A comparative study on deep-learning methods for dense image matching of multi-angle and multi-date remote sensing stereo-images

Hessah Albanwan\textsuperscript{1,2} | Rongjun Qin\textsuperscript{1,2,3,4}

\textsuperscript{1}Geospatial Data Analytics Laboratory, The Ohio State University, Columbus, Ohio, USA
\textsuperscript{2}Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio, USA
\textsuperscript{3}Department of Electrical and Computer Engineering, The Ohio State University, Columbus, Ohio, USA
\textsuperscript{4}Translational Data Analytics Institute, The Ohio State University, Columbus, Ohio, USA

Correspondence
Rongjun Qin, Geospatial Data Analytics Laboratory, The Ohio State University, 218B Bolz Hall, 2036 Neil Avenue, Columbus, OH 43210, USA.
Email: qin.324@osu.edu

Funding information
Office of Naval Research, Grant/Award Number: N000142012141

Abstract
Deep-learning (DL) stereomatching methods gained great attention in remote sensing satellite datasets. However, most of these existing studies conclude assessments based only on a few/single stereo-images lacking a systematic evaluation on how robust DL methods are on satellite stereo-images with varying radiometric and geometric configurations. This paper provides an evaluation of four DL stereomatching methods through hundreds of multi-date multi-site satellite stereopairs with varying geometric configurations, against the traditional well-practiced Census-semi-global matching (SGM), to comprehensively understand their accuracy, robustness, generalisation capabilities, and their practical potential. The DL methods include a learning-based cost metric through convolutional neural networks (MC-CNN) followed by SGM, and three end-to-end (E2E) learning models using Geometry and Context Network (GCNet), Pyramid Stereo Matching Network (PSMNet), and LEASTereo. Our experiments show that E2E algorithms can achieve upper limits of geometric accuracies, while may not generalise well for unseen data. The learning-based cost metric and Census-SGM are rather robust and can consistently achieve acceptable results.
All DL algorithms are robust to geometric configurations of stereopairs and are less sensitive in comparison to the Census-SGM, while learning-based cost metrics can generalise on satellite images when trained on different datasets (airborne or ground-view).

**KEYWORDS**
convolutional neural network, deep learning, dense image matching, Geometry and Context Network, LEAStereo, Pyramid Stereo Matching Network

**Résumé**
Les méthodes de mise en correspondance stéréo par apprentissage profond (DL) ont suscité une grande attention pour les données de télédétection par satellite. Cependant, la plupart des études existantes conduisent à des évaluations basées sur peu d’images stéréo, sans évaluation systématique de la robustesse des méthodes par DL sur des images stéréo acquises avec des configurations radiométriques et géométriques variables. Cet article présente une évaluation de quatre méthodes d’appariement stéréo par DL, basée sur des centaines de couples stéréo multi-dates et multi-sites avec des configurations géométriques variées, en comparaison avec la méthode traditionnelle largement utilisée Census-SGM (Semi-global matching), afin de bien comprendre leur précision, leur robustesse, leur capacité de généralisation et leur potentiel d’application pratique. Les méthodes par DL comprennent une métrique de coût basée sur l’apprentissage utilisant des réseaux neuronaux convolutifs (MC-CNN) avant SGM, et trois modèles d’apprentissage de bout en bout (E2E) utilisant Geometry and Context Network (GCNet), Pyramid Stereo Matching Network (PSMNet) et LEAStereo. Nos expérimentations montrent que les algorithmes E2E peuvent atteindre les limites supérieures de la précision géométrique, mais qu’ils peuvent être difficiles à généraliser à des données n’ayant pas servi à l’apprentissage. La métrique de coût basée sur l’apprentissage et le Census-SGM sont plutôt robustes et donnent des résultats acceptables. Tous les algorithmes par DL sont robustes vis-à-vis des configurations géométriques des couples stéréo et sont moins sensibles que le Census-SGM, tandis que la métrique de coût basée sur l’apprentissage peut être généralisée aux images satellite même entrainée sur des données aériennes ou terrestres.

**Zusammenfassung**
Deep Learning (DL)-Stereo-Matching-Methoden haben in Fernerkundungssatellitendatensätzen große Aufmerksamkeit erlangt. Die meisten dieser bestehenden Studien schließen jedoch Bewertungen ab, die nur auf wenigen/einzelnen Stereobildern basieren, und es fehlt eine systematische Bewertung, wie robust DL-Methoden auf
Satellitenstereobildern mit unterschiedlichen radiometrischen und geometrischen Konfigurationen sind. Dieses Papier bietet eine Bewertung von vier DL-Stereo-Matching-Methoden durch Hunderte von Satelliten-Stereopaares mit mehreren Daten und mehreren Standorten mit unterschiedlichen geometrischen Konfigurationen, gegen das traditionelle Geübte Census-SGM (Semi-globales Matching), um ihre Genauigkeit, Robustheit, Verallgemeinerungsfähigkeit und ihr praktisches Potenzial umfassend zu verstehen. Die DL-Methoden umfassen eine lernbasierte Kostenmetrik durch Convolutional Neural Networks (MC-CNN), gefolgt von SGM, und drei End-to-End (E2E)-Lernmodelle mit Geometry and Context Network (GCNet), Pyramid Stereo Matching Network (PSMNet). ) und LEAStereo. Unsere Experimente zeigen, dass E2E-Algorithmen die oberen Grenzen der geometrischen Genauigkeit erreichen können, aber möglicherweise nicht gut für ungewöhnliche Daten verallgemeinern. Die lernbasierte Kostenmetrik und Census-SGM sind recht robust und können durchweg akzeptable Ergebnisse erzielen. Alle DL-Algorithmen sind robust gegenüber geometrischen Konfigurationen von Stereopaaaren und weniger empfindlich im Vergleich zum Census-SGM, während die lernbasierte Kostenmetrik auf Satellitenbildern verallgemeinert werden kann, wenn sie auf verschiedenen Datensätzen (Luft- oder Bodenanansicht) trainiert wird.

**Resumen**

Los métodos de correspondencia estéreo de aprendizaje profundo (DL) en conjuntos de datos satelitales de detección remota han ganado gran interés. Sin embargo, la mayoría de los estudios existentes, basadas en evaluaciones de solo pocos pares de imágenes estereoscópicas, concluyen que carecen de una evaluación sistemática sobre la robustez de los métodos de DL en imágenes estereoscópicas satelitales con configuraciones radiométricas y geométricas variables. Para comprender de manera integral la precisión, robustez, capacidades de generalización y el potencial práctico, en este documento se proporciona una evaluación de cuatro métodos de correspondencia estéreo DL a través de cientos de pares estéreo de satélites de múltiples fechas y localizaciones con configuraciones geométricas variables, contra el tradicional Census-SGM (coincidencia semiglobal) ampliamente empleado. Los métodos DL incluyen una métrica de costos basada en el aprendizaje a través de redes neuronales convolucionales (MC-CNN) seguidas de SGM, y tres modelos de aprendizaje de extremo a extremo (E2E) que utilizan Geometry and Context Network (GCNet), Pyramid Stereo Matching Network (PSMNet). ) y LEAStereo. Nuestros experimentos muestran que los algoritmos E2E pueden alcanzar límites superiores en precisión geométrica, mientras que pueden no generalizarse bien para datos no vistos. La métrica de costos basada en el aprendizaje y el Census-SGM son bastante robustos y pueden alcanzar
INTRODUCTION

Stereo dense image matching (DIM) has been an active area of study over the years, as it offers a cost-effective and efficient approach to generate digital surface models (DSMs) for applications such as 3D modeling, forestry mapping, and change detection (Albanwan & Qin, 2020; Albanwan et al., 2020; Furukawa & Hernández, 2015; Navarro et al., 2018; Nebiker et al., 2014; Tian et al., 2014; Wang et al., 2017). This is especially relevant when these applications are fueled by satellite-based 3D reconstructions due to their wide data coverage and consistent data collection over time (Albanwan & Qin, 2020; Qin, 2017). However, it has been noted (Brown et al., 2003; Qin, 2019; Seitz et al., 2006) that the quality of the DSM not only depends on a specific DIM algorithm, but also on various factors of stereopairs including: (1) sensor and image characteristics (e.g., spatial and radiometric resolutions), (2) acquisition conditions including the atmosphere and position and orientation of the sun, camera, and objects, and (3) scene structure and texture across different geographical regions, for instance, different land patterns including urban, suburban, or forest areas, water surfaces, and parallaxes formed by stereopairs, etc. Although there have been studies evaluating the achievable quality of these satellite-based reconstructions using
different DIM algorithms, most of them only take one or a few pairs for quality assessment, thus the resulting conclusions are often not comprehensive to cover data with various stereo configurations appearing in practice.

A typical stereo DIM algorithm performs disparity estimation on rectified stereo-images (i.e., epipolar images), which generally follows several key steps, including cost matching, aggregation, disparity optimisation, and filtering (Scharstein et al., 2001). Cost matching is the main step for measuring feature similarities for disparity computation. For the last decade, Census has been known as the classical cost-matching metric for stereo DIM problems and has been studied intensively due to its robustness to radiometric differences, at the same time it can be computed efficiently (Chen et al., 2019; Ma et al., 2013; Xia et al., 2018; Zabih & Woodfill, 1994). On the other hand, recently developed deep-learning (DL) algorithms have shown a promising performance that is assumed to surpass classical cost metrics like Census (Chen et al., 2019; Hamid et al., 2020). There are mainly two ways that DL algorithms are implemented in stereo DIM, for example, (1) learning-based cost metric and (2) end-to-end (E2E) learning. The learning-based cost metric learns similarity from image patches given examples of similar and dissimilar patches, such methods were first introduced by Žbontar and LeCun (2015) through convolutional neural networks (MC-CNN) as learnable feature extractors. Like Census, such patch-based cost metrics do not handle texture-less regions well, and often require multi-stage optimisation after cost computation, followed by cost aggregation for regularisation and smoothness of the disparity map (Bobick & Intille, 1999; Hirschmuller, 2005, 2008; Kolmogorov & Zabih, 2001). There are plenty of cost aggregation methods ranging from local, semi-global, and global approaches. A well-known example is the semi-global matching (SGM) algorithm (Hirschmuller, 2005), which is known to effectively leverage both accuracy and speed well, thus nowadays it is used as a standard cost aggregation approach. However, even with cost aggregation, studies found that MC-CNN still suffers from poor performance in ill-posed regions with occlusion, high-reflective surfaces, lack of texture, and repetitive patterns (Chen et al., 2019; Ma et al., 2013; Xia et al., 2018; Zabih & Woodfill, 1994) demanding further post-processing and refinement. Alternatively, E2E learning methods directly generate disparity from stereopair rectified images without additional optimisation steps. They have become a popular line of stereo DIM algorithms because they can directly predict highly accurate disparity maps through learning from geometry and context (e.g., cues such as shading, illumination, objects, etc.) rather than low-level features (Chang & Chen, 2018; Cheng et al., 2020; Gu et al., 2020; Kendall et al., 2017; Xu & Zhang, 2020; Zhang, Prisacariu, et al., 2019). Examples of this type of work are Geometry and Context Network (GCNet), Pyramid Stereo Matching Network (PSMNet), and LEAStereo, which due to their high accessibility and performances (i.e., winning the best rank in the KITTI 2012 and 2015 leaderboards (Chang & Chen, 2018; Cheng et al., 2020; Kendall et al., 2017)), have become popular in the field.

However, DL algorithms often suffer from generalisation problems, at the same time, face challenges to process large-volume and large-format remote sensing images, limiting their practical values (Pang et al., 2018; Song et al., 2021; Zhang, Qi, et al., 2019). The degree of generalisation may vary with both DL models and tasks. For instance, studies indicate that most E2E approaches enjoy deep feature representations, which, however, may encode scene-specific information (Pang et al., 2018; Song et al., 2021; Zhang, Qi, et al., 2019), thus may poorly perform on unseen scenes or data (Song et al., 2021). Most of the existing DL stereo DIM algorithms are trained using Computer Vision (CV) benchmark datasets such as KITTI or Middlebury, etc. (Geiger et al., 2012; Hamid et al., 2012; Geiger et al., 2012; Hamid et al., 2020), which mainly consist of images of everyday scenes. These images are distinctively different from satellite images in terms of scene content, view perspectives, and object granularity, thus DL models trained from these CV datasets may not perform well when applied to satellite datasets. However, acquiring a large number of high-quality satellite datasets for training is a challenge mainly due to its high cost. Additionally, the scene content of the earth’s surface is extremely diverse across the entire globe, thus it is difficult to create a comprehensive dataset by only using data from a few regions, although these regions can be as large as an entire city. A few remote sensing benchmarks such as the IARPA (Intelligence Advanced Research Projects Activity) Multiview stereo 3D mapping challenge (Bosch et al., 2016) and the 2019 Institute of Electrical and Electronics Engineers (IEEE) Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest (DFC) (Le Saux, 2019) provide the ground-truth dataset in the form of lidar point clouds and raster DSMs, which must be post-processed and converted to
the ground-truth disparity for training. Unfortunately, this step may generate undesired errors in the training data due to geometric errors in the orientation parameters, inconsistencies in the temporal and spatial resolutions, or other operations such as triangulation, projection, and interpolation (Bosch et al., 2016; Cournet et al., 2020; Patil et al., 2019; Wu et al., 2021).

In addition, a significantly under-evaluated criterion for DIM algorithms on satellite images is their robustness to varying stereo configurations such as sun illuminations, intersection angles, different sensors, etc. A robust DIM algorithm pertaining to these factors is extremely important because this will allow us to make full use of the already limited satellite images (as compared to everyday images). Ultimately, we want these algorithms to be agnostic and less selective to stereo configurations of data. It has been reported that the result of a typical stereo algorithm, i.e., Census with SGM, directly correlates with geometrical acquisition factors such as sun angle difference and intersection angle (Qin, 2019). Unfortunately, the same analysis has not been covered for DL algorithms. Since DL methods (especially E2E ones) can learn context, it is of interest to understand their performance under varying stereo configurations.

In this paper, we aim to comprehensively explore the limitations and strengths of recent stereo DIM algorithms for satellite datasets. We consider representative DIM algorithms of three main categories: (1) traditional approaches (e.g., Census cost metric with SGM), (2) DL-based cost metrics (e.g., MC-CNN cost-matching metric with SGM), and (3) three E2E learning methods (e.g., GCNet, PSMNet, and LEAStereo). Since “Census + SFGM” has been well studied and widely used in satellite stereophotogrammetry (Qin, 2019), it serves as a baseline method in this study. DL-based cost metric is regarded as a simpler task than E2E learning for DIM because it learns patch-level similarity as a binary classification problem (similar or not similar). The one algorithm in this category is MC-CNN, which applies a Siamese network to learn the similarities (Žbontar & LeCun, 2015). Despite the very many E2E methods in the CV community (Laga, 2019), we choose three state-of-the-art (SOTA) methods that are frequently used by the community, and well-performed in the leaderboard (Chang & Chen, 2018; Cheng et al., 2020; Kendall et al., 2017), and have well-organised codes available. To achieve a comprehensive evaluation and analysis, we use nine satellite datasets from different locations, and each dataset contains ~100–500 stereopairs with their respective ground-truth lidar data. We train and test the models on the same and different datasets, and analyze the results to understand their performance, robustness, and generalisation. To be more specific, this paper presents three contributions:

1. We comprehensively evaluate five stereo DIM algorithms (including four DL approaches) on satellite stereo-images using hundreds of pairs from nine test sites, to inform the community of the performance of such DIM algorithms under varying configurations.
2. We analyze the accuracy of the evaluated DL methods and study their robustness against stereo configurations of data that were reported to be critical for the resulting accuracy of DSMs for traditional methods (Qin, 2019).
3. We study the generalisation capability (or transferability) of these DL stereo DIM methods trained on and applied to datasets across different geographical regions and resolution/sensors (including satellite, airborne, and ground-view images).

The remainder of the paper is organised as follows: the section “Related Works” introduces relevant work including a brief review of stereo DIM algorithms and existing comparative studies. The section “Dataset, Preprocessing, and Experimental Setup” describes the dataset, preprocessing methods, and experimental data and setup as used in this work. The section “The Dense Stereomatching Methods for Evaluation” describes in detail the dense stereo DIM algorithms and the evaluation method. The section “Experimental Results and Discussion” presents the results, evaluation, and discussion. Finally, the conclusion, limitations, and potential future directions are discussed in the “Conclusion” section.
RELATED WORKS

Stereo DIM algorithms

There has been a tremendous development in stereo DIM algorithms over the years; they are broadly classified into traditional and DL methods (Zhou et al., 2020). Traditional methods are the very early algorithms with basic cost-matching metrics as the sum of absolute differences (SAD), normalised cross-correlation (NCC), mutual information (MI), and Census transformation (Brown et al., 2003; Seitz et al., 2006). With the development of DL methods, the cost-matching pipeline was replaced by convolutional neural networks (CNN) (Žbontar & LeCun, 2015). DL-based algorithms have attracted great attention due to their superior performances in benchmark testing (Geiger et al., 2012). Depending on the task of learning, DL stereo methods can be further categorised into learning-based cost metrics (Žbontar & LeCun, 2015, 2016) and E2E learning (Chang & Chen, 2018; Cheng et al., 2020; Kendall et al., 2017; Xu & Zhang, 2020; Zhang, Prisacariu, et al., 2019). Learning-based cost metric was first introduced by Žbontar and LeCun (2015) to learn similarities from image patches. Both traditional and learning-based algorithms process low-level features as intensity or gradient patches to indicate similarity. As a result, their performance is limited to repetitive patterns and texture-less regions. Thus, post-processing like cost aggregation, optimisation, and refinement based on these metrics is necessary to enhance the quality of the disparity map (Hirschmüller, 2005, 2008; Scharstein et al., 2001). DL methods rapidly evolved to E2E learning algorithms, where their main contribution is to replace the classical multi-stage optimisation with a trainable network to directly predict the disparity from stereo-images (Hamid et al., 2020; Laga, 2019). The underlying concept is that these neural networks can directly capture more global features, hence, they may better perform (Chang & Chen, 2018; Cheng et al., 2020; Kendall et al., 2017). In addition to these intensively studied methods, there are a few methods that perform context learning for part of the traditional pipeline but do not fully fall into either of these DL categories, for example, SGM-Net (Seki & Pollefeys, 2017) learns the per-pixel smoothness penalty and GA-Net (Zhang, Prisacariu, et al., 2019) learns networks to guide the cost-aggregation process.

Existing comparative studies of stereo dense matching algorithms

Most of the existing review papers on stereo DIM algorithms take a single to a few stereopairs for evaluation (Hamid et al., 2020; Laga, 2019; Xia et al., 2020; Zhou et al., 2020), which may be insufficient to provide an accurate and conclusive evaluation. There are a few but limited studies concerning the use of more pairs to study DIM algorithms: they indicate that the quality of the DSM is significantly correlated with the radiometric and geometric characteristics and the configurations of the stereopairs (Facciolo et al., 2017; Qin, 2019; Yan et al., 2016; Zhou & Boulanger, 2012). For instance, Facciolo et al. (2017) found that selecting stereopairs based on specific heuristics as minimum seasonal differences improves DIM and reduces the uncertainties in the DSM. In Qin (2019), the author observed a direct relation between the geometric configurations at the time of acquisition as the sun angle difference and intersection angle (base–height ratio) and the accuracy of DSMs. In addition, the spatial distribution of objects in space (e.g., building density, sizes, and distances) and land cover types (e.g., trees, roads, water surfaces, etc.) are highly diverse across different test sites (Chi et al., 2016), hence, may impact the performance of stereo DIM algorithms. With satellite images being rich in content and acquisition configurations, it is necessary to analyze the performance of stereo DIM algorithms on a large number of datasets covering a variety of regions, complexities, and configurations to understand their practical values in various applications.
Deep-learning training models and generalisation

Despite the superiority of DL algorithms in DIM, the generalisation issue remains a major challenge. Learning across different domains such as data collected from different sensors, data of different locations, data with different spatial resolutions, etc., often leads to a deep drop in the accuracy of DIM because of their inability to predict disparity from unseen data (Pang et al., 2018; Song et al., 2021). One possible solution is to encapsulate a large number of training datasets covering all scenarios and instances that a network may encounter (Najafabadi et al., 2015). However, in practice, obtaining a large training dataset with their ground-truth (such as lidar) is costly and often unavailable for large-scale areas (Chi et al., 2016), it also requires manual or post-processing to convert to the ground-truth disparity which may produce systematic errors (Cournet et al., 2020; Patil et al., 2019; Song et al., 2021). Although some benchmark datasets are available (Bosch et al., 2016; Le Saux, 2019; Rottensteiner et al., 2012), existing evaluations are mostly performed on a dataset-by-dataset basis. Moreover, these datasets, although seem to be large in data volume, mostly present typical urban scenes of a few major cities and remain very sparse when considering the generalisation in terms of scene contexts around the globe, sensor types, and resolutions.

DATASET, PREPROCESSING, AND EXPERIMENTAL SETUP

In this section, we will first describe the datasets used in this work. Then, we will describe in detail the preprocessing methods and experimental setup applied to comprehensively train and evaluate the DL-based stereomatching algorithms.

Dataset description

In this work, we have two types of data that are used for training and evaluation purposes. All our data are collected from publicly available benchmarks. First, we will describe our evaluation datasets and then we will describe the training datasets.

The evaluation is based on satellite images from IARPA (Bosch et al., 2016) and the 2019 DFC (track 3) (Le Saux, 2019) benchmarks. They provide stereo-images from the WorldView-3 satellite sensor with a spatial resolution of 0.3 m. In addition, they provide the airborne lidar which we convert to the ground-truth DSM and use for evaluation. The IARPA benchmark provides 50 overlapping images for a 100 km² area near San Fernando, Argentina (ARG) collected between January 2015 and January 2016. The 2019 DFC benchmark provides 16–39 overlapping images for 100 km² of Omaha, NE, USA (OMA) and Jacksonville, FL, USA (JAX) collected from September 2014 to November 2015. We select three sub-regions as test sites from the OMA, JAX, and ARG datasets. Our selection covers test sites with a variety of densities, complexities, land types, and covers. The 16–50 stereo-images provided for all test sites, yield approximately 6278 stereopair images. After omitting stereopairs with extremely small intersection angles (i.e., <2°) and those that fail during feature matching for relative orientation (Kuschk et al., 2014) due to large radiometric variations and clouds, we kept 2861 stereopairs for the experimental comparison. The stereopair images have a wide range of geometric configurations. The sun angle difference ranges between 0 and 50° and the intersection angle ranges between 0 and 67°. Figure 1a provides detailed information about the evaluation dataset and test sites.

The training dataset is used to train the DL algorithms. We include a variety of datasets from satellite, airborne, and ground-view sensors. For training with a satellite dataset, we use another set from the 2019 DFC benchmark known as track 2. The dataset provides 4293 rectified stereopair images of size 1024 x 1024 with their ground-truth disparities derived from lidar. We also use datasets that are distinctly different from satellite images including ground-view images from the KITTI dataset (Geiger et al., 2012) and airborne dataset from the Toronto, Canada ISPRS (International Society for...
Photogrammetry and Remote Sensing) benchmark (Rottensteiner et al., 2012). The airborne dataset contains 13 images from an aerial block, captured using the UltraCam-D camera covering an area of 1.45 km$^2$ at a ground sampling distance (GSD) of 15 cm. Each image has a size of 11,500 x 7500 pixels and an intersection angle between neighboring images ranges from 15 to 30°. The provided images have a varying overlap, thus in total, we have eight stereopair images. The ISPRS airborne dataset also provides lidar point clouds from Optech’s airborne laser scanner (ALS), which we converted to ground-truth disparity and DSM for training and testing. The Toronto dataset includes a mixture of small to high-rise buildings, as well as other classes such as roads, trees, grass, etc. For more details refer to Figure 1b.

### Ground-truth data derivation and preprocessing

The ground-truth disparities for the satellite and airborne datasets are derived from lidar point clouds following three steps (Patil et al., 2019). First, the point clouds are aligned to the original unrectified images using correlating
methods to register the images and the lidar intensity data, such as using the MI. The offsets between these two sources are further used to adjust their corresponding RPC parameters of the satellite images; second, the stereopair images are rectified using their adjusted RPC models to the epipolar image space; third, the 3D lidar point clouds are projected to the satellite images and mapped to the rectified left and right images; finally, the disparity map is computed for both the left and right rectified image based on the projected 3D points. Obtaining disparity maps with sub-pixel accuracy from the lidar on the satellite images is very challenging. According to Wu et al. (2021), the transformation between lidar and ground-truth disparity may produce additional errors. Whereas, Patil et al. (2019) report that the ground-truth disparity in the 2019 DFC benchmark (Bosch et al., 2019) lacks adequate validation, thus may include errors that have not been acknowledged. A common source of error is the change of scene content between lidar and satellite images due to different acquisition times. As can be noticed in the example in Figure 2, the yellow rectangles show some trees in the ground-truth disparity that are not recorded in the image. Other sources of errors include missing points in the ground-truth disparity due to occlusions or shadows (see the blue rectangle in Figure 2), random noise and artifacts (see red and green rectangles in Figure 2), or systematic errors from the image rectification process. These errors will inevitably impact on the training of the DL models. Therefore, we minimise some of these errors by pre-filtering the disparity map using a 3-sigma rule, and we fill the small holes according to the nearest neighbors to create a smooth disparity map for training purposes.

**Experimental data and setup**

**Training data**

The satellite-training dataset is from the 2019 DFC (track 2) benchmark. It has 4293 rectified stereo-images of size $1024 \times 1024$ pixels with their ground-truth disparity maps. Because of limited GPU memory, we crop the images to small patches with $W = 1248$ pixels and $H = 384$ pixels, yielding a total number of 25,705 training patches. We then split the input data into 80% training and 20% testing. The airborne training dataset includes the Toronto dataset from the ISPRS benchmark. There are eight stereo-images with sufficient overlap, each with a size of $11,500 \times 7500$ pixels. We rectify these images and crop them to small patches of size $W = 1248$ pixels and $H = 384$ pixels. The total number of patches is around 3000.

There are two types of DL methods based on our taxonomy (i.e., learning-based and E2E learning), and both are trained using two different datasets. MC-CNN is supposed to be trained using both the satellite and the airborne dataset. However, considering that the ground-level KITTI dataset should be more distinctively different, and it has been well-trained already as shared in its original work (Žbontar & LeCun, 2016), we therefore alternatively use the KITTI dataset and satellite dataset as more challenging testing. The same was attempted for E2E, while this overwhelmed the E2E algorithms to yield any meaningful results. Therefore, we only train the E2E networks using satellite and airborne datasets.

![Figure 2](image-url)  
**Figure 2**  Examples of the errors in the ground-truth disparity derived from lidar.
Training setup

The training parameters and settings for each DL network are shown in Table 1. The settings include preprocessing steps such as color normalisation and random crop of the input images to smaller patch size, in addition to the parameters selected prior to training such as the type of optimiser, learning rate, batch sizes, etc.

THE DENSE STEREOMATCHING METHODS FOR EVALUATION

In this section, we describe the main stereomatching approaches used in our analysis. First, we will briefly present the selected stereo DIM algorithms including: (1) Census + SGM, (2) MC-CNN + SGM, and (3) E2E methods using GCNet, PSMNet, and LEAStereo. Then, we will describe the evaluation method as applied in our analysis. In particular, MC-CNN has two architectures called MC-CNN-fst and MC-CNN-acc (Žbontar & LeCun, 2016); the former uses a product of two feature vectors to decide the similarity and the latter uses fully connected layers following the concatenated feature vectors. Although the latter was reported to be slightly more accurate, we use the former version (MC-CNN-fst) in our evaluation, given its good leverage of performance and speed for large-sized satellite images. Without specifically stated, MC-CNN mentioned hereafter refers to the MC-CNN-fst version.

Stereo DIM algorithms

As mentioned in the previous "Stereo DIM Algorithms" section, we evaluate two types of DL approaches, (1) DIM methods following a classic multi-stage paradigm including feature extraction, cost matching, and cost aggregation for disparity computation, where the DL component can be one of these stages, for example, the learning-based cost metric in our evaluation learns the feature extraction, or (2) DL methods with E2E learning which use trainable networks to learn from stereo data and ground-truth disparities, and directly predict disparity maps. All approaches require rectified epipolar images as inputs, which we generate using RPC Stereo Processor (RSP) software (Qin, 2016). As mentioned in the Introduction, we use SGM for cost aggregation following Census and MC-CNN cost metrics as proposed by Hirschmuller (2005, 2008). In this subsection, we briefly introduce these stereo DIM algorithms used for our evaluations.

Cost metrics for classic multi-stage stereo DIM methods

Census maps a window of pixels across left and right images to find similar features with minimum cost (Hirschmuller, 2005; Zabih & Woodfill, 1994). First, a fixed window \( w \) is set around every pixel in the left image and a sliding window scans the same epipolar line in the right image. The pixels in \( w \) are transformed into a vector of binary strings. Depending on their values relative to the central pixel, a value of 1 is assigned if a pixel is larger than the central pixel and 0 otherwise. The binary vectors are processed by a hamming distance and summed to compute the cost and indicate the number of variant pixels. A match is selected based on the minimum cost. A detailed example of the steps is shown in Figure 3a.

MC-CNN uses a Siamese convolutional neural network to compute the matching cost as shown in Figure 3b. The training is performed using small patches of size \( 9 \times 9 \) extracted from grayscale left and right images and their ground-truth disparity maps. The testing, on the other hand, takes the entire image as input. During training, the network learns from each pair of patches by classifying them into similar or dissimilar. The patches are processed separately in two sub-networks that include four convolutional layers with a rectified linear
| Input parameters                  | MC-CNN               | GCNet\textsuperscript{a} | PSMNet\textsuperscript{a} | LEASTereo\textsuperscript{a} |
|----------------------------------|----------------------|---------------------------|-----------------------------|-----------------------------|
| Color normalisation              | Grayscale patches    | RGB normalisation         | RGB normalisation           | RGB normalisation           |
| Patch size                       | 9×9                  | 512×256                   | 512×256                     | 192×384                     |
| Optimiser                        | Gradient Descent     | Adam                      | Adam                        | Stochastic Gradient Descent |
| Optimisation parameters          | $\beta = 0.9$        | $\beta_1 = 0.9, \beta_2 = 0.999$ | $\beta_1 = 0.9, \beta_2 = 0.999$ | $\beta = 0.9$ |
| Learning rate ($\lambda$)        | $\lambda = 0.002$    | $\lambda = 0.001$        | $\lambda = 0.001$          | Cosine $\lambda = 0.025–0.001$ |
|                                  | Decay: 0.9 (after 10 epochs) |                        |                             | Decay: 0.0003               |
| Maximum disparity                | –                    | 192                       | 192                         | 192                         |
| Batch size                       | 1                    | 1                         | 1                           | 1                           |
| # Epochs for training using satellite dataset | 10                   | 35                        | 10                          | 57                          |
| # Epochs for training using airborne dataset | –                    | 500                       | 20                          | 300                         |

\textsuperscript{a}E2E learning algorithms are implemented in Pytorch and trained on Nvidia GeForce RTX 2080 Ti.
activation function (ReLU), a convolutional layer, and a normalisation layer. The sub-networks generate two normalised feature vectors that are fed into a cosine dot product to compute their similarity and indicate a good or bad match. In terms of model parameters, MC-CNN has a small model capacity with a total number of parameters of 0.3 million (M).

**E2E learning stereo DIM methods**

Three recently published E2E networks (with available codes) are described in this subsection. These methods vary in their model architecture, backbone building blocks, and model complexity (model capacity). Some of the architectures are designed to readily model the conceptual nature of depth sensing and recognition, such as the pyramid architectures (Chang & Chen, 2018) that process the data in a coarse-to-fine fashion. Understanding the model capacity is useful to determine the degree of generalisation and applicability in practice. Generally, a network with a large number of parameters reduces the training error, increases the computational cost, and most likely has a high testing error, thus, leading to overfitting and poor generalisation problems (Huang et al., 2021).

**GCNet** involves several steps to generate the disparity map. It begins with unary feature extraction from the left and right images using 2D convolutional layers (CNNs) and eight residual blocks. Then, it matches these features at every disparity level by concatenating both feature maps to build the 4D cost volume. The cost volume is then regularised via cost aggregation using 3D CNNs to learn contextual information and refine the cost volume. Finally, it applies regression to predict the final disparity map (Kendall et al., 2017). The total number of model parameters for GCNet is 3.5 M, which is the highest among the three selected E2E algorithms. For more details on the network’s architecture, see Figure 4.

**PSMNet** starts by extracting feature maps using three CNN layers and four residual blocks. These features are then processed hierarchically on multiple scales through the Spatial Pyramid Pooling (SPP) module to generate the 4D cost volume. The SPP module can enhance the visibility and feature matching of deformed objects (He et al., 2014). The cost volume is then processed by a stacked 3D hourglass module for cost aggregation and regularisation, which can provide better learning of the context information. The final step is to apply regression to estimate the disparity map (Chang & Chen, 2018). PSMNet has a total number of model parameters of 2.8 M. For more details on the network’s architecture, see Figure 4b.
A COMPARATIVE STUDY ON DL METHODS FOR SATELLITE DATASETS

LEAStereo uses a hierarchical neural architecture search (NAS) pipeline to find the optimal stereomatching network parameters (Cheng et al., 2020). The optimal parameters determined by NAS may include the filter size of the convolutional layer, strides, etc. The stereomatching network consists of two sub-networks, first, the feature net, which is used to extract the features from the left and right image to generate the 4D feature volume, which is then processed by the matching net to compute the matching cost (using 3D convolutions) to generate the 3D cost volume. Finally, regression is applied to compute the final estimates of the disparity map. LEAStereo has the least number of model parameters about 1.81 M. For more details on the network's architecture, see Figure 4.

FIGURE 4 E2E learning stereo DIM methods (a) GCNet, (b) PSMNet, and (c) LEAStereo.

Evaluation of the stereo DIM

For evaluation and comparison of the stereo DIM algorithms, we triangulate the predicted disparity to DSMs. We assess each DSM $I(x,y)$ against the ground-truth DSM from lidar $GT(x,y)$ using the root mean square error (RMSE) as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N \times M} (I(x,y) - GT(x,y))^2}{N \times M}}$$

where $(x,y)$ are the pixel positions and $(N,M)$ are the DSM dimensions.

EXPERIMENTAL RESULTS AND DISCUSSION

Based on the DSMs generated using different stereo DIM algorithms for the entire nine test sites, we assess each class of the algorithms in the following three aspects:

1. Overall performance: we evaluate the RMSE of the resulting DSMs of each algorithm on each of the few hundred pairs based on the average RMSE, error distribution, completeness rate, and visual analysis.
2. The robustness with respect to varying acquisition configurations: we assess for each algorithm, the impact of two critical geometric/radiometric parameters, i.e., sun angle difference and intersection angle, on the resulting DSM through correlation analysis using hundreds of stereopairs.
3. Generalisation capability of DL algorithms: we evaluate for each of the DL methods, their RMSE against the ground truth by training the model on one dataset and applying it to another dataset (of a different location and/or different sensor).

Note that co-registration can be a critical factor for this evaluation as well as many other applications such as change detection and multimodal data fusion (Lv et al., 2021, 2022). To achieve reasonable co-registration, we first aligned the DSMs to a reference DSM. Following the method proposed by Gruen and Akca (2005), we applied a simple version of least-squares matching to optimise the shifts in all three dimensions. Blunders are eliminated progressively by using a predefined threshold of 5 m, computed as a factor of 10 times the GSD. DSMs are considered well-registered and unbiased when their mean difference to the reference DSM is smaller than a pixel-size, i.e., 0.5 m.

The overall performance analysis

Performance analysis is the key factor to assess the reliability of stereo DIM algorithms. We use standard statistical measures such as the average RMSE and distribution of errors to analyze the DSMs from the 2681 stereopairs over the nine test sites. We are particularly interested in the achievable upper and lower bounds of the accuracy for all stereopairs. We also provide an analysis of the completeness rate and visual results. Based on knowledge from prior literature, we expect DL algorithms to outperform traditional methods (i.e., Census+SGM) since they are more complex and trained on examples of ground-truth data.

Table 2a presents the average errors (RMSE) of the DSMs from all stereopairs and test sites. We can notice that E2E algorithms can achieve the minimum errors. This can be obvious from several aspects. First, they have minimum overall average errors as shown from PSMNet and LEAStereo where their values are 4.73 and 4.83 m, respectively (see the last row in Table 2a). Second, eight of nine test sites have the lowest errors in one of GCNet, PSMNet, and LEAStereo (indicated in bold in Table 2a). LEAStereo in particular has the most frequent lowest average RMSE among all test sites. We can also notice that E2E algorithms can achieve the highest errors in some test sites. Their performance varies drastically across different test sites. Their average errors can range from 2 to 18 m. For instance, the average RMSE for GCNet, PSMNet, and LEAStereo in JAX II is 2.73, 2.76, and 4.39 m, respectively, while for OMA II it is as high as 17.56, 8.99, and 8.84 m, respectively. This implies that E2E methods do not predict disparity maps well from unseen data. Census+SGM and MC-CNN+SGM show opposite performance to E2E methods. They have more consistent average errors across different test sites close to the overall average errors (see Table 2a). In addition, Census+SGM outperforms MC-CNN+SGM, as it achieves a lower overall average RMSE of 5.50 m, as compared to 6.43 m achieved by MC-CNN+SGM (see Table 2a). This result reveals a different conclusion from the original MC-CNN (Žbontar & LeCun, 2015), which may be due to the sensitivity of DL algorithms towards noise in the ground-truth disparity maps used for training the networks (as indicated in the “Ground-Truth Data Derivation and Preprocessing” section).

While E2E methods can achieve the lower and upper bounds of the average errors, their average standard deviation in Table 2b indicates higher consistency for the error distribution of stereopairs at the same test site. GCNet, PSMNet, and LEAStereo have the lowest average standard deviations (less than 2.73 m) as can be seen in the last row in Table 2b. This is because of their ability to learn similar context and geometry, which makes them invariant to changing radiometric properties of the stereopairs. In contrast, Census+SGM and MC-CNN+SGM have higher standard deviations with averages of 4.08 and 3.40 m, respectively. This implies that similarity-learning algorithms are sensitive to the radiometric properties of the stereopair images.

In general, E2E methods are able to achieve the absolute minimum errors. Their ability to learn from context makes them robust to varying radiometric properties of the stereopairs. However, they can also have the absolute
TABLE 2  The overall performance of the stereo DIM algorithms over nine test sites presented by (a) average RMSE, (b) standard deviation of the RMSE, and (c) completeness rates of the DSMs.

| Test site ID | (a) Average RMSE (m) | (b) Standard deviation of the RMSE (m) | (c) Average completeness (%) |
|--------------|---------------------|---------------------------------------|-----------------------------|
|              | Census + SGM | MC-CNN + SGM | GCNet | PSMNet | LEAStereo | Census + SGM | MC-CNN + SGM | GCNet | PSMNet | LEAStereo | Census + SGM | MC-CNN + SGM | GCNet | PSMNet | LEAStereo |
| OMA I        | 6.55          | 5.94          | 18.11  | 4.35    | 5.20      | 6.32        | 4.16          | 9.63    | 3.19    | 3.49      |
| OMA II       | 6.52          | 7.46          | 17.56  | 8.99    | 8.84      | 4.71        | 4.31          | 7.61    | 5.22    | 4.77      |
| OMA III      | 3.32          | 3.96          | 3.30   | 2.55    | 2.14      | 2.20        | 2.04          | 1.47    | 0.75    | 0.50      |
| JAX I        | 5.75          | 10.60         | 6.79   | 5.46    | 5.97      | 1.35        | 4.52          | 2.31    | 1.48    | 1.47      |
| JAX II       | 5.19          | 7.32          | 3.31   | 3.30    | 3.23      | 4.72        | 3.03          | 0.19    | 0.21    | 0.19      |
| JAX III      | 4.29          | 7.07          | 2.73   | 2.76    | 4.39      | 3.92        | 5.08          | 0.31    | 0.33    | 0.46      |
| ARG I        | 5.31          | 5.07          | 4.25   | 5.03    | 3.88      | 4.14        | 3.39          | 1.16    | 4.08    | 2.21      |
| ARG II       | 5.84          | 5.01          | 5.13   | 5.00    | 4.84      | 4.63        | 1.27          | 0.80    | 3.99    | 3.44      |
| ARG III      | 6.69          | 5.42          | 4.78   | 5.08    | 4.98      | 4.72        | 2.76          | 1.00    | 1.73    | 1.94      |
| Overall average error | 5.50          | 6.43          | 7.33   | 4.73    | 4.83      | 4.08        | 3.40          | 2.72    | 2.33    | 2.05      |

Note: The bold numbers indicate the lowest values in (a-b) and the highest values in (c).
highest errors when learning from new unseen test sites. In contrast, traditional and learning-based algorithms may have higher errors but can provide more consistent performance across different test sites.

The completeness rate is another assessment metric for performance. It refers to the number of filled pixels over the entire test site. It indicates two main components: first, the degree to which an algorithm is able to infer disparity values in all situations like the existence of insufficient texture, occlusions, radiometric inconsistencies, etc.; second, the percentage of correctly matched points relative to the number of filled pixels. The completeness rates are shown in Table 2c. We can notice that E2E learning algorithms have the highest completeness rates (>91%), with LEAStereo in specific having the absolute highest completeness rates with an average of around 95%. On the other hand, traditional and learning-based methods have lower completeness rates in the lower 80% range. The visual illustration in Figure 5 presents sample DSMs from the ARG I test site, where the white pixels represent missing values. We can see that the DSMs from Census + SGM and MC-CNN + SGM have more missing values and less distinctive objects. Whereas, the DSMs from E2E algorithms have more filled pixels and a higher level of details with better visibility of objects as indicated by the boundaries of the building.

Robustness analysis using varying geometric configurations of stereo-image pairs

The robustness measures the ability of the stereo DIM algorithm to maintain stable performance across stereopairs with varying acquisition configurations. As reported by Qin (2019), geometric configuration parameters like the sun angle difference and intersection angle have a critical impact on the accuracy of DSMs generated using Census + SGM. Similarly, we aim to study their impact on the results of the DL stereo DIM algorithms. We use robust regression to fit a model between the RMSE of the DSMs and the geometric parameters (for all stereopairs) to measure the correlation and eliminate potential outliers. The value of $R^2 (\in [0, 1])$ represents the goodness of fit and the degree of impact. Thus, we use $R^2$ to indicate the amount of impact the geometric parameters have on the accuracy of DSMs as follows:

$$R^2 = \frac{S_0^2 - S_1^2}{S_0^2}$$

(2)

With $S_0^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$ and $S_1^2 = \sum_{i=1}^{n} r_i^2$, where $S_0^2$ is the sum of squares between the input observations $y_i$ and the mean $\bar{y}$, and $S_1^2$ is the sum of residuals $r_i$ computed from the difference between the RMSE fitted and predicted from the regression model and the actual computed RMSE. Given previous observations in Qin (2019), we expect DL algorithms to be impacted by varying acquisition configurations.

![Figure 5](visual-illustration-for-sample-dsms-from-different-stereo-dim-algorithms-for-the-arg-i-test-site-the-white-pixels-indicate-missing-values)
| Test site ID | $R^2$ with respect to the sun angle difference | $R^2$ with respect to the intersection angle |
|--------------|------------------------------------------------|---------------------------------------------|
|              | Census + SGM | MC-CNN + SGM | GCNet | PSMNet | LEAStereo | Census + SGM | MC-CNN + SGM | GCNet | PSMNet | LEAStereo |
| OMA I        | 0.80         | 0.50         | 0.02  | 0.53   | 0.39      | 0.82         | 0.50       | 0.23   | 0.55   | 0.44      |
| OMA II       | 0.27         | 0.26         | 0.03  | 0.13   | 0.08      | 0.30         | 0.27       | 0.53   | 0.25   | 0.28      |
| OMA III      | 0.31         | 0.20         | 0.13  | 0.17   | 0.11      | 0.53         | 0.29       | 0.40   | 0.10   | 0.06      |
| JAX I        | 0.35         | 0.26         | 0.01  | 0.00   | 0.00      | 0.33         | 0.18       | 0.74   | 0.72   | 0.73      |
| JAX II       | 0.53         | 0.03         | 0.21  | 0.08   | 0.09      | 0.48         | 0.12       | 0.21   | 0.19   | 0.43      |
| JAX III      | 0.63         | 0.41         | 0.07  | 0.07   | 0.03      | 0.66         | 0.42       | 0.11   | 0.27   | 0.32      |
| ARG I        | 0.62         | 0.67         | 0.11  | 0.50   | 0.58      | 0.60         | 0.57       | 0.14   | 0.49   | 0.61      |
| ARG II       | 0.41         | 0.23         | 0.52  | 0.68   | 0.67      | 0.42         | 0.17       | 0.43   | 0.67   | 0.66      |
| ARG III      | 0.76         | 0.71         | 0.70  | 0.80   | 0.79      | 0.62         | 0.69       | 0.67   | 0.81   | 0.77      |
| Average      | 0.52         | 0.37         | 0.20  | 0.33   | 0.30      | 0.53         | 0.36       | 0.38   | 0.45   | 0.48      |

Note: Bold indicates the highest $R^2$ values. All DL models are trained using the satellite dataset from the 2019 DFC benchmark.
Table 3 presents $R^2$ showing the relationship between the RMSE of the DSMs and the geometric parameters for all stereopair images in the test sites. Census + SGM shows the highest $R^2$ scores between the RMSE of DSMs and each of the sun angle difference and the intersection angle with an average of 0.52 and 0.53, respectively. In addition, five to six test sites have $R^2$ values between 0.53 and 0.82. This implies that traditional DIM methods are impacted by the geometric parameters. As for all DL algorithms, $R^2$ scores with respect to the sun angle difference are very small and their average does not exceed 0.37. While for $R^2$ with respect to the intersection angle, the learning-based MC-CNN is the least impacted by the intersection angle where it has the minimum average $R^2$ score of 0.36. On the other hand, the $R^2$ for E2E algorithms slightly increase to an average of 0.38, 0.45, and 0.48 for GCNet, PSMNet, and LEAStereo, respectively. Despite this minor increase, it is still insignificant in comparison to Census + SGM. This indicates that DL stereo DIM algorithms can be more robust to varying geometric configuration parameters that are associated with the stereopair images.

The model generalisation analysis for the DL stereo DIM algorithms

DL algorithms have difficulties generalising to new unseen scenes. This is because of the significant differences in the characteristics of images such as intensities, illuminations, scene content, textures, noise level, etc. DL algorithms may have different generalisation capabilities. To evaluate their generalisation, we use similar and different training and testing datasets, i.e., from different sensors and regions.

As mentioned in the “Experimental Data and Setup” section, due to the different architectures of the DL algorithms, we train and test the learning-based and E2E methods on similar and different datasets. Figure 6 illustrates the training and testing pipelines. For simplicity, we are going to refer to the 2019 DFC (track 2) dataset as the "satellite-training" dataset and the OMA, JAX, and ARG datasets as the "satellite-testing" dataset. We train MC-CNN using KITTI and satellite-training datasets separately and test it on the satellite-testing dataset from the nine test sites. While for E2E algorithms, we train them separately on airborne and satellite-training datasets and test them on the airborne dataset. Given that MC-CNN is a binary cost metric, we expect it will have good generalisation. As for E2E algorithms, since we use similar training and testing datasets (i.e., similar scene content as roads, buildings, etc.), we expect them to generalise well.

In Table 4a, we compare the DSMs from MC-CNN + SGM trained independently on KITTI and satellite-training datasets and applied to the satellite-testing datasets from the nine test sites. We observe a slight difference between the overall average RMSE of the two training models of 0.29 m, with MC-CNN + SGM trained using the KITTI dataset performing slightly better. In addition, five test sites have a lower average RMSE for the model trained using the KITTI dataset than the satellite-training dataset. This may be due to some imperfections in the ground-truth disparity derived from lidar as indicated in the “Ground-Truth Data Derivation and Preprocessing” section. Census + SGM seems to outperform MC-CNN + SGM (trained using KITTI and satellite-training dataset) in
| Test site ID | (a) Average RMSE (m) | (b) Standard deviation for the RMSE (m) |
|-------------|----------------------|-----------------------------------------|
|                      | Census + SGM         | Trained using satellite-training dataset | Trained using KITTI dataset | Census + SGM | Trained using satellite-training dataset | Trained using KITTI dataset |
| OMA I        | 6.55                 | 5.94                                    | 7.55                        | 4.16          | 4.31                                    | 6.77                        |
| OMA II       | 6.52                 | 7.46                                    | 7.28                        | 4.71          | 4.31                                    | 5.02                        |
| OMA III      | 3.32                 | 3.96                                    | 3.81                        | 2.20          | 2.04                                    | 2.53                        |
| JAX I        | 5.75                 | 10.60                                   | 6.44                        | 1.35          | 4.52                                    | 1.70                        |
| JAX II       | 5.19                 | 7.32                                    | 5.38                        | 4.72          | 3.03                                    | 3.68                        |
| JAX III      | 4.29                 | 7.07                                    | 5.41                        | 3.92          | 5.08                                    | 4.26                        |
| ARG I        | 5.31                 | 5.07                                    | 6.84                        | 4.14          | 3.39                                    | 5.90                        |
| ARG II       | 5.84                 | 5.01                                    | 6.32                        | 4.63          | 1.27                                    | 3.57                        |
| ARG III      | 6.69                 | 5.42                                    | 6.21                        | 4.72          | 2.76                                    | 3.96                        |
| Overall average | 5.50                 | 6.43                                    | 6.14                        | 4.08          | 3.40                                    | 4.15                        |

| Stereopair ID | Census + SGM | MC-CNN + SGM | Trained using airborne dataset | Trained using satellite-training dataset |
|---------------|--------------|-------------|-------------------------------|------------------------------------------|
|               |              |             | GCNet | PSMNet | LEAStereo | GCNet | PSMNet | LEAStereo |
| 1             | 15.34        | 18.84       | 15.89 | 17.39  | 15.63     | 34.52 | 19.01  | 65.72     |
| 2             | 14.99        | 17.82       | 14.66 | 14.86  | 14.42     | 35.20 | 18.00  | 67.65     |
| 3             | 8.13         | 11.44       | 9.08  | 9.87   | 11.49     | 36.48 | 12.04  | 74.71     |
| 4             | 11.37        | 11.32       | 12.40 | 12.40  | 19.72     | 28.47 | 19.72  | 56.49     |
| 5             | 21.49        | 29.00       | 12.96 | 16.52  | 12.50     | 29.52 | 13.85  | 56.21     |
| 6             | 6.17         | 7.78        | 6.51  | 6.98   | 7.65      | 34.52 | 41.24  | 73.56     |
| 7             | 30.41        | 29.58       | 24.32 | 28.42  | 23.34     | 29.46 | 24.62  | 29.46     |
| 8             | 23.98        | 22.62       | 20.35 | 20.72  | 20.15     | 55.41 | 22.02  | 55.41     |
| Average RMSE (m) | 16.49        | 18.55       | 14.52 | 15.90  | 15.61     | 35.45 | 21.31  | 59.90     |

Note: Bold indicates the minimum values. The satellite-training dataset refers to the 2019 DFC satellite benchmark. MC-CNN + SGM is trained using the 2019 DFC satellite benchmark. The satellite-training and airborne datasets used to train E2E models refer to the 2019 DFC and Toronto ISPRS benchmarks.
terms of the number of test sites with the minimum average RMSE. This implies that Census-SGM is more stable than MC-CNN across different test sites. Figure 7 shows an example for DSMs from both training models for OMA I test sites. It is obvious there is a large resemblance in the DSMs, however in OMA I, we can notice some errors randomly scattered over the DSM generated from the KITTI trained model. The results show a good generalisation of MC-CNN + SGM and that it can perform well on different datasets. Furthermore, it indicates the importance of the quality of the training data which can directly impact the results and must be given greater attention.

In Table 4b, we compare DSMs from E2E algorithms trained on each airborne and satellite-training dataset separately and applied to the airborne dataset. It is obvious that there is a significant jump in the average RMSE for GCNet and LEAStereo of about 20–45 m difference between the results of airborne and satellite-training models, where PSMNet shows about 6 m difference. This indicates that the generalisation degree varies based on the model architectures, and PSMNet shows significantly better results. Part of the reason is that its pyramid architecture can better adapt data with different resolutions and scales than the other two. Additionally, E2E algorithms trained on and applied to airborne datasets outperform traditional and learning-based algorithms, as they can achieve the lowest average errors (RMSE) and the most frequent minimum errors. The results from Table 4b indicate that the source of the training data plays an important role. Example results are shown in Figure 6 for GCNet, PSMNet, and LEAStereo trained using airborne and satellite-training datasets and applied to the airborne dataset. It shows LEAStereo and GCNet perform poorly when trained using the satellite-training datasets. The PSMNet reaches better results but presents many noises and outliers (see red circles in Figure 8). This implies that E2E algorithms have a poor generalisation capability and require domain-specific training datasets to achieve acceptable performance.

**CONCLUSION**

This work presents a comparative study of the traditional and DL stereo DIM algorithms, particularly for satellite datasets. We perform a comprehensive evaluation over a large volume of satellite data of 2681 stereopairs.
A COMPARATIVE STUDY ON DL METHODS FOR SATELLITE DATASETS

covering nine different regions around the world. We evaluate the stereo DIM algorithms on three bases, the overall performance, robustness to varying acquisition configurations, and generalisation of the DL models. We study three classes of algorithms including: (1) traditional algorithms (e.g., Census + SGM), (2) DL-based algorithms (MC-CNN + SGM), and (3) E2E algorithms (e.g., GCNet, PSMNet, and LEAStereo). The results of the experiments suggest the following conclusions:

1. Overall, E2E methods outperform other types of stereo methods as they can achieve the lowest and most frequent minimum errors. They have also proven to have the highest completeness rates (of greater than 92%) of DSMs and the most consistent performance within individual test sites. However, they can also achieve the highest errors (>10 m) in some test sites, possibly due to the lack of generalisation. We found LEAStereo as the top-performing algorithm, as it achieved the lowest RMSE for half of the tested pairs and has the highest completeness rates (as seen in the “Robustness Analysis Using Varying Geometric Configurations of Stereo-Image Pairs” section). In addition, E2E methods proved to be robust with respect to the acquisition configurations, as compared with traditional methods, and this may be largely due to their ability to learn contextual correlations. However, their poor performance on some of the pairs shows that these types of methods lack generalisation. Although they achieve the best results in some of the pairs, their average RMSE over all pairs being lower than Census + SGM, suggests that E2E methods are still unpredictable when processing a large number of datasets.

2. Census + SGM as a traditional method, to our surprise, outperforms MC-CNN in terms of the overall average error and the number of test sites having the minimum average RMSE. This shows an opposite conclusion to the original work of MC-CNN, the reason for which might be due to the noisy training data projected from the lidar point clouds.

3. MC-CNN + SGM (similarity-based training) proved to be robust towards acquisition configurations of the stereopair images. In addition, due to its nature of learning similarity, it has indicated experimentally in our work that it has superior generalisation capability, as long as the training datasets are noise-free.

Although a well-generalised stereomatcher with superior performance is ideal, our experiments show that in practice, the choice of stereomatcher can be case-specific. It is likely a well-trained E2E network produces superior quality results, provided that the training datasets are available, and the testing datasets are of the same type as the training datasets. The similarity-learning-based method has good generalisation capability, and if well-trained by datasets, may outperform a traditional stereomatcher, although may not be as competitive as the E2E methods. Future research may improve the weaknesses of DL stereo DIM algorithms, and potentially consider domain adaptation methods to address some of the challenges. As for the quality of the training data, since it is imperative to the performance of DL DIM algorithms, we recommend further investigation on its impact to understand their implications and provide potential solutions.

ACKNOWLEDGEMENTS

The study was partially supported by an ONR grant (Award No. N000142012141). Hessah Albanwan was sponsored by Kuwait University. The authors would like to thank Johns Hopkins University Applied Physics Lab for their support of the imagery, the IARPA for organising the 3D challenge that drove this work, the provision of the Downtown Toronto dataset by Optech Inc., First Base Solutions Inc., GeoICT Lab at York University, and ISPRS WG III/4.

ORCID

Rongjun Qin 🐘 https://orcid.org/0000-0002-5896-1379
REFERENCES

Albanwan, H. & Qin, R. (2020) Enhancement of depth map by fusion using adaptive and semantic-guided spatiotemporal filtering. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 3, 227–232. Available form: https://doi.org/10.5194/isprs-annals-V-3-2020-227-2020

Albanwan, H., Qin, R., Lu, X., Li, M., Liu, D. & Guldfmann, J.-M. (2020) 3D iterative spatiotemporal filtering for classification of multitemporal satellite data sets. Photogrammetric Engineering & Remote Sensing, 86(1), 23–31. Available form: https://doi.org/10.14358/PERS.86.1.23

Bobick, A.F. & Intille, S.S. (1999) Large occlusion stereo. International Journal of Computer Vision, 33(3), 181–200. Available form: https://doi.org/10.1023/A:1008150329890

Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G.D. & Brown, M. (2019) Semantic stereo for incidental satellite images. IEEE Winter Conference on Applications of Computer Vision (WACV), 2019, 1524–1532. Available form: https://doi.org/10.1109/WACV.2019.00167

Bosch, M., Kurtz, Z., Hagstrom, S. & Brown, M. (2016) A multiple view stereo benchmark for satellite imagery. IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 2016, 1–9. Available form: https://doi.org/10.1109/AIPR.2016.8010543

Brown, M.Z., Burschka, D. & Hager, G.D. (2003) Advances in computational stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(8), 993–1008. Available form: https://doi.org/10.1109/TPAMI.2003.1217603

Chang, J.-R. & Chen, Y.-S. (2018) Pyramid stereo matching network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5410–5418. Available form: https://doi.org/10.1109/CVPR.2018.00567

Chen, B., Qin, R., Huang, X., Song, S. & Lu, X. (2019) A comparison of stereo-matching cost between convolutional neural network and census for satellite images. ArXiv, 7, 1905.09147.

Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Drummond, T., Li, H., & Ge, Z. (2020) Hierarchical neural architecture search for deep stereo matching. NeurIPS 2020: Advances in Neural Information Processing Systems, 33, 22158–22169.

Chi, M., Plaza, A., Benediktsson, J.A., Sun, Z., Shen, J. & Zhu, Y. (2016) Big data for remote sensing: challenges and opportunities. Proceedings of the IEEE, 104(11), 2207–2219. Available form: https://doi.org/10.1109/PROC.2016.2598228

Cournet, M., Sarrazin, E., Dumas, L., Michel, J., Guinet, J., Youssefi, D., Defonte, V., & Fardet, Q. (2020) Ground truth generation and disparity estimation for optical satellite imagery. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2, 127–134. Available form: https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-127-2020

Facciolo, G., Franchis, C.D. & Meinhardt-Llopis, E. (2017) Automatic 3D reconstruction from multi-date satellite images. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, 1542–1551. Available form: https://doi.org/10.1109/CVPRW.2017.198

Furukawa, Y. & Hernández, C. (2015) Multi-view stereo: a tutorial. Foundations and Trends in Computer Graphics and Vision, 9(1–2), 1–148. Available form: https://doi.org/10.1561/0600000052

Geiger, A., Lenz, P. & Urtasun, R. (2012) Are we ready for autonomous driving? The Kitti vision benchmark suite. IEEE Conference on Computer Vision and Pattern Recognition, 2012, 3354–3361. Available form: https://doi.org/10.1109/CVPR.2012.6248074

Gruen, A. & Akca, D. (2005) Least squares 3D surface and curve matching. ISPRS Journal of Photogrammetry and Remote Sensing, 59, 151–174.

Gu, X., Fan, Z., Dai, Z., Zhu, S., Tan, F. & Tan, P. (2020) Cascade cost volume for high-resolution multi-view stereo and stereo matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2495–2504.

Hamid, M.S., Manap, N.A., Hamzah, R.A. & Kadmin, A.F. (2020) 3D iterative spatiotemporal filtering for classifica-

tion of multitemporal satellite data sets. Photogrammetric Engineering & Remote Sensing, 86(1), 23–31. Available form: https://doi.org/10.14358/PERS.86.1.23

Hirschmuller, H. (2005) Accurate and efficient stereo processing by semi-global matching and mutual information. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2, 807–814. Available form: https://doi.org/10.1109/CVPR.2005.56

Hirschmuller, H. (2008) Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2), 328–341. Available form: https://doi.org/10.1109/TPAMI.2007.1166

Huang, Z., Gu, J., Li, J. & Yu, X. (2021) A stereo matching algorithm based on the improved PSMNET. PLoS One, 16(8), e0251657. Available form: https://doi.org/10.1371/journal.pone.0251657

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., & Bry, A. (2017) End-to-end learning of geometry and context for deep stereo regression. IEEE International Conference on Computer Vision (ICCV), 2017, 66–75. Available form: https://doi.org/10.1109/ICCV.2017.17
Kolmogorov, V. & Zabih, R. (2001) Computing visual correspondence with occlusions using graph cuts. Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, 2, 508–515. Available form: https://doi.org/10.1109/ICCV.2001.937668

Kuschk, G., d’Angelo, P., Qin, R., Poli, D., Reinartz, P. & Cremers, D. (2014) DSM accuracy evaluation for the ISPRS Commission I image matching benchmark. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-1, 195–200. Available form: https://doi.org/10.5194/isprsarchives-XL-1-195-2014

Laga, H. (2019) A survey on deep learning architectures for image-based depth reconstruction. ArXiv, 1906.06113 [Cs, Eess].

Le Saux, B. (2019) Data fusion contest 2019 (DFC2019). IEEE DataPort. Available form: https://doi.org/10.21227/C6TM-VW12

Lv, Z., Li, G., Jin, Z., Benediktsson, J.A. & Foody, G.M. (2021) Iterative training sample expansion to increase and balance the accuracy of land classification from VHR imagery. IEEE Transactions on Geoscience and Remote Sensing, 59(1), 139–150. Available form: https://doi.org/10.1109/TGRS.2020.2996064

Lv, Z., Yang, X., Zhang, X. & Benediktsson, J.A. (2022) Object-based sorted-histogram similarity measurement for detecting land cover change with VHR remote sensing images. IEEE Geoscience and Remote Sensing Letters, 19, 1–5. Available form: https://doi.org/10.1109/LGRS.2022.3163169

Ma, L., Li, J., Ma, J. & Zhang, H. (2013) A modified census transform based on the neighborhood information for stereo matching algorithm. Seventh International Conference on Image and Graphics, 2013, 533–538. Available form: https://doi.org/10.1109/ICIG.2013.113

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. & Muharemagic, E. (2015) Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1). Available form: https://doi.org/10.1186/s40537-014-0007-7

Navarro, J.A., Fernández-Landa, A., Tomé, J.L., Guillén-Climent, M.L. & Ojeda, J.C. (2018) Testing the quality of forest variable estimation using dense image matching: a comparison with airborne laser scanning in a Mediterranean pine forest. International Journal of Remote Sensing, 39(14), 4744–4760. Available form: https://doi.org/10.1080/0143161.2018.1471551

Neibiker, S., Lack, N. & Deuber, M. (2014) Building change detection from historical aerial photographs using dense image matching and object-based image analysis. Remote Sensing, 6(9), 8310–8336. Available form: https://doi.org/10.3390/rs6098310

Pang, J., Sun, W., Yang, C., Ren, J., Xiao, R., Zeng, J., & Lin, L. (2018) Zoom and learn: generalizing deep stereo matching to novel domains. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 2070–2079. Available form: https://doi.org/10.1109/CVPR.2018.00221

Patil, S., Comandur, B., Prakash, T. & Kak, A.C. (2019) A new stereo benchmarking dataset for satellite images. ArXiv 1907.04404 [Cs].

Qin, R. (2016) RPC stereo processor (RSP)—a software package for digital surface model and orthophoto generation from satellite stereo imagery. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, III-1, 77–82. Available form: https://doi.org/10.5194/isprs-annals-III-1-77-2016

Qin, R. (2017) Automated 3D recovery from very high resolution multi-view satellite images. ASPRS Conference (IGTF) 2017, Baltimore, Maryland, USA.

Qin, R. (2019) A critical analysis of satellite stereo pairs for digital surface model generation and a matching quality prediction model. ISPRS Journal of Photogrammetry and Remote Sensing, 154, 139–150. Available form: https://doi.org/10.1016/j.isprsjprs.2019.06.005

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., & Breitkopf, U. (2012) The ISPRS benchmark on urban object classification and 3D building reconstruction. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, I-3(1), 923–928. Available form: https://doi.org/10.5194/isprs-annals-I-3-923-2012

Scharstein, D., Szeliski, R. & Zabih, R. (2001) A taxonomy and evaluation of dense two-frame stereo correspondence algorithm. Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), 47, 131–140. Available form: https://doi.org/10.1109/SMBV.2001.988771

Seltz, S.M., Curless, B., Diebel, J., Scharstein, D. & Szeliski, R. (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), 1, 519–528. Available form: https://doi.org/10.1109/CVPR.2006.19

Seki, A. & Pollefeys, M. (2017) SGM-Nets: semi-global matching with neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 6640–6649. Available form: https://doi.org/10.1109/CVPR.2017.703

Song, X., Yang, G., Zhu, X., Zhou, H., Wang, Z. & Shi, J. (2021) AdaStereo: a simple and efficient approach for adaptive stereo matching. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 10323–10332. Available form: https://doi.org/10.1109/CVPR46437.2021.01019
Tian, J., Cui, S. & Reinartz, P. (2014) Building change detection based on satellite stereo imagery and digital surface models. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1), 406–417. Available from: https://doi.org/10.1109/TGRS.2013.2240692

Wang, H., Jin, L., Gong, X. & Ai, J. (2017) Multi-view stereo 3D reconstruction of medical implants. *Zhongguo Yi Liao Qi Xie Za Zhi = Chinese Journal of Medical Instrumentation*, 41, 84–88. Available form: https://doi.org/10.3969/j.issn.1671-7104.2017.02.002

Wu, T., Vallet, B., Pierrot-Deseilligny, M. & Rupnik, E. (2021) A new stereo dense matching benchmark dataset for deep learning. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2, 405–412. Available form: https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-405-2021

Xia, Y., d’Angelo, P., Tian, J. & Reinartz, P. (2020) Dense matching comparison between classical and deep learning based algorithms for remote sensing data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2, 521–525. Available form: https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-521-2020

Xia, Y., Tian, J., d’Angelo, P. & Reinartz, P. (2018) Dense matching comparison between census and a convolutional neural network algorithm for plant reconstruction. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2, 303–309. Available form: https://doi.org/10.5194/isprs-annals-IV-2-303-2018

Xu, H. & Zhang, J. (2020) AANet: adaptive aggregation network for efficient stereo matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1959–1968.

Yan, L., Fei, L., Chen, C., Ye, Z. & Zhu, R. (2016) A multi-view dense image matching method for high-resolution aerial imagery based on a graph network. *Remote Sensing*, 8(10), 799. Available form: https://doi.org/10.3390/rs8100799

Zabih, R. & Woodfill, J. (1994) Non-parametric local transforms for computing visual correspondence. *Proceedings of the Third European Conference—Computer Vision*, II, 151–158.

Žbontar, J. & LeCun, Y. (2015) Computing the stereo matching cost with a convolutional neural network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, 1592–1599. Available form: https://doi.org/10.1109/CVPR.2015.7298767

Žbontar, J. & LeCun, Y. (2016) Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1), 2279–2318.

Zhang, F., Prisacariu, V., Yang, R. & Torr, P.H.S. (2019) GA-Net: guided aggregation net for end-to-end stereo matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 185–194.

Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B. & Torr, P. (2019) Domain-invariant stereo matching networks. ArXiv, 1911.13287 [Cs]. Available form: http://arxiv.org/abs/1911.13287

Zhou, K., Meng, X. & Cheng, B. (2020) Review of stereo matching algorithms based on deep learning. *Computational Intelligence and Neuroscience*, 2020, e8562323. Available form: https://doi.org/10.1155/2020/8562323

Zhou, X. & Boulanger, P. (2012) Radiometric invariant stereo matching based on relative gradients. *19th IEEE International Conference on Image Processing*, 2989–2992. Available form: https://doi.org/10.1109/ICIP.2012.6467528

---

How to cite this article: Albanwan, H. & Qin, R. (2022) A comparative study on deep-learning methods for dense image matching of multi-angle and multi-date remote sensing stereo-images. *The Photogrammetric Record*, 37, 385–409. Available from: https://doi.org/10.1111/phor.12430