URBAN CLASSIFICATION BASED ON TOP-VIEW POINT CLOUD AND SAR IMAGE FUSION WITH SWIN TRANSFORMER

R. Xue [1,2], X. Zhang [2]*, U. Soergel [2]

[1] National Lab of Radar Signal Processing, Xidian University, 710071 Xi’an, China - rhxue@stu.xidian.edu.cn
[2] Institute for Photogrammetry, University of Stuttgart, 70174 Stuttgart, Germany - (ruihang.xue, xinlong.zhang, uwe.soergel) @ifp.uni-stuttgart.de

KEY WORDS: Deep Learning, Transformer, Feature Fusing, Urban Classification, Synthetic Aperture Radar, Point Cloud.

ABSTRACT:

Urban areas are complex scenarios consisting of objects with various materials. This variety poses a challenge to single-data classification schemes. In this paper, we propose a feature fusion and classification network on RGB top-view point cloud and SAR images with swin-Transformer. In this network, the heterogeneous features are learned separately by an asymmetric encoder, and then they are concatenated along the channel dimension and fed into a fusing encoder. Finally, the fused features are decoded by an UperNet for generating the semantic labels. As data we use the subset of high-resolution 3D point cloud provided by Hessigheim benchmark which are complemented by TerraSAR-X images. The overall precision and the mean intersection over union (mIoU) achieves 87.25% and 73.56%, respectively, which outperforms the single-data swin-Transformer by 4.08% and 1.91%, respectively.

1. INTRODUCTION

Thanks to their high-resolution and altitude information, large-scale point clouds play an important role in urban planning, ecosystem monitoring, and land resources analysing. Urban areas are complex scenarios consisting of large numbers of objects with various materials. This variety poses a challenge to any single-data classification scheme, for example, solely based on RGB point clouds. Synthetic aperture radar (SAR) images reflect the backscattering intensity of objects and show strong contrast between artificial materials and natural objects. Therefore, the RGB point clouds and the SAR images provide complementary features which is advantageous especially in urban areas. Fusing such data by deep networks could effectively improve the performance for urban classification (Kahraman and Bacher, 2021).

The available multi-source data fusion methods can be grouped into pixel-level fusion, feature-level fusion and decision-level fusion from low to high. Pixel-level fusion (Kulkarni and Rege, 2020) firstly maps heterogeneous data from the natural domain to the high-dimensional fusion domain. After superposition or component substitution, the fused data are mapped back to the natural domain. Feature-level fusion (Zhang et al., 2021) extracts features from the respective data and then combines them by summation, splicing, etc. For decision-level fusion (Waske and van der Linden, 2008), a series of classifier are designed according to the characteristics of multi-source data, and the classification results are obtained by voting or filtering on each output of the classifier. Among them, the pixel-level fusion requires to define elaborate pre-mapping functions. Although decision-level fusion is simpler and computationally efficient, the spatial information interaction between heterogeneous data is ignored. Therefore, in this paper, we explore feature-level fusion and propose a data-driven classification network based on swin-Transformer.

A Transformer is a deep neural network aiming at global perception governed by a so-called self-attention mechanism; such schemes achieved state-of-the-art performance in computer vision (Dosovitskiy et al. 2020). The self-attention mechanism calculates the attention score by key-value querying to abstract the essential relationship among all elements. Such elements could be feature vectors or a time-step of the sequential data. Compared with traditional convolution networks of strong inductive bias, Transformer do not rely on the prior of distance-based receptive field, but focus on the correlation of elements. Therefore, Transformer could show more adaptive to multi-source data. However, the global and fixed perception range of Transformer requires more learnable parameters, especially on high-resolution images. Such parameters may lead to complex network and heavy calculational burden (Li et al., 2018). In addition, a Transformer is a scale-sensitive network where the variation of the object size is regarded as new patterns. To reduce the parameter redundancy, swin-Transformer (Liu et al., 2021) limits self-attention calculation to non-overlapped windows. Such windows are cycling shifted among adjacent layers to trade off the calculation burden and perception range. Furthermore, the multi-scale feature extraction is realized by cascading swin-Transformer blocks with patch-merging layers.

The primary contribution of this paper is on data fusion and classification of RGB top-view point clouds and SAR images. We propose a data-driven swin-Transformer network, which firstly registers the SAR image to the RGB top-view point cloud of the same scene manually. Then several swin-Transformer blocks are connected as an asymmetric encoder and a fusing encoder to realize individual feature extraction and feature fusion, respectively. Finally, the features of each block are fed into a unified perceptual parsing network (UperNet) to obtain semantic labels. The proposed network only requires that the input images belong to the same region, and has high adaption to the sensor and image resolution. In the experiments

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-559-2022 | © Author(s) 2022. CC BY 4.0 License.

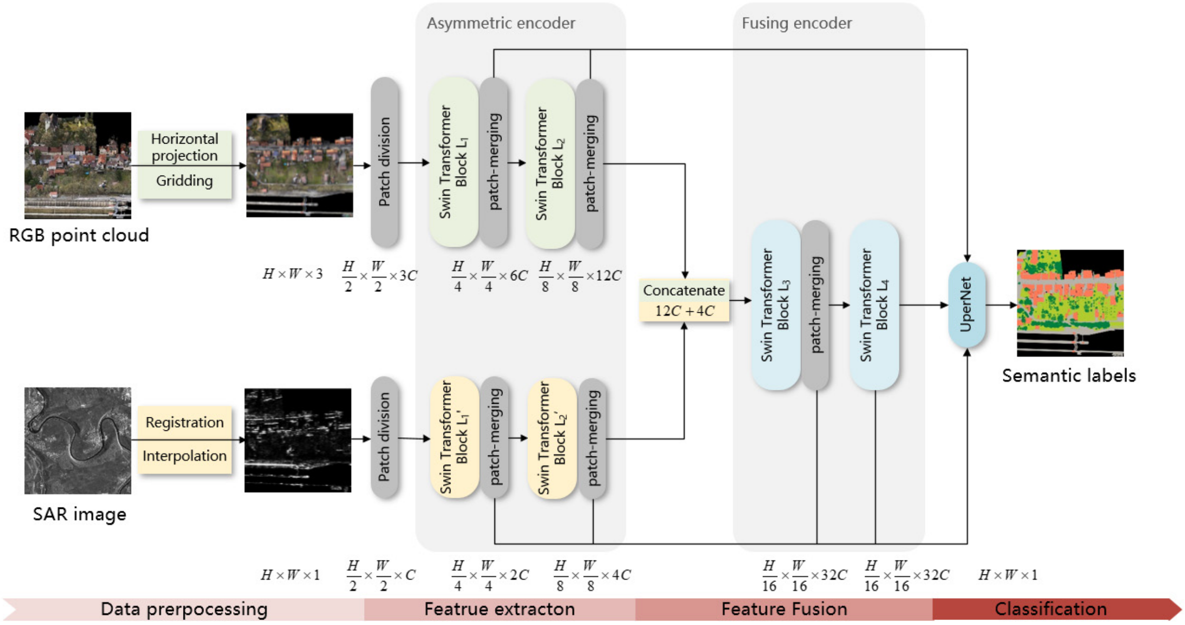


Figure 1. Structure of the proposed network.

of urban classification on a subset of the Hessigheim high-resolution 3D point cloud (H3D) benchmark data and a TerraSAR-X image dataset, the proposed network has better accuracy and mean intersection over union (mIoU) than fully convolutional network (FCN) and a data-level fusion method.

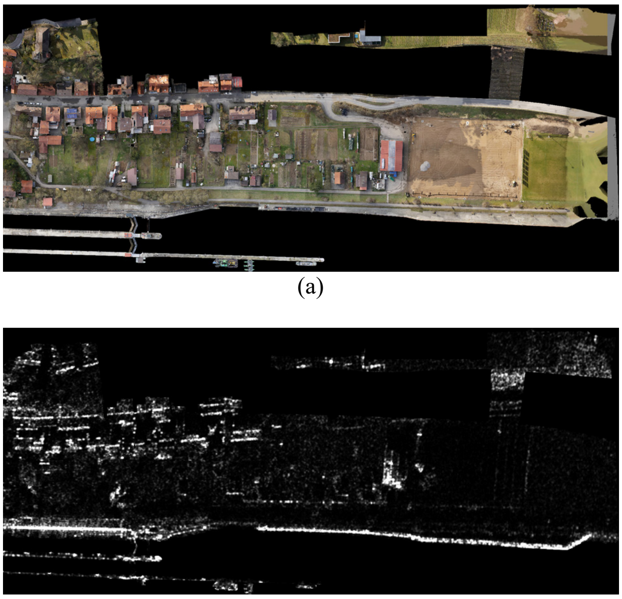
2. THE PROPOSED NETWORK

The structure of proposed network is shown in Figure 1, which can be separated to 4 stages: data pre-processing, feature extraction, feature fusion and classification. Firstly, the pre-processing on point cloud and SAR image data generates two images with the same size corresponding to one region. Secondly, the top-view image together with the SAR registration image is fed into an asymmetric encoder for individual feature extraction. Thirdly, the features extracted from top-view image and SAR image are concatenated along the channel dimension and fed into a fusing encoder. The asymmetric encoder and the fusing encoder are both realized by a 2-block swin-Transformer. Finally, the fused features are decoded by an UperNet for generating semantic labels. The detailed process of the network is given below.

2.1 Data Pre-processing

As a dual-input network, the input 3D point cloud and SAR image need to be regulated as two images corresponding to the same site and size, i.e. a 3-channel RGB Top-view point cloud image, and a single-channel SAR intensity image. For the 3D point cloud data, it is firstly projected to a horizontal plane, and the whole scene is divided to grids with 0.05m×0.05m. Then the RGB value corresponding to the top point in each grid is regarded as the RGB value of a pixel. If there is no point in the grid, we set the pixel value to (0,0,0). The prepossessed RGB top-view image is shown in Figure 2(a).

For the SAR image, routine transformations including filtering, radiation correction and geocoding (Roth et al., 2004) are performed on a large-scale SAR image. Then a sub-image corresponding to the region of point cloud data is captured. Although the RGB top-view image and the SAR image have different resolutions, imaging planes and visual characteristics, we can still identify some homologous points belong to the specific targets, for example, the docks, bridges, and intersections (Yamamoto et al., 2015). Picking these points from the RGB top-view image and the SAR image, the two images could be registered by affine transformation. Finally, the SAR image with identity size of the RGB top-view image is obtained by bilinear interpolation, which is shown in Figure 2 (b). It can be observed that because of the hand selected key points, there are some misplacements on image registration. Whereas, the swin-Transformer blocks could tolerance such registration errors within the confines of a window, which is realized by the position-insensitive of key-value query calculation.


(a)

(b)

Figure 2. Data obtained by pre-processing for (a) top-view image, and (b) SAR image.

2.2 Swin-Transformer Block

As part and parcel of the proposed network, the swin-Transformer block acquires the correlation among feature elements by self-attention. Single-layer swin-Transformer block can be divided into 4 steps: window division, scaled dot-product attention calculation, window shifting and residual connection, as shown in Figure 3.

Compared with the classic Transformer, the swin-Transformer block divides the input feature maps to non-overlapped windows along image length and height. The self-attention calculation within these windows can effectively reduce the difficulty of relation estimation, so that make the network training easier to converge. Considering a feature map with the size of $X \times Y$ and the window size of $D \times D$, we can get local windows with the number of XY / D^2 for each channel. The standard multi-head scaled dot-product attention is applied in each window. For the l th layer, the features $z^l \in \mathbb{R}^{D^2 \times d}$ are mapped to the value $v^{l,h}$, key $k^{l,h}$, and query $q^{l,h}$ with the size of $D^2 \times d_h$ as

$$\begin{aligned} v^{l,h} &= \mathrm{LN}\left(z^l\right) W_v^h \\ k^{l,h} &= \mathrm{LN}\left(z^l\right) W_k^h , \\ q^{l,h} &= \mathrm{LN}\left(z^l\right) W_q^h \end{aligned} \tag{1}$$

where $W_v^h, W_k^h, W_q^h \in \mathbb{R}^{d \times d_h}$ are the conversion matrix, d is the feature dimension, d_h is the head dimension, h is the head index, and $\mathrm{LN}(\cdot)$ denotes the layer normalization. Then the attention score $a^{l,h}$ is calculated by scaled dot-product as

$$a^{l,h} = \mathrm{softmax}\left(\frac{q^{l,h} \cdot \left(k^{l,h}\right)^T}{\sqrt{d_h}} \right) \cdot v^{l,h} . \tag{2}$$

The attention scores of all heads can be concatenated and obtain the attention output $m^l \in \mathbb{R}^{D^2 \times d}$ satisfies

$$m^l = \mathrm{Concat}\left(a^{l,1},\ldots,a^{l,He}\right) W_t + z^{l-1} , \tag{3}$$

where $Concat(\cdot)$ denotes the concatenation along feature dimension, $W_t \in \mathbb{R}^{d_h He \times d}$ is the conversion matrix, He is the total number of the attention heads, and $z^{l\text{-}1}$ is added as a residual to accelerate training.

The attention output m^l successively can be further fused by a feedforward network (FFN), which is

$$z^l = \mathrm{LN}\left[\mathrm{FFN}\left(m^l\right)\right] + m^l , \tag{4}$$

where the FFN is comprised of linear connections and rectified linear unit (ReLU).

The window-based self-attention limits the receptive field to the range of $D \times D$, but it pays the price of global perception ability. The local information cannot be shared among different windows. To introduce the correlation learning among non-overlapping windows, the window shifting technique is adopted, which re-divides the windows between two self-attention calculations. The centre coordinate of each window is added an offset of half window size from (x, y) to $(x + D/2, y + D/2)$, so that the inter-window information is obtained. However, such window shifting will lead to the different window size at the edge of feature maps. The cycle-shift technique is applied to pad the upper left edge windows to the lower right edge, then the number and size of windows are consistent with the original window. Repeating the self-attention calculation of (1)-(4), and reversing the cycle-shift of the windows, the output of a single-layer swin-Transformer block is obtained. For a swin-Transformer block with L layers, we stack $2L$ times of self-attention calculations to ensure that the window is shifted at least once.

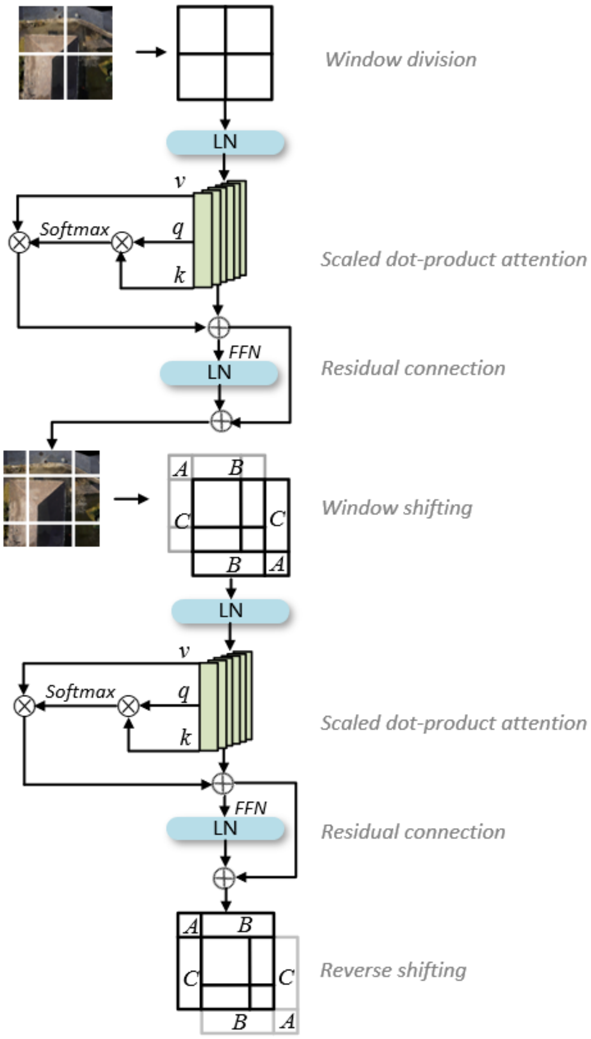


Figure 3. Structure of a single-layer swin-Transformer block.

2.3 Feature Extraction and Feature Fusion

In the proposed network, feature extraction and feature fusion are fulfilled by asymmetric encoder and fusing encoder, respectively. The asymmetric encoder has two individual flows for RGB top-view image and SAR image. Considering the object properties and resolution of the original data, such flows share similar structure but have different number of parameters. Consider a top-view image $\mathbf{X}_{RGB} \in \mathbb{R}^{3\times H\times W}$, where the size of each image is $H \times W$. Firstly, each input image is decomposed into patches without overlapping, each of size 2×2, so the patches totalled $3HW/4$. Then the 2×2 patches are projected to $3C$, where C is called embedding dimension.

In Transformer, a learnable position encoding is essential for relation learning, because the self-attention calculation is position independent. Hence, we add the position encoding to the patches and get the embedding $\mathbf{Z} \in \mathbb{R}^{H/2\times W/2\times 3C}$. For the top-view image flow, the embedding $\mathbf{Z}$ is fed into two stacked swin-Transformer blocks, where each block learns the relationship among patches. To realize multi-scale feature extraction, each swin-Transformer block is connected by a patch-merging layer except the output layer, where the feature dimension is doubled while the patch number is halved along height and width. The size of output feature map of the top-view image flow is $H/8 \times W/8 \times 12C$ as shown in Figure 1. The SAR image flow has the similar procedure with 1/3 of the embedding dimension, so that the size of output feature map of the SAR image flow is $H/8 \times W/8 \times 4C$.

After obtaining the feature maps corresponding to the two flows, the feature fusion is performed by a fusing encoder. Firstly, the two feature maps are concatenated along the feature dimension, i.e., a larger feature map sized $H/8 \times W/8 \times 16C$ is obtained. The fusing encoder is stacked by two swin-Transformer blocks with a patch-merging layer. Afterwards, the feature maps of such blocks have the size of $H/16 \times W/16 \times 32C$.

2.4 Classification

As a backbone network, swin-Transformer cannot realize pixel-level semantic outputs without a feature decoder. Here we perform UperNet as a feature decoder to classify each pixel and generate the semantic labels (Xiao et al. 2018). UperNet is a general scene, semantic and texture classification pipeline, which is realized based on the feature pyramid. For urban classification, only the semantic segmentation head is utilized for learning and evaluating.

The UperNet is connected to each output of the swin-Transformer blocks, where the last block is passed through a pyramid pooling module (Zhao et al. 2017)) to obtain multi-scale features. The other blocks are up-sampled and added to the multi-scale features, and finally obtaining a feature map with the same size of input image. The semantic labels are evaluated by a convolutional layer with kernel 1×1 on that feature map.

3. EXPERIMENTS

3.1 Experimental Configuration

The proposed network is evaluated on a subset of the H3D data set and a TerraSAR-X image. Such data combinations can be obtained from any labelled LiDAR point clouds and SAR images in the same region. Obviously, urban areas with rich ground objects will provide more complementary information for data fusing. The original H3D dataset is a high-precision point cloud segmentation benchmark based on the unmanned aerial vehicle platform (Kölle et al. 2021). The point could data with 800pts/m^2 is obtained by Riegl VUX-1LR LiDAR scanner at the region of Hessigheim, Germany. Besides, two oblique Sony Alpha 6000 cameras are applied to colorize the point cloud, and then manually mark the point cloud to 11 categories of semantic labels. Since the semantic labels of the original H3D dataset include mobile targets such as the vehicle, the small target chimney, and the targets are only valid in 3D processing such as façade and vertical surface, we remove some semantic labels to obtain a subset of H3D dataset with 4 labels, including low vegetation, impervious surface, roof and tree. The validation set is merged into the training set to get as complete an image of the city area as possible.

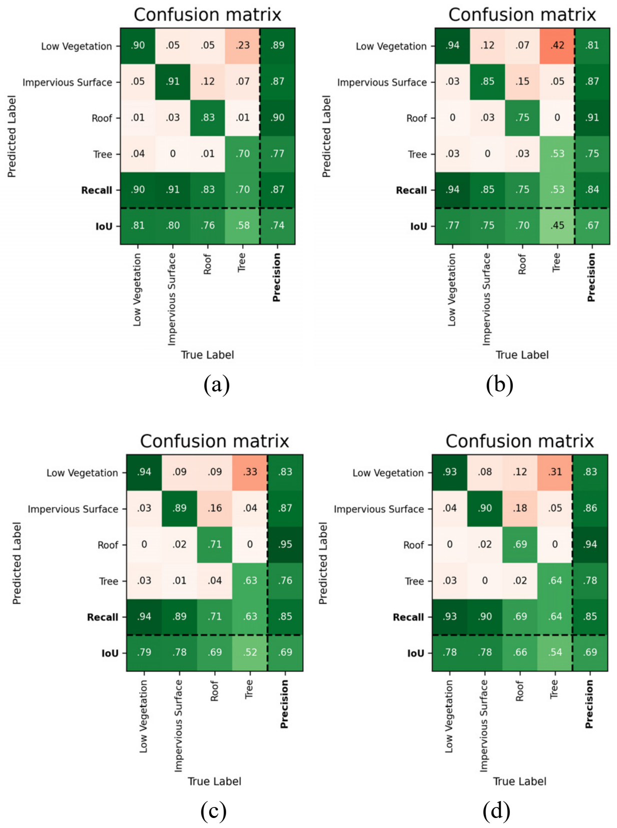
TerraSAR-X is a remote sensing satellite project led by the German aerospace center (DLR) (Roth et al. 2005), which could implement multiple SAR imaging modes based on active electronically steered array, and could obtain multi-polarization and multi-resolution SAR images in many areas of the world. In our experiments, an HH polarimetric SAR image of the Hessigheim region on March 21, 2018 is acquired, which is imaged in descending orbit with high resolution spotlight mode, and the spatial resolution is 1m.

After data pre-processing introduced in Section 2.1, the RGB top-view image and registered SAR image share the label space from the subset of H3D dataset. In order to ensure the input images having a fixed size, a sampling window with the size of 384×384 pixels slides on the training and test images with the step size of 192×192 pixels, and the samples without any objects are removed. Hence, 397 training and 254 test samples are obtained for both RGB top-view images and SAR images. In the training set, the input image is normalized, and then performed data augmentation such as random flipping, random brightness and contrast adjustment to enhance the network generalization.

The proposed network is implemented on a NVIDIA RTX3090 GPU with the framework of Pytorch 1.9. The image size is set as $H = W = 384$, the embedding dimension $C = 30$ and the window length $D = 7$. As for the asymmetric encoder, the swin-Transformer blocks of top-view image flow and SAR image flow both have layer number $[L_1, L_2] = [L_1', L_2'] = [1,1]$, and the number of attention head $[He_1, He_2] = [He_1', He_2'] = [3,6]$. For the fusing encoder, the layer number of two swin-Transformer blocks $[L_3, L_4] = [3,1]$, and the number of attention head $[He_3, He_4] = [12,24]$. The network loss is mixed by cross-entropy and DiceLoss (Milletari et al. 2016) in a ratio of 1:4, which improve the training bias caused by the unbalanced number of samples. The loss is optimized by an AdamW optimizer (Loshchilov et al. 2019) with the learning rate 0.0001. The training phase has 20000 iterations, where the mini-batch size is set to 2.

3.2 Classification Results

The classification result and confusion matrix of the proposed network are shown in Figure 4(a), where the overall accuracy achieves 87.25% and the mIoU achieves 73.56%. The visualization of the corresponding result on test set is shown in Figure 5(a). It can be observed that the tree and low vegetation have the most confusion, because they share similar plant textures and the sharp edges of the tree crowns are hard to be estimated.



(a) (b)

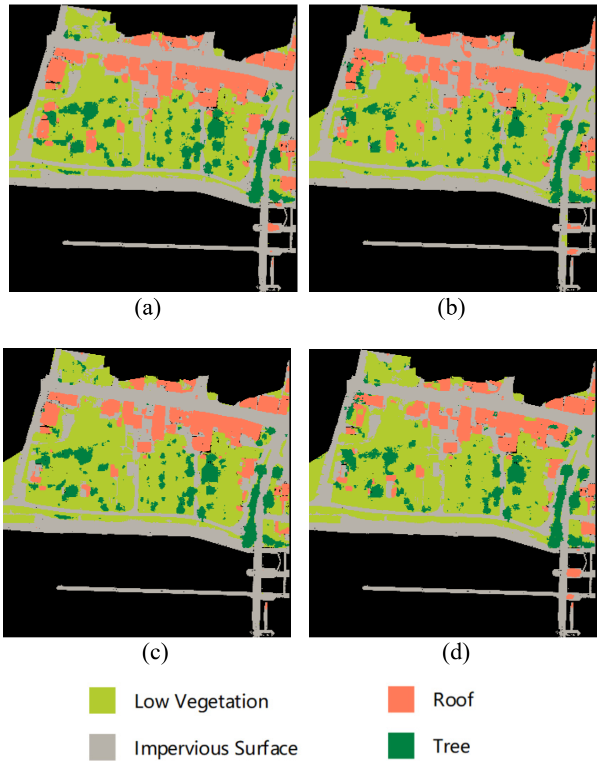
(c) (d)

Figure 4. Confusion matrices on (a) the proposed network (b) FCN, (c) Single-ST, and (d) PLF-ST.

3.3 Model Comparations

In order to verify the effectiveness of the swin-Transformer blocks and feature-level fusing methodology, we compare the proposed network with some available models: fully convolutional network (FCN), swin-Transformer with single image features (Single-ST), and pixel-level fusion swin-Transformer (PLF-ST). FCN is a classic framework for semantic segmentation of image data (Shelhamer et al. 2016), which utilizes ResNet50 as a backbone and generates the high-resolution semantic labels by deconvolutions. The input of FCN is a single-source data, i.e. the RGB top-view image. Single-ST is based on the proposed network but the flow of SAR image is removed. It has 4 swin-Transformer blocks and only extracts the features from RGB top-view image. PLF-ST is a pixel-level fusion method. It has the same network structure as Single-ST, but PLF-ST regards the SAR image as an accessory channel of the RGB top-view image.

The overall accuracy and mIoU of the proposed network and other models are shown in Table 1. The detailed confusion matrixes are shown in Figure 4(b)-(d) and the comparisons of the test visualizations are shown in Figure 5(b)-(d). For models with single data, the Single-ST performs better than the classic FCN model because of the global perception of self-attention calculation. For data-fusion models, however, the PLF-ST do not demonstrate obvious advantages of data fusion, which has approximate performance as the Single-ST. Although acquiring more information, the naïve pixel-level fusion method without feature extraction is still difficult to handle heterogeneous data.



(a) (b)

(c) (d)

Figure 5. Visualization of the test results on (a) the proposed network, (b) FCN, (c) Single-ST, and (d) PLF-ST.

Models	Accuracy	mIoU
FCN	83.94%	66.83%
Single-ST	85.34%	69.48%
PLF-ST	85.13%	69.14%
Proposed network	**87.25%**	**73.56%**

Table 1. Performance comparison between proposed network and other models.

The proposed network is equipped with asymmetric encoder to extract features of two different data individually, and then fuses the extracted features rather than the raw data. The asymmetric encoder is equivalent to a feature selector, which can adaptively adjust the middle-layer features according to the propagated errors. Through this adaptive data fusion processing, the respective characteristics of SAR image and RGB point cloud can be implicitly learned, so as to complement each other and improve classification accuracy and mIoU. It is observed that the mIoU has boosted for 4.08% than Single-ST, especially on the tree and roof classes, which is consistent with the intuitive estimates on the SAR image, i.e. the stronger contrast between artificial and natural objects.

4. CONCLUSION AND OUTLOOK

To fully exploit the information of RGB top-view point clouds and SAR images, a feature-level fusion network based on swin-Transformer is proposed for urban classification. Firstly, the point cloud is projected to a horizontal plane, and the RGB value of each point is interpolated to a pixel in the grid to obtain an RGB top-view image. Afterwards, some homologous points in the RGB top-view image and the SAR image are extracted for registration. The feature extraction on the top-view image flow and SAR image flow is performed separately in the asymmetric encoder. These extracted features are fused by the fusing encoder and decoded by a standard UperNet. Compared with the classic models, the proposed network achieves higher accuracy and mIoU on the subset of H3D data set and the TerraSAR-X data set. The experiments also verify the effectiveness of both the swin-Transformer blocks and the feature-fusion methodology.

However, the proposed network still has the limitation of unpredictable error from manual image registration. Moreover, during horizontal projection of the point cloud data, all points are squeezed into the horizontal plane and the overlapped points are ignored. The future work will be focused on fusing raw point cloud data and other remote sensing data in 3D space, such as generating 3D point clouds of scattering centres by tomography SAR.

ACKNOWLEDGEMENTS

Considerable parts of the research presented in this paper were supported by the China Scholarships Council No. 202006960021.

REFERENCES

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. 2020: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint,* arXiv:2010.11929.

Kahraman, S., Bacher, R., 2021: A comprehensive review of hyperspectral data fusion with LiDAR and SAR data. *Annual Reviews in Control*, 51(1), 236-253. doi.org/10.1016/j.arcontrol.2021.03.003.

Kulkarni, S. C., Rege, P. P., 2020: Pixel level fusion techniques for SAR and optical images: A review. *Information Fusion*, 59(1), 13-29. doi.org/10.1016/j.inffus.2020.01.003.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J. D., Ledoux, H., 2021: The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and multi-view-stereo. ISPRS Open Journal of Photogrammetry and Remote Sensing, 1, 11. doi.org/10.1016/j.ophoto.2021.100001

Li, X., Grandvalet, Y., Davoine, F., 2018: Explicit inductive bias for transfer learning with convolutional networks. *Proc. Int. Conf. Mach. Learn.*, 80, 2825-2834. proceedings.mlr.press/v80/li18a.html.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B., 2021: Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint*, arXiv:2103.14030.
Loshchilov, I., Hutter, F., 2019: Decoupled weight decay regularization. *International Conference on Learning Representations*, arxiv.org/abs/1711.05101.

Milletari, F., Navab, N., Ahmadi, S. A., 2016: V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV),* 565-571. doi: 10.1109/3DV.2016.79.

Roth, A., Hoffmann, J., Esch, T., 2005: TerraSAR-X: How can high resolution SAR data support the observation of urban areas? *Proceedings of the ISPRS WG VII/1 "Human Settlements and Impact Analysis" 3rd International Symposium Remote Sensing and Data Fusion Over Urban Areas (URBAN 2005).* elib.dlr.de/46692.

Roth, A., Huber, M., Kosmann, D. 2004: Geocoding of TerraSAR-X data. *Proc. Int. 20th ISPRS Congr.*, 7, 840-844. elib.dlr.de/467.

Shelhamer, E., Long, J., Darrell, T., 2016: Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4), 640-651. 10.1109/CVPR.2015.7298965.

Waske, B., van der Linden, S., 2008: Classifying multilevel imagery from SAR and optical sensors by decision fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5), 1457-1466. doi.org/10.1109/TGRS.2008.916089.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018: Unified perceptual parsing for scene understanding. *Proceedings of the European Conference on Computer Vision (ECCV)*. 418-434. doi.org/10.1007/978-3-030-01228-1_26.

Yamamoto, T., Nakagawa, M., 2015: Merging airborne LIDAR data and satellite SAR data for building classification. *The International Archives of Photogrammetry*, *Remote Sensing and Spatial Information Sciences*, 40(4), 227. doi:10.5194/isprsarchives-XL-4-W5-227-2015.

Zhang, M., Li, W., Tao, R., Li, H., Du, Q., 2021: Information fusion for classification of hyperspectral and LiDAR data using IP-CNN. *IEEE Transactions on Geoscience and Remote Sensing*. doi.org/10.1109/TGRS.2021.3093334.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017: Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2881-2890. doi.org/10.1109/CVPR.2017.660