Active Learning Based 3D Semantic Labeling From Images and Videos

Mengqi Rong, Hainan Cui, Zhanyi Hu, Hanqing Jiang, Hongmin Liu, Shuhan Shen

***Abstract*—3D semantic segmentation is one of the most fundamental problems for 3D scene understanding and has attracted much attention in the field of computer vision. In this paper, we propose an active learning based 3D semantic labeling method for large-scale 3D mesh model generated from images or videos. Taking as input a 3D mesh model reconstructed from the image based 3D modeling system, coupled with the calibrated images, our method outputs a fine 3D semantic mesh model in which each facet is assigned a semantic label. There are three major steps in our framework: 2D semantic segmentation, 2D-3D semantic fusion, and batch image selection. A limited annotation image set is first used to fine-tune a pre-trained semantic segmentation network for obtaining the pixel-wise semantic probability maps. Then all these maps are back-projected into 3D space and fused on the 3D mesh model using Markov Random Field optimization, thus yield a preliminary 3D semantic mesh model and a heat model showing each facet's confidence. This 3D semantic model is used as a reliable supervisor to select the parts that are not well segmented for manual annotation to boost the performance of the 2D semantic segmentation network, as well as the 3D mesh labeling, in the next iteration. This Training-Fusion-Selection process continues until the label assignment of the 3D mesh model becomes steady. By this means, we significantly reduce the amount for annotation but not the labeling quality of 3D semantic models. Extensive experiments demonstrate the effectiveness and generalization ability of our method on a wide variety of datasets.**

***Index Terms*—Semantic Segmentation, Geometric Constraint, 3D semantic mesh model, Active Learning**

I. Introduction

With the rapid development of big data, computer technology, and photogrammetry, tremendous achievements have been made in the field of image-based 3D geometric reconstruction. Detailed Large-scale 3D model could be reconstructed from massive images or videos captured by aerial or ground cameras with the help of off-the-shelf commercial [1]–[3] and open-source 3D reconstruction softwares [4]–[7]. But at the same time, the continuous emergence of a series of fresh technologies such as autonomous driving, high-precision maps, smart cities, and virtual reality have also put forward higher requirements for the expression and understanding of 3D scenes, leading to the appearance of 3D semantic labeling with strong vitality. Since then, many scholars have begun to devote themselves to study this challenging issue, which can be formulated as inputting a 3D representation of a scene and finding the way to obtain the semantic label of each geometric primitive point in a point cloud or facet in a mesh model.

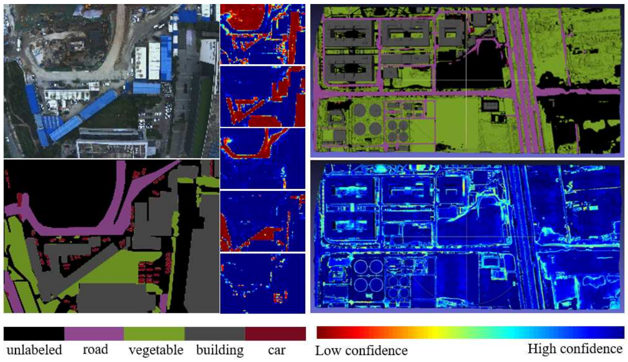


Fig. 1: An example of the inputs and outputs in our method. The left column is the input RGB image and its semantic segmentation result with its associated class-probability maps, and the right column is the generated semantic mesh model and its corresponding heat model showing the confidence of the semantic label for each facet in the mesh.

As is known to all, there are two main ideas for 3D semantic modeling. The first that comes into our mind is to segment the basic geometric primitives for extracting and classifying features directly on the 3D model, such as some 3D convolutions based method. At present, many new network frameworks have been proposed to solve the problem of 3D semantic segmentation, either they directly manipulate the point cloud, or convert the irregular, unstructured and disordered point cloud into a new data structure. PointNet [8], PointNet++ [9], and PointCNN [10] are all representative works for feature extraction directly on the point cloud. Wang *et al.* [11], [12] proposed an octree-based convolutional neural network OCNN to analyze 3D shapes and has achieved significant improvements in computational space and time consumption. In view of the complexity of high-dimensional data analysis and processing, as well as the enormous consumption of calculation and space required, the design of 3D neural networks is still much more complicated than that of 2D. What's more,

This work was supported by the National Natural Science Foundation of China under Grants 61991423, 61873265, 62073320 and 61632003. (Corresponding authors: S. Shen and H. Liu)

M. Rong, H. Cui, Z. Hu and S. Shen are with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, are also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and are also with CASIA-SenseTime Research Group (e-mail: {mengqi.rong, hncui, huzy, shshen}@nlpr.ia.ac.cn).

H. Jiang is with SenseTime Research, Hangzhou, Zhejiang, China (e-mail: jianghanqing@sensetime.com).

H. Liu is with School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: hmliu_82@163.com).

these networks can merely be trained with supervision on 3D datasets, and the acquisition of training data is also a tricky problem. Especially for large-scale urban scenes, the number of points or facets in the model can reach millions or even more, which seems to be an impossible labeling workload. And there is currently no effective interactive software to annotate on 3D space directly.

For the reason that the 3D meshes are generated by image based modeling systems, the intrinsic (focal length, principle point, distortions) and extrinsic (camera 6dof poses) calibration parameters of each camera are also available during the geometric modeling process. And inspired by the brilliant success of the 2D semantic segmentation network, another 3D labeling way is first using the 2D semantic segmentation network to predict semantic labels of each image and subsequently back-projecting all them onto 3D reconstructed models for fusion [13]–[16]. By this means, 2D semantic segmentation and 3D geometric are both taken into consideration, and the problems that are difficult to handle in high-dimensional are transformed into low-dimensional ones. Nevertheless, there is still a stumbling block: the performance of the 2D semantic segmentation network is crucial as it will greatly affect the quality of the final 3D semantic model and a large amount of training data is still necessary for robustness. Though fine-tuning a pre-trained network instead of training from scratch can alleviate the dependence on annotated images to a certain extent, it does not solve the problem well because large-scale scene 3D reconstruction usually requires a large number of images, and it is difficult to decide which images should be chosen for labeling. To solve this problem, Active Learning provides us with a way to select limited number of critical and difficult samples for annotation to maximally boost the classifier's performance iteratively, and the key to its effectiveness lies in how to reliably measure the prediction quality of the sample.

In this paper, we propose an active learning based method for semantic labeling of large-scale 3D scenes reconstructed from massive images and videos. The key idea is to use the 3D semantic model and heat model generated in each iteration as a reliable instructor to select the unannotated images corresponding to the worst segmentation results, which can maximally improve the performance of the semantic segmentation CNN in a targeted manner and also improve the quality of the 3D semantic model accordingly. In the proposed framework, we first obtain the semantic segmentation for each image using a trained CNN and back-project them to the 3D mesh model by ray casting, and then the geometric consistencies of adjacent 3D facets are used to constrain and optimize the allocation of semantic labels for the 3D model.

Since this 3D labeling process takes both 2D semantic segmentation and 3D geometric into consideration, the generated 3D semantic model could work as a reliable guider to select the images for the next training. The above procedures are repeated until the 3D semantic model becomes stable. In summary, our contributions are:

1) A 2D-3D semantic fusion algorithm based on Markov Random Field that takes into consideration both the quality of 2D semantic segmentation and 3D geometric consistency.

2) A novel annotation suggestion strategy based on active learning and the generated 3D semantic model, which significantly reduces the amount for annotated training data while preserving 3D labeling quality.

3) The proposed method has good generality and could be used for various types of scenes captured by images or videos. Experiments on datasets containing different types of urban scenes and different types of shooting methods demonstrate the effectiveness and robustness of our method.

II. Related Work

In this section, we review related works with respect to three major aspects of the proposed pipeline: 2D semantic segmentation, 3D semantic segmentation and active learning.

A. 2D Semantic Segmentation

Semantic segmentation is a challenging but essential task in the field of computer vision and the goal is to assign a most likely semantic label to each pixel in the image. In the past few decades, with the high-performance computing units and the popularization of deep learning, a great quantity of significant progress has been achieved, and some large-scale datasets [17]–[19], network architectures and pre-trained models have also been provided. What is generally considered to be the pioneering work of deep learning for semantic segmentation is the full convolutional network (FCN) proposed by Long et al [20], which discards the fully connected layer of the CNN and is replaced by the fully convolutional layer. Since then, almost all advanced methods have adopted this structure, such as SegNet [21] and U-Net [22]. In addition, the algorithm based on integrating context information adopted by PSPNet [23] and DeepLab [24] has also shown great success. PSPNet proposes a pyramid pooling module to aggregate background information while DeepLab replaces polling with atrous convolutions, which fuses the features at different scales and increases the receptive field of multi-resolution. The up-to-data version of DeepLab is V3+ [25] that combines the spatial pyramid and encoder-decoder.

In recent years, with the development of aerial photography, there are also many researchers devoted themselves to adapting CNN-based methods to remote sensing image classification or segmentation and have achieved significant breakthroughs [26]–[28], a comprehensive review can be found in Cheng *et al.* [29]. Among them, one of the most representative studies is the discriminative CNN model proposed by Cheng *et al.* [30], which imposed a metric learning regularization term into the off-the-shelf CNN features, which effectively solved the problem of rich diversity within class and high similarity among different classes. Soon after, in order to reduce the information loss caused by scale differences, Xie *et al.* [31] proposed a scale-free CNN. They first convert the fully connected layer in CNN into convolutional layers and then followed by a global average pooling layer, which makes it possible to accept images to be of arbitrary sizes.

B. 3D Semantic Segmentation

In the past few decades, a large number of solutions for assigning semantic labels to the 3D reconstruction have been proposed. Typically, there are two ways to achieve this goal. One of the processing ideas is to benefit from the mature 2D semantic segmentation network [20], [24], [32] and the multi-view perspective [13], [33], [34]. First rely on the 2D semantic segmentation network to predict the semantic results for the 2D images, and then back-project them onto 3D reconstructed surfaces for fusion via the camera parameters. Hane *et al.* [35] introduce a method to joint semantic and dense 3D reconstruction. The method employs a pre-trained decision tree for image segmentation. And then the semantic model is reconstructed with label images and depth maps together. Valentin *et al.* [36] use a cascaded classifier with image, geometric and context features as inputs to obtain semantic labels, and the final label of the 3D mesh model is yield by optimizing the output of the classifier on a conditional random field. Blaha *et al.* [37] also propose a method for jointly refining the geometry and semantic segmentation of 3D surface meshes. They first get the semantic labels of image pixels by a pre-trained classifier and then project the per-image class scores onto the mesh surface. In their work, they define a set of priors (e.g., surface shape serves, class-specific priors) and model the labeling problem as a Markov Random Field (MRF) over the mesh facets. On this basis, Romanoni *et al.* [13] propose a novel MRF formulation that does not require any additional knowledge of the environment except the 3D model and the image segmentation.

Another intuitive method is to construct the classifier for 3D point clouds [8]–[10], [38]–[40] or surface meshes [41], [42], and the emerging 3D convolutions is the main way to realize this idea. Point-based networks perform computation in continuous 3D space and can thus directly accept point clouds as input. Qi *et al.* [8] proposed PointNet based on recurrent neural network to extract the characteristics for each 3D point, which solved the unstructured problem of point cloud for the first time. On this foundation, they later proposed PointNet++ [9], which made it possible to extract local features at different scales by introducing sampling and combination. Soon after, PointCNN [10] was put forward, which uses $\mathcal{X}$-transform to maintain the shape of the point set while eliminating the influence of the order of input points. At the same time, it uses hierarchical convolution to extract features at different scales. Those methods have achieved decent segmentation accuracy, but the preprocessing calculations are too large and the memory consumption is high. Just recently, a new lightweight and efficient semantic segmentation network for large-scale 3D point cloud scenes is proposed by Hu *et al.*, called RandLA-Net [40], consisting of two parts: random point sampling and a novel local feature aggregation module increasing the receptive field for each 3D point. Polygon meshes are also an effection representation for 3D shapes, so the mesh-based networks were born. Hanocka *et al.* [41] designed a specific convolution neural network MeshCNN for triangular meshes, which performs specialized convolution and pooling on the mesh edges. Coincidentally, Huang *et al.* [42] also introduced a neural network architecture TextureNet to extract features from high-resolution signals associated with 3D surface meshes. The main bottleneck of these 3D network based methods is that it is usually difficult to obtain a large amount of labeled 3D data, so its generalization ability is generally weak. In contrast, the 2D segmentation network research has accumulated a large amount of labeled data and effectively pre-trained models.

C. Active Learning

When it comes to deep learning, it is usually accompanied by a large number of high-quality annotated samples, which is not feasible in the fields that require high professional knowledge. At present, methods based on active learning [43]–[45] are emerging in an endless stream, which aims to maximize the performance of the system with as few labelled training samples as possible. A general introduction to active learning and a survey of the classical literature can be found in Settles *et al.* [46]. The earlier selection strategy in active learning relies on uncertainty sampling [47], [48]. In detail, the active learner queries for the most uncertain areas for annotation [46], which works remarkably well in many cases [49]–[51]. But for a specific scene, what is limited is that simply using uncertainty sampling will result in duplicated selections of annotation areas, thus ignoring the geometric constraints in the real world. After that, representativeness [52]–[54] is designed around the idea that the selected areas should carry useful features of the unannotated images as much as possible. Yang *et al.* [55] presented a deep active learning framework that utilizes uncertainty and similarity information provided by DNN and formulates a generalized version of the maximum set cover problem to determine the most representative and uncertain areas for annotation. Xie *et al.* [56] proposed the semantic difficulty branch, using a pixel-level probability attention module to learn the semantic difficulty scores of different semantic areas.

Most of the traditional active learning methods are designed for image classification or semantic segmentation, and some of them extract useful information from networks for uncertainty estimation. As we all know, a single 2D image is only a partial projection of the 3D world under a specific perspective and has lost the spatial geometric information from a higher dimension. Therefore, nowadays, there are few explorations using the 3D mesh model to guide active learning. Such methods can not only make full use of the constraint relationship between the multi-view images of the objects but also the local geometric information between primitives, such as the normal of the adjacency facets and the histogram distribution among the normal of the facets in a local area. Zhou *et al.* [57] proposed an active learning based method for fine-level semantic segmentation of 3D models reconstructed from images. The observation uncertainty and the observation divergence constitute the criterion for image selection, which is the first time that 3D geometric information has been introduced into candidate queries for active learning. Siddiqui *et al.* proposed ViewAL [58], a novel active learning strategy for semantic segmentation, in which viewpoint entropy and view divergence scores were

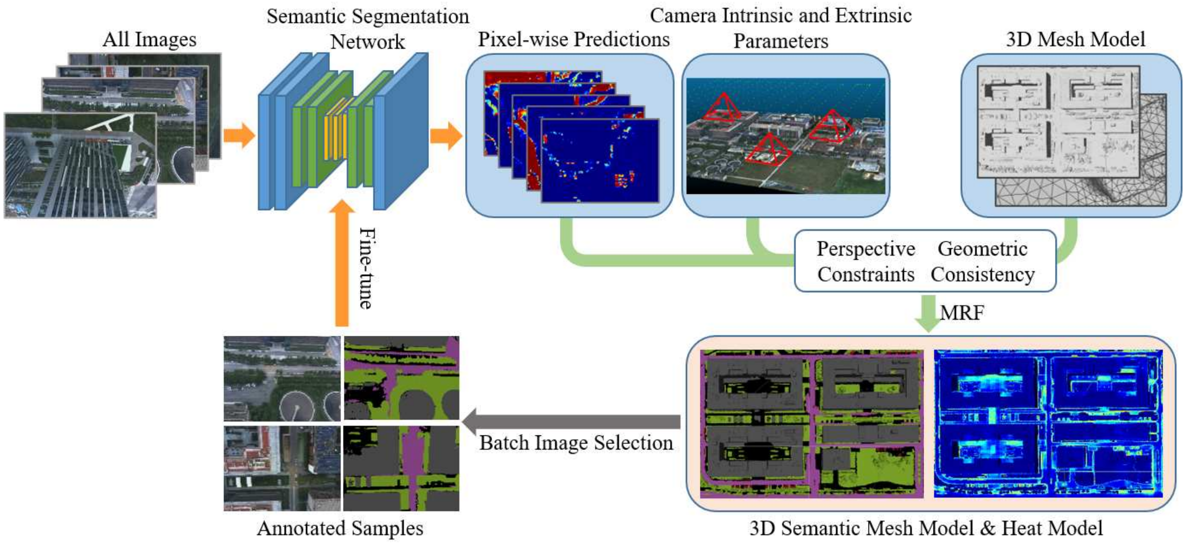


Fig. 2: The pipeline of our proposed method consists of three main steps: fine-tuning the 2D semantic segmentation network with an ever-enlarging annotated image set (orange stream), back-projecting the pixel-wise predictions onto 3D mesh model for semantic fusion based on visibility constraint and geometric consistency (green stream), selecting a batch of images for annotation and adding them into the training set for the next iteration (grey stream).

introduced and executed on a super-pixel level. In this paper, we draw lessons from [57] and concentrate on the semantic mesh labeling not only from images but also videos-based data, and proposed a new measurement method of uncertainty (low confidences) and divergence (high coverage).

III. Proposed Method

The overall pipeline of the proposed method is shown in Fig. 2. Taking as input a 3D mesh model generated from image based reconstruction algorithm, as well as the images with their calibration parameters (camera intrinsic and extrinsic parameters), our method outputs a detailed 3D semantic mesh model whose each facet is assigned a most likely semantic label and a heat model showing the overall confidence. Our method can be roughly divided into three phases: 2D semantic segmentation, 2D-3D semantic fusion, and batch images selection.

Since the 3D reconstruction of different scenes may require different acquisition methods, such as aerial capture v.s. ground capture, oblique images v.s. videos, images of different scenes are significantly different in many aspects, like viewing angle, object distance and redundancy. Therefore, it is generally necessary to fine-tune a pre-trained network for each scene individually. Note that to meet the needs of accurate image based 3D modeling, the captured images are usually very redundant, i.e. there are a lot of overlapping fields of view between images. So there is no need to manually annotate too many images for the fine-tuning, instead we just first randomly select a few images for annotation to fine-tune the pre-trained semantic segmentation network and then use it to segment all the images for their pixel-wise probability maps.

Next, all those probability predictions are back-projected onto the 3D mesh models by ray casting according to the camera calibration parameters. Afterwards, visibility constraints and geometric consistency are applied within the Markov Random Field optimization to get a 3D semantic model.

There is no doubt that the quality of the 3D semantic mesh model still largely depends on the performance of 2D semantic segmentation. So it is especially important to select the best representing images to improve the segmentation accuracy. Inspired by active learning, we take the current 3D semantic mesh model with its confidence model as a reliable supervisor to measure the segmentation quality of each image and select those poor segmented for annotation, and then add these annotated images into the fine-tuning of the semantic segmentation network for the next iteration.

The above processes are repeated until the semantic labels of the mesh model no longer change so much, and at that time, the heat model will become glossy enough. In the following sections, these steps will be described in detail.

A. Images Semantic Segmentation

Since our intention is to use the results of 2D semantic segmentation to guide the semantic labeling for 3D meshes, a semantic segmentation network with good performance is crucial. However, good performance is often accompanied by a large amount of annotated training data, which usually requires professionals with specific knowledge to annotate manually and is time-consuming. Therefore, how to achieve the best results with as little data as possible has become a growing concern today. To this end, we incorporate the idea of active learning into the 3D labeling process, and select the most worthwhile images for annotation according to the 3D semantic labeling results generated by each iteration. Thus with an ever-expanding annotated image set, the 2D segmentation network is improved gradually with minimal manual annotation costs.

In the beginning, all images are with no ground truth and several images are randomly selected for annotation to fine-tune a semantic segmentation network to convergence. The first subset of the dataset is not so important as active learning will gradually select the most worthy images later. Here, we used DeepLab V3+ [25] with xception65 as feature extractor and pre-trained on cityscapes [19]. In principle, any semantic segmentation network can be applied, but a well trained network will help reduce the number of iterations and obtain a better 3D fusion result. We made this choice as it is one of the top performing networks on benchmarks. And in the following iterations, the selected images will also be sent to experts for annotation and re-fine-tune the semantic segmentation network together with the previous training set.

Note that for semantic segmentation tasks as well as many other classification or regression tasks, using a large amount of training data to improve the generalization performance of the network is the most straightforward way, but this is unrealistic in many cases. And in our scheme, what we want is a delicate 3D semantic labeling result for each 3D representation. So it is more reasonable to train a single segmentation model for each specific scene separately due to the difference in flight attitude, terrain conditions, and image content across scenes. Additionally, we also found through experiments that combing the images from several different scenes into a training dataset to fine-tune a segmentation network makes it difficult to achieve the best recognition accuracy for both scenes at the same time. Therefore, it's necessary to train a semantic segmentation network for each reconstructed scene to make it to best adapt to the images of the current scene, which makes our annotation suggestion even more significant because we can reduce the dependence on annotated data.

B. 2D-3D Semantic Fusion

In this section, the trained classifier is first utilized to segment all the images for their class-probability maps and then they are back-projected onto the 3D mesh by ray casting through calibrated intrinsic and extrinsic parameters. Furthermore, a more stringent visibility constraint is introduced to optimize the label assignments, which behaves well and removes some potential errors near the edge of the image. Besides, considering from 3D space, the geometric relationship between the adjacent facets also provides a constraint for 3D semantic labelling. With the two constraints above, the problem can be formulated as an energy minimization over a MRF taking into consideration both 2D semantic segmentation and 3D geometric consistency. Finally, each facet of the mesh model can be assigned a preliminary semantic label.

1) Back-Projection: Once the semantic segmentation network is fine-tuned, it can be used to segment all the images. We modify the last layer of the network with softmax and use it to predict all the images for the probability of each pixel belongs to the different semantic classes, which can be described as a vector and for a pixel p at the position (r,c) of i-th image is:

$$d_i^{(r,c)} = \left(p_i^{(r,c)}(1), p_i^{(r,c)}(2), \cdots, p_i^{(r,c)}(L)\right)^T \tag{1}$$

where $L(l \in L)$ is the number of semantic classes.

Then, using the camera parameters of each image I_i and a point x on the image, a unique ray can be determined. Each pixel can only correspond to one facet at most, once it intersects any facet of the mesh model, its semantic information will play a part, and vice versa. But each facet $f(f \in F)$ in the mesh model F is a triangular region, which means not only multiple images but also a small block of pixels of a single image can intersect with it. Thus traversing all the pixels of the images, the corresponding relationship between those 2D images features and the 3D mesh model can be easily obtained. Finally, we calculate the average of all likelihood probabilities of these visible pixels and obtain the fused semantic features for each facet. And the probability distribution d_f of the facet f can be expressed as

$$d_f = \frac{\sum_{i=1}^{I} \sum_{(r,c)\in\Omega_{i,f}} d_i^{(r,c)}}{\sum_{i=1}^{I} \sum_{(r,c)\in\Omega_{i,f}} 1} \tag{2}$$

whew I is the entire image set, $\Omega_{i,f}$ is the visible area on image I_i of the facet f.

The jth entry of d_f represents the probability of assigning the semantic label j to the facet f, denoted as $p_{(l_f=j)}$. And the final semantic label of the facet f obtained by just back-projection is the position with the highest probability value, namely $argmax(d_f)$.

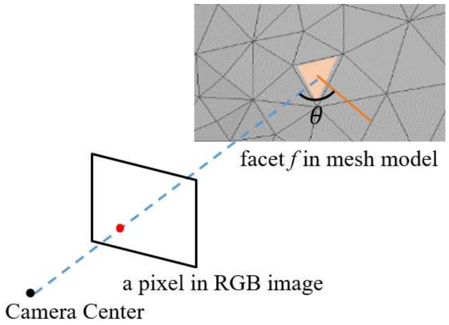


Fig. 3: The angle between the ray passing through the camera's center and a point on the 2D image and its intersecting facet of the mesh model.

2) Visibility Constraint: When the scene is captured by video cameras, the extracted images from the video, even at low frame rate (like 1fps), are still very redundantly distributed throughout the space. The advantage of redundant images is that the reconstructed 3D model could be more complete and accurate, but it will also cause inappropriate visible images to be selected when calculating the visibility of the mesh facet, and further lead to errors in 3D semantic fusion.

In order to reduce the negative impact of these potential incorrect semantic segmentation as much as possible, visibility constraints are introduced. As shown in Fig. 3, a ray determined by a pixel in an image and its camera center intersects with one facet of the mesh model and the angle between the ray and the normal vector of the facet is represented by θ. Only the pixels whose angle θ is within the set threshold are allowed to participate in the 3D semantic labeling, in other words, the codomain space of $\Omega_{i,f}$ is reduced, as:

$$\Omega_{i,f} = \left\{ p_i^{(r,c)} \mid \arccos\left(\left(K_i, R_i, T_i, p_i^{(r,c)} \right), n_f \right) < \delta \right\} \quad (3)$$

where K_i, R_i, T_i represents the camera parameters (internal parameter, rotation matrix and translation matrix respectively) of image I_i while $(K_i, R_i, T_i, p_i^{(r,c)})$ is the ray determined, n_f is the normal vector of the intersection facet, δ is the set threshold.

In this way, the number of pixels for fusion decreased but the percentage of pixels with fine semantic labels is increased, which not only has improved the robustness of the later 3D semantic model, but also the fusion efficiency.

3) Geometric Consistency: It is generally considered that 2D semantic segmentation does not perform so perfectly at the boundary of two similar objects, such as buildings and roads, and can also make mistakes when back-projected into the 3D mesh model. What's more, a large object may even have some individual pixels or small areas are misclassified. Therefore, the 3D semantic model obtained by back-projecting is much more coarse.

As we all know, 2D images lose spatial information compared with 3D presentation. Here, the relationship between adjacent facets of 3D mesh model works as a prior to optimize label assignment, which is called geometric constraint, also known as spatial smoothing constraint.

Generally speaking, there are two considerations that need to be settled. One is that for two adjacent facets in the same plane, they are expected to have the same semantic labels, or they will be severely penalized. However, for facets that lie in the places where the local differential geometry changes greatly, such as the junction of two connected planes, even if the two facets are identified as adjacent, they should also be assigned different semantic labels. Referring to the method proposed in [59], the distance between two adjacent facets with the same semantic labels is measured by their principal directions and curvatures. The closer are the adjacent facets, the smaller is the penalty. Given a facet f, another facet j is one of its adjacent facets, and the semantic labels of the two facets are l_f and l_j respectively. Then the smoothness term is defined as:

$$E_{smooth}(l_f, l_j) = \begin{cases} 1 & if\ l_f \neq l_j \\ \min\left(1, \alpha \left\| W_f - W_j \right\|_2\right) & if\ l_f = l_j \end{cases} \quad (4)$$

where α is a scale factor, W_f and W_j are 6×1 vectors [60] combing the principal curvature k_{min}, k_{max} and their principal direction vectors w_{min}, w_{max}:

$$W = \begin{pmatrix} k_{\min} \cdot w_{\min} \\ k_{\max} \cdot w_{\max} \end{pmatrix} \quad (5)$$

4) Energy Minimization: The facet labeling assignment can be regarded as an energy minimization in a MRF and the energy function is of the form:

$$E(l) = \sum_{f \in F} E_{data}(l_f) + \beta \sum_{(f,j) \in A} E_{smooth}(l_f, l_j) \quad (6)$$

where F is the collection of all facets on the mesh model, the neighborhood relationship between facets is given by A, and β is a constant to balance the contribution of the two terms.

The first term E_{data} is known as the likelihood data term, which aggregates the likelihoods from multiple perspectives estimated by the semantic classifier. And as we have already known the probability of assigning semantic label l_f to the facet f according to the segmentation network output, thus here,

$$E_{data}(l_f) = 1 - d_f^{l_f} \quad (7)$$

The second term E_{smooth} is the smoothness term, which stands for the pairwise interaction potential between adjacent facets f and j, as defined in Eq.4. Ultimately, to find the label configuration l to minimize the energy function, we use the α-expansion algorithm [61].

After optimization, each facet in the mesh model can be given a most likely semantic label with its confidence, a value from 0 to 1 representing the reliability of the assigned label. To visualize these confidences, they can be easily converted into RGB channels under the guidance of a specific criterion, thus yield a 3D heat model, an example of the heat model is shown at the bottom right of Fig. 1, and the bottom is the color bar where blue areas are with high confidence while relatively low in red (badly segmented).

C. Active Learning Based Annotation Suggestion

Once the 3D semantic model is obtained, active learning is used to query the most informative samples from all images, which form the next batch of training data together with those current state. As our 3D semantic mesh model takes both 2D semantic segmentation results and 3D geometry consistency into consideration, it is a more reliable supervisor than directly measured by the 2D semantic segmentation network. Thus, we propose a new annotation suggestion method based on view uncertainty which reflects the reliability of 3D semantic labeling on each 2D image and view divergence which measures the area coverage by the selected image set on the 3D model.

1) Uncertainty Scores: One natural choice for worse samples is the uncertainty based sample selection, which tends to select those samples whose categories are most ambiguous to be determined by the current classifier.

The so-called uncertainty for each pixel or image is not judged by the performance of the classifier. Instead, the 3D semantic mesh model that integrates multi-view class-probability distribution and 3D geometric consistency is re-projected back onto all images, and the reprojection semantic information is used as an indicator to measure the uncertainty of each image. The less likely the facet belongs to a certain category, the greater the uncertainty, and the uncertainty for a pixel p on the image i is defined as:

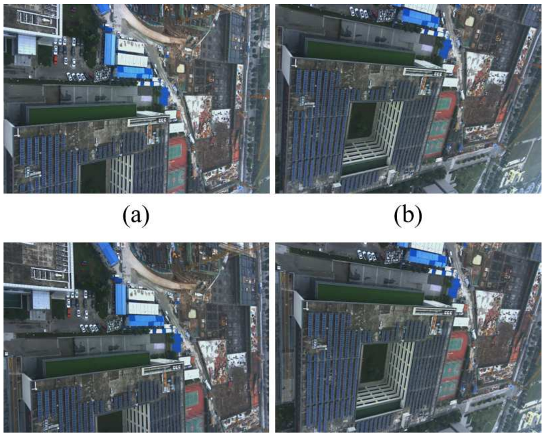
$$u_{p,i} = \begin{cases} 1 - d_f^{l_f} & p \cap F = f \\ 0 & otherwise \end{cases} \quad (8)$$

where, F is the 3D semantic model, $p \cap F = f$ means p is visible in F and the ray interaction between p and F is on facet f. Concretely speaking, for a pixel who is visible on a certain facet of the 3D mesh model, its uncertainty is measured as one minus confidence of the facet's semantic label.

Then, given a batch of unannotated images $I_s (I_s \in I)$, the overall uncertainty is computed as the sum uncertainty of all pixels,

$$US_{I_s} = \sum_{I_i \in I_s} \sum_{p \in I_i} u_{p,i} \tag{9}$$

Then US_{I_s} is normalized by dividing by the total number of pixels, and the normalized uncertainty is defined as $\widetilde{US}_{I_S}$.



(a) (b)

(c) (d)

Fig. 4: The top four images in terms of uncertainty scores have many duplicated selections of annotation areas.

2) Divergence Scores: Only with the uncertainty of the candidate as a metric, a batch of unannotated images which are clustered together and have high similarities are often selected, and an example is shown in Fig. 4. For this annotation, the information obtained is so one-sided and redundant that it should be avoided as possible. On the contrary, the selected areas are expected to carry as many useful characteristics and features of the unannotated images as possible. In our method, we calculate the number of facets that are visible by the subset of the collected images and use its ratio to the whole 3D semantic model F as another metric, also called coverage rate, which approximates the problem to a maximum subset coverage problem,

$$DS_{I_s} = \frac{\sum_{I_i \in I_s, p \in I_i, p \cap F = f} 1}{F} \tag{10}$$

But in our problem, for a small batch of images, only a small part of the scene can be seen. It seems that the coverage rate is far less than one, which results in the imbalance between the two measurement indices. For the measurement consistency, which is also proven to be efficient, we redefine this term DS_{I_s} as the ratio of the intersection and union of the visible facet by the selected images subset,

$$\widetilde{DS}_{I_s} = \frac{\widetilde{F}_\cap}{\widetilde{F}_\cup} \tag{11}$$

where $\widetilde{F}_\cap$ is the intersection, representing the number of facets that is visible by more than one image in the subset I_s; while $\widetilde{F}_\cup$ is the union, which refers to the number of facets can be seen by at least one image.

3) Annotation Suggestion: Combining the uncertainty and divergence scores, the criterion to suggest a batch of images for annotation can be expressed as

$$\arg\max_{I_s} (1-\lambda)\widetilde{US}_{I_s} + \lambda \left(1 - \widetilde{DS}_{I_s}\right) \tag{12}$$

where λ is a Lagrange multiplier that represents the weight of the two different potential energies.

In each batch image selection stage, for the sake of efficiency for training network, we intend to select several images with both high-uncertainty and large-coverage. However, the optimization problem is actually NP-hard and one of its possible solutions is a simple greedy method, which is to choose one at a time to make the objective value maximum iteratively until the selected images reach the number we set. Note that the first time is special, initially, the subset I_s is empty and we can choose the one image with the highest uncertainty regardless of the coverage.

IV. Experimental Results

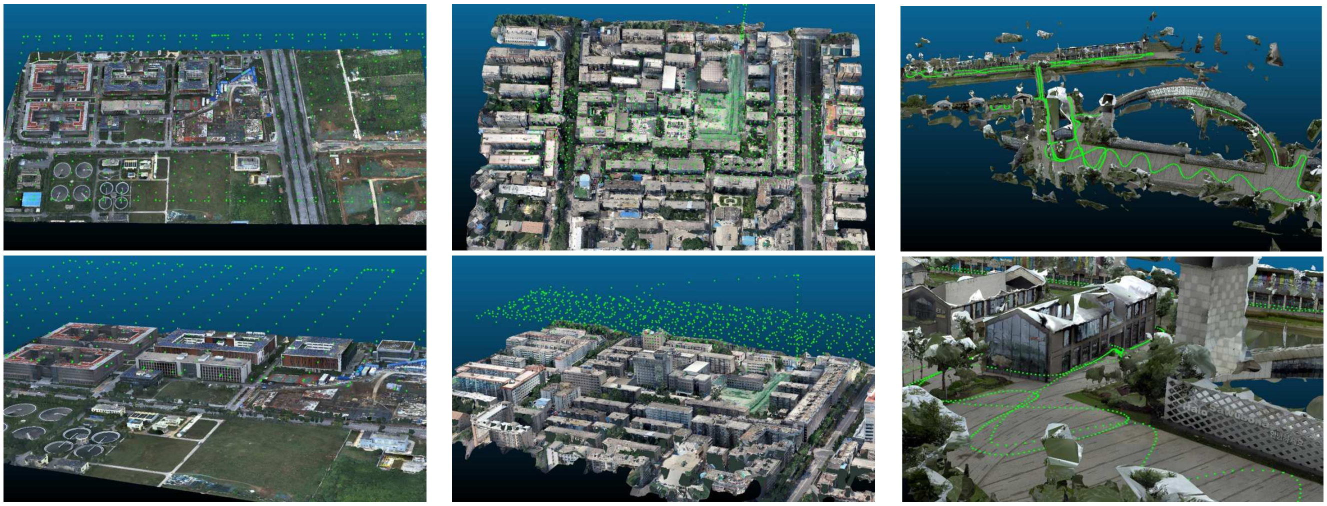
A. Datasets

For large scene image based 3D reconstruction, images and videos are all commonly used in applications. Discrete oblique photography images have the characteristics of high resolution and clear imaging, but they require professional cameras and drone equipment, while video capture is low in cost but with high efficiency, and small drones can handle it. In order to evaluate our proposed method on different datasets thoroughly, three outdoor large-scale scenes are used: one is reconstructed from oblique aerial images and the other two are reconstructed from videos. The three datasets, called Urban1, Urban2 and Urban3, and some pivotal parameters of these scenes are listed in TABLE.I. The 3D mesh models of these three datasets are generated by the state-of-the-art 3D reconstruction system, openMVG [5] and openMVS [6], and the 3D models and the camera trajectories are shown in Fig. 5.

Urban1 is a city scene captured by a professional drone with a five-lens oblique photography camera, its mesh model contains 2,999,393 facets covering an area of $0.57km^2$ and is reconstructed from 2820 oblique images with a resolution of 3688*5168. The scene is mainly composed of trunk roads, office buildings, factories, construction sites and plenty of vegetation.

Urban2 is a residential area captured by a small drone with a single video camera. A total of 951 images with a resolution of 3648*5472 are extracted from this video at a frame rate of 1fps. The generated 3D mesh model contains 2,999,824 facets. Compared to Urban1, Urban2 has a lower flying height during data collection and with a smaller coverage area of $0.21km^2$. Besides, due to the influence of inclination and vegetation coverage, dense buildings are difficult to be separated.

Urban3 is a street scene reconstructed from handheld video camera shooting, which seems not to be so regular and stable as the aerial scene. As there are 9492 images with a resolution of 1200*1600 for reconstruction, the visible area of this scene can be completely covered, and the number of facets in the mesh model has reached 4,998,712, which is approximately twice that of the previous two aerial scenes.



(a) Urban1 (b) Urban2 (c) Urban3

Fig. 5: The 3D texture models of the three scenes and their respective camera trajectories (green dots), in which (a) is an urban scene captured by a professional five-lens oblique photography camera, (b) is a residential scene captured by a single aerial video camera, and (c) is a business district captured by a handhold ground video camera.

TABLE I: The overview of the three different scenes

Datasets	Areas	Facets	View	Type	Images	Resolution	Labels
Urban1	0.57km^2	3M	aerial	image	2820	3688*5168	5
Urban2	0.21km^2	3M	aerial	video	951	3648*5472	5
Urban3	0.31km^2	5M	ground	video	9492	1200*1600	4

Considering the diversity of objects in the scene and the necessity for its semantic labeling, we define five categories: road, vegetation, building, car and those unlabelled in the aerial scene while four semantic classes in Urban3 expect for cars. One sample image and its color-coded annotation result, as well as the correspondence between semantic labels and color ribbons, can be seen in Fig. 1. It is difficult to construct the ground-truth semantic segmentation of large-scale 3D models, because it is hard to label in 3D space. In our experiments, we carefully labelled Urban1, facet by facet, from its textured 3D model manually, and use it as the ground truth for qualitative evaluation, and use Urban2 and Urban3 for qualitative evaluation.

B. Results of Our Method

1) Qualitative Evaluation: In this section, we carry out a series of experiments that focus on verifying the feasibility of our proposed method.

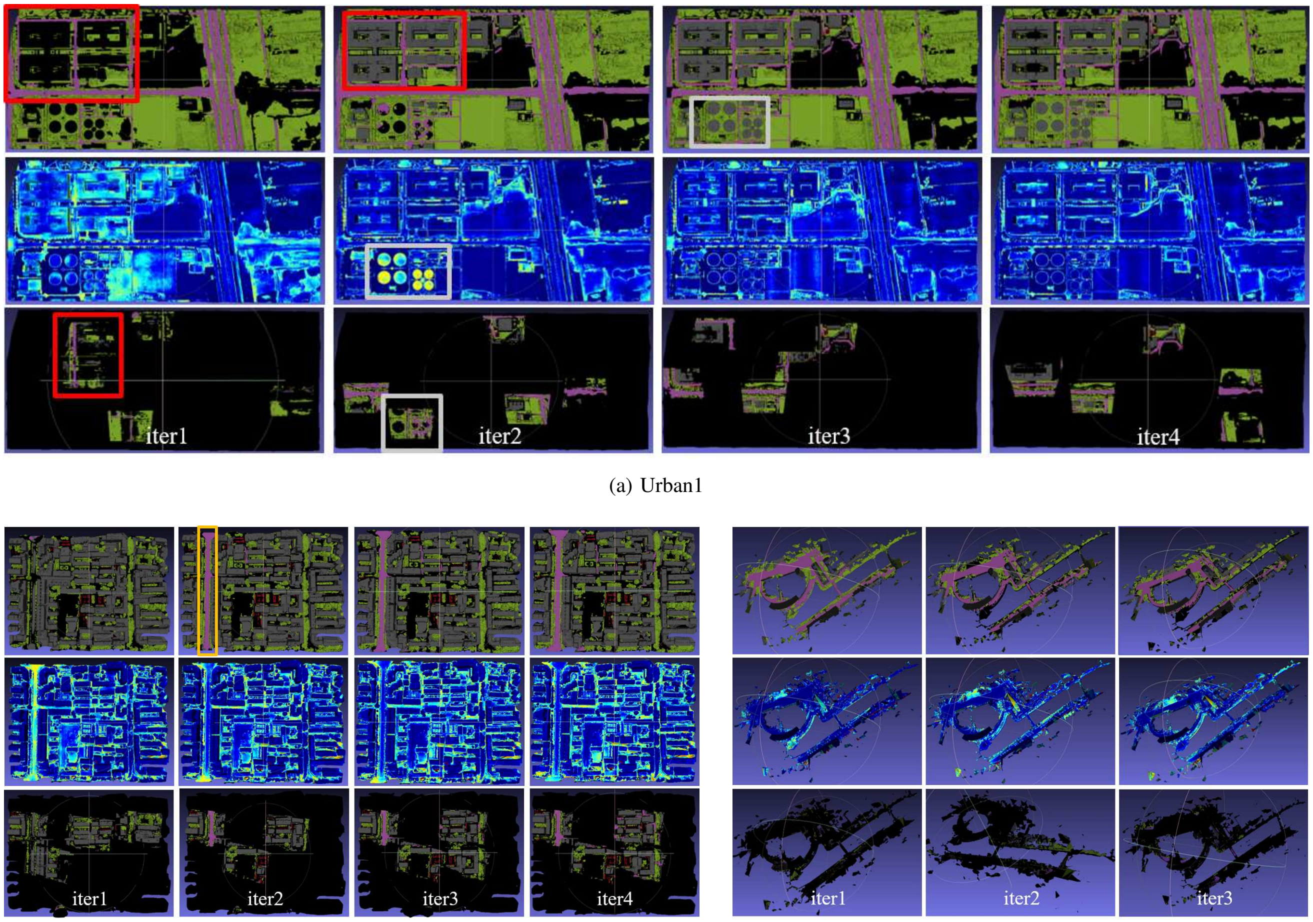
For Urban1, we started with 5 randomly selected images, and in each subsequent iteration, 5 most worthy images are selected for annotation to improve the segmentation network. After four iterations, the semantic label of the mesh model has reached a relatively stable stage and the heat model also became glossy enough, as shown in Fig. 6 (a). Taking office buildings in the red box and circular buildings in the gray box as examples, these regions that were assigned to incorrect semantic labels and accompanied by lower confidences, are selected by our annotation suggestion approach and revised in the next iteration. In order to better display the segmentation results, especially the boundary area, we conduct semantic segmentation on the textured mesh model, and Fig. 8 (a). shows the segmentation results of each category with texture in the last iteration, in which the boundaries between the urban objects are well aligned.

For Urban2, the area of the scene covered by a single image is relatively large (low flying altitude), hence we instead select four images for annotation in each iteration while the other steps and settings are consistent with that Urban1. The results of its four iterations are shown in Fig. 6 (b) and the final segmentation results with texture for each category are shown in Fig. 8 (b), where the 3D semantic model in iter2 is largely improved compared to the first one in terms of the main road. Nevertheless, there is no obvious change in that iter3 and iter4, which is mainly because the images selected by iter2 and iter3 coincide with iter1 greatly and also demonstrates that those areas are indeed the most difficult to segment for the 2D semantic segmentation network.

TABLE II: The percentage of the number of facets whose semantic label changed accounted for the whole facet sets respectively between two adjacent iterations.

datasets	iter1 vs iter2	iter2 vs iter3	iter3 vs iter4
Urban1	0.2216	0.0468	0.0246
Urban2	0.3142	0.0313	0.0241
Urban3	0.1351	0.0270	—

For Urban3, since images are captured on the ground, the objects in the distance of the street images are usually tiny and incomplete due to occlusion, which are troublesome for the network to learn. Typically, projecting such an image to a 3D mesh model, the visible facets are also relatively scattered and sparse, as shown in the last row of Fig. 6 (c), which is the



(a) Urban1

(b) Urban2

(c) Urban3

Fig. 6: Qualitative comparison of the 3D semantic segmentation results among iterations on the three datasets. In each scene, the first row is the detailed views of the 3D semantic mesh model obtained at each iteration with its heat models in the second row. And the last row is the several images chosen according to our proposed active learning based image selection criterion.

projection results of the several images selected by our batch image selection criterion on the 3D model. Furthermore, the scene is extremely large and the image set extracted from the video is dense. So here we set 10 images to be selected in each iteration. And the 3D semantic mesh model and heat model generated during each iteration are shown in Fig. 6 (c). Besides, a partial enlarged view of the texture model and the semantic segmentation is provided in Fig. 7. A facet on the texture model that should be divided into a building is incorrectly labeled with vegetation because of occlusion, which will also be misclassified once only using the projection method to get the semantic labels of such facets. But they have been corrected by our method during semantic segmentation. This is mainly the credit of our geometric constraints.

TABLE. II shows the ratio of the number of facets whose semantic labels changed to the whole facet sets between two adjacent iterations on the three datasets and when this index has been reduced to less than 3%, the semantic model is considered to be converged and the iteration is terminated. And it can be seen from the table that the biggest changes all occur between iter1 and iter2 and followed by some minor and detailed adjustments. This gives a support that our method can efficiently select the hard-to-separate samples that have the greatest impact on the segmentation quality, and can make the model converge faster.

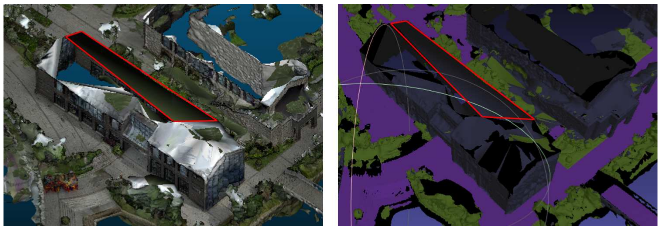
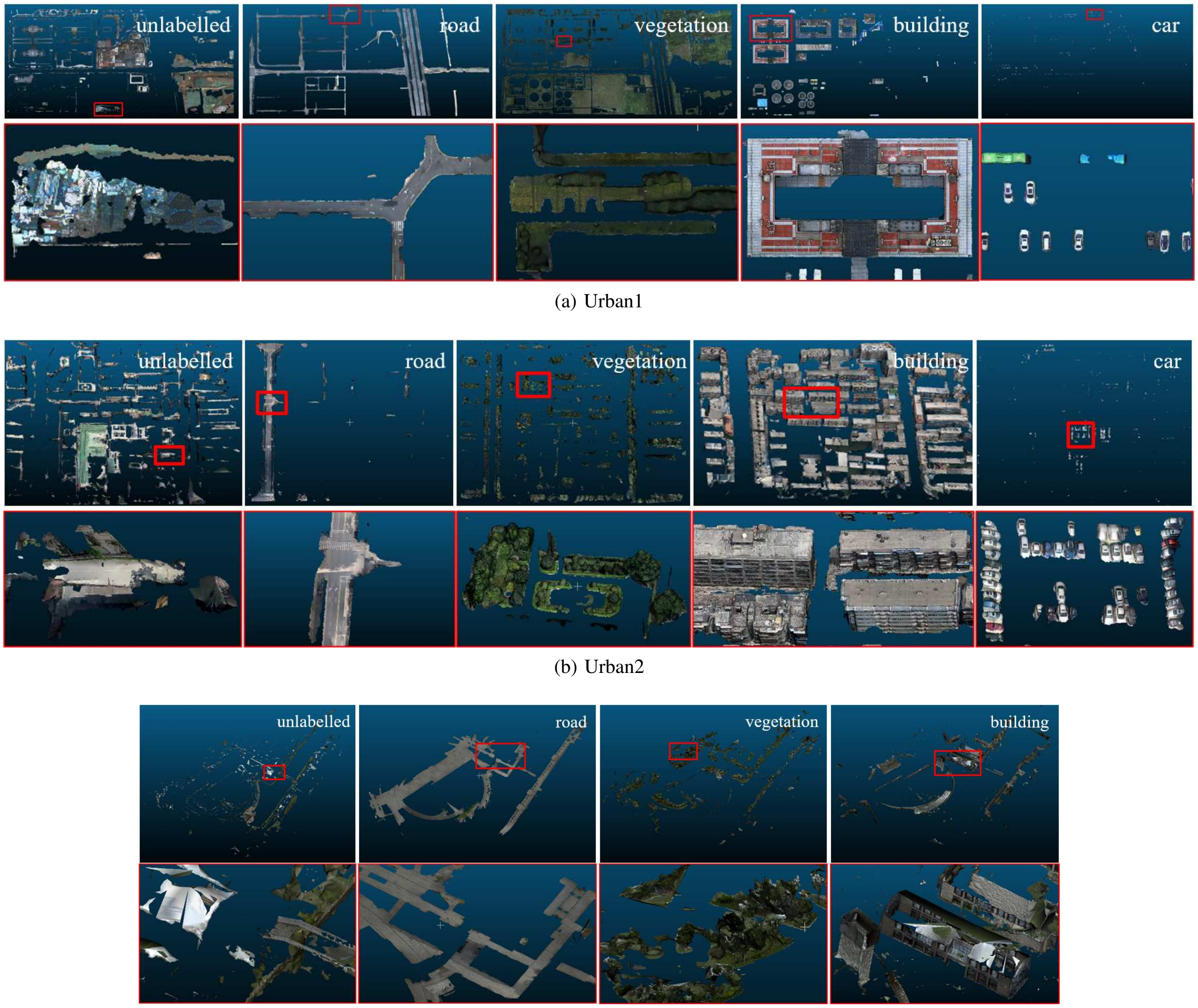


Fig. 7: A partial close-up of the texture model (left) on Urban3 and the semantic segmentation result of this part (right).

2) *Quantitative Evaluation:* In this section, we show the quantitative results of our active learning based method on Urban1 in TABLE. III, where the ”iter” represents the number of iterations and the second column indicates the overall accuracy (*oAcc*) and mean class intersection-over-union (*mIoU*) on the entire model respectively, then followed by the evaluation on each category. The meaning and detailed calculation of *oAcc* and *mIoU* are as follows,



(a) Urban1

(b) Urban2

(c) Urban3

Fig. 8: The segmented texture model of each single semantic category on the three scenes. The first row is the segmentation results of the whole scene while the second is the local enlarged presentations corresponding to the red box above. For (a) and (b), unlabeled, road, vegetation, building and car are in order from left to right, and for (c) there are only the first four categories.

Suppose there are a total of $k+1$ categories, p_{ij} represents the number of facets that belong to category i but are predicted to be category j. Then oAcc is expressed as the ratio of the number of correctly classified facets to all the facets while mIoU is expressed as the ratio of the intersection and union of the two sets of true samples and predicted value.

$$oAcc = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{13}$$

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} + p_{ii}} \tag{14}$$

It can be seen that with the increase of the number of iterations, the two evaluation indicators of both the individual categories and the whole semantic model are improved.

Ablation Study. In order to show that our proposed visibility constraints (VC) and geometric constraints (GC) are all valuable techniques for 3D semantic labelling, we have performed ablation experiments on Urban1 and made quantitative comparison. Here, we mainly use oAcc for evaluation. We first evaluate the labelling accuracy on the 3D mesh model of just back-projecting for semantic fusion as a baseline, as shown in the purple line of Fig.9. For the baseline method, the semantic label for each facet of the mesh model is obtained by averaging the softmax probabilities of all visible pixels from the images. Afterwards, visibility constraint is first to be included for measuring its influence on performance. And as a result, it leads to 1.66% improvement on 3D semantic segmentation as it removes the negative effects of those pixels whose intersection angle with 3D mesh is poor. Finally, geometric constraints are thrown into, giving us our complete scheme. Based on the theory that adjacent patches should have the same semantic labels, geometric constraints are endowed with an ability to correct some single or partial segmentation errors,

TABLE III: Quantitative results of our active learning method on the Urban1 datasets.

	overall	unlabelled	road	vegetation	building	car
iter1	0.7263 / 0.4005	0.7502 / 0.3299	0.803 / 0.7328	0.9595 / 0.8965	0.0434 / 0.0433	0 / 0
iter2	0.8912 / 0.6746	0.7871 / 0.6945	0.8875 / 0.7598	0.9576 / 0.9204	0.8427 / 0.8056	0.1925 / 0.1925
iter3	0.9315 / 0.8403	0.8697 / 0.8066	0.9013 / 0.8600	0.9457 / 0.9332	0.9121 / 0.8887	0.7487 / 0.7129
iter4	0.9436 / 0.8977	0.9231 / 0.8783	0.9255 / 0.9115	0.9575 / 0.9469	0.9334 / 0.9293	0.8393 / 0.8227

which does perform well and further improves the performance of our method by a considerable margin.

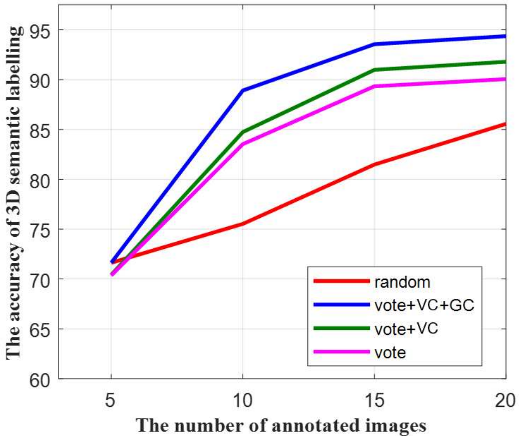


Fig. 9: Quantitative comparison of the two constraints on Urban1, as well as the comparison between random sampling and our active learning based method. The labeling accuracy is expressed in percentage.

3) The efficiency of active learning: Next, in order to verify the effectiveness of our active learning based batch image selection method, we compared the results of our method with that completely random sampling on Urban1. The accuracy curves of random sampling method and our method against the number of annotated images transmitted to the 3D semantic labeling system are shown in Fig. 9 and our method significantly outperforms the random method. Fig. 10 shows the semantic labeling results generated by 20 randomly selected images. It can be seen that there are many areas that are segmented incorrectly with an overall accuracy of only 85.56%, even lower than that in our iter2, such as the circular buildings are wrongly classified as unlabelled.

Active learning helps to pick out those representative regions which are difficult to segment for the semantic segmentation network, and random sampling method can not cover these features when the number for training is limited. As a result, our method minimizes the amount for annotation and improves efficiency.

C. Comparisons with Other Methods

First we compared our method with the state-of-the-art 3D geometry feature based method proposed by Lafarge *et al.* [62], [63], which learns the discriminative features to distinguish between different classes such as local non-planarity, elevation, scatter and regular grouping. In [63], a neighboring relationship is defined to create spatial dependencies between the 3D-points and is optimized by the Graph-Cut based algorithm, and now is included in the CGAL [64] library. This method needs an input data with partial ground truth to be segmented, so we manually marked some points in each scene. And the segmentation results are shown in Fig.11, which are much rougher than ours. For Urban1, the result seems to be similar to ours, but there still are more mistakes in details with an accuracy of 75.02%, which is much lower than ours in Fig. 9, such as those areas in the box, and the buildings in Urban2 cannot be distinguished from each other.

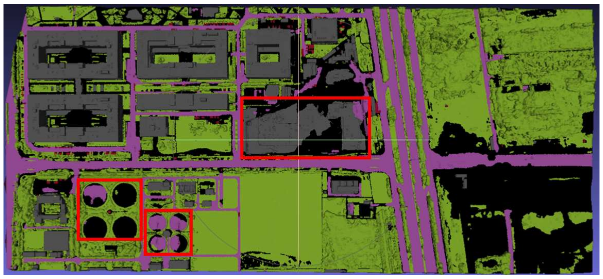
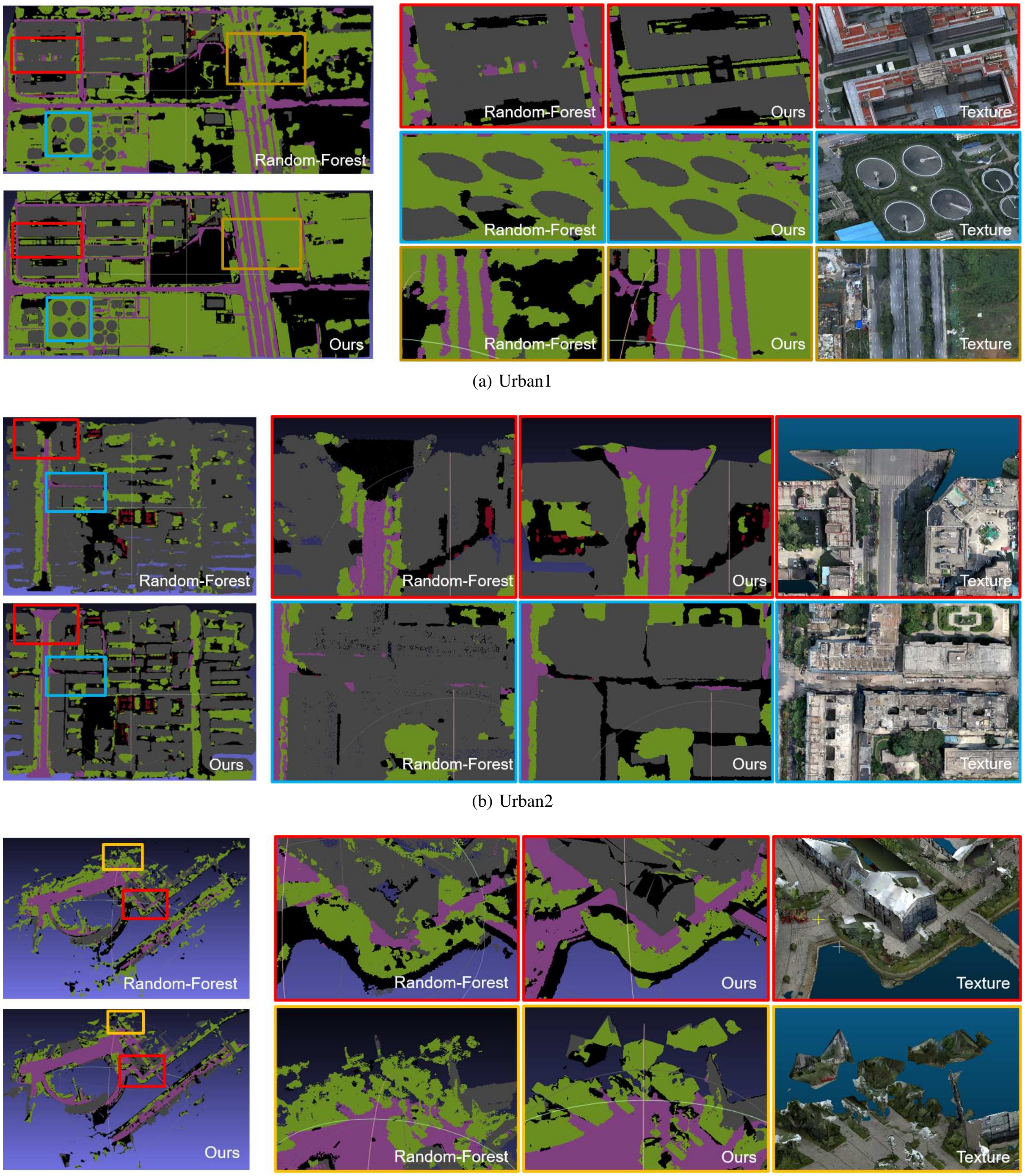


Fig. 10: The semantic mesh model generated by 20 randomly chosen images.

We also evaluate an end-to-end 3D segmentation method, RandLA-Net [40], on these evaluated datasets. Different from previously widely used 3D points segmentation method, such as PointNet [8], PointNet++ [9] and PointCNN [10], which are not specifically designed for large scenes and may lose the overall geometry information, RandLA-Net is an effective and lightweight semantic segmentation network for large-scale 3D point clouds, which introduces a random point sampling module and a local feature aggregation module. The segmentation results of RandLA-Net on the three urban scenes using its pre-trained model on Semantic3D [65] are shown in Fig. 12, and each scene is inclined to be divided into roads and buildings, with very little green vegetation, not as rich as its predefined classes. Fig. 12 shows that the results are far from perfect, for example, the roofs of most buildings and most of the vegetation are incorrectly classified as roads in Urban1 and Urban2, and a lot of vegetation is wrongly classified as buildings in Urban3. We think the reason for this result is mainly because the generalization ability of 3D network is relatively weak due to the lack of 3D training data, thus can only be effective for a certain scene, which is also the key problem of 3D CNN based method. In contrast, the training data of 2D segmentation network and the generalization ability of the pre-trained model are much more abundant and mature, and this is why we choose combining 2D segmentation and 3D geometric for 3D semantic segmentation. However, we also believe that with the continuous increase of 3D labeled



(a) Urban1

(b) Urban2

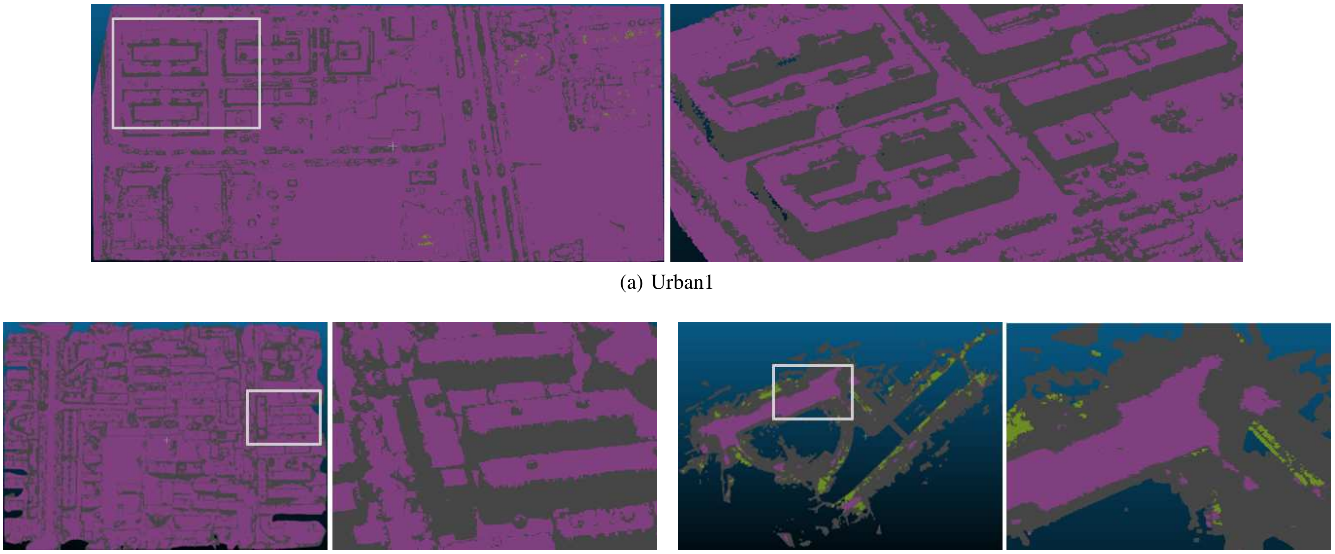
(c) Urban3

Fig. 11: Qualitative comparison between the semantic mesh model generated by the random forest-based method (the top left of each sub-fig) and ours (the bottom left one) on the three datasets. For each urban scene, the right is the enlarged view of the rectangular area in the overall result on the right, which is distinguished by the color of the picture border.

data, the accuracy and generalization ability of these 3D segmentation networks will be rapidly improved.

V. Conclusion

In this paper, we proposed a novel active learning based framework for 3D semantic labelling generated from images or videos, which minimizes the labeling workload while keeping the quality of 3D labelling. Considering making full use of the mature 2D semantic segmentation network and the unique geometric information provided by 3D models, we propose a 2D-3D semantic fusion algorithm and use Markov Random Field to optimize the labels. Besides, inspired by active learning, we use the fused 3D semantic model as a supervisor to select the most effective images for annotation. This segmentation-fusion-selection iterative process makes full use of 2D semantic information and 3D geometric information,



(a) Urban1

(b) Urban2

(c) Urban3

Fig. 12: The results generated by RandLA-Net on the three evaluated urbans. In each sub-fig, the one of the right is an enlargement of the gray box on the left.

and can minimize the workload of image annotation while ensuring the accuracy of 3D scene segmentation. Experimental results on three large-scale outdoor 3D scenes captured by different shooting mode show that the proposed method works effectively and robustly in real scenes, and outperforms two state-of-the-art semantic labeling methods.

Although our proposed method did achieve satisfactory results, there are still some problems that should not be ignored. The first is that this method cannot accurately assign semantic labels to the moving objects in the scene, as dynamic objects like moving people or moving cars exist only in images but not in the 3D model. In addition, the ground-truth on the 2D images used to fine-tune the segmentation network are labeled by ourselves, they may be rough and the diversity of categories is relatively limited, thus the performance still remains to be verified in the case of fine categories. The other is that the segmentation network must be re-fine-tuned in every iteration as the training set enlarges, which is a complicated and time-consuming process as it takes sophisticated skills to adjust the parameters of the network. These are all problems that need to be further solved to build a practical system.

REFERENCES

[1] ContextCapture, https://www.bentley.com/en/products/brands/contextcapture.

[2] PhotoScan, https://www.agisoft.com.

[3] Pix4D, https://www.pix4d.com.

[4] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[5] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "Openmvg: Open multiple view geometry," in *International Workshop on Reproducible Research in Pattern Recognition*. Springer, 2016.

[6] openMVS, https://github.com/cdcseacave/openMVS.

[7] S. Shen, "Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes," *IEEE transactions on image processing*, vol. 22, no. 5, pp. 1901–1914, 2013.

[8] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[9] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in neural information processing systems*, 2017, pp. 5099–5108.

[10] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in *Advances in neural information processing systems*, 2018, pp. 820–830.

[11] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong, "O-cnn: Octree-based convolutional neural networks for 3d shape analysis," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[12] P.-S. Wang, C.-Y. Sun, Y. Liu, and X. Tong, "Adaptive o-cnn: A patch-based deep representation of 3d shapes," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–11, 2018.

[13] A. Romanoni and M. Matteucci, "A data-driven prior on facet orientation for semantic mesh labeling," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 662–671.

[14] M. Rouhani, F. Lafarge, and P. Alliez, "Semantic segmentation of 3d textured meshes for urban scene analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 123, pp. 124–139, 2017.

[15] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*. IEEE, 2017, pp. 4628–4635.

[16] A. Kundu, X. Yin, A. Fathi, D. Ross, B. Brewington, T. Funkhouser, and C. Pantofaru, "Virtual multi-view fusion for 3d semantic segmentation," *arXiv preprint arXiv:2007.13138*, 2020.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[18] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE*

transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[26] G. Wang, B. Fan, S. Xiang, and C. Pan, "Aggregating rich hierarchical features for scene classification in remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 9, pp. 4104–4115, 2017.

[27] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6712–6722, 2018.

[28] F. Ghazouani, I. R. Farah, and B. Solaiman, "A multi-level semantic scene interpretation strategy for change interpretation in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8775–8795, 2019.

[29] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[30] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE transactions on geoscience and remote sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[31] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6916–6928, 2019.

[32] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[33] H.-Y. Chiang, Y.-L. Lin, Y.-C. Liu, and W. H. Hsu, "A unified point-based framework for 3d segmentation," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 155–163.

[34] M. Jaritz, J. Gu, and H. Su, "Multi-view pointnet for 3d scene understanding," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[35] C. Häne, C. Zach, A. Cohen, and M. Pollefeys, "Dense semantic 3d reconstruction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1730–1743, 2016.

[36] J. P. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. Torr, "Mesh based semantic modelling for indoor and outdoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2067–2074.

[37] M. Blaha, M. Rothermel, M. R. Oswald, T. Sattler, A. Richard, J. D. Wegner, M. Pollefeys, and K. Schindler, "Semantically informed multiview surface refinement," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3819–3827.

[38] Q.-H. Pham, T. Nguyen, B.-S. Hua, G. Roig, and S.-K. Yeung, "Jsis3d: joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8827–8836.

[39] Z. Hu, M. Zhen, X. Bai, H. Fu, and C.-l. Tai, "Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds," *arXiv preprint arXiv:2007.06888*, 2020.

[40] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 108–11 117.

[41] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "Meshcnn: a network with an edge," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[42] J. Huang, H. Zhang, L. Yi, T. Funkhouser, M. Nießner, and L. J. Guibas, "Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4440–4449.

[43] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[44] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9368–9377.

[45] R. Mackowiak, P. Lenz, O. Ghori, F. Diego, O. Lange, and C. Rother, "Cereals-cost-effective region-based active learning for semantic segmentation," *arXiv preprint arXiv:1810.09726*, 2018.

[46] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[47] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[48] K. Chitta, J. M. Alvarez, and A. Lesnikowski, "Large-scale visual active learning with deep probabilistic ensembles," *arXiv preprint arXiv:1811.03575*, 2018.

[49] W. Luo, A. Schwing, and R. Urtasun, "Latent structured active learning," in *Advances in Neural Information Processing Systems*, 2013, pp. 728–736.

[50] Q. Sun, A. Laddha, and D. Batra, "Active learning for structured probabilistic models with histogram approximation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3612–3621.

[51] A. Casanova, P. O. Pinheiro, N. Rostamzadeh, and C. J. Pal, "Reinforced active learning for image segmentation," *arXiv preprint arXiv:2002.06583*, 2020.

[52] E. Elhamifar, G. Sapiro, A. Yang, and S. Shankar Sasrty, "A convex optimization framework for active learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 209–216.

[53] M. Hasan and A. K. Roy-Chowdhury, "Context aware active learning of activity recognition models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4543–4551.

[54] A. Mosinska-Domanska, R. Sznitman, P. Glowacki, and P. Fua, "Active learning for delineation of curvilinear structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5231–5239.

[55] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 399–407.

[56] S. Xie, Z. Feng, Y. Chen, S. Sun, C. Ma, and M. Song, "Deal: Difficulty-aware active learning for semantic segmentation," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[57] Y. Zhou, S. Shen, and Z. Hu, "Fine-level semantic labeling of large-scale 3d model by active learning," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 523–532.

[58] Y. Siddiqui, J. Valentin, and M. Nießner, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9433–9443.

[59] F. Lafarge, R. Keriven, and M. Brédif, "Insertion of 3-d-primitives in mesh-based representations: towards compact models preserving the details," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1683–1694, 2010.

[60] S. Rusinkiewicz, "Estimating curvatures and their derivatives on triangle meshes," in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.* IEEE, 2004, pp. 486–493.

[61] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[62] F. Lafarge and C. Mallet, "Building large urban environments from unstructured point data," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1068–1075.

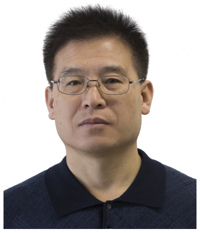
[63] ——, "Creating large-scale city models from 3d-point clouds: a robust approach with hybrid representation," *International journal of computer vision*, vol. 99, no. 1, pp. 69–85, 2012.

[64] CGAL, https://doc.cgal.org/latest/Classification/index.html.

[65] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3d. net: A new large-scale point cloud classification benchmark," *arXiv preprint arXiv:1704.03847*, 2017.

Mengqi Rong is currently pursuing her Ph.D. degree in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and are also with School of Artificial Intelligence, University of Chinese Academy of Sciences. She received the B.S. degree from School of Automation and Electrical Engineering, University of Science and Technology Beijing in 2018. Her research interests include weakly supervised learning, and 3D semantic segmentation.

Hainan Cui received the B.S. Degree from the South China University of Technology in 2011, and the Ph.D. Degree from the University of Chinese Academy of Sciences in 2016. Since 2016, he has been with the National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences, where he is now an associate professor. His research interest is computer vision, with a particular interest in using vast amounts of imagery to reconstruct and visualize the world in 3D.



Zhanyi Hu received the B.S. degree from the North China University of Technology and the Ph.D. degree from the University of Liege, Belgium. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He has published more than 150 peer-reviewed journal articles, including IEEE T-PAMI, IEEE T-IP, IJCV, and PR. His current research interests include biology-inspired vision and large-scale 3D scene reconstruction from images. He was the Organization Committee Co-Chair of ICCV 2005, and the Program Co-Chair of ACCV 2012. He is the Deputy Editor-in-Chief for Chinese Journal of CAD and CG, and an Associate Editor for Science China and Journal of Computer Science and Technology.

Hanqing Jiang received his PhD degree from Zhejiang University, after which he was a postdoctoral researcher in the State Key Lab of CAD&CG, Zhejiang University. He is currently an Associate Research Director in Sensetime Group Ltd, China. His research interests focus on computer vision, including video enhancement, multi-view stereo, 3D reconstruction, and augmented reality.

Hongmin Liu received the B.S. degree from Xidian University, China, in 2004 and the Ph.D degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, 2009. She is currently a Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. Her research is focused on computer vision, image processing and deep learning.

Shuhan Shen received the B.S. and M.S. degrees from Southwest Jiaotong University and the Ph.D. degree from Shanghai Jiaotong University. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests are in 3D computer vision, which includes image-based 3D modeling of large-scale scenes, 3D perception for intelligent robot, and 3D semantic reconstruction.