	29.36	15	17.805	73

Deep Learning, Cost Volume CNN
Deep Learning, Disparity CNN
Deep Learning, Forests
handcrafted, Whole Cost Curve
handcrafted, Disparity Map Analysis
handcrafted, Discrete Cost Curve Sampling
handcrafted, Left-right Consistency
handcrafted, Self Matching
handcrafted, Reference Image Analysis

Table 1. AUC*100 ranking of handcrafted and learned confidence metrics from [32]

3. CONFIDENCE METRICS

Confidence estimation produces an ordinal [31, 6, 16] or binary [35, 32] quality metric. A confidence metric is considered to have performed perfectly if, when all the matches are sorted by their confidence, the good and bad matches are segregated. Poggi *et al.* [32] grouped and evaluated 73 confidence metrics. The 49 handcrafted metrics were categorized as functions of discrete cost curve samples, functions of the entire cost curve, left-right consistency metrics, disparity map analysis, reference image analysis, self-matching, or semi-global matching measures. The 24 learned metrics were categorized as random forests, disparity CNNs, and cost volume CNNs. Each of these subgroups will be introduced and top performers will be cited in the following sections.

3.1 HANDCRAFTED CONFIDENCE METRICS

Poggi *et al.* [32] evaluated 49 handcrafted metrics and ranked them according to performance on the benchmarks above. handcrafted confidence metrics are simple functions of data in the cost volume, local image area, or disparity map. They would be the simplest to implement because we have the necessary input data readily available in the current pipeline. They also performed well, even in comparison to the best learned methods. Six of the top ten confidence metrics in [32] were handcrafted, and the top handcrafted metric only lagged the top learned metric by 1.8%

3.1.1 Discrete Cost Curve Sampling

Confidence metrics in this category are functions of discrete values sampled from the cost curves. Five of the ten top performing metrics are in this category, which is notable because these are among the simplest and most intuitive metrics. The best performing metric in this category was Maximum Margin (MM) ranked 2^{nd} out of 49 handcrafted metrics and 5^{th} overall [32]. MM's performance was 2.1% behind the best handcrafted metric and 3.9% behind the best learned metric. It is defined as the difference between the second local minimum, $lm_2(p)$, and the minimum cost $mc(p)$. The metric thus quantifies how distinct the global best is compared to the runner-up.

$$MM(p) = lm_2(p) - mc(p) = lm_2(p) - lm_1(p)$$

Non-linear Margin (NLM) was listed as the next best performing metric [32]. Its performance was identical to MM because NLM's ordinal ranking of matches was identical. This can be observed because NLM is a monotonically increasing function of MM. MM will be favored because it has a lower computation cost.

Peak Ratio (PKR) was the 5^{th} ranked handcrafted metric and 10^{th} overall. PKR's performance was 5.6% behind the best handcrafted metric and 7.5% behind the best learned metric. It's defined as the ratio between the second best local cost, $lm_2(p)$, and the best local cost, $lm_1(p)$.

$$PKR(p) = \frac{lm_2(p)}{lm_1(p)}$$

Average Peak Ratio (APKR) sums PKR in a neighborhood around the active point, $N(p)$. The conceptual justification for the metric is that the stability of neighboring pixels have a bearing on the accuracy of the active pixel due to the semi-global matching function employed. Poggi *et al.* tested a 5x5 window size, and it improved on PKR by 2.2%–which was enough to move it from 10^{th} to 7^{th} overall [32].

$$APKR(p) = \sum_{q \in N(p)} PKR(q)$$

3.1.2 Entire Cost Curve

Metrics in this category are functions of the entire cost curve or segments over a disparity range. This category had only one metric, Maximum Likelihood Measure (MLM), in the top 20 performers. MLM was ranked 6^{th} among handcrafted metrics and 14^{th} overall. MLM was 7.4% behind the best handcrafted metric and 9.3% behind the best learned metric. MLM attempts to generalize the concept of distinctness used in MM. Instead of comparing the minimum cost, $mc(p)$, to the second best it compared $mc(p)$ to a summation across the entire disparity range, D. The summation used an exponential decay function to give more weight to cost curves values, $C_d(p), d \in D$, that approach the best. Thus the metric penalizes having more of the cost curve near the minimum. Note that this metric varies with the cost curve standard deviation, $\sigma(p)$, and the magnitude of the disparity range, $|D|$. meaning that even within the same image the metric is not directly comparable pixel to pixel (because $\sigma(p)$ certainly varies, and $|D|$ may too). This makes an ordinal sorting suspect and a general threshold, δ, between good and bad matches a particularly elusive quantity for this metric.

$$MLM(p) = \frac{e^{-\frac{mc(p)}{2\sigma}}}{\sum_{d \in D} e^{-\frac{c_d(p)}{2\sigma}}}$$

3.1.3 Left-Right Consistency

Left-Right consistency refers to the comparison of left-right dense matching solutions to right-left dense matching solutions. The idea is that the left disparity, $d_L(p_L)$, implies a matching pixel in the right image, $p_R = p_L - d_L(p)$. The reverse dense matching solution from right to left, $d_R(p_R)$, should close back to $p_L = p_R - d_R(p_R)$. The circle often doesn't close due to mismatches, occlusions, and perspective differences. According to [32], these metrics performed poorly in detecting bad disparity matches. The best performer in this category was 35^{th} overall, and all the others were in the bottom 17. However, Poggi *et al.* [32] made no accounting for identifying occlusions in any of their comparisons. The basic form of this metric, Left-Right Consistency (LRC), has long been recommended to identify occlusions [8], and may be the most useful metric for that application [11, 28].

$$LRC(p) = -\left|d_L(p_L) - d_R(p_r)\right|$$

However, if occlusions and mismatches can be identified without left-right consistency checks, then we can get nearly a factor of two speed improvement over our current pipeline design.

3.1.4 Disparity Map Analysis

Disparity map based metrics are matching algorithm agnostic because they are functions of only the output from the cost volume. Thus, they are easy, non-invasive, additions to existing pipelines. One group of metrics was based on a point disparity's agreement with other local disparities [37, 30]. Another, hypothesized that reliability would be a function of the distance to the nearest disparity discontinuity [37]. The most successful metric, Variance of Disparity (VAR) [29], was in the category that based their metric on the variability of disparities within a local neighborhood, $N(p)$. The idea was with more variability in disparity came less reliability [7, 37, 30, 29]. VAR, defined using a 19x19 disparity map window, ranked 1^{st} among handcrafted confidence metrics and 3^{rd} overall (1.8% worse than the best learned metric). Interestingly, none of the other disparity based metrics performed nearly as well. VAR is simply the variance of disparity of all the points q in $N(p)$.

$$VAR(p) = \frac{1}{\left|N(p)\right|} \sum_{q \in N(p)} [d(q) - \mu(d(q \in N(p)))]^2$$

3.1.5 Reference Image Analysis

Reference image analysis metrics propose ideas like matching confidence is related to the proximity of image edges or the variance of the intensity image. All five metrics were in the bottom ten performers in the Poggi *et al.* review [32]. The best performer in the category was Distance from Left Border (DLB) [29]. As the name implies, DLB is the distance in pixels from the left image edge.

3.1.6 Self-Matching

These metrics are calculated from the input intensity images, and are based on the idea that repeating textures and featureless image regions should correlate with bad matching. All three metrics in this category tested by [32] performed poorly. The best among them was Distinctive Similarity Measure (DSM) [38]. The DSM metric penalized similar pixels along the epipolar lines in both left and right images. It ranked 43^{rd} overall.

3.2 LEARNED CONFIDENCE METRICS

The deep learning metrics are described in three groups: random forests, disparity CNNs, and cost volume CNNs. Random forests operated by combining a set of handcrafted descriptors into a description vector of the matching conditions around a point. Thus, the random forest approach offers the potential to build on the effectiveness of each of its constituent parts. Ensemble Learning (7 features) (ENS7) is a simple example built on three poorly-ranked metrics (56^{th}, 59^{th}, and 69^{th} [32]) and improved the performance by 35% compared to the best performing component metric [32]. Unfortunately, ENS7, was the only random forest built on separately ranked components that outperformed each of its component metrics, and it was only ranked 50^{th} in overall performance [32]. For example, Leveraging Stereo Confidence (LEV) used a more comprehensive 50-element description vector, $\overrightarrow{V_{LEV}}$, that included 25 different handcrafted metrics including MM, MLM, PKR, and VAR. Five of LEV's metrics (including VAR) depended on a sample window size, and were evaluated for 6 different windows (shown as superscripts below). Unfortunately, LEV accuracy was 11.8% worse than its best performing component metric (VAR).

$$\overrightarrow{V_{LEV}} = \langle MM, MLM, PKR...VAR^{5}, VAR^{9}, VAR^{11}, VAR^{15}, VAR^{21}, VAR^{31}... \rangle$$

14^{th} overall-ranked SGM Forest (SGMF) was the best random forest metric. SGMF performed 10.9% worse than the best learned metric and 9.0% worse than the best handcrafted metric [32]. The components of SGMF's description vector were values related to the consistency and cost of the individual SGM scan lines. The component values were not individually performance ranked, and, thus, we cannot say if SGMF performed better than its component parts.

CNN are not built from component handcrafted descriptors but rather are trained on either the cost volume (3D convolution) or the disparity map (2D convolution). Three of four top performing metrics were cost-volume CNNs: Matching Probability Network (MPN) [17], Unified Confidence Network (UCN) [20], and LAFNet [18] [32]. This is a logical result because the cost volume contains a comprehensive picture of the matching conditions. Cost volumes are an intermediate output in SGM dense matching that, to avoid memory transfer costs, are declared, computed, and used exclusively on the GPU. It is possible this type of metric could be implemented as a step inside the SGM processing kernel without paying heavy processing costs. However, their current performance is only 2-5% better than the best ranked handcrafted metrics [32].

It is interesting that the 3^{rd}-ranked VAR [29], a simple handcrafted metric that was operated on a small window of disparity map, outperformed seven different disparity map CNNs [32]. This maybe a sampling bias or a result of poor training data, but it does imply that the disparity CNNs are not ready for production.

3.3 IMPLEMENTATION CONSIDERATIONS

Uncertainty metrics directly estimate a useful quantity (the standard errors of matches). Unfortunately, the body of research is thin and dedicated to a different domain. In contrast, there is an extensive body of work on confidence metrics. However, even a perfectly performing confidence metric may not be practically useful. Consider that, for a perfect confidence metric, there is some value, δ, that separates all the good matches from all the bad matches. However, δ isn't required, or expected, to be consistent from pair to pair. For non-perfect confidence metrics, the AUC evaluations imply conditional probabilities of a match being acceptable, but, again, this varies from pair to pair. In sort, the research was framed academically and neglected practical considerations for using confidence metrics.

The shift in research focus away from occlusion detection [11, 28] to optimizing AUC [32] (or other deep learning metrics) may have a downside. The latest research using learned methods lumped occlusions into disparity errors and attempted to correct all errors whether they were occluded or not [17, 18, 19]. AUC is computed excluding occluded pixels as they are labeled in the ground truth dataset [17, 18, 19]. This can lead to incorrect biased estimates of performance. It may also explain a reported poor performance of deep-learning in challenging image regions such as close to depth discontinuities, occlusions, thin objects, and weakly textured areas [13]. A combination of deep-learning and handcrafted methods may address this weakness [13].

Finally, it should be noted that learned methods cannot be readily retrained in the remote sensing domain. Training requires extensive stereo image data with truth disparity maps. In satellite remote sensing, this requires near simultaneous collection of images and dense ground truth 3D data (e.g., aerial lidar). The expense, and difficulty coordinating, the multi-modal collection make in-domain training impractical (at least in the short term).

3.4 CONCLUSION

First, there are interesting implications from the review paper [32]. To start, it is interesting that the left-right consistency metrics performed so poorly despite their common use. Based on this evidence there is potential to improve matching time by a factor of two and get better results by using different confidence metrics. The solid performance of many simple handcrafted metrics compared to the learned methods is also interesting. Particularly the random forest methods failure to improve on their component parts implies the training data is corrupt or inadequate. Interpretation is further compounded by the fact the AUC performance metric is not designed for statistical summary or to promote production utility because interpretation is dependent on the context of each matching pair. Finally, it is curious that the metrics that sought to directly capture the phenomena commonly listed as sources of matching error (featureless image regions and repeating textures) performed so poorly. These confidence metrics are in the reference image analysis and self-match categories, and none of them tested well. The poor performance of these metrics, despite their expected utility, may indicate the problem is counter-intuitive or that the whole testing procedure is flawed. Theses observation lead us to question if the the confidence metric body of research is mature.

The only practical choice for bench mark testing among the uncertainty estimators is SEDNet because it is the only one with published source code[1]. Among the learned confidence metrics, LAFNet is the highest ranked learned option with published source code. LGC-Net[2] may also be a good option because it's within 10% of the top performing metric. There are 5 handcrafted metrics in the top-10 confidence estimators including the 3^{rd}-ranked VAR_{21} [29]. These handcrafted metrics are simple to implement in our pipeline because of their simplicity and the availability of the required data. Based on our lack of confidence in the maturity of the confidence metric research, we should also test some that didn't perform well in the reviews.

[1] https://github.com/lly00412/SEDNet

[2] https://github.com/fabiotosi92/LGC-Tensorflow.

4. REFERENCES

[1] Filippo Aleotti et al. "Neural Disparity Refinement for Arbitrary Resolution Stereo". In: *2021 International Conference on 3D Vision (3DV)*. 2021, pp. 207–217. DOI: 10.1109/3DV53792.2021.00031.

[2] Liyan Chen, Weihan Wang, and Philippos Mordohai. "Learning the Distribution of Errors in Stereo Matching for Joint Disparity and Uncertainty Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 17235–17244.

[3] 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.

[4] Xiaoyang Guo et al. *Group-wise Correlation Stereo Network*. 2019. arXiv: 1903.04025 [cs.CV].

[5] Ian Hacking. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 1975.

[6] Ralf Haeusler and Reinhard Klette. "Evaluation of stereo confidence measures on synthetic and recorded image data". In: *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE. 2012, pp. 963–968.

[7] Ralf Haeusler, Rahul Nair, and Daniel Kondermann. "Ensemble learning for confidence measures in stereo vision". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 305–312.

[8] H. Hirschmuller. "Accurate and efficient stereo processing by semi-global matching and mutual information". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. 2005, 807–814 vol. 2. DOI: 10.1109/CVPR.2005.56.

[9] Xiaoyan Hu and Philippos Mordohai. "A Quantitative Evaluation of Confidence Measures for Stereo Vision". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2121–2133. DOI: 10.1109/TPAMI.2012.46.

[10] Xiaoyan Hu and Philippos Mordohai. "A Quantitative Evaluation of Confidence Measures for Stereo Vision". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2121–2133. DOI: 10.1109/TPAMI.2012.46.

[11] Xiaoyan Hu and Philippos Mordohai. "A Quantitative Evaluation of Confidence Measures for Stereo Vision". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2121–2133. DOI: 10.1109/TPAMI.2012.46.

[12] Xiaoyan Hu and Philippos Mordohai. "Evaluation of stereo confidence indoors and outdoors". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 1466–1473. DOI: 10.1109/CVPR.2010.5539798.

[13] Waseem Iqbal, Jens-Andr/'e Paffenholz, and Max Mehltretter. "Guided Deep Learning with Expert Knowledge for Dense Stereo Matching". In: *PFG* 91.6 (2024), pp. 365–380. DOI: https://doi.org/10.1007/s41064-023-00252-0.

[14] Alex Kendall and Yarin Gal. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

[15] Kwonyoung Kim et al. "PointFix: Learning to Fix Domain Bias for Robust Online Stereo Adaptation". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Springer Nature Switzerland, 2022, pp. 568–585. URL: https://arxiv.org/abs/2207.13340.

[16] Sanghun Kim, Dong-gon Yoo, and Young Hwan Kim. "Stereo confidence metrics using the costs of surrounding pixels". In: *2014 19th International Conference on Digital Signal Processing*. IEEE. 2014, pp. 98–103.

[17] Sunok Kim et al. “Deep stereo confidence prediction for depth estimation”. In: *2017 ieee international conference on image processing (icip)*. IEEE. 2017, pp. 992–996.

[18] Sunok Kim et al. “LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 205–214. DOI: 10.1109/CVPR.2019.00029.

[19] Sunok Kim et al. “Stereo Confidence Estimation via Locally Adaptive Fusion and Knowledge Distillation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5 (2023), pp. 6372–6385. DOI: 10.1109/TPAMI.2022.3207286.

[20] Sunok Kim et al. “Unified confidence estimation networks for robust stereo matching”. In: *IEEE Transactions on Image Processing* 28.3 (2018), pp. 1299–1313.

[21] *LAFNet CVRP19*. https://github.com/seungryong/LAF. 2019.

[22] Chuang-Wei Liu et al. “Stereo Matching: Fundamentals, State-of-the-Art, and Existing Challenges”. In: *Autonomous Driving Perception: Fundamentals and Applications*. Ed. by Rui Fan, Sicen Guo, and Mohammud Junaid Bocus. Singapore: Springer Nature Singapore, 2023, pp. 63–100. ISBN: 978-981-99-4287-9. DOI: 10.1007/978-981-99-4287-9_3. URL: https://doi.org/10.1007/978-981-99-4287-9_3.

[23] M. Mehltretter. “Joint Estimation of Depth and its Uncertainty from Stereo Images Using Baysian Deep Learning”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-2-2022 (2022), pp. 69–78. DOI: isprs-annals-V-2-2022-69-2022. URL: https://isprs-annals.copernicus.org/articles/V-2-2022/69/2022/.

[24] Max Mehltretter. “Uncertainty estimation for end-to-end learned dense stereo matching via probabilistic deep learning”. In: *arXiv preprint arXiv:2002.03663* (2020).

[25] Max Mehltretter and Christian Heipke. “Aleatoric uncertainty estimation for dense stereo matching via CNN-based cost volume analysis”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 171 (2021), pp. 63–75.

[26] Moritz Menze, Christian Heipke, and Andreas Geiger. “Joint 3D Estimation of Vehicles and Scene Flow”. In: *ISPRS Workshop on Image Sequence Analysis (ISA)*. 2015.

[27] *Official Implementation of SEDNet*. https://github.com/lly00412/SEDNet. 2023.

[28] Min-Gyu Park and Kuk-Jin Yoon. “Learning and Selecting Confidence Measures for Robust Stereo Matching”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.6 (2019), pp. 1397–1411. DOI: 10.1109/TPAMI.2018.2837760.

[29] Min-Gyu Park and Kuk-Jin Yoon. “Leveraging Stereo Matching with Learning-based Confidence Measures”. In: *CVPR2015* (2015). DOI: https://openaccess.thecvf.com/content_cvpr_2015/papers/Park_Leveraging_Stereo_Matching_2015_CVPR_paper.pdf.

[30] Matteo Poggi and Stefano Mattoccia. “Learning a general-purpose confidence measure based on o (1) features and a smarter aggregation strategy for semi global matching”. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE. 2016, pp. 509–518.

[31] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. “Quantitative evaluation of confidence measures in a machine learning world”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5228–5237.

[32] Matteo Poggi et al. “On the confidence of stereo matching in a deep-learning era: a quantitative evaluation”. In: *CoRR* abs/2101.00431 (2021). URL: https://arxiv.org/abs/2101.00431.

[33] Daniel Scharstein et al. “High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth”. In: *Pattern Recognition*. Ed. by Xiaoyi Jiang, Joachim Hornegger, and Reinhard Koch. Cham: Springer International Publishing, 2014, pp. 31–42. ISBN: 978-3-319-11752-2.

[34] Thomas Schöps et al. “A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[35] Amit Shaked and Lior Wolf. "Improved Stereo Matching With Constant Highway Networks and Reflective Confidence Learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[36] Zhelun Shen et al. "Digging Into Uncertainty-Based Pseudo-Label for Robust Stereo Matching". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.12 (2023), pp. 14301–14320. DOI: 10.1109/TPAMI.2023.3300976.

[37] Aristotle Spyropoulos, Nikos Komodakis, and Philippos Mordohai. "Learning to detect ground control points for improving the accuracy of stereo matching". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1621–1628.

[38] C Tomasi and R Manduchi. "Distinctiveness Maps for Image Matching". In: (1999).

[39] Yan Wang et al. *Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving*. 2020. arXiv: 1812.07179 [cs.CV].

[40] Zeyun Zhong and Max Mehltretter. "Mixed probability models for aleatoric uncertainty estimation in the context of dense stereo matching". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2 (2021), pp. 17–26.