Exploiting SAR Visual Semantics in TomoSAR for 3D Modeling of Buildings

<table>
<thead>
<tr>
<th>Journal:</th>
<th>National Science Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>NSO20230067.R1</td>
</tr>
<tr>
<td>Manuscript Type</td>
<td>Research Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>25-Jan-2024</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Wang, Wei; Zhoukou Normal University
Wang, Haixia; Shandong University of Science and Technology; Chinese Academy of Sciences
Yu, Liankun; Chinese Academy of Sciences Institute of Automation
Dong, Qiulei; Chinese Academy of Sciences
Hu, Zhanyi; Chinese Academy of Sciences |
| Keywords:      | SAR visual semantics, TomoSAR, box-like building modeling |

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

NSO20230067_Figures.zip
Exploiting SAR Visual Semantics in TomoSAR for 3D Modeling of Buildings

Wei Wang 1,3#, Haixia Wang 2,3#, Liankun Yu 2,3#, Qiulei Dong 3,* and Zhanyi Hu 3,*

1School of Network Engineering, Zhoukou Normal University, Zhoukou, 466000, China; 2College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China; 3State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

#Contributed equally to this work.

*Corresponding authors (emails: qldong@nlpr.ia.ac.cn (Qiulei Dong); huzy@nlpr.ia.ac.cn (Zhanyi Hu))

Abstract: Recently a new paradigm is emerging in SAR (Synthetic Aperture Radar) 3D imaging technology where the imaging performance is enhanced by exploiting SAR visual semantics. Here by “SAR visual semantics”, we mean primarily the scene conceptual structural information extracted directly from SAR images. Under this paradigm, a paramount open problem lies in what and how the SAR visual semantics could be extracted and used at different levels associated with different structural information. This work is a tentative attempt to tackle the above what-and-how problem, and it mainly consists of the following two parts: The first one is a sketchy description of how three-level (low, middle, and high) SAR visual semantics could be extracted and used in SAR Tomography (TomoSAR), including an extension of SAR visual semantics analysis (e.g., façades and roofs) to sparse 3D points initially recovered via traditional TomoSAR methods. The second part is a case of study on two open source TomoSAR datasets to illustrate and validate the effectiveness and efficiency of SAR visual semantics exploitation in TomoSAR for box-like 3D building modeling. Due to the space limit, only main steps of the involved methods are reported, and we hope, such neglects of technical details will not severely compromise the underlying key concepts and ideas.

Keywords: SAR visual semantics, TomoSAR, box-like building modeling.

1 INTRODUCTION

Synthetic aperture radar tomography (TomoSAR) can reconstruct high-resolution 3D structures of targets from a stack of coregistered SAR images and has the advantage of being unaffected by weather conditions, time, and terrain limitations. Recently, it has been applied to city modeling, geological exploration, environmental monitoring, target detection, military reconnaissance, and other fields [1-4].

TomoSAR 3D reconstruction methods can roughly be divided into three categories: Fourier transformation methods [5], spectral estimation methods, and compressed sensing methods [6]. In practice, the methods in the first category are sensitive to the sampling manners (Nyquist) and the number of tracks, and can only obtain low-resolution results. Relatively, the methods in the second category can obtain results in nonparametric or/and parametric manners such as Capon [7], singular value decomposition (SVD) [8], multiple signal classification (MUSIC) [9] and weighted subspace fitting (WSF) [10]. However, such methods usually need to estimate a variance matrix based on multi-look tracks, which might degrade the azimuth-range resolutions. Specifically, for coherent scatterers encountered in urban scenes, such methods may produce inferior results [11]. Moreover, though the parametric methods can produce better elevation
resolutions than the nonparameteric methods, they usually require the prior information on scatters to be estimated. In contrast, the methods in the third category can obtain super-resolution results from the data with sparse or non-uniform baseline along the elevation direction based on the compressed sensing theory. However, the delicate hyper-parameters tuning usually hampers its applicability in addition to the involved heavy computational load.

To improve the accuracy and reliability of the Tomography 3D reconstruction, a variety of spatial priors and constraints have been considered in the tomographic inversion. For example, Rambour et al. [12] proposed an iterative method to enhance spatial regularization (i.e., the natural sparsity of targets in urban environments and the spatial smoothness such as the spatial proximity of scatters) based on variable splitting and the augmented Lagrangian technique. Aghababaee et al. [13] formulated the reconstruction problem as an energy minimization problem by introducing a regularization term to include the prior information about scene height variation in the array processing chain.

![Figure 1](https://engine.scichina.com/doi/10.1360/nso/20230067)

Figure 1 Problems in traditional TomoSAR methods. (a)An input SAR image; (b) 3D points produced by conventional methods; (c) Close-up of the 3D points in the rectangle in (b).

In fact, urban buildings with different geometric information (e.g., height and footprint) are generally arranged in an irregular layout, occluding from each other. As shown in Figure 1, the side-looking imaging geometry and numerous interference factors (e.g., the layover region where the façades of a building are overlaid with the ground and other buildings) pose significant challenges for effectively reconstructing reliable building structures from SAR images.

Recently, a new paradigm has been advocated in the literature [14, 15, 16, 17, 18] in which SAR visual semantics is emphasized and exploited to improve the SAR 3D imaging quality, instead of by collecting more observations as done in the traditional approaches. Our current work is a small exploring step in this field and is concentrated on what and how SAR visual semantics could be extracted and used in TomoSAR for modeling 3D buildings. Our main contributions are two-fold:

1. The three-level (low, middle, and high) SAR visual semantics are investigated in TomoSAR, and SAR semantics analysis is extended to the sparse 3D points initially recovered by traditional TomoSAR techniques.

2. The efficiency and effectiveness of exploiting SAR visual semantics in TomoSAR are demonstrated for 3D modeling of buildings on a real TomoSAR dataset.

### 2 SAR VISUAL SEMANTICS EXTRACTION

Here “SAR visual semantics” refer to the scene conceptual information extracted directly from SAR images, in particular structural and geometrical information. The main structure of an urban building comprises footprints, façades, and roofs, and the imaging regions (e.g., layover and shadow) contain pixels of different
intensities corresponding to different structures. As shown in Figure 2, different structures (e.g., the patches centered at points $a$, $b$ and $c$ corresponding to the terrain plane, façade, and roof, respectively) with the same slant range are imaged in the same region, and the pixels in the double-bounce regions have the highest intensities because the corresponding signals are reflected twice. Moreover, double-bounce regions have salient structures such as single- or double-line segments due to the different directions of the façades, and a small surface patch can be modeled as a planar one. Overall, different geometric primitives exist in the SAR images.

Based on the complexity and importance of geometric primitives (e.g., pixels, line segments, and planes), SAR visual semantics is divided into three levels in the following: low, middle, and high. Next, each level of visual semantics is separately discussed.

2.1 Low-level SAR visual semantics

Low-level visual semantics typically exist at the pixel level and are characterized by differences in intensities and structure types between neighboring pixels and the intensity distributions of the pixels in different regions. Generally, the main applications of low-level visual semantics include (1) detecting the initial position or shape of targets (e.g., facades and roofs corresponding to double-bounce regions), (2) improving the accuracy of pixel-level structure inference using the structural similarity between neighboring pixels, and (3) jointly processing neighboring pixels to improve the efficiency of structure inference.

2.1.1 Intensity distribution of different regions

In a SAR image, different regions have different intensity distributions which usually represent the characteristics of different scene structures. Therefore, constructing intensity distributions for specified regions (e.g., double-bounce regions) is helpful for reconstructing or recognizing the corresponding buildings [19]. For example, as shown in Figure 3, there is a considerable difference between the intensity distributions of double-bounce regions (that can be detected using line segment detection methods based on the results produced by the methods in Section 2.1.3) and other regions (e.g., the former and the latter have high probabilities of producing high- and low-intensity values, respectively).

Formally, intensity distributions can be represented as a multivariate Gaussian distribution, Fisher distribution, or skewed distribution estimated using polynomial fitting, deep learning, or other parameter estimation methods. For the examples shown in Figure 3(b), a five-layer neural network (in which three fully
connected layers are used to capture complex patterns in the data points) can produce better results, and a six-order polynomial fitting can also produce similar results but is not robust to noise and outliers.

Figure 3  Intensity distributions of different regions. (a) Example regions (red: double-bounce regions; yellow: other regions); (b) Fitted intensity distributions corresponding to double-bounce regions (blue) and other regions (black).

2.1.2 Geometric constraints between neighboring pixels

The intensities and structure types (e.g., face and roof) of two neighboring pixels are likely to be similar. In practice, neighboring pixels can be handled using image segmentation, morphological operations, and global optimization to improve the performance of 3D reconstruction or to highlight middle-level visual semantics. Specifically, image segmentation can be used to cluster multiple neighboring pixels with similar intensities into a single region or superpixel. According to the scene piecewise planar assumption, the 3D surface patch corresponding to a superpixel can be modeled as a planar one, which can help produce complete building structures by inferring the plane for each superpixel. Different image segmentation methods generally yield different results. For example, as shown in Figure 4(a) and (b), the superpixels produced by the SLIC [20] have approximately the same size and can be subsequently handled similarly (e.g., with the same parameter settings in the corresponding algorithms). The superpixels produced by the mean-shift method [21] are irregular but more consistent with real structures (e.g., double-bounce regions). Moreover, different parameters settings produce different results for the same image segmentation method. In general, small-scale image segmentation is unfavorable for extracting constraints from more pixels, and large-scale image segmentation may lead to superpixels that are inconsistent with real structures (e.g., one superpixel corresponds to two or more planes). Therefore, different applications require different image segmentation methods, and multiscale image segmentation could be considered if needed.

In addition, since pixels in multiple-bounce regions usually have high intensity values, as shown in Figure 4(c), these pixels can be highlighted after binarization. Furthermore, neighboring pixels can be connected to construct regions with salient structures (e.g., single- or double-line segments) through morphological operations (e.g., opening and closing operations), while eliminating isolated pixels or noise (Figure 4(d)).

2.1.3 Structure constraints between neighboring pixels

To infer the underlying structure (different from geometric constraints that denote the location relation between neighboring pixels, the structure is referred in particular to the component primitive of a building, such as façade and roof, and the structure constraint denotes that the component primitives corresponding to
neighboring pixels are likely to be similar according to the piecewise planar assumption) for each pixel and improve the global structure consistency between neighboring pixels, the structure can be globally optimized under the Markov Random Field (MRF) framework by encouraging two neighboring pixels to have the same structure. To this end, the common energy function is defined as

\[
E(f) = \sum_{p \in I} E_p(f_p) + \rho \cdot \sum_{q \in N(p)} E_{pq}(f_p, f_q)
\]

(1)

where \(f_p\) denotes the structure label assigned to the current pixel \(p\) in the SAR image \(I\), \(N(p)\) denotes the set of pixels neighboring to pixel \(p\), \(E_p(\cdot)\) and \(E_{pq}(\cdot)\) denote the unary and pairwise potentials, respectively, and \(\rho\) is the weight.

In Eq.(1), the unary potential measures the cost of assigning the structure label \(f_p\) to the current pixel \(p\) and is constructed according to the intensity distribution (e.g., the negative logarithm of the intensity distribution for detecting double-bounce regions). Moreover, the pairwise potential is used to encourage two neighboring pixels to have the same structure label and is generally constructed according to the intensity difference between the two neighboring pixels (e.g., a large penalty will be generated while assigning different structure labels to two neighboring pixels with a small intensity difference). As shown in Figure 4(e), based on the results of OTSU binarization, the MRF optimization can produce better results (e.g., more salient line structures and fewer small regions) than morphological operations.

2.2 Middle-level SAR visual semantics

Middle-level visual semantics typically exist in the form of specific geometric primitives (e.g., line segments and rectangles) to represent the local shape and size of the target. The main applications of middle-level visual semantics include (1) detecting target categories and spatial relationships, and (2) inferring the underlying structures of targets.

2.2.1 Double-bounce region detection

As shown in Figure 2(b), double-bounce regions [22] correspond to the intersection locations between the façade and terrain plane, and they are helpful for determining the locations and sizes of the buildings. Formally, double-bounce regions can be represented as structures with single- or double-line segments and are defined as follows:

\[
L = (L_1, L_2) = (p; \varphi_1, \varphi_2, l_1, l_2, w_1, w_2)
\]

(2)

where \(L_i\) denote the component line segments, \(\varphi_i\), \(l_i\), and \(w_i\) denote the angle between line segment \(L_i\) and the azimuth direction, length, and width, respectively, and \(p\) represents the intersection point of two
component line segments.

In practice, double-bounce regions are frequently contaminated by many interfering factors (e.g., noise and surface roughness), which leads to problems such as irregular edges, inaccurate lengths, and the presence of several small sub-regions. Therefore, to improve the reliability of double-bounce region detection, it is necessary to simultaneously adopt the following processing steps:

Step 1: Highlight the double-bounce regions through binarization, morphological operations, or MRF optimization.

Step 2: Detect initial line segments using existing line segment detection methods.

![Figure 5 Double-bounce region detection. (a) Initial line segments (red) and boundary sweeping directions (arrow); (b) Double-bounce regions represented by minimum area bounding rectangles; (c) Structural regions corresponding to façades, footprints, and roofs; (d) Partitioning candidate regions into equally sized sub-regions.](a) (b) (c) (d)

Step 3: Cluster the initial line segments to generate the directions corresponding to façades and regularize the initial line segments using the resulting directions.

Step 4: Sweep the boundaries along the two directions vertical to the regularized line segments until the candidate boundary no longer contains any pixels.

Step 5: Determine the double-bounce regions using the minimum area bounding rectangles based on the resulting boundaries.

As shown in Figures 5(a) and (b), based on the initial line segments detected in the double-bounce regions, the corresponding lengths and widths are determined by sweeping the boundary based on the number of pixels in each candidate boundary.

2.2.2 Structure region detection

Based on the constraints constructed by the double-bounce regions, the structural regions corresponding to façades, footprints, and roofs can be detected to improve the reliability of the subsequent 3D reconstruction. As shown in Figure 5(c), these regions frequently exhibit the following features:

(1) Footprints are located on the side of the double-bounce regions away from the radar, and their lengths and widths can be estimated using the size of the double-bounce regions.

(2) The façades and roofs are located between the radar and double-bounce regions, and the roofs are far away from the double-bounce regions.

Therefore, these building structure regions can be detected using the following steps:

Step 1. Candidate regions that are likely to contain façades and roofs are partitioned into superpixels or equally sized sub-regions (Figure 5(d)).
Step 2. The resulting superpixels or sub-regions are parsed into different structures based on the intensity distribution. To improve reliability, the elevations of a few pixels can be incorporated into the intensity distribution, and the corresponding measure can be defined as follows:

\[ M(r) = D(r) + \epsilon \cdot E(r) \]  

(3)

where \( D(r) \) denotes the probability of region \( r \) belonging to different structures (the correlation between intensity and structure can be modeled using deep learning methods), and \( E(r) \) denotes the elevation distribution generated using clustering or density estimation methods (i.e., K-means and Gaussian mixed model) based on the elevations corresponding to the randomly sampled pixels from region \( r \), and \( \epsilon \) is the weight.

Step 3. If the double-bounce region is a double-line segment structure, the footprint can be determined using the parallelogram containing the component line segments; otherwise, the width or length is determined using the width or length of the single-line structures and the length or width of the roofs detected using Steps 1 and 2.

2.3 High-level visual semantics

High-level visual semantics refers to the category and global geometric information (e.g., height and layout) of single or multiple targets. The main applications of high-level visual semantics include (1) understanding the global structure with different target positions and sizes, and (2) higher-order constraints constructed to improve the global accuracy and efficiency of the pixel-level 3D reconstruction.

2.3.1 Building layout prior

In practice, urban buildings usually have the following structure and layout priors:

1. Single building: The main structure of a building is composed of basic geometric shapes (e.g., cubes and cones), which can be further decomposed into multiple planes intersecting at specific angles (e.g., 90°).

2. Multiple buildings: Multiple buildings are arranged according to specific plans, and a variety of potential geometric relationships (e.g., coplanarity and parallelism) frequently arise across different planes associated with different buildings (specifically for two neighboring buildings).

Therefore, building layout priors can be detected through the following steps:

1. Collinearity detection and line segment regularization. Detect the collinearity of the initial line segments in double-bounce regions and utilize the resulting line segments to regularize the initial line segments.

2. Structure exploration. Detect the lines which intersect the lines associated with double-bounce regions at specified angles, and further explore the potential line segments corresponding to facades and roofs.

3. Region-based structure regularization. According to the piecewise planar assumption that a small 3D surface patch can be modeled as a planar one, the structure (e.g., façade and roof) types associated with the pixels in the corresponding region in the SAR image are likely to be the same. Therefore, the problem can be solved by assigning an optimal plane to each region in the SAR image under the energy minimization framework by incorporating building layout priors. Generally, given the set of regions \( (R) \) generated by image segmentation methods and the set of lines \( (L) \) corresponding to colinear line segments detected in double-bounce regions, the energy function can be formulated as

\[ E(f) = \sum_{r \in R} E_r(f_r) + \alpha \cdot \sum_{s \in N(r)} E_{s,r}(f_r, f_s) + \beta \cdot E_L(f_L) \]  

(4)

where \( f_r \) denotes the structure label assigned to the current region \( r \in R \), \( f_L \) denotes the building complexity measured by the number of the lines associated with structure labels (i.e., the larger the \( f_L \) value, the more
complex the buildings), and $E_r(\cdot)$, $E_{s,r}(\cdot)$ and $E_L(\cdot)$ represent the unary term, pairwise, and high-order potentials, respectively, $\alpha$ and $\beta$ are the penalty factors.

In Eq. (4), the unary potential measures the cost of assigning the structure label $f_r$ to the current region $r$, and can be formulated based on the intensity distribution of the regions and the distance between the structure $f_r$ and the elevations of the randomly sampled pixels in the current region $r$. In contrast, the pairwise potential is used to encourage two neighboring regions to have the same structure label and are generally constructed based on the average intensity difference between two neighboring regions. Moreover, the high-order potential is used to encourage the regions lying on the same line to have the same structure label and can be defined based on the variance of the average intensities of these regions (i.e., the smaller the variance, the greater the probability that these regions belong to the same structure).

### 2.3.2 Building height estimation

In a SAR image, different types of regions (e.g., layovers and shadows) contain different structural information about a building. Therefore, the correlations between these regions can be simultaneously considered to infer the height of the building. For the imaging regions of the buildings, the vertical sides of the façades are parallel to the range direction, and the building height $h$ is related to the length $L$ [23] of the layover containing the radar-visible façades by:

$$h = \frac{L}{\cos\theta}$$

(5)

where $\theta$ is the incidence angle, as shown in Figure 2(a).

![Figure 6 Building height estimation. (a) Double-bounce regions (red) and different regions $M$ (yellow) corresponding to different $L$ (thine dashed lines with different colors). (b) Final $M$ and $L$ generated by Res2Net framework and the projected points (yellow) from initial 3D points.](image)

To estimate the building height, taking a box-like building as an example, let $G(h, w, l)$ ($h$, $w$, and $l$ denote the height, width, and length) and $M$ represent the geometric shape and imaging region of a building, respectively. The reconstruction of the building can be formulated as the computation of the following probability $P(G|M)$.

$$P(G|M) \propto C(w, l) \cdot P(h|M) = C(w, l) \cdot C(\theta) \cdot P(L|M)$$

(6)

where $C(w, l)$ and $C(\theta)$ denote the constants related with the footprint and SAR incidence angle,
respectively, and $L$ denotes the length along the range direction of the building.

As shown in Figure 6(a), the correlation between imaging region $M$ and length $L$ corresponding to the height $h$ is clear. Therefore, when a reliable imaging region $M$ is determined, the length $L$ or height $h$ can be obtained. Generally, the probabilities $P(L|M)$ with different candidates $M$ and $L$ can be computed based on the output of the logistic model, which takes the features extracted by multiscale convolutional neural networks as input. Then, the optimal $M$ and $L$ with the largest probability $P(L|M)$ can be selected from the candidates $M$ and $L$. Figure 6(b) shows the optimal $M$ generated under the Res2Net [24] framework, and the corresponding length $L$ is consistent with the ground truth length. Note that length $L$ can be normalized using $w*l$ to address the scalability problem and set to a predefined range to improve the overall efficiency.

3. SAR VISUAL SEMANTIC EXTENSION

SAR visual semantics could be extended to the sparse 3D points recovered by the traditional techniques, such as the orthogonal matching pursuit (OMP) [25] and iterative shrinkage thresholding algorithm (ISTA) [26], and they could again be used to further infer the complete structures (e.g., façades and roofs) of the buildings. As shown in Figure 7(a), a 2D projection map can be generated by projecting initial 3D points on the terrain plane, where the “intensity” value of each projected point corresponds to the number of the 3D points that are projected into it. Based on this, the line segments and regions corresponding to the façades and roofs can be further detected using the aforementioned methods.

![Figure 7 Façade detection from initial 3D points. (a) Projection map generation; (b) Line segment detection; (c) Façade generation.](image)

3.1 Façade detection

In reality, since the majority of initial 3D points are distributed on radar-visible facades, the projected points corresponding to these facades usually have high “intensity” values. Therefore, after binarization and morphological operations of the projection map, these projected points can be connected as salient regions with either single or double-line segment structures. Note that façades can be rather reliably determined by detecting line segments in these regions as shown in Figures 7(b) and (c). In addition, the angle prior (e.g., 90°) can be used in this process to look for potential line segments by searching the regions along the vertical direction to reliably detect line segments.

3.2 Roof detection

Based on the detected line segments corresponding to façades, the roofs (including heights and boundaries) can be detected via a region-growing procedure by incorporating 2D and 3D cues, and the key steps are described as follows:

Step 1. Candidate seed point generation. For the current line segment, as shown in Figure 8(a), two
candidate seed points are first selected according to the two conditions: (1) lying on both sides of the midpoint of the line segment and (2) having the same vertical distance to the line segment.

Step 2. Height estimation. The mean height of the 3D points corresponding to the points around each candidate seed point is computed, and the final seed point with the maximum height is selected to perform the subsequent region-growing.

Step 3. Roof boundary detection. Starting with the final seed point, as shown in Figure 8(b), the roof boundary detection is performed by evaluating whether its neighboring points can be classified as the roof point, and if yes, these neighboring points are further taken as new seed points to repeat the same region-growing process. As shown in Eq(7), three 2D and 3D cues are jointly used in the stopping rule of this process: (1) the current point has low “intensity” values because the roof is generally parallel to the terrain plane; (2) its corresponding 3D point has the same height as that of its neighboring seed point; (3) the “intensity” values of the points beyond the roof boundary are usually close to zero.

\[ M(S, p) = (D(S, p) < \omega) \land (H(S, p) = \text{True}) \land (I(p) \neq 0) \]  \hspace{1cm} (7)

where \( S \) and \( p \) denote the current seed point and its neighboring points obtained using the Delaunay triangulation method, respectively, \( D(S, p) \) denotes the intensity difference between points \( S \) and \( p \), \( H(S, q) \) denotes whether points \( S \) and \( p \) are on the same roof (or have the same roof height), and \( I(p) \) denotes the intensity value of point \( p \).

Step 4. After region growing, as shown in Figure 8(c), the resulting roof points are delineated using the minimum area bounding rectangle as the final roof boundary.

![Figure 8](https://engine.scichina.com/doi/10.1360/nso/20230067)

Figure 8 Roof detection. (a) Candidate seed point (white and black points) generation according to the current line segment (red); (b) Roof boundary detection by region-growing (red points: the projected points around the selected seed point; green points: the projected points produced after region growing); (c) Roof boundary indicated by the minimum area bounding rectangle.

The above process is summarized in the following Algorithm 1.

**Algorithm 1. Roof point detection and height optimization.**

Input: an initial seed roof point.

Output: roof points and optimal roof height.

Initialization: add the initial seed roof point to the list of seed points (\( \Phi \)), and set its label to true.

1: If \( \Phi \) contains the points with true labels:
2: Select a point (\( S \)) with true label from \( \Phi \) as the current seed point, and set its label to false.
3: Determine the set of the points (\( P \)) neighboring to the current seed point.
4: For each point \( p \in P \) that is not visited:
4.1: If the condition \( M(S, p) \) is met:
4.2: Add the point \( p \) to \( \Phi \) and set its label to true.
4.3: Update the roof height using the mean height of the 3D points corresponding to the points in \( \Phi \).
4.4: Else: discard the point \( p \).
4.5: End If
4.6: End For
5: Else: Terminate the roof point detection and height optimization.
6: End If
7: Output the points in \( \Phi \) (roof point) and roof height.

4 A CASE OF STUDY FOR BOX-LIKE BUILDING MODELING

For TomoSAR 3D reconstruction, the three levels of visual semantics are frequently interdependent and they collectively provide strong constraints for improving the reconstruction completeness, accuracy, and efficiency. This section introduces an efficient TomoSAR 3D reconstruction method that incorporates different visual semantics (e.g., superpixels and line segments) to produce box-like models of buildings. In the proposed method, only a few pixels (5%) are randomly selected to compute the elevations using the OMP method, in which the sparsity is set to 1. This largely alleviates the problem encountered in traditional methods, wherein the number of targets needs to be known in advance. A flowchart of the proposed method is shown in Figure 9.

![Figure 9 Flowchart of the proposed 3D modeling method.](https://engine.scichina.com/doi/10.1360/nso/20230067)

Next, each component step will be elaborated in the subsequent sections.

4.1 Dataset

Experiments are conducted on the Emei dataset collected by the airborne TomoSAR system proposed in [27]. The dataset contains 12 SAR co-registered images with 3600×1800 resolution, and the topography of the observed areas mainly contain box-like urban buildings situating at a relatively flat terrain plane.

4.2 Double-bounce detection

Based on the binarized image generated using MRF optimization as outlined in Section 2, small regions were first filtered out. For the remaining connected regions with salient line segment structures, the line segments were detected using the method proposed in our previous study [28], which performed line segment detection in the following three steps: (1) detecting the dominant line segments in each connected region; (2) exploring potential line segments based on dominant line segments using structure priors (e.g., the intersection angles between line segments); and (3) globally optimizing line segments by incorporating geometric constraints and structure priors. Finally, as shown in Figure 10, the resulting line segments (DB line segments) rather faithfully represent the double-bounce regions, which in turn provide key indicators for subsequent facade and roof detection.
4.3 Façade reconstruction

To facilitate the structure inference process, the two systems, the radar system (azimuth-range-elevation) system $C_{are}$ and the geodetic systems $C_{xyz}$, are alternatingly engaged in this work. For a DB line segment, to estimate the location of the corresponding façade in the geodetic system, we first provided a unified representation of the point/line transformation between two systems $C_{are}$ and $C_{xyz}$. Specifically, let the matrix $M_{3\times 3} = \{m_{ij}\} (i, j = 1,2,3)$ (that is constructed according to the TomoSAR imaging geometry) denote the transformation from the point $P(a, r, e)^T \in C_{are}$ to the point $Q(x, y, z)^T \in C_{xyz}$ as,

$$M_{3\times 3} \cdot P_{3\times 1} = Q_{3\times 1}$$ (8)

Thus, the relationship between elevation and pixel $(a, r)$ can be represented as follows:

$$s = (z - (m_{31} \cdot a + m_{32} \cdot r))/m_{33}$$ (9)

Then, the coordinates $x$ and $y$ can be computed using the following equation:

$$\begin{align*}
x &= m_{11} \cdot a + m_{12} \cdot r + m_{13} \cdot s \\
y &= m_{21} \cdot a + m_{22} \cdot r + m_{23} \cdot s
\end{align*}$$ (10)

According to Eq. (10), when the terrain plane $H \in C_{xyz}$ is known, the points in it can be directly computed from the pixel $(a, r)$. Furthermore, for the experimental DB line segment $l \in L_{are}$ shown in Figure 10, the corresponding line segment $\ell \in H$ can be computed using its two endpoints. As a result, as shown in Figure 11(a), line segment $\ell$ is consistent with the projected points on the terrain plane from the 3D points on the façade. Therefore, the façade can be directly generated because it is perpendicular to the terrain plane (Figure 11(b)). Note that, the terrain plane can be set according to the scene and SAR imaging geometry. In our experiments, the terrain plane was estimated by fitting 3D points produced from the DB line segments under the RANSAC (random sample consensus) framework [29] for robustness concern.

4.4 Superpixel-based roof detection
To effectively detect the roof, as shown in Figure 11(c), the layover regions with respect to the current DB line segment were first partitioned into a set of superpixels. Then, because each superpixel usually corresponds to one or more planes, including the terrain plane, façade, and roof, the roof was validated using the constraints constructed by the reconstructed façades and the elevations (or 3D points) computed from the randomly sampled pixels in the superpixel.

Specifically, from the initial 3D points (in the geodetic system) computed from the randomly sampled pixels, the possible 3D points belonging to the roofs can be determined by excluding the 3D points belonging to façades and terrain planes (according to the distance of the initial 3D points to the two planes). Generally, the 3D points belonging to the roof are not empty only when the current superpixel is associated with the layover region including the roof. Therefore, the roof can be determined using the mean height of these 3D points. Accordingly, Figure 11(d) shows the 3D points corresponding to all the pixels in the superpixel on the generated roof. In addition, because other buildings were imaged in the current layover region, the results could also contain false roofs. To address this problem, these 3D points were projected onto the terrain plane, and the roof regions were further detected by sweeping the boundary of the projected points along the direction vertical to the line segment corresponding to the façade, and the roof is indicated by the rectangle in Figure 12(a).

Finally, the detected roof is shown in Figure 12(b) by removing false roofs points, and the result looks satisfactory.

4.5 Box-like model generation

Based on the detected façade and roof (including height and boundary), a box-like model can be generated that represents the complete structure of the building. Note that, for double-bounce regions with double-line segments, only the long DB line segment is selected to reconstruct the box-like models for improving efficiency. Moreover, to quantitatively evaluate the accuracy at each stage of the proposed method, the following criteria are defined:

1. $M_1$: The criterion is used to evaluate the accuracy of the line segment detected in double-bounce regions. For the detected line segment $x$ and ground truth line segment $y$, let $L(x,y) = \left( \sum_{m=1}^{2} d(x_m,y) + \sum_{n=1}^{2} d(y_n,x) \right) / 4$ where $x_m$ and $y_n$ denote the endpoints of line segments $x$ and $y$, respectively, and $d(\cdot)$ denotes the distance between a point and a line, the criterion is defined as the average of the $L(x,y)$ values with respect to all the detected line segments.

2. $M_2$: The criterion is used to evaluate the accuracy of the reconstructed façade. For the reconstructed façade $x$ and the set of the corresponding initial 3D points $y$, let $P(x,y) = \left( \sum_{i \in y} (D(i,x)/D_{\text{max}}) \right) / |y|$ where $D(i,x)$ and $D_{\text{max}}$ denote the distance between the 3D point $i$ and plane $x$, and the maximum of all the
\(D(i, x)\) values, respectively, the criterion is defined as the average of the \(P(x, y)\) values with respect to all the reconstructed facades.

(3) \(M_3\): The criterion is used to evaluate the accuracy of the estimated roof height. For the estimated roof height \(x\) and ground truth roof height \(y\), let \(H(x, y) = \left| x - y \right| / y\), the criterion is defined as the average of the \(H(x, y)\) values with respect to all the estimated roof heights.

(4) \(M_4\): The criterion is used to evaluate the accuracy of the detected roof boundary. For the detected roof boundary \(x\) and ground truth roof boundary \(y\), let \(I(x, y)\) denote the Intersection over Union (IoU) value that is commonly used in image semantic segmentation, and the criterion is defined as the average of the \(I(x, y)\) values with respect to all the detected roof boundaries.

(5) \(M_5(\alpha, \beta)\): The criterion is represented as the number of reliably reconstructed box-like models where a box-like model is considered reliable when its \(H(x, y)\) value is smaller than the threshold \(\alpha\), and \(I(x, y)\) value is larger than the threshold \(\beta\).

In the five criteria, the ground truth line segments and roof boundaries are manually annotated in the SAR image or the projected map on the terrain plane from initial 3D points, and the ground truth roof height is set to the average height of the 3D points that are projected in the annotated roof boundary.

Overall, as shown in Table 1, each stage of the proposed method can produce reliable results, which indicates SAR visual semantics play some important roles in improving the accuracy and completeness of TomoSAR 3D reconstruction. More importantly, the process is substantially faster than time-consuming pixel-by-pixel elevation computation. In our experiments, the proposed method (implemented in Matlab code) takes approximately 10 seconds to generate box-like models for all buildings.

<table>
<thead>
<tr>
<th>ID</th>
<th>Stages</th>
<th>Accuracy</th>
<th>Time</th>
<th>Number of primitives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Double-bounce detection</td>
<td>(M_1)</td>
<td>1.46</td>
<td>3.5</td>
</tr>
<tr>
<td>2</td>
<td>Façade reconstruction</td>
<td>(M_2)</td>
<td>0.64</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>Roof detection</td>
<td>(M_3 &amp; M_4)</td>
<td>0.55&amp;0.87</td>
<td>4.7</td>
</tr>
<tr>
<td>4</td>
<td>Box-like model generation</td>
<td>(M_5(0.2, 0.8))</td>
<td>9</td>
<td>0.4</td>
</tr>
</tbody>
</table>

**Table 1.** Accuracy and running time (second) on the Emei dataset.

**Figure 13** Box-like model generation. (a)-(b) Single box-like model (two views); (c) All box-like models; (d) 3D points produced by the OMP method; (e) Ground truth box-like models.
Figure 14 Number of box-like models with thresholds \( \alpha \) and \( \beta \). (a) Threshold \( \alpha \); (c) Threshold \( \beta \).

Figure 13 shows the box-like models corresponding to single DB line segment (Figure 13(a)-(b)) and all DB line segments (Figure 13(c)). Compared with the sparse 3D points produced by the OMP method (Figure 13(d)), the resulting box-like models are more reliable for representing the complete structure of the buildings; moreover, they are consistent with the ground truth box-like models within the permissible range of error (e.g., the errors indicated by the rectangles). To this point, the number of box-like models with different thresholds \( \alpha \) and \( \beta \) are shown in Figure 14. Obviously, the proposed method achieves better results when \( \alpha \geq 0.2 \) and \( \beta \leq 0.8 \), demonstrating that it could perform steadily.

To further validate the effectiveness of the proposed method, it is also tested on Yuncheng dataset where a stack of 8 coregistered SAR images are provided. As shown in Figure 15 and Table 2, similar satisfactory results are obtained, and details are omitted due to the space limit.

Table 2. Accuracy and running time (second) on the Yuncheng dataset.

<table>
<thead>
<tr>
<th>ID</th>
<th>Stages</th>
<th>Accuracy</th>
<th>Time</th>
<th>Number of primitives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Double-bounce detection</td>
<td>( M_1 )</td>
<td>1.72</td>
<td>6.3</td>
</tr>
<tr>
<td>2</td>
<td>Façade reconstruction</td>
<td>( M_2 )</td>
<td>0.49</td>
<td>5.7</td>
</tr>
<tr>
<td>3</td>
<td>Roof detection</td>
<td>( M_3 &amp; M_4 )</td>
<td>0.75&amp;0.82</td>
<td>11.2</td>
</tr>
<tr>
<td>4</td>
<td>Box-like model generation</td>
<td>( M_5 (0.2,0.8) )</td>
<td>21</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Overall, different visual semantics play important roles for reconstructing the box-like models of the buildings. For example, the low-level visual semantics are used to highlight the double-bounce regions, and the middle- and high-level visual semantics are used to determine the façade locations and the roof
boundaries, respectively. Consequently, the proposed method performs well in term of accuracy and efficiency.

5 CONCLUSION AND LIMITATION

This study introduces some detection and exploitation methods of SAR three-level visual semantics in TomoSAR 3D reconstruction. Specifically, low-level visual semantics are characterized by the intensity distribution, location relationship, and structural constraints between neighboring pixels, which are the foundation of mid-level visual semantics represented by geometric primitives (e.g., line segments and rectangles). Relatively high-level visual semantics refer to building types and geometric information (e.g., height and angle). In practice, different visual semantics are interdependent, complementary, and should ideally be jointly used to improve the accuracy and efficiency of TomoSAR 3D reconstruction. To validate the effectiveness of visual semantics, a box-like building modeling method is proposed by incorporating different visual semantics such as line segments and superpixels. The experimental results confirm that SAR visual semantics can indeed provide strong constraints to efficiently guide the process of TomoSAR 3D reconstruction and produce reliable results.

As we said at the very beginning, this work is only a tentative attempt on the exploitation of SAR visual semantics in TomoSAR technology, its limitations are numerous in both technical level and semantic representational level. From technical level, clearly the accuracy of the roof detection could be affected by the size of the superpixels, and building modeling accuracy is limited by the accuracies of the estimated elevations and DB line segments. From semantic representational level, since various facades and roofs exist, detecting such entities from SAR images are difficult per se. If too much efforts are put into the SAR semantics extraction, its desired role in SAR 3D imaging will be largely compromised, hence some sort of balance between 3D imaging and semantics extraction should be explored. Another point seems a more intriguing one. Currently, deep-learning based representations are invariably in the form of “implicit representation”, how to exploit such implicit representations of the scene semantics in TomoSAR imaging framework seems a promising but difficult task. Currently for our immediate future work, we would test our proposed 3D modeling method on more complicated scenes, for example, cluttered buildings with several possible roof structures.

Acknowledgements

We are grateful to Journal of Radars for sharing the data for this study.

Funding

This work was supported by the National Natural Science Foundation of China (61991423, 62376269) and the Key Scientific and Technological Project of Henan Province (232102321068).

Author contributions

W.W., Q.D., and Z.H. designed the research; W.W., H.W., and L.Y. carried out the experiments; W.W., H.W., and L.Y. analyzed the data. W.W., Q.D., and Z.H. wrote the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Data availability

https://engine.scichina.com/doi/10.1360/nso/20230067
The original data are available from corresponding authors upon reasonable request.

References


23. Sun Y, Mou L, Wang Y, Montazeri S, Zhu X. Large-scale building height retrieval from single SAR imagery based on...


