


## Article

# Low-Cost and Efficient Indoor 3D Reconstruction through Annotated Hierarchical Structure-from-Motion

Youli Ding <sup>1</sup>, Xianwei Zheng <sup>1,\*</sup>, Yan Zhou <sup>1</sup> , Hanjiang Xiong <sup>1</sup> and Jianya Gong <sup>1,2</sup>

<sup>1</sup> State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China; whu\_dyl@whu.edu.cn (Y.D.); zhouyan9103@whu.edu.cn (Y.Z.); xionghanjiang@163.com (H.X.); gongjy@whu.edu.cn (J.G.)

<sup>2</sup> School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

\* Correspondence: zhengxw@whu.edu.cn; Tel.: +86-181-714-18527

Received: 13 November 2018; Accepted: 24 December 2018; Published: 29 December 2018



**Abstract:** With the widespread application of location-based services, the appropriate representation of indoor spaces and efficient indoor 3D reconstruction have become essential tasks. Due to the complexity and closeness of indoor spaces, it is difficult to develop a versatile solution for large-scale indoor 3D scene reconstruction. In this paper, an annotated hierarchical Structure-from-Motion (SfM) method is proposed for low-cost and efficient indoor 3D reconstruction using unordered images collected with widely available smartphone or consumer-level cameras. Although the reconstruction of indoor models is often compromised by the indoor complexity, we make use of the availability of complex semantic objects to classify the scenes and construct a hierarchical scene tree to recover the indoor space. Starting with the semantic annotation of the images, images that share the same object were detected and classified utilizing visual words and the support vector machine (SVM) algorithm. The SfM method was then applied to hierarchically recover the atomic 3D point cloud model of each object, with the semantic information from the images attached. Finally, an improved random sample consensus (RANSAC) generalized Procrustes analysis (RGPA) method was employed to register and optimize the partial models into a complete indoor scene. The proposed approach incorporates image classification in the hierarchical SfM based indoor reconstruction task, which explores the semantic propagation from images to points. It also reduces the computational complexity of the traditional SfM by avoiding exhausting pair-wise image matching. The applicability and accuracy of the proposed method was verified on two different image datasets collected with smartphone and consumer cameras. The results demonstrate that the proposed method is able to efficiently and robustly produce semantically and geometrically correct indoor 3D point models.

**Keywords:** indoor mapping; 3D reconstruction; semantic classification; 3D modeling; hierarchical SfM

## 1. Introduction

Indoor 3D models deliver precise geometry and rich scene knowledge about indoor spaces, which have great potential in object tracking and interaction, scene understanding, virtual environment rendering, indoor localization and route planning, etc. [1–3]. Given the rapid development of location-based services (LBS) and indoor applications, fast acquisition and high-fidelity reconstruction of complete indoor 3D scenes has become an important task [4]. Most of the current model acquisition technologies are based on light detection and ranging (LiDAR) surveys [5,6], Kinect depth cameras [7,8], or image-based approaches such as robot simultaneous localization and mapping (SLAM) [9]. Despite the improvements that have been achieved, methods that rely on professional instruments and

operation result in high capital and logistical costs [10]. In addition, outdoor reconstruction systems can usually efficiently output a city-scale model from one sampling, for example, from long-range photographs taken by unmanned aerial vehicles (UAVs) or street images captured by moving survey vehicles. However, indoor survey methods can only obtain a short-range model in a limited space, which limits the reconstruction efficiency of indoor models. Hence, in contrast to outdoor 3D models, indoor 3D model coverage remains insufficient. In order to satisfy the requirements for low-cost and large-scale indoor modeling, reconstruction methods such as Structure-from-Motion (SfM) [11], which recover 3D scene points from any unmanned images, can supplement the existing methods.

The SfM algorithm has made significant progress in city-scale model reconstruction [12]. It exploits the scale-invariant feature transform (SIFT) features, epipolar geometry, and bundle adjustment to determine the metric information and produces a point cloud model, without making any assumptions of the input image or the acquisition framework [13]. The main SfM approaches are the incremental SfM algorithms [14], which start with an image pair and then expand to the whole scene by sequentially adding related cameras and scene points. However, these incremental methods are limited by their computational efficiency, and they involve exhaustive pair-wise image matching and repeated bundle adjustment calculation. This is usually alleviated by adopting parallel computation [12], multi-core optimization [11], or by removing the redundant images to form a skeletal subset graph [15]. Other algorithms such as the revised bundle adjustment method can be used to speed up the optimization [16], and the spanning tree algorithm can be used to optimize the image connection [17], improving the efficiency of the computation. However, errors tend to be propagated in an incremental manner with the visual connections [18].

Globally optimized SfM has been one solution to this problem [19,20]. Instead of simultaneously involving all the images in the pair-wise matching, globally optimized SfM independently estimates the relative camera rotations between pair-wise views, and then uses these separate rotations to solve the camera translations and structure in the global optimization step [21]. The global pose registration approach is less sensitive to drift but is not robust to noise, and it is prone to being compromised by a bad initialization [22,23]. Another alternative solution, which is robust to drift and initialization, is to exploit a hierarchical reconstruction [24,25]. By partitioning the image dataset into smaller and more tractable components, these methods construct a tree-structured SfM where the reconstructions are executed independently and merged together along the tree into a global framework [26]. With a compact and balanced tree, these methods outperform their counterparts because they distribute the errors evenly throughout the reconstruction and bridge over degenerate configurations [27]. These methods also reduce the computational complexity by one order of magnitude [13].

With the advent of rapid and low-cost image data acquisition technologies such as smartphone cameras and crowdsourcing platforms [28], SfM has revealed its potential in indoor spaces. However, as a result of the incomplete indoor model reconstruction, the set of disconnected 3D pieces recovered from SfM has been laid on a 2D floor plan to assist with indoor sightseeing [29]. Furthermore, due to poor texture images the model develops defects in the form of disconnected parts or unwanted indentations that require the use of volumetric depth map fusion to achieve a dense reconstruction [30]. Despite the achievements made, these approaches are incapable of producing satisfactory indoor models. Unlike exterior mapping, which focuses on the flat surfaces of building facades [26], indoor reconstruction faces many challenges, including highly cluttered scenes, occlusions, and diverse structural layouts [31]. This implies a need for reconstruction approaches that not only can recover the structural layout of the indoor scenes, but also the complex semantic objects that are abundant indoors.

To fulfill these requirements, recent indoor reconstruction methods have aimed at not only recovering well-regularized 3D polygon models [32,33], but have also emphasized dense object reconstruction and semantic annotation [34,35] since it is the widespread semantic objects that define how people occupy an indoor space and how location-based services are provided. Furthermore, semantic regularities have been proven to be an efficient means for determining the geometrical structure of incomplete models [36,37] and recognizing the objects of interest and the objects'

contextual relationships in the reconstruction [38]. However, the point clouds obtained by LiDAR or RGB-D cameras are intrinsically blind to recognition and require laborious per-frame semantic labeling or additional semantic recognition. In contrast, the SfM pipeline has the advantage of one-to-one correspondence between the images and points, where the semantic information can be propagated directly.

Based on the above observations, a novel semantically guided hierarchical SfM indoor reconstruction approach is proposed in this paper, which integrates image clustering, object segmentation, and 3D point model reconstruction into the same pipeline. Firstly, a classification scheme combining bag-of-visual-words (BOVW) and the support vector machine (SVM) was applied to cluster the image dataset into classes containing a particular object. In this study, we did not need to add the extra step of employing deep learning methods for image recognition and classification since feature extraction (i.e. SIFT features) is an essential step in the SfM system, and the BOVW and SVM can make full use of SIFT features to accelerate the reconstruction process. To propagate semantic information from 2D images to 3D models, the image clusters were then arranged in an annotated hierarchical tree with each one independently reconstructed using SfM. Finally, an improved random sample consensus (RANSAC) generalized Procrustes analysis (RGPA) algorithm [26] was exploited to register and optimize the separate reconstructions into an integrated, semantically and geometrically complete 3D model. The proposed method inherits the computational efficiency and robust properties of hierarchical SfM, with further improvements that incorporate image semantic information in the data partitioning and model reconstruction. As a result, the proposed method efficiently and robustly recovers a complete indoor point model with coarse level objects and annotations from image collections.

The main contributions of the proposed method are as follows. (1) We present a low-cost and efficient indoor 3D reconstruction method using unordered images collected with widely available smart phones or consumer-level cameras, which alleviates the dependence on professional instruments and operation. (2) Unlike traditional SfM methods, we integrate image clustering, object segmentation (coarse-level), and 3D point model reconstruction into the same pipeline. (3) We perform the SfM in an annotated hierarchical manner, whereby the cluttered images are independently classified and reconstructed along a hierarchical scene tree, thus improving the computational efficiency while balancing the distribution of error. (4) We present a strategy to search for matching points while running the RGPA to align point clouds during atomic point cloud registration, which improves the efficiency and robustness of the registration process.

## 2. Methodology

In this part, we detail the annotated hierarchical SfM approach based on image classification and RGPA, which can quickly and robustly identify the widespread objects in an indoor environment as well as recover the complete 3D scene. The workflow of the proposed annotated hierarchical SfM approach is illustrated in Figure 1. The cluttered objects are recognized and reconstructed independently along a hierarchical scene tree, which recovers the indoor space. Starting with the semantic annotation of the images, the images sharing the same object are detected and classified utilizing visual words and the SVM algorithm. The SfM method is then applied to hierarchically recover the atomic 3D point cloud of each object with the semantic information from the images attached. Finally, RGPA is used to merge the separate models into a complete structure.



**Figure 1.** The workflow of the proposed annotated hierarchical SfM approach.

### 2.1. Semantic Information Extraction and Image Classification

The traditional indoor point cloud segmentation and annotation approach is based on structural inference about the “blind” points, which ignores the semantic information that inherently exists in raw image collections. While every point in the reconstructed indoor model has a corresponding pixel in the raw image, the model semantic recognition can be reformulated by assigning labels to an image according to its semantic category. Therefore, we exploited the image classification strategy to extract the semantic information in the indoor images, and we propagated this information to the point cloud of the indoor model.

Image classification is usually conducted by extracting locally invariant features with SIFT. However, due to the inherent object clutter and variation, as well as changes in the viewpoints, lighting, and occlusion of the images, indoor scene classification cannot be performed satisfactorily using pure SIFT features. This is because the SIFT descriptors are low-level local features that are not capable of characterizing a particular class. To robustly characterize the indoor scene features for classification, we combined the Fisher vector (FV)-encoded BOVW model and SVM to recognize and classify the images. The BOVW algorithm clusters similar features as a visual word, and counts the occurrence of each word in the image to form the feature vector, which improves the semantic level and enhances the expression of class-level features. The FV encoding was used in the BOVW model to encode the visual words with Gaussian mixture model (GMM) gradient vectors and derive visual words (clusters of feature descriptors) with an extended dimension, thereby reducing the number of words needed and improving their generalization, and consequently outputting more efficiently and effectively to the classifier.

Suppose that  $X = \{x_t\}, t = 1, \dots, T$  represents the feature sets of an image that contains  $T$  SIFT descriptors, then the Fisher kernel for this image is the summation of each normalized gradient vector of the local feature descriptors [39]:

$$\sum_{t=1}^T F_{\lambda}^{-1/2} G_{\lambda}^X = \sum_{t=1}^T F_{\lambda}^{-1/2} \nabla_{\lambda} \log p(X|\lambda) \quad (1)$$

where  $F_{\lambda} = E_x [G_{\lambda}^X G_{\lambda}^{X'}]$  is the Fisher information matrix [40];  $G_{\lambda}^X$  is the gradient vector of one local feature descriptor; and  $p$  is the probability density function, with the parameters denoted by  $\lambda$ .

The Fisher kernels are related to the visual vocabularies by means of the GMM, i.e.,  $\lambda = \{w_k, u_k, \Sigma_k\}, k = 1, \dots, K$ . Each Gaussian corresponds to a visual word, where the weight  $w_k$  is the number of times word  $k$  occurred,  $u_k$  represents the mean of the words, and the covariance matrix  $\Sigma_k$  is the variation around the mean.

Therefore, the occurrence probability  $\gamma_t(i)$  can be denoted by:

$$\gamma_t(i) = p(i|x_t, \lambda) = \frac{w_i p_i(x_t|\lambda)}{\sum_{j=1}^N w_j p_j(x_t|\lambda)} \quad (2)$$

While  $x_t$  is the probability observation generated by the  $i$ -th Gaussian,  $L(X|\lambda) = \log p(X|\lambda)$ ,  $\sigma_i^2 = \text{diag}(\sum_i)$ , and the subscript  $d$  denotes the  $d$ -th dimension of a vector. Then, the resulting derivation is:

$$\begin{aligned} \frac{\partial L(X|\lambda)}{\partial w_i} &= \sum_{t=1}^T [\frac{\gamma_t(i)}{w_i} - \frac{\gamma_t(1)}{w_1}] \text{ for } i \geq 2 \\ \frac{\partial L(X|\lambda)}{\partial u_i^d} &= \sum_{t=1}^T \gamma_t(i) [\frac{x_t^d - u_i^d}{(\sigma_i^d)^2}] \\ \frac{\partial L(X|\lambda)}{\partial \sigma_i^d} &= \sum_{t=1}^T \gamma_t(i) [\frac{(x_t^d - u_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d}] \end{aligned} \quad (3)$$

From these equations, it is clear that the FV-encoded visual word approach is superior to the BOVW model, as it not only considers the gradient with respect to the weight parameters, i.e., the occurrences of the  $i$ -th word, but also the means and standard deviations. By incorporating the gradient statistics into the feature vector representation, the FV approach can achieve competitive results in both efficiency and effectiveness, whereas the BOVW model would otherwise require a large quantity of words.

After obtaining the FV-encoded image feature vectors, they immediately serve as the input to SVM for the classification, which attempts to achieve the optimal separating hyperplane between two classes in order to minimize the desired classification error [41]. In detail, suppose  $D = \{(x_1, y_1), \dots, (x_m, y_m)\}$  is a set of training samples,  $x_i \in R^n$  and the corresponding decision values  $y_i \in \{1, -1\}$ ,  $i = 1, \dots, m$ , SVM aims to find the best separating hyperplane  $w^T x + b$  with the largest distance between the two classes. The problem can be equivalently formulated as:

$$\min_{w,b} \frac{1}{2} w^T w, \text{ s.t. } y_i(w^T x_i + b) \geq 1, i = 1, \dots, m \quad (4)$$

For the non-separable data, SVM handles the problem by utilizing the slack variable  $\xi_i$ . The optimization problem can then be reformulated as:

$$\begin{aligned} \min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^m \xi_i \\ \text{s.t. } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \dots, m \end{aligned} \quad (5)$$

where  $C$  is the punishment factor for the regularization, balancing the margin and the tolerance of noise. A larger value of  $C$  will assign a higher penalty to errors. In the proposed approach, the  $K$ -class problem is transferred into  $K2$ -class problems to make the approach more practical. The output of FV is passed for the training in SVM as sample  $x_i$ . We used the Gaussian kernel, which shows the best performance for SVM. The radial basis function is given by:

$$K(x_i, x_j) = \exp(-\gamma \|x_j - x_i\|^2) \quad (6)$$

In the process, principal component analysis (PCA) is exploited to compress the dimension of the feature descriptors. PCA can achieve dimensionality reduction using linear projection, while preserving the property of linear separability.

We exploited the FV-encoded BOVW model to extract more appropriate feature descriptions for the indoor space, and we classified the originally unordered mass of images with SVM. The result is a set of well-categorized images, which accordingly depict the diverse indoor objects.



## 2.2. Object Oriented Partial Scene Reconstruction

We now have well-classified images of the indoor scene. In the next step, the SfM algorithm is exploited to reconstruct the object models separately from the classified images. Indoor model reconstruction has long been limited by the scattered objects in indoor spaces. However, it is exactly these objects that play a significant role in the reconstruction. This is because the type and style of objects reflect how people occupy the indoor space. For example, computers and desks imply that an indoor space is an office; beds imply a bedroom, etc. Fortunately, the object information has already been provided in the above step, which classifies the images with semantics and indicates which object the captured image belongs to. Based on this, we constructed a tree structure that hierarchically divides and reconstructs the indoor space, with each leaf node representing the image patch of a particular object. The SfM algorithm is then applied to reconstruct the indoor object models separately and in parallel, from leaves to roots. The proposed approach combines semantic annotation, object recognition, and reconstruction in a collaborative process, which recovers the indoor model with a compact pipeline while maintaining semantic and geometric completeness.

Furthermore, the traditional incremental SfM tends to suffer from extremely high computational complexity as the image sets grow larger. On the other hand, dividing the images into smaller patches improves the computational efficiency and balances the error distribution at the same time. The SfM algorithm reconstructs the scene structure and the camera pose by extracting and matching the feature correspondences to recover a feature track from different views. Bundle adjustment, which estimates the optimal 3D structure and the calibrated parameters by minimizing a least square function, is adopted to optimize the camera parameters and feature locations.

$$\arg \min \sum \|x_{ij} - f_j P(O_j(X_i - c_j))\|^2 \quad (7)$$

where  $P$  is the projection function:  $P(x, y, z) = (x/z, y/z)$ .  $X_i (i = 1, \dots, N)$  denotes the 3D points and  $N$  is the number of points.  $O_j, c_j, f_j (j = 1, \dots, M)$  denote the orientation, position, and the focal length of the  $j$ -th camera, respectively. The SfM problem is to infer  $X_i, O_j, c_j, f_j$  from the observation  $x_{ij}$ . According to the SfM projection function  $x'_{ij} = f_j P(X_i - c_j)$ , the re-projected coordinates of the 3D point on the corresponding image can be calculated from the camera parameters, represented by  $x'_{ij} = f_j P(O_j(X_i - c_j))$ . Therefore, the BA problem can be solved by minimizing the sum of distances between the re-projected coordinates  $x'_{ij}$  of each 3D point and its corresponding image feature point  $x_{ij}$ . We solved this non-linear least squares minimization problem with the Levenberg-Marquardt algorithm [42]. Accurate initial camera estimates are the starting point when adding cameras to solve the minimum distances. First, the set of images with the largest number of matching key points are selected as the initial camera parameter, and the intrinsic parameter from image EXIF tags is used to initialize the focal length of the camera [14]. The external parameter of one of the initial cameras is set to  $[I|0]$ , the other is set to  $[R|t]$ . Next, we optimize the camera parameters by adding one camera per iteration. The camera that observed the highest number of key points that match those observed by one of the initial camera pairs is added incrementally. Finally, we optimize camera parameters by matches observed by the added new camera. The procedure is repeated until all the cameras are used for 3D reconstruction.

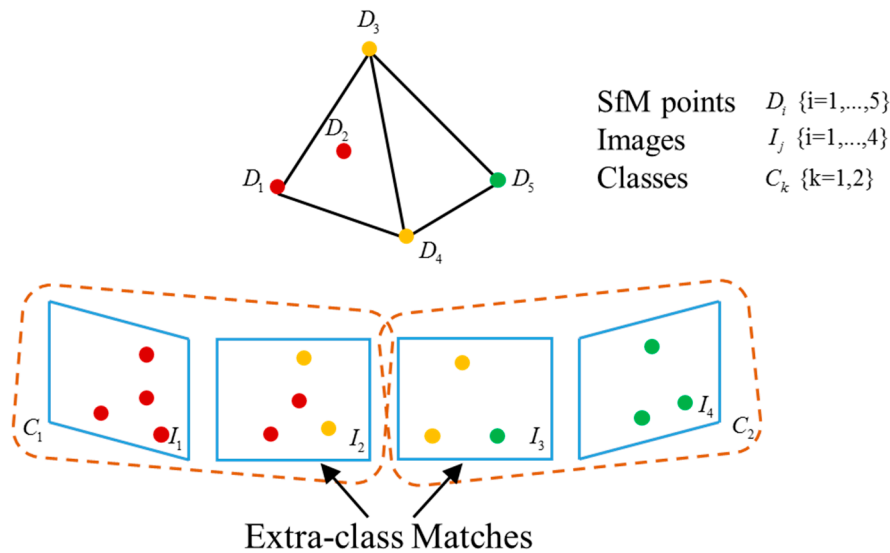
In summary, we introduced the SfM algorithm into indoor model reconstruction. Based on the concept of semantic division of the indoor image sets in the above step, the diverse objects in the indoor space are recognized and annotated to guide the object model reconstruction. In this way, we can obtain the geometrically and semantically complete point cloud model of each object existing in the indoor space in a joint semantic annotation, object recognition, and reconstruction framework.

### 2.3. Point Cloud Registration and Optimization

After obtaining the separate object models in the last step, we then merge the separate point cloud models of the obtained indoor objects into a complete indoor model using the RGPA algorithm. This complete model is used during point cloud registration to find the set of  $n$  similarity transformations  $T \triangleq \{R_1, t_1, \alpha_1, \dots, R_n, t_n, \alpha_n\}$  between the cloud points and the reference shape  $F = (F_1, \dots, F_m)$  that minimizes the cost function [26,43].

$$\varepsilon(T, F) = \sum_{i=1}^n \sum_{j=1}^m u_{i,j} \|F_j - \alpha_i R_i D_{i,j} - t_i\|_2^2 \quad (8)$$

where  $R_i$  represents the rotation matrices,  $t_i$  represents the translation vectors, and  $\alpha_i$  represents the scale factors that define the seven degrees of freedom similarity transformation. The input point clouds are represented by matrices  $D_1, \dots, D_n$ . Each  $D_i$  is composed of  $m$  three-dimensional points  $D_i = (D_{i,1}, \dots, D_{i,m})$ .  $n, m$ , are the number of point clouds and reconstructed 3D points in the point cloud, respectively.  $u_{i,j} \in \{0,1\}$  is a binary indicator that is only active when the matched pairs were detected between the cloud points. In the RANSAC generalized Procrustes analysis (RGPA) algorithm, all the models are aligned successively by alternating computation of similarity transformation and reference updating. The model to be aligned is first matched with the reference to obtain the similarity transformation based on the matched points per iteration, and then the transformed model is aligned with the reference to update the reference. RANSAC [44] is used in estimating the similarity transformation by choosing the transformation with the most inliers. The whole algorithm terminates when all the models have been aligned. The selection of matching points between point clouds is illustrated in Figure 2. Since these matching points are reconstructed from the SIFT features extracted from the images, the points are selected reversely from the images to identify the reconstructed points that match. These matched points between images are chosen as matched points between image classes. Based on the matching points, the similarity transformation matrices are computed.



**Figure 2.** An example of matched point selection.  $D_1 - D_4$  are cloud points recovered from class one, which are marked in red.  $D_3 - D_5$  are cloud points recovered from class two, which are marked in green.  $D_3$  and  $D_4$  are matched points that can be recovered from both classes, which are marked in yellow. The matched points are selected by conducting extra-class matches between images  $I_2$  and  $I_3$ . The matched feature points between images are chosen as the matched points between classes.

Since the alignment is based on the matched points between models, the problem becomes how to determine the matching points, which has not been detailed in previous work. We introduce an automatic matched points searching algorithm based on images, which reversely identify the matched points in images that reconstruct the SfM models. This intrinsically accords with the SfM pipeline in which the 3D points are re-projected by the 2D features in images. Consequently, the matched point search problem can be reformulated as feature matching of images. The process is illustrated in Figure 2. Extra-class feature matching is first conducted on the marginal images, which are defined by the images with the lowest number of matched features in the class. Extra-class matching obtains the matched connection points between atomic point clouds from the matching relationship within extra-class images and the correspondences between image pixels and 3D points determined by the SfM system. The matched features of each class then search for their corresponding points in the point clouds and finally obtain the matched 3D points between models.

After detecting the matched points between point clouds, the RGPA algorithm merges all the models through an alternating alignment and reference updating process. An arbitrary model is chosen as the reference  $F_r$  in initialization. In the alignment step, the similarity transformations are calculated by aligning each point cloud  $D_i$  with the reference  $F_r$  using RANSAC. Only the minimum matched points exceeding the number of 20 are input for similarity transformation estimation. The iteration of RANSAC trials are set to 250, which guarantees a success probability of over 99% under a conservative estimate of 40% outliers. The transformed errors below the threshold are treated as inliers per iteration. The transformation with the most inliers is selected as the result. After obtaining the similarity transformation, a new reference is updated by superimposing all the aligned models. The matched points from multiple models are averaged as the new reference points. In order to counteract noise in the point cloud and limit the convergence error of the reference to an acceptable range, the iteration is executed three times. The RGPA algorithm that aligns all the models in a group is summarized in Algorithm 1.

In contrast with other approaches that require a large overlap between the to-be-aligned point clouds, such as iterative closest point (ICP) [45], the RGPA algorithm can cope with situations with sparse overlap by only requiring moderate coverage of images. This further avoids extra local reconstruction between the point clouds or exhaustive pair-wise matching of all images to search for matching points. By inversely searching for the matching points from images, the RGPA algorithm automatically and efficiently aligns all the models. Another advantage of the RGPA algorithm is ability to counteract noise, which is achieved by the dynamic selection of matching points in the inlier estimation during the RANSAC trials and multiple iterations in the reference updating process. Specifically, the randomly distributed inliers allow the model to be resistant to outliers and avoid local optimization. In addition, the multiple iterations effectively remove outliers and maintain the accuracy of the updated reference by detecting unreliable points that fail to converge to a steady point. The outliers that are difficult to detect in the bundle adjustment in the SfM pipeline are easily detected by the point clouds merging using the iterative RGPA. Furthermore, RGPA benefits from aligning multiple cloud points with the reference shape and the constructed low-depth tree, which is computationally efficient.

Through an iterative alignment and reference update module, the RGPA algorithm can obtain a registered indoor model despite moderate noise, with reliable points gradually updated until convergence; erroneous points are rejected by crosschecking the corresponding sets of point clouds. With the above process, the separate object models are merged into a uniform and complete indoor point model.



---

**Algorithm 1.** RANSAC generalized Procrustes analysis for point cloud alignment.

---

**Input:** Group of point clouds with matched points  $S = \{D_1, D_2, \dots, D_n\}$ .

**Initialization:** Choose the reference shape  $F_r^0 = F_r$ , set iteration times numIter = 3

**for**  $i = 1 \dots \text{numIter}$

**for**  $j = 1 \dots n$

Extra-class image feature matching between the model  $D_j$  and the reference  $F_r^{i-1}$ ;

Search the matched 3D points in the point cloud based on the matched image features;

Align  $D_j$  to  $F_r^{i-1}$  using RGPA to obtain the similarity transformation  $T_j^i$  and the transformed model  $D_j^i$ .

**end**

Superimpose the aligned model to obtain the new reference  $F_r^i = \text{align}(F_r^{i-1}, D_j^i)$ .

**end**

Align all the models by using the converged reference:  $S = F_r^i$

**Output:** merged structure  $S$ .

---

### 3. Experiments

To test the proposed annotated hierarchical SfM approach for indoor scenes, two sets of experiments were conducted to reconstruct the indoor structure of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) building of Wuhan University. We first evaluated the accuracy of the indoor semantic classification using the bag-of-words based SVM classification. Then, based on the classified images, we reconstructed the semantically annotated point cloud model of the indoor scenes with SfM and the GPA algorithm, and compared the efficiency with that of the state-of-the-art algorithm. Finally, the semantically annotated model is presented.

The image datasets used in the experiments were collected by widely available smartphone cameras (iPhone 7 and XIAOMI), and Cannon EOS 6D SLR. The first dataset is a meeting room in the LIESMARS building, which includes 304 images that differ in viewpoint, size, illumination, and time. The image dataset is divided into eight predefined classes, consisting of board, elevator, door, stair, table, TV, window, and furniture. Each class contains different number of images. Exemplar images are shown in Figure 3.



**Figure 3.** (a–h) Exemplar images for each class.

We first tested the performance of the bag-of-words based SVM classification methods with the above dataset. This experiment was conducted in MATLAB based on the LIBSVM package on a Lenovo ThinkPad X240 laptop. To accurately classify the images into eight classes, three kinds of encoding

methods were used to encode the image features: BOVW [46], vector of locally aggregated descriptor (VLAD) [47], and FV [39]. Since the images belonging to a particular class were further fed into the SfM pipeline for reconstruction, retrieval precision was also a significant indicator in our experiments. We chose the RBF kernel for the SVM classification, and the one-to-all extension classification strategy. To obtain a reasonable classification result, the number of training images was no less than 15% of the whole dataset.

Table 1 reports the performance of the three classification methods. It is clear that all the bag-of-words based classification approaches achieve satisfactory results, while the FV-encoded approach outperforms the others. For the BOVW-based classification methods, the classification accuracy improves with the increase of the number of words. However, the computation time also increases accordingly. Compared to BOVW, which represents the feature vector with the original 128-dimension SIFT descriptors, the VLAD and FV methods convert the images into  $K \times D$  dimensional vectors. In FV,  $K$  represents the number of GMMs, and  $D$  is the length of the local feature descriptor; the dimension was reduced by principal components analysis (PCA) to achieve higher efficiency. In VLAD,  $D$  was reduced to 100, while in FV, the dimensionality was reduced to 80. Thus, given the same number of words, the dimensionality of the histogram obtained by FV surpasses that of BOVW. In other words, to obtain the same dimensions for the histogram, FV requires fewer words, while achieving the highest classification accuracy and the highest retrieval accuracy.

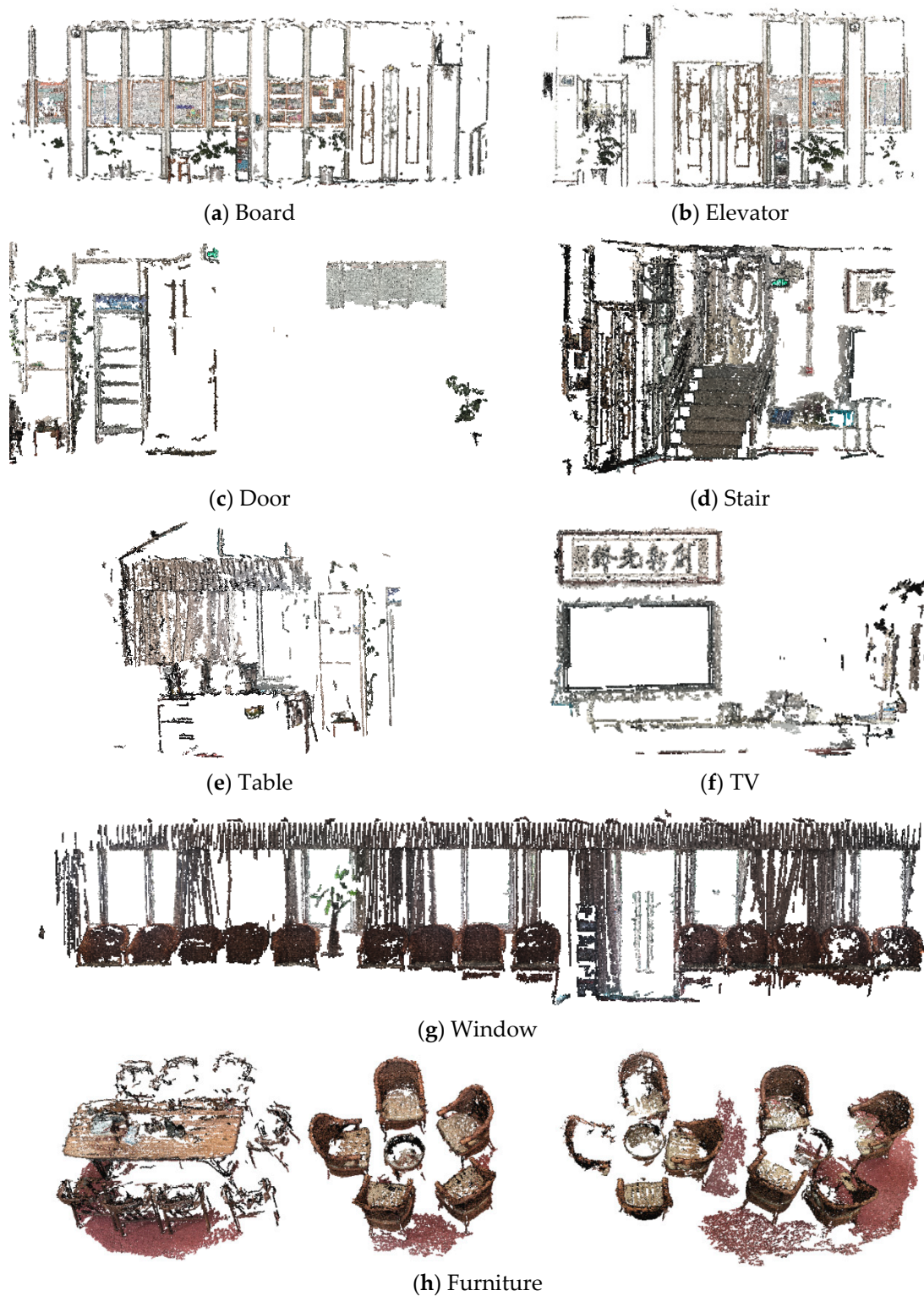
**Table 1.** The results of the three kinds of classification algorithms.

Encoding Method	Number of Words	Classification Accuracy	Mean Average Retrieval Precision	Classification Time (s)
BOVW	1000	0.942857	98.21%	62.45
BOVW	2000	0.957143	98.69%	220.97
VLAD	25,600	0.957143	99.25%	21.02
FV	20,480	0.985714	99.46%	57.19

After obtaining the image clusters belonging to the same class, we then independently reconstructed the atomic point cloud models. The proposed annotated hierarchical SfM was implemented using C++ on the basis of Bundler [14]. Table 2 lists the average reprojection errors of the recovered models (Reproj. error); the time required for the SIFT-based image matching (matching time), SfM reconstruction (construction time), and bundle adjustment (BA time); the number of images involved (No. of images); and the classified images that were successfully recovered with camera pose and points (Recovered views and Recovered points). The reconstructed atomic models of each class are shown in Figure 4, where it can be seen that each model is reconstructed accurately and with high efficiency.

**Table 2.** The construction information for the separate point clouds (time units: s).

Dataset	Board	Elevator	Door	Stair	Table	TV	Window	Furniture
Reproj. error (pixel)	2.81	2.19	2.65	2.63	3.07	1.86	1.94	2.43
Matching time	36	12	22	47	34	11	248	987
Construction time	56	22	20	44	26	30	87	336
BA time	2	2	1	2	2	2	2	4
No. of images	28	17	22	45	19	20	41	112
Recovered points	8317	3349	3949	5923	4973	2785	19,383	49,286
Recovered views	28	17	22	45	19	20	41	112



**Figure 4.** (a–h) The reconstructed atomic models of each class.

After obtaining the atomic scene models, we then hierarchically aligned the atomic models into a whole model. For simplicity, we use the category name of the classified images to denote a reconstructed local point model. For example, the “board” point cloud model is an atomic point cloud model reconstructed from the images belonging to the category of “board”. To give a quantitative analysis of the proposed algorithm point for cloud registration, the “board” point cloud model was

randomly chosen as the reference registration model, and the error was measured based on the “board” results. After the registration process, the point-to-point distance was calculated between the “board” point cloud model and its adjacent point model. The RMSE was derived from the point-to-point distance and referred to as the registration error. The point model adjacent to the “board” point cloud model was then used as the reference to calculate registration error for the neighboring point cloud. The registration error for the rest of the atomic point models was analyzed in the same way. Table 3 lists the registration error of the proposed method and Figure 5 shows the model. It can be observed that the alignment error is small, and the whole model is correctly reconstructed.

**Table 3.** The registration error.

Point Cloud	Elevator	Door	Stair	Table	TV	Window	Furniture
Error(pixel)	0.19	0.24	0.11	0.0001	0.04	0.18	0.16



**Figure 5.** The reconstructed model of the meeting room.

The second dataset used for reconstruction was the lobby of the LIESMARS building, which is a challenging scene containing widespread objects and repetitive textures. We first classified the scene into eight classes, and then reconstructed the atomic models. Table 4 lists the classification accuracy. By dividing the whole image sets into smaller and more tractable ones, the atomic models could be reconstructed correctly and efficiently, as shown in Table 5. Finally, the independent models were aligned into a complete indoor scene, without misalignment and discrepancy. Table 6 and Figure 6 show the registration error (the reference is “front door”) and the reconstructed scene, respectively.

**Table 4.** The classification accuracy for the lobby dataset.

Encoding Method	Number of Words	Classification Accuracy	Mean Average Retrieval Precision	Classification Time (s)
BOVW	1000	0.924793	98.54%	58.34
BOVW	2000	0.938769	98.34%	240.76
VLAD	25,600	0.977849	99.24%	30.87
FV	20,480	0.985714	99.47%	65.63

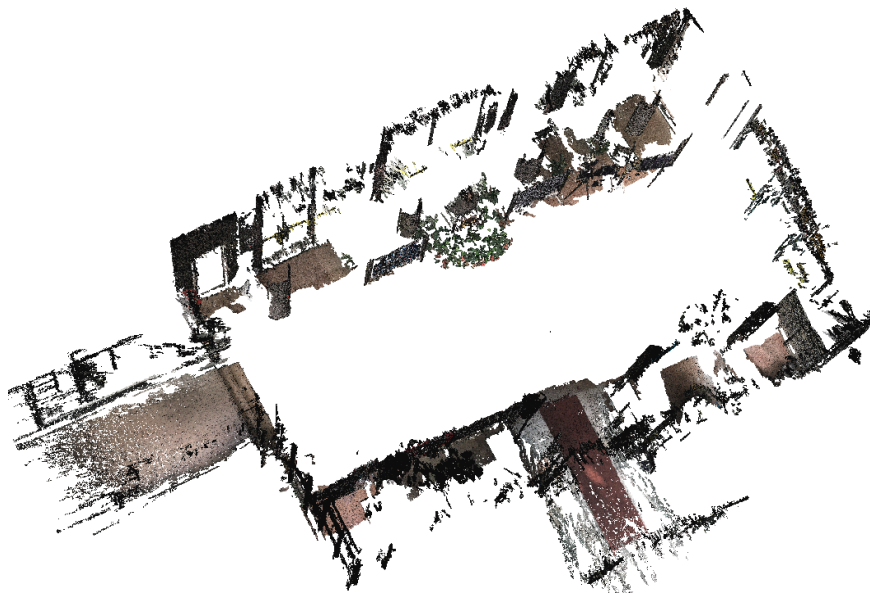


**Table 5.** The reconstruction information for the atomic models of the lobby dataset (time units: s).

Dataset	Front Door	Status	Board	Office	Corner	Corridor	Model	Meeting Room
Reproj. Error (pixel)	2.79	1.94	2.47	2.48	2.04	2.00	2.66	2.39
Matching time	43	33	129	65	171	137	247	183
Constr. time	46	16	48	34	64	130	42	57
BA time	1	2	2	1	3	2	1	5
No. of images	20	13	27	27	42	45	34	35
Recovered points	19,688	19,408	99,459	26,328	51,552	63,545	72,999	92,713
Recovered views	20	13	27	27	42	45	34	35

**Table 6.** The registration error for the lobby dataset.

Point Cloud	Status	Board	Office	Corner	Corridor	Model	Meeting Room
Error(pixel)	0.03	0.01	0.14	0.002	0.18	0.09	0.036

**Figure 6.** The reconstructed model of the lobby dataset.

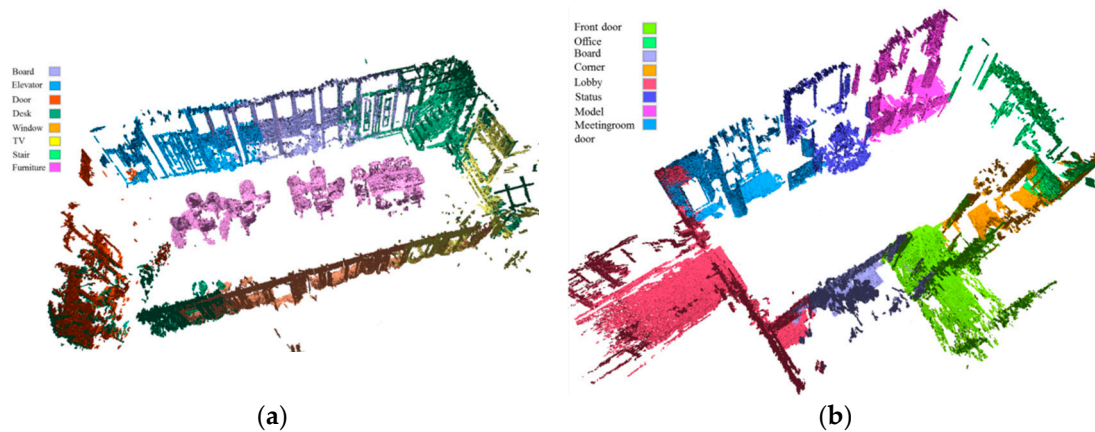
To demonstrate that the proposed approach can obtain an annotated point cloud model with a high efficiency, we compared the proposed method with the state-of-the-art SfM reconstruction method of VisualSfM (VSFM) [16]. The initial focal lengths were extracted from EXIF. Table 7 compares the results of both methods on the two datasets. For the meeting room dataset, although both methods yield correct structures, the camera pose accuracy of the incremental method is inferior to the proposed approach. What is more, the computational time is exponentially larger than that of the proposed method. VSFM recovers 287 images but fails to recover 17 images because of the insufficient inlier projections, while the proposed method successfully recovers all of the images. For more cluttered datasets, the problems related to VSFM could be severe. The second dataset has many repetitive structures and textures, which usually cause mismatching, or incorrect epipolar geometry. The results show that VSFM can recover five separate parts in the relative coordination but cannot merge them into an entire scene. The proposed method successfully recovers the full scene, with a greatly reduced computation time. From this result, we can conclude that the advantage of the proposed method over VSFM in speed is evident. By semantically partitioning the whole dataset, the algorithm gains robustness in reconstructing the complete scene. Furthermore, the semantic annotation from the images is propagated to the point cloud model, and produces the semantic augmented indoor scenes (see Figure 7). The semantic obtained in the proposed annotated hierarchical SfM pipeline is at a coarse



level. However, it is meaningful for some specific indoor applications. For example, they can help to accelerate the feature searching and matching process for real-time visual localization.

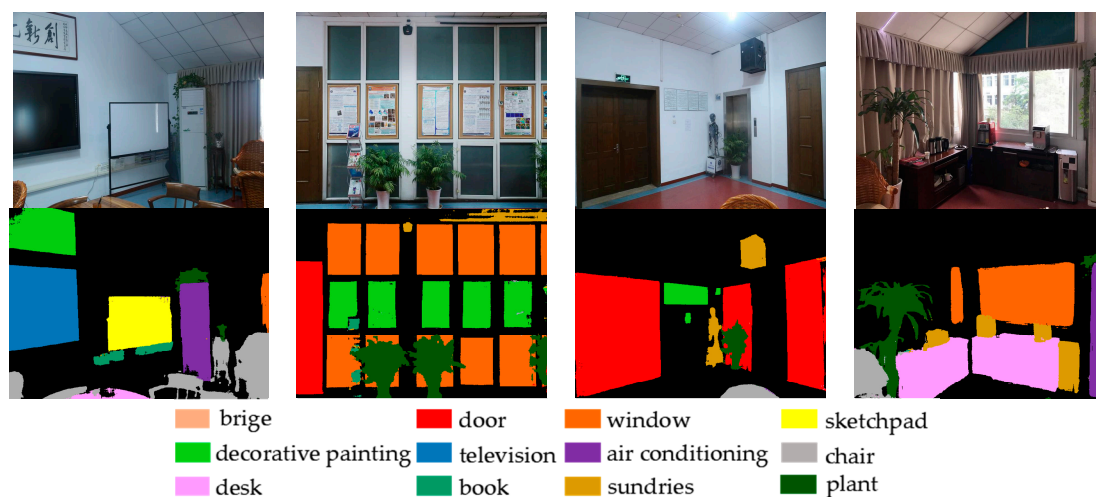
**Table 7.** Comparison between VSFM and the proposed method.

Dataset/Method	Meeting Room Dataset			Lobby Dataset		
	Error (Pixel)	Time (s)	No. of Views Recovered	Error (Pixel)	Time (s)	No. of Views Recovered
VSFM (Wu, 2013)	2.641	18,735	287	2.293	13,987	235
The proposed method	2.454	2025	304	2.040	1526	243



**Figure 7.** The semantically annotated models: (a) the semantically annotated model for the meeting room dataset; (b) the corresponding semantically annotated model for the lobby dataset.

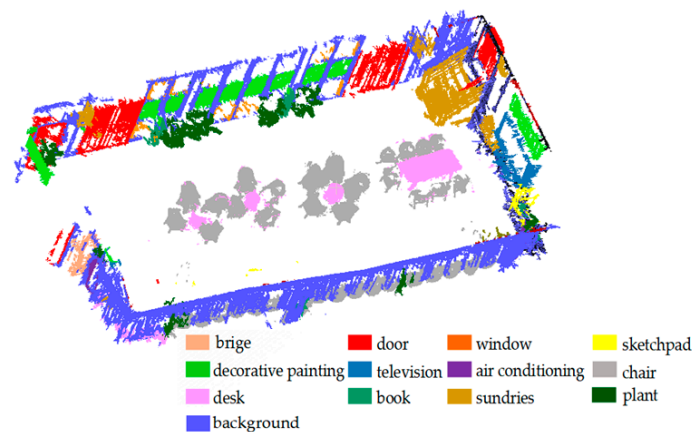
Different level of semantics can serve for different indoor applications. Generally, fine-level point model segmentation is performed as a next step to 3D scene reconstruction [48,49]. To show that the reconstructed point model can achieve precise semantic labeling, an additional experiment was also conducted as an example. We trained a deep network on the NYUDv2 RGBD indoor dataset by combining the two popular 2D segmentation network Deeplab v3+ [50] and Densenet [51]. A local dataset with 24 manually annotated images from the meeting room was used to fine-tune the net. Figure 8 shows some of the segmentation results after fine-tuning on the local dataset, while Table 8 shows the accuracy performance of each category. From the obtained pixel-wise segmentation results on 2D, we propagated the fine-level semantics from the images to the point model just as the coarse-level semantics did. The final labeled point model for the 3D scene of the meeting room can be seen in Figure 9, which has a performance of 90.38% in labeling 12 object classes.



**Figure 8.** The partial semantic segmentation result on 2D images of the meeting room. The first row shows the input images, while the second row are the corresponding semantic segmentation results.

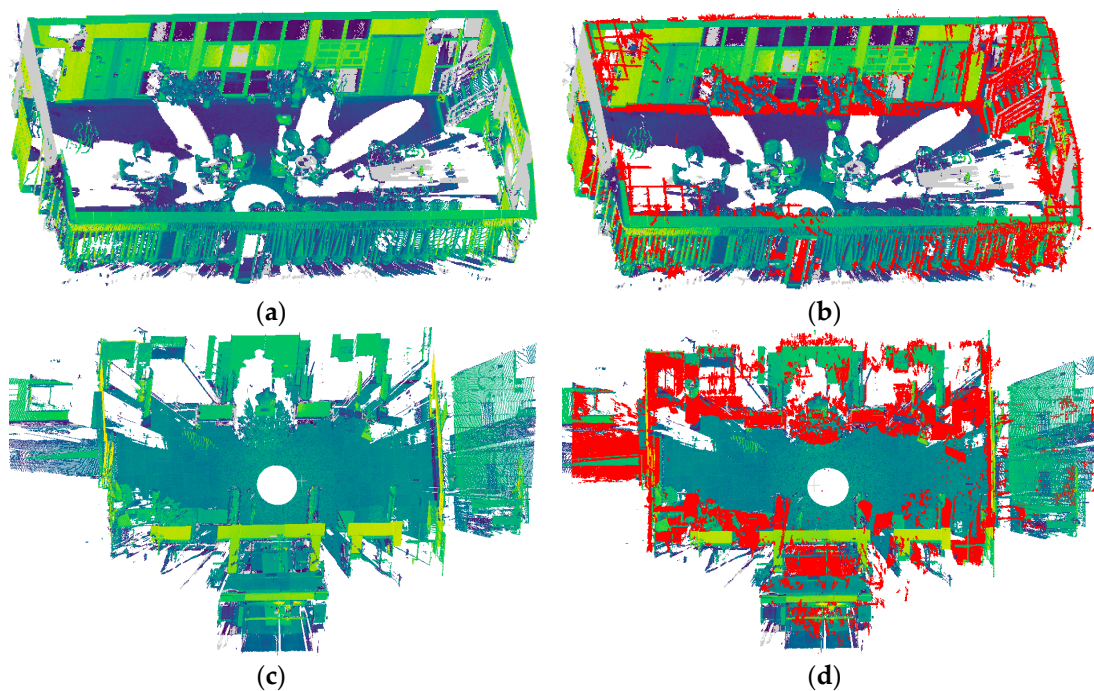
**Table 8.** The semantic segmentation performance on 2D images.

	Door	Window	Bridge	Skatchboard	Painting	Television	Chair	Air-Conditioning	Desk	Book	Plant	Sundries
Accuracy	0.96	0.98	0.95	0.98	0.95	0.99	0.96	0.94	0.92	0.87	0.85	0.83



**Figure 9.** The fine-level labeling results of the point model for the meeting room.

A quantitative evaluation of the two reconstructed point cloud models was also made for reference. The point clouds collected with a terrestrial laser scanner (TLS) for the meeting room and lobby were used as the ground truth, as seen in Figure 10a,c. The visual effects obtained by registering the two SfM reconstructed point clouds to the TLS point clouds are shown in Figure 10b,d. The differences between the SfM reconstructed point clouds and the reference TLS point clouds in terms of point-to-point distance were statistically analyzed. Table 9 lists the accuracy measurements for the two SfM reconstructed point clouds, that is, the RMSE of the registration error (the Euclidean distance between these two types of points) [52]. The low registration errors shown in Table 9 indicate that the annotated hierarchical SfM reconstructed point models created with the proposed algorithm are comparable to the TLS point models, which demonstrates the effectiveness of the improved method.



**Figure 10.** Comparison of hierarchical SfM reconstructed point cloud models (in red color) with the TLS point cloud models. (a,c) are the TLS point clouds of the meeting room and lobby, respectively. Figures (b,d) show the annotated hierarchical SfM reconstructed point models registered to the TLS points.

**Table 9.** Registration error of the proposed SfM reconstructed point clouds and TLS point clouds (cm).

Dataset	Meeting Room Dataset	Lobby Dataset
RMSE	2.14	3.24

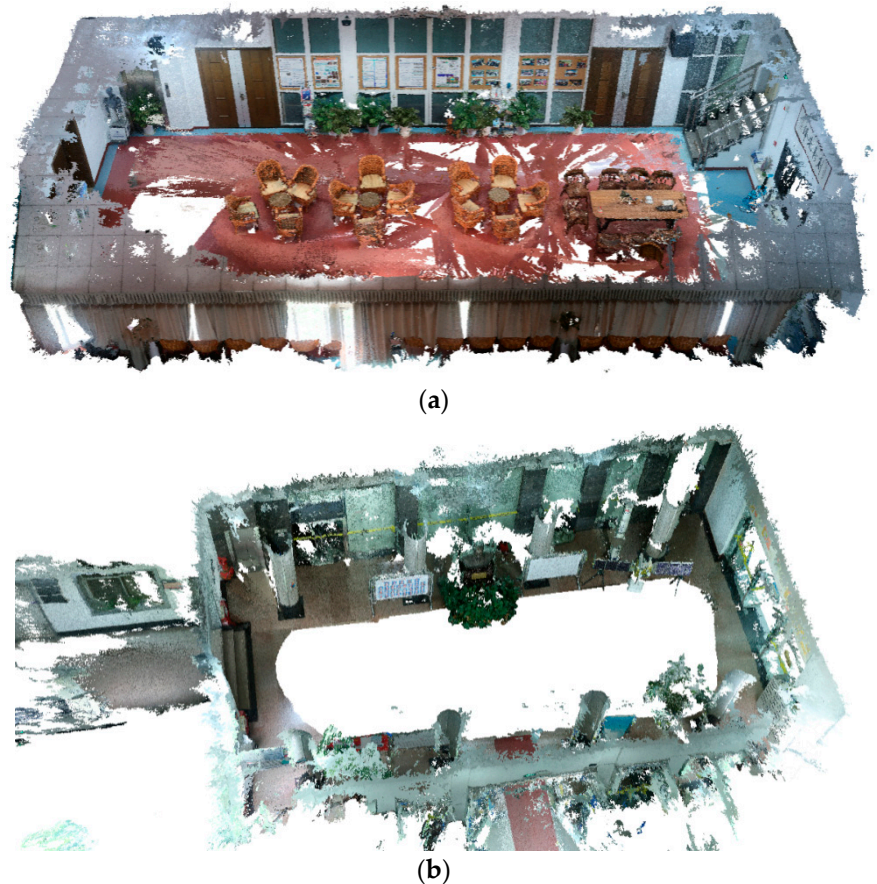
#### 4. Discussion

The proposed method can obtain an annotated indoor 3D point model from unordered images collected by low-cost cameras with high-efficiency; nevertheless, there are still some outstanding issues to consider. Based on the experimental results described above, it can be clearly seen that the models reconstructed from most of image-based methods are negatively affected by the density of the detected features. In the feature-sparse areas, such as the white wall without decoration and the glass in the windows, only the skeleton can be accurately recovered, with the holes remaining unfilled. Extension of the size of the datasets and improvements in more robust feature extraction methods could promote the model quality to a certain degree. However, the most effective solution would be to exploit shape priors in the dense model reconstruction to recover the complete model [53,54]. For example, shape regulation that encodes the normal surface as identical could be applied to the glass in the windows, the white wall orthogonal to the floor, and the floor connecting to the four orthogonal walls, to fill in the missing parts and remove erroneous points in the original model. Recent achievements in convolutional neural networks allow for predicting 3D geometry from a single image [55], which can be used to repair the defects in the image-based 3D reconstruction.

Another factor that affects the completeness of the model is the weakly or indirectly observed surfaces hidden in the input data, such as a floor underneath furniture or a wall facade behind decoration. It is hard to recover partially occluded models, especially when they are observed by very few images. To accurately recover the partly occluded objects, semantic priors and geometry priors can be combined to determine the dense point cloud [36]. The semantic information is used to segment the independent objects, and additional geometry optimization is carried out to fill in the holes. In this way, partially hidden facades behind the decoration can be fully reconstructed. Exploring



the deep visual image features and camera geometries to infer the depth maps is a practical way to fill in some of the non-hidden openings, such as windows. Benefiting from the pioneering work of [56], deep convolutional neural networks [57,58] now enable sophisticated depth estimation even for unstructured images with ill-posed regions. By employing the method presented by [51] to compute the depth maps for the two image datasets with the calibrated camera poses and sparse 3D points recovered from the proposed SfM pipeline, we further reconstructed the dense 3D point models, as can be seen in Figure 11. These results yielded a better visual effect, revealing that the proposed solution is capable of reconstructing dense and well-represented indoor 3D point models.



**Figure 11.** Dense 3D point model reconstruction: (a) Meeting room; (b) Lobby.

Data quality is another important issue that directly affects the quality of the reconstructed 3D models. In particular, sparse coverage between images can often cause discrepancy in the models. The division of images in the proposed hierarchical SfM naturally deals with the problem by reconstructing separate models and merging them into a complete one. Model discrepancy due to missing images can be solved by first reconstructing the models from the available images and then filling the gaps when new images are captured. This strategy can also be applied in model updating. Only the partial models that need to be updated are replaced while the unchanged parts remain the same, avoiding redundant capturing and reconstructing. Consequently, the annotated hierarchical SfM approach is an appropriate scheme in data management and updating. Based on the discussion above, and despite the additional improvements needed to obtain a denser model, the proposed method is viewed as efficient and effective in reconstructing indoor scenes and it could serve as a complementary approach for ubiquitous indoor reconstruction.

## 5. Conclusions

In this paper, we proposed an annotated hierarchical SfM algorithm that detects objects, labels the semantics, and seamlessly reconstructs the model from unmanned images. Compared with the existing methods, the proposed method has many advantages. (1) By exploiting the semantic propagation from images to the point cloud, we can simplify the semantic labeling procedure through image classification in the preprocessing. (2) By organizing an entire indoor scene with a compact hierarchical tree, we can reconstruct separately, and in parallel, the atomic point cloud with a reduced complexity. (3) Using the improved RGPA algorithm to align and update the multiple point clouds into an entire model, we can simultaneously merge the point clouds with a high accuracy. The experiments confirmed that the proposed method is highly efficient and robust in indoor 3D scene modeling. However, we do not propose replacing the existing LiDAR surveying or other methods with our approach. We instead, consider the proposed annotated hierarchical SfM as a supplementary solution to the existing methods. Given the rapid development of crowdsourcing platforms, the low-cost and ubiquitous nature of SfM could enable the public to participate more fully in indoor 3D collection, thus alleviating the dependence on professional instruments and operation.

The proposed method does have some limitations. Compared with the dense and regular point clouds obtained by Kinect or LiDAR, the reconstructed image models may contain poorly recovered parts in feature-sparse places. This missing structure needs additional dense reconstruction by further exploiting the structure and semantic regulators. Furthermore, semantics are incidentally detected in our proposed SfM pipeline, which is still at a coarse level and is limited by the number of predefined classes. For more advanced indoor applications, this could be extended to include more detailed semantics given the development of deep learning algorithms.

**Author Contributions:** Y.D. and Y.Z. implemented the methods, analyzed the data, and wrote the manuscript. X.Z. designed the field experiment, realized this idea, and reviewed the manuscript. H.X. and J.G. provided suggestions about the field design and contributed to all phases of the investigation. All authors read and approved the final manuscript.

**Funding:** This research was funded by the National Key Research and Development Program of China under Grant 2018Y-FB0505401, the National Natural Science Foundation of China Project under Grant 41701445 and the LIESMARS Special Research Funding.

**Acknowledgments:** This work is supported by the National Natural Science Foundation of China Project under Grant 41701445.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Li, L.; Su, F.; Yang, F.; Zhu, H.; Li, D.; Zuo, X.; Li, F.; Liu, Y.; Ying, S. Reconstruction of Three-Dimensional (3D) Indoor Interiors with Multiple Stories via Comprehensive Segmentation. *Remote Sens.* **2018**, *10*, 1281. [[CrossRef](#)]
2. Zhou, Y.; Zheng, X.; Chen, R.; Xiong, H.; Guo, S. Image-Based Localization Aided Indoor Pedestrian Trajectory Estimation Using Smartphones. *Sensors* **2018**, *18*, 258. [[CrossRef](#)] [[PubMed](#)]
3. Hermans, A.; Floros, G.; Leibe, B. Dense 3d semantic mapping of indoor scenes from rgb-d images. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014; pp. 2631–2638.
4. Jamali, A.; Abdul Rahman, A.; Boguslawski, P. A hybrid 3D indoor space model. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Istanbul, Turkey, 16–17 October 2016.
5. Fan, H.; Yao, W.; Fu, Q. Segmentation of sloped roofs from airborne LiDAR point clouds using ridge-based hierarchical decomposition. *Remote Sens.* **2014**, *6*, 3284–3301. [[CrossRef](#)]
6. Henn, A.; Gröger, G.; Stroh, V.; Plümer, L. Model driven reconstruction of roofs from sparse LIDAR point clouds. *Isprs J. Photogramm. Remote Sens.* **2013**, *76*, 17–29. [[CrossRef](#)]



7. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A.J.; Kohi, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A. KinectFusion: Real-time dense surface mapping and tracking. In Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, Nantes, France, 9–13 October 2017; pp. 127–136.
8. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334. [[PubMed](#)]
9. Afanasyev, I.; Sagitov, A.; Magid, E. ROS-Based SLAM for a Gazebo-Simulated Mobile Robot in Image-Based 3D Model of Indoor Environment. In *Advanced Concepts for Intelligent Vision Systems*; Springer: Cham, Switzerland, 2015.
10. Westoby, M.J.; Brasington, J.; Glasser, N.F.; Hambrey, M.J.; Reynolds, J. ‘Structure-from-Motion’ photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* **2012**, *179*, 300–314. [[CrossRef](#)]
11. Wu, C.; Agarwal, S.; Curless, B.; Seitz, S.M. Multicore bundle adjustment. In Proceedings of the Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2011; pp. 3057–3064.
12. Agarwal, S.; Snavely, N.; Simon, I.; Seitz, S.M. Building Rome in a day. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 72–79.
13. Gherardi, R.; Farenzena, M.; Fusiello, A. Improving the efficiency of hierarchical structure-and-motion. In Proceedings of the Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1594–1600.
14. Snavely, N.; Seitz, S.M.; Szeliski, R. Photo Tourism: Exploring Photo Collections In 3D. *ACM Trans. Graph.* **2006**, *25*, 835–846. [[CrossRef](#)]
15. Snavely, N.; Seitz, S.M.; Szeliski, R. Skeletal graphs for efficient structure from motion. In Proceedings of the CVPR 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
16. Wu, C. Towards Linear-time Incremental Structure from Motion. In Proceedings of the International Conference on 3d Vision, Seattle, WA, USA, 29 June–1 July 2013; pp. 127–134.
17. Yin, L.; Snavely, N.; Gehrke, J. MatchMiner: Efficient Spanning Structure Mining in Large Image Collections. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 45–58.
18. Moulon, P.; Monasse, P.; Marlet, R. Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 3248–3255.
19. Arie-Nachimson, M.; Kovalsky, S.Z.; Kemelmacher-Shlizerman, I.; Singer, A.; Basri, R. Global Motion Estimation from Point Matches. In Proceedings of the International Conference on 3d Imaging, Liege, Belgium, 13–14 December 2016; pp. 81–88.
20. Sinha, S.N.; Steedly, D.; Szeliski, R. A multi-stage linear approach to structure from motion. In Proceedings of the European Conference on Trends and Topics in Computer Vision, Heraklion, Greece, 10–11 September 2010; pp. 267–281.
21. Jiang, N.; Cui, Z.; Tan, P. A Global Linear Method for Camera Pose Registration. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 481–488.
22. Crandall, D.J.; Owens, A.; Snavely, N.; Huttenlocher, D.P. SfM with MRFs: Discrete-Continuous Optimization for Large-Scale Structure from Motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2841–2853.
23. Hartley, R.; Trampf, J.; Dai, Y.; Li, H. Rotation Averaging. *Int. J. Comput. Vis.* **2013**, *103*, 267–305. [[CrossRef](#)]
24. Corsini, M.; Dellepiane, M.; Ganovelli, F.; Gherardi, R.; Fusiello, A.; Scopigno, R. Fully automatic registration of image sets on approximate geometry. *Int. J. Comput. Vis.* **2013**, *102*, 91–111. [[CrossRef](#)]
25. Furukawa, Y.; Curless, B.; Seitz, S.M.; Szeliski, R. Towards Internet-scale multi-view stereo. In Proceedings of the Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1434–1441.
26. Chen, Y.; Chan, A.B.; Lin, Z.; Suzuki, K.; Wang, G. Efficient tree-structured SfM by RANSAC generalized Procrustes analysis. *Comput. Vis. Image Underst.* **2017**, *157*, 179–189. [[CrossRef](#)]
27. Ni, K.; Dellaert, F. HyperSfM. In Proceedings of the 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), Zurich, Switzerland, 13–15 October 2012; pp. 144–151.
28. Fan, H.; Zipf, A.; Wu, H. Detecting repetitive structures on building footprints for the purposes of 3D modeling and reconstruction. *Int. J. Digit. Earth* **2017**, *10*, 785–797. [[CrossRef](#)]

29. Martin-Brualla, R.; He, Y.; Russell, B.C.; Seitz, S.M. *The 3D Jigsaw Puzzle: Mapping Large Indoor Spaces*; Springer International Publishing: Cham, Switzerland, 2014; pp. 1–16.
30. Furukawa, Y.; Curless, B.; Seitz, S.M.; Szeliski, R. Reconstructing building interiors from images. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 80–87.
31. Kim, Y.M.; Mitra, N.J.; Yan, D.M.; Guibas, L. Acquiring 3D indoor environments with variability and repetition. *ACM Trans. Graph.* **2012**, *31*, 1–11. [[CrossRef](#)]
32. Choi, S.; Zhou, Q.-Y.; Koltun, V. Robust reconstruction of indoor scenes. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 5556–5565.
33. Turner, E.; Cheng, P.; Zakhori, A. Fast, Automated, Scalable Generation of Textured 3D Models of Indoor Environments. *IEEE J. Sel. Top. Signal Process.* **2015**, *9*, 409–421. [[CrossRef](#)]
34. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Niebner, M. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In Proceedings of the CVPR, Honolulu, Hawaii, 21–26 July 2017.
35. Koppula, H.S.; Anand, A.; Joachims, T.; Saxena, A. Semantic labeling of 3D point clouds for indoor scenes. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 5–10 December 2013; pp. 244–252.
36. Haene, C.; Zach, C.; Cohen, A.; Pollefeys, M. Dense Semantic 3D Reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1730–1743. [[CrossRef](#)] [[PubMed](#)]
37. Xiong, X.; Adan, A.; Akinci, B.; Huber, D. Automatic creation of semantically rich 3D building models from laser scanner data. *Autom. Constr.* **2013**, *31*, 325–337. [[CrossRef](#)]
38. Ikehata, S.; Yang, H.; Furukawa, Y. Structured Indoor Modeling. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1323–1331.
39. Perronnin, F.; Dance, C. Fisher Kernels on Visual Vocabularies for Image Categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
40. Jaakkola, T.S.; Haussler, D. Exploiting Generative Models in Discriminative Classifiers. *Adv. Neural Inf. Process. Syst.* **1998**, *11*, 487–493.
41. Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [[CrossRef](#)]
42. Wright, S.; Nocedal, J. *Numerical Optimization*; Springer Science: Berlin, Germany, 1999; Volume 35, p. 7.
43. Pizarro, D.; Bartoli, A. Global optimization for optimal generalized procrustes analysis. In Proceedings of the Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2409–2415.
44. Fischler, M.A.; Bolles, R.C. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*; ACM: New York, NY, USA, 1981; pp. 726–740.
45. Chetverikov, D.; Stepanov, D.; Krsek, P. Robust Euclidean alignment of 3D point sets: The trimmed iterative closest point algorithm. *Image Vis. Comput.* **2005**, *23*, 299–309. [[CrossRef](#)]
46. Quelhas, P.; Monay, F.; Odobez, J.M.; Gatica-Perez, D.; Tuytelaars, T. A Thousand Words in a Scene. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1575. [[CrossRef](#)]
47. Jegou, H.; Perronnin, F.; Douze, M.; Sánchez, J.; Perez, P.; Schmid, C. Aggregating Local Image Descriptors into Compact Codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1704–1716. [[CrossRef](#)] [[PubMed](#)]
48. Zhou, Y.; Shen, S.; Hu, Z. Fine-Level Semantic Labeling of Large-Scale 3D Model by Active Learning. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 523–532.
49. Boulch, A.; Guerry, J.; Le Saux, B.; Audebert, N. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Comput. Graph.* **2018**, *71*, 189–198. [[CrossRef](#)]
50. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv*, 2018; arXiv:1802.02611.
51. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. Available online: [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Huang\\_Densely\\_Connected\\_Convolutional\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.pdf) (accessed on 26 December 2018).
52. Chen, C.; Yang, B.; Song, S.; Tian, M.; Li, J.; Dai, W.; Fang, L. Calibrate Multiple Consumer RGB-D Cameras for Low-Cost and Efficient 3D Indoor Mapping. *Remote Sens.* **2018**, *10*, 328. [[CrossRef](#)]

53. Cabral, R.; Furukawa, Y. Piecewise Planar and Compact Floorplan Reconstruction from Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 628–635.
54. Xiao, J.; Furukawa, Y. Reconstructing the world's museums. *Int. J. Comput. Vis.* **2014**, *110*, 243–258. [[CrossRef](#)]
55. Häne, C.; Tulsiani, S.; Malik, J. Hierarchical Surface Prediction for 3D Object Reconstruction. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 412–420.
56. Zbontar, J.; LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* **2016**, *17*, 2.
57. Chang, J.-R.; Chen, Y.-S. Pyramid Stereo Matching Network. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 5410–5418.
58. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *arXiv*, **2018**, arXiv:1804.02505.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).