Generalized Zero-Shot Vehicle Detection in Remote Sensing Imagery via Coarse-to-Fine Framework

Hong Chen[1] , **Yongtan Luo**[1] , **Liujuan Cao**[1*] , **Baochang Zhang**[2] , **Guodong Guo**[3,4] , **Cheng Wang**[1] , **Jonathan Li**[1] , **Rongrong Ji**[1,5]

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, China

[2]School of Automation Science and Electrical Engineering, Beihang University, China

[3]Institute of Deep Learning, Baidu Research

[4]National Engineering Laboratory for Deep Learning Technology and Application

[5] Peng Cheng Laboratory, China

{hongc, luoyongtan}@stu.xmu.edu.cn, {caoliujuan, cwang, junli, rrji}@xmu.edu.cn, bczhang@buaa.edu.cn, guoguodong01@baidu.com

Abstract

Vehicle detection and recognition in remote sensing images are challenging, especially when only limited training data are available to accommodate various target categories. In this paper, we introduce a novel coarse-to-fine framework, which decomposes vehicle detection into segmentation-based vehicle localization and generalized zero-shot vehicle classification. Particularly, the proposed framework can well handle the problem of generalized zero-shot vehicle detection, which is challenging due to the requirement of recognizing vehicles that are even unseen during training. Specifically, a hierarchical DeepLab v3 model is proposed in the framework, which fully exploits fine-grained features to locate the target on a pixel-wise level, then recognizes vehicles in a coarse-grained manner. Additionally, the hierarchical DeepLab v3 model is beneficially compatible to combine the generalized zero-shot recognition. To the best of our knowledge, there is no publically available dataset to test comparative methods, we therefore construct a new dataset to fill this gap of evaluation. The experimental results show that the proposed framework yields promising results on the imperative yet difficult task of zero-shot vehicle detection and recognition.

1 Introduction

Alongside the recent progress of remote sensing technology, vehicle detection has experienced significant developments, resulting in fruitful results [Soleimani *et al.*, 2018; Cao *et al.*, 2016]. These methods often serve as an essential step in intelligent transportation system, thus leading to numerous real-world applications. Mainstream detection approaches that are popularly applied rely heavily on fully supervised learning schemes, which poses prohibitive demands on comprehensive training data covering all categories, which are often hard to come by. As a work-around, approaches capable of unseen vehicle detection are required [Zhu *et al.*, 2018; Demirel *et al.*, 2018].

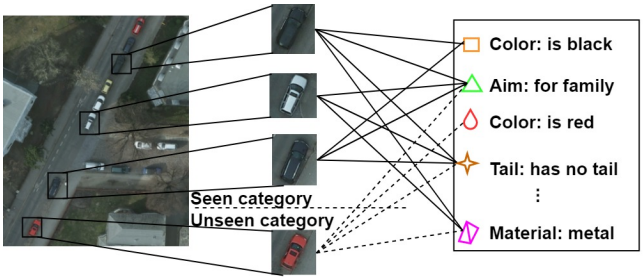

Figure 1: Vehicle patches extracted from remote sensing images have different shapes, types, materials. However, they share the same semantic space. As is shown above, visual-semantic mapping function is utilized to obtain the semantic descriptions of test instances in the testing phase for zero-shot recognition.

Zero-shot learning (ZSL) [Lampert *et al.*, 2014] is based on the assumption that both the *seen classes* (with training examples) and *unseen classes* (with no training examples) share the same semantic space, where visual models for *seen classes* are transferred to the *unseen classes* by exploiting semantic relationships between the two. As is shown in Fig. 1, although vehicles exhibit different appearance with verified color and shape, yet they share the same semantic space such as *'Tail: has no tail'*, *'Material: metal'* and so on. Accordingly, such learning scheme enables the capacity of recognizing *unseen* vehicle categories.

To introduce the power of ZSL, we decompose vehicle detection in remote sensing images into two phases, *i.e.*, the vehicle localization and vehicle category label prediction. Previously, ZSL is merely used for recognition task, which means that it is destined to work for simple cases where only a single dominant object is present in an image [Xian *et al.*,

*Corresponding author

2017; Romera-Paredes and Torr, 2015; Xian *et al.*, 2016; Jiang *et al.*, 2017]. Such a setting works fine in most cases. Nonetheless, in the context of vehicle detection, different kinds of objects from *unseen classes* and *seen classes* appear in one image at the same time, wherein the standard recognition setting no longer stands. This is especially true for detection in remote sensing images that contain objects of different kinds and with different scales. Under such circumstance, typical ZSL methods meets its end. Notably, our goal is to simultaneously detect each individual instance of vehicle classes, even in the absence of any visual examples of those classes during the training phase. To make it more suitable for the task of zero-shot detection in remote sensing images, task-specific ZSL strategy is of great importance. Consequently, we propose to learn latent attributes for generalized zero-shot vehicle detection, in which the test vehicle instance can not only come from *unseen classes* but *seen classes*.

As another evidence in the traditional detection task and the newly developed manuscripts on zero-shot detection of natural images, category labels with fine-grained bounding boxes annotations should be provided for training [Redmon and Farhadi, 2017; Demirel *et al.*, 2018]. This inevitably introduce additional annotation burdens. Furthermore, in remote sensing images, sufficient bounding box annotations for object detection is unavailable. We are thus motivated and propose to make full use of the given annotation of existing remote sensing datasets for semantic segmentation task. Thus we propose to locate vehicles of original remote sensing images in a pixel-wise level and then feed them to subsequent generalized zero-shot vehicle recognition phase for fine-grained vehicle classification.

In general, the contributions of this paper are three-fold:

- A hierarchical DeepLab v3 is introduced with hierarchical connections, which fully capture the global and local features for the pixel-wise level vehicle localization.
- Generalized zero-shot recognition with latent attributes learning are introduced to handle the challenging task of fine-grained vehicle classification even when they are not trained. We lead the latent attributes that are both discriminative and semantic-preserving.
- A new dataset is constructed based on the ISPRS 2D semantic labeling contest dataset, which provides extensive experimental studies for the task of generalized vehicle detection.

The remainder of this paper is organized as follows: related works are introduced in Sec.2. Detailed descriptions of the proposed framework is well illustrated in Sec. 3. In Sec. 4, we conduct quantitative experiments to verify the effectiveness of the proposed framework. Finally, we conclude this paper in Sec.5.

2 Related Work

We note that a new study of generalized zero-shot vehicle detection is provided for remote sensing images. In this section, we will first discuss the models of semantic segmentation and zero-shot detection of natural images and then give a brief review of the application of ZSL in remote sensing images.

2.1 Semantic Segmentation

Semantic segmentation presents pixel-wise level classification that has many applications, *e.g.*, automatic driving. FCN [Long *et al.*, 2014] brings the first work that utilizes full convolution network for such a dense classification task, which supports the feeding images with different resolutions. DeepLab v1 and v2 [Chen *et al.*, 2018a] proposes atrous convolution to not only enlarge the field of view but cost limited extra computation. To further capture the information of each layer, and is inspired by the SPPNet [Zhao *et al.*, 2016], DeepLab v3 [Chen *et al.*, 2017] adopts pyramid network with atrous convolution. To alleviate the information decay, our network employs hierarchical connections with global average pooling to deliver detail descriptions from the low-level layers, which helps the detection of small objects.

2.2 Zero-Shot Detection

Zero-shot detection is a new concept that has come into the community. It is proposed to detect objects even when they have not been trained before. Recently, there appear some manuscripts on zero-shot detection in the community of natural images. Zhu *et.al* and Demirel *et.al* retain the efficiency and effectiveness of YOLO [Redmon and Farhadi, 2017] for objects seen during training, while improving its performance for novel and unseen objects [Zhu *et al.*, 2018; Demirel *et al.*, 2018]. Ankan[Ankan Bansal, 2018] introduces background-aware approaches that use a fixed background class and iterative latent assignments that are based on RCNN [Girshick *et al.*, 2016] framework. However, these models require bounding box annotations in the training phase, which is a limitation in the remote sensing community. Our framework is different from the detection models that need well-annotated bounding boxes.

2.3 Zero-Shot Learning in Remote Sensing Imagery

Various ZSL methods have been developed in the general field of natural images analysis. However, to the best of our knowledge, there are limited works focusing on exploring zero-shot recognition in remote sensing. Among these works, Li *et al.* first introduced zero-shot recognition into remote sensing research with a label refinement phase to classify novel scenes in high-resolution remote sensing images [Li *et al.*, 2017]. Later on, Sumbul *et al.* introduced a new dataset for zero-shot tree classification [Sumbul *et al.*, 2017]. Different from the aforementioned methods, we study the zero-shot vehicle detection task of remote sensing images, which is the first in the literature.

3 Generalized Zero-Shot Vehicle Detection Framework

We formulate the vehicle detection task in remote sensing images with only pixel-wise level annotations by a hierarchical DeepLab v3 model and a generalized zero-shot recognition phase. The proposed framework is shown in Fig. 2. It consists of two components, *i.e.*, a hierarchical DeepLab v3 for coarse-grained vehicle recognition to locate the vehicle in

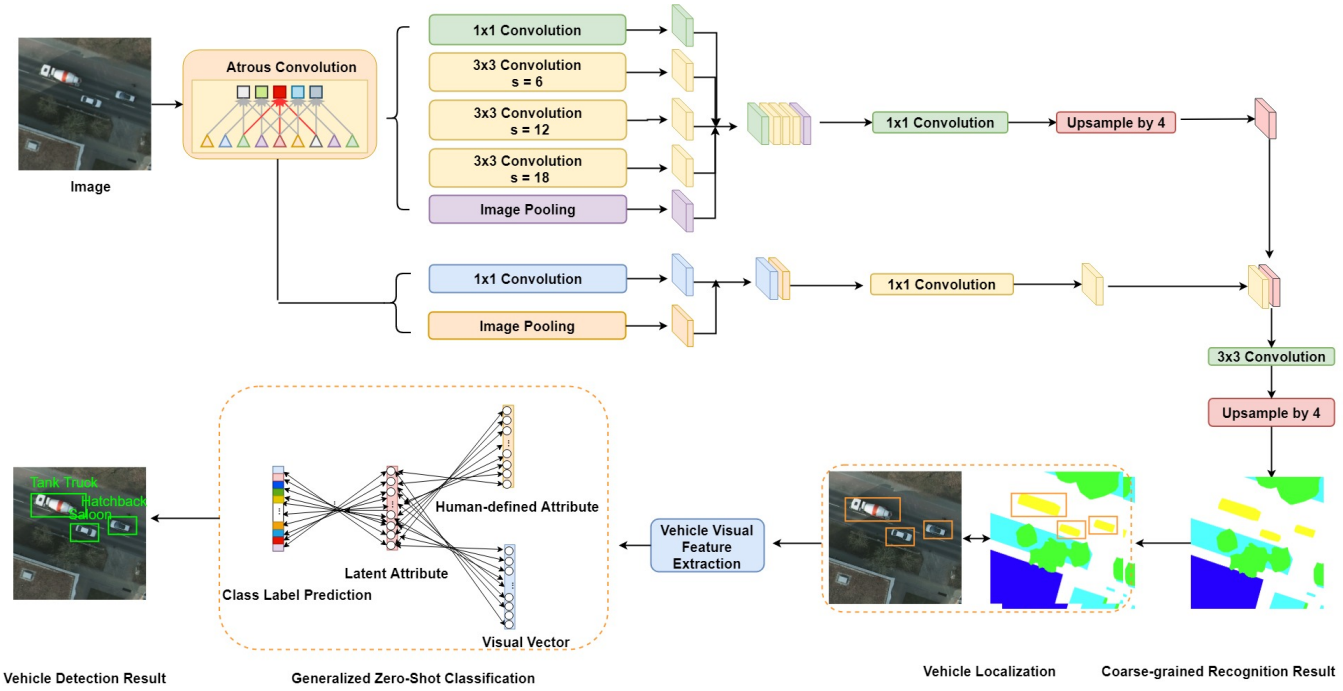

Figure 2: Flowchart of the proposed framework. Original remote images are fed into this framework directly. A hierarchical DeepLab v3 model is utilized to localize the general locations of vehicles in the images. Then, vehicle pathes are fed into generalized zero-shot recognition phase to predict their category labels. Generalized zero-shot recognition with latent attribute learning process enables the posibility of correctly classifying the vehicle patches even they are not trained during the training phase.

a pixel-wise level and a generalized zero-shot vehicle fine-grained classification phase to learn latent attributes for vehicle categorization not matter they have been trained or not.

3.1 The Hierarchical DeepLab v3 Model for Coarse-Grained Vehicle Recognition

Note that vehicle patches in remote sensing images suffer low-resolution, the consecutive striding in previous segmentation models is harmful to segmentation procedure because of the signal decimation. We propose to improve the DeepLab v3 model with atrous convolution and hierarchical connections to alleviate such signal decimation.

Atrous Convolution.

Atrous convolution serves as an important tool that allows us to explicitly enlarge the filter's field-of-view [Chen *et al.*, 2017; 2018b]. Multi-scale information can be captured through performing atrous convolution with different atrous rates. *i.e.*, we sample the input signal with different strides. Consider two-dimensional signals, for each location i on the output $Feat$ and a filter w, atrous convolution is applied over the input feature map x with an atrous rate of s, the formulation is:

$$Feat = \sum_{k} x[i + s * k]w[k], \tag{1}$$

where s denotes the atrous rate that is equivalent to convolving the input x with upsampled filters produced by inserting $s - 1$ *'hole'*s between two consecutive filters.

The Hierarchical DeepLab v3 for Coarse-Grained Vehicle Recognition.

Atrous convolution has been adopted in DeepLab [Chen *et al.*, 2018a; 2017] and caught much attention. However, the spatial resolution of the final feature maps is usually $32\times$ smaller than the original input images in the task of semantic segmentation, which implicitly indicates that information of small objects is discarded. To this end, as shown in Fig. 2, we propose a hierarchical DeepLab v3 with hierarchical connections to concatenate low-level features with higher-level features of the same spatial resolution. 1×1 convolution and global image pooling are applied to reduce the number of channels and obtain more informative details of the input image, respectively. This extra hierarchical connections help to alleviate the information reduction thus improves the performance of pixel-wise classification. These coarse-grained vehicle patches are further fed into the generalized zero-shot fine-grained recognition phase to classify the vehicle categories.

3.2 Generalized Zero-Shot Vehicle Fine-Grained Classification

Typical supervised classification methods can do nothing when there comes an instance of new species. In this section, we introduce generalized zero-shot recognition that aims to recognize vehicles in remote sensing images even when they haven't been trained.

Formulation

Given a training set $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n^s}$, *i.e.*, the *seen classes*, where $x_i^s \in X$ is a d-dimensional column vector for the i-th training image from $|Y^s|$ seen classes. $y_i^s \in Y^s$ is the label of x_i. Similarly, the *unseen classes*, is defined as $\mathcal{U} = \{(x_j^u, y_j^u)\}_{j=1}^{n^u}$, where x_j^u is a d-dimensional column vector of the j-th test image from $|Y^u|$ unseen classes and $y_j^u \in Y^u$ is its corresponding label. We utilize the human-defined attributes as semantic information for knowledge transfer between *seen classes* and *unseen classes*, which can be denoted as $A = \{a_i\}_{i=1}^{m}$. Here, a_i indicates human-defined attribute vector for the i-th class. We generally learn a visual-semantic mapping in the training process. Then, the mapping function is used for the *unseen* instances. Typical ZSL method predicts the category label of *unseen* instance x_j^u from label sets of Y^u.

Typical ZSL setting assumes that test instances come from the *unseen classes* so that each of them is categoried to one of the labels in Y^u. Label sets of *seen classes* and *unseen classes* are disjointed, *i.e.*, $Y^s \bigcap Y^u = \varnothing$. In generalized ZSL scenario, the resource of test images is flexible and they can come from either *seen classes* or *unseen classes*. Each of them is categorized to labels in $Y = Y^s \bigcup Y^u$.

Generalized Zero-Shot Vehicle Fine-Grained Recognition

[Jiang *et al.*, 2017; Yan *et al.*, 2018] have demonstrated that it's important to learn discriminative attributes, which helps to obtain specific perspective for the classification task. Therefore, we propose to learn latent attributes for generalized zero-shot vehicle recognition. The loss is defined as:

$$\mathcal{L}_1 = \min_{W,P} \|XW - YP\|_F^2,\ s.t.,\ \|w_i\|_2^2 \leq 1,\ \forall i, \quad (2)$$

where $\|.\|_F^2$ denotes the *Frobenius* norm, w_i is the *i*-th column of W. W is a dictionary that performs projection from visual features to latent discriminative attributes. P can be viewed as classifiers for the latent discriminative attributes. Here we have $Y = [y_1, y_2, ..., y_s]$ and $y_i = [0, 0, 1..., 0]$ is a one-hot encoded vector corresponding to category x_i.

Human-defined attributes present category-level descriptions directly and involves real-world meanings. To make the learned latent attributes semantic-preserving, a linear transform is designed to connect human-defined attributes with the latent attributes. Especially, the function is formulated as:

$$\mathcal{L}_2 = \min_{W,Q} \|XW - AQ\|_F^2,\ s.t.,\ \|q_i\|_2^2 \leq 1,\ \forall i, \quad (3)$$

where A denotes the human-defined attributes. Q is a linear function to be learned, which guarantee the semantic meanings of the latent distinctive attributes.

To further make the learned visual-semantic mapping both discriminative and semantic-preserving, a function Q correlates the human-defined attributes and the latent attributes, P is utilized for the classification task. It can be inferred that it will help to implicitly combine strongly correlated attributes and prefers discriminative attributes.

$$\mathcal{L}_3 = \min_{P,Q} \|AQ - YP\|_F^2,\ s.t.,\ \|p_i\|_2^2 \leq 1,\ \forall i, \quad (4)$$

Algorithm 1 Training procedure of the proposed model.

Require:
X: training images from *seen classes*;
Y: corresponding labels for training images;
A: human-defined attributes;
α: hyper-parameter;
β : hyper-parameter;
γ: hyper-parameter;
Ensure: Projection function, $i.e.$, W, P and Q ;
Initialize W, P, Q randomly;
Choose vehicle patches for training;
while not converge **do**
    optimize W while fix P and Q ;
    optimize P while fix W and Q;
    optimize Q while fix W and P;
end while
return W, P and Q ;

To be more adaptable for fine-grained vehicle recognition task, we propose to consider these factors simultaneously and the objective function is formulated as:

$$\mathcal{L}_{full} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3 \quad (5)$$

α, β and γ are hyper-parameters that control the strength of each constrain. Eq. 5 is convex for W, P and Q individually, but is not convex for them simultaneously. Therefore, we solve it using an alternating, which circularly fixes the other parameters and optimizes one parameter one time. In each optimization, the sub-problem is transferred to a conventional least square minimization problem that can be optimized by the Lagrange dual, which has a closed-form solution. The overall optimization procedure is shown in Algorithm 1.

Given a test remote sensing image, it is first fed into the semantic network to obtain the locations of the vehicle category and to extract the features of vehicle patches, *i.e.*, $\phi(x_i)$. These patches are subsequently fed into the generalized zero-shot recognition module and get the latent attribute representation through mapping function W. The Nearest Neighbor (NN) algorithm is utilized to perform generalized zero-shot recognition. In general, the object function for zero-shot inference phase is defined as follows:

$$label_{x_i} = \arg\min_A \|\phi(x_i)W - AQ\|_F^2, \quad (6)$$

where $\phi(x_i)$ indicates the feature vector of x_i. A denotes the human-defined attributes. W and Q are the mapping functions to project the visual feature vectors and human-defined attributes to latent discriminative attributes, respectively.

4 Experiments

4.1 Datasets and Settings

Datasets. Note that there is no dataset available in the task of generalized zero-shot vehicle detection of remote sensing images. We introduce a new dataset based on images from ISPRS WG III/4 2D Semantic Labeling Contest[1]. Each of

[1]http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

the original images is with the size of 6,000×6,000 pixels, consisting of the true orthophoto of Potsdam with ground sampling distance of 5 cm and together with pixel-wise level semantic annotation. We automatically crop vehicle patches from the images and produce 12,495 vehicle patches to construct the training set for generalized zero-shot recognition. They are categorized into 16 classes.

Data Split. We train the hierarchical DeepLab v3 on the ISPRS 2D Semantic Labeling Contest dataset introduced above with the provided pixel-wise level annotations. This contest is to classify image pixels of remote sensing images into 6 categories, *e.g.*, vehicle, tree and building. The zero-shot classification network is trained through the cropped vehicle patches. 12 categories of them are randomly selected for *seen classes* and the remaining 4 categories for *unseen classes*.

Implement Details. VGG-16 [Simonyan and Zisserman, 2014] and Xception [Chollet, 2017] are employed as the backbone of hierarchical DeepLab v3. For all experiments, we use SGD optimizer with a momentum of 0.9 for training. Weight decay rate is fixed as 0.0005. α and β are tuned using five-fold cross-validation, while γ is fixed as 1. We tuned from 100 to 700 and fix the size of latent dictionary as 400 for better performance. The located vehicle patches are resized into 321×321 to get a fixed size of $\phi(x)$.

Evaluation Metrics. For the convenience of evaluation, we use the same metric, *i.e.*, overall accuracy for coarse-grained vehicle recognition. IoU is further adopted for the comparison. We adopt the standard evaluation metrics of ZSL, *i.e.*, the multi-class classification accuracy (MCA) to evaluate the performance of generalized zero-shot fine-grained recognition.

$$MCA = \frac{1}{|N|}\sum_{i=1}^{|N|} class_i, \tag{7}$$

where $class_i$ is the prediction accuracy of i-th class. $|N|$ corresponds the total number of vehicle categories.

4.2 Experimental Results

In this section, we conduct comparisons to verify the effectiveness of the proposed framework. For the lack of bounding box annotations of remote sensing datasets, traditional detection models or the zero-shot detection models are not allowed for the comparison, they require bounding boxes. Fig. 3 shows several results of our framework, from which we can observe that our framework is able to automatically detect *seen* and *unseen* vehicles in a given remote sensing image.

Several classical approaches in the literature of zero-shot recognition are re-implemented to adapt to the more challenging task of generalized zero-shot vehicle recognition of remote sensing images. Tab. 1 shows several results of the proposed framework of generalized zero-shot vehicle detection. From Tab. 1, we can draw the following conclusions: 1) all generalized zero-shot vehicle recognition methods surpass the randomly guessing (1/16, *i.e.*, 6.25%), which demonstrates the effectiveness of generalized zero-shot vehicle recognition and 2) we present an effective framework in the task of generalized zero-shot vehicle detection of remote sensing images that outperforms all the comparisons with VGG and Xception backbones. The performance can serve as a baseline method for the subsequent research of generalized zero-shot vehicle detection.

Method	Feature	MCA (%)
ESZSL	hierarchical DeepLab v3 +VGG-16 backbone	7.65
LatEm	hierarchical DeepLab v3 +VGG-16 backbone	11.46
DEM	hierarchical DeepLab v3 +VGG-16 backbone	9.38
Ours	hierarchical DeepLab v3 +VGG-16 backbone	**16.89**
ESZSL	hierarchical DeepLab v3 +Xception-65 backbone	16.87
LatEm	hierarchical DeepLab v3 +Xception-65 backbone	18.12
DEM	hierarchical DeepLab v3 +Xception-65 backbone	9.38
Ours	hierarchical DeepLab v3 +Xception-65 backbone	**21.25**
ESZSL	hierarchical DeepLab v3 +Xception-71 backbone	19.37
LatEm	hierarchical DeepLab v3 +Xception-71 backbone	20.63
DEM	hierarchical DeepLab v3 +Xception-71 backbone	9.38
Ours	hierarchical DeepLab v3 +Xception-71 backbone	**23.75**

Table 1: Comparisons on generalized zero-shot fine-grained recognition setting.

Method	VGG-16$^\natural$	VGG-19$^\natural$	ResNet-34$^\natural$	**Ours**$^\flat$
MCA(%)	45.60	46.87	46.87	**39.58**

Table 2: Comparison to supervised learning methods. $\natural$: supervised learning, $\flat$: zero-shot learning.

We also include a comparison between supervised learning and unsupervised learning. VGG-16, VGG-19 [Simonyan and Zisserman, 2014] and ResNet-34 [He *et al.*, 2015] are selected as the backbone of supervised learning models. The results are illustrated in Tab. 2. Note that our baseline framework is based on VGG-16 backbone, our framework obtains an MCA of 39.58% under ZSL settings, which is only 5.62% lower than the supervised classification method that based on a VGG-16 backbone (45.60%), demonstrating the success of the proposed framework.

4.3 Ablation Study

Hierarchical Connection. Tab. 3 presents the comparison of coarse-grained recognition, from which we find that our baseline model achieves an accuracy of 90.1% that surpasses other recently proposed comparisons of RITL7 [Liu *et al.*, 2017], KLab3 [Kemker *et al.*, 2018] and DeepLab v3 [Chen *et al.*, 2017]. This implicitly indicates that we provide more accurate vehicle localizations. The only difference between the proposed hierarchical DeepLab v3 and DeepLab v3 is that our model employs extra hierarchical connections to alleviate signal decimation. It shows in Tab. 4 that our model brings more accurate result of vehicle localization and feature extraction, which helps a promising improvement in the subsequent vehicle fine-grained classification task. With Xception-65 and Xception-71 backbone, our framework achieves more improvements. Note that there are no IoU results available of other methods, we present in Tab. 5 the IoU of our models.

Latent Attribute Learning. Among all the methods presented in Tab. 1, ESZSL [Romera-Paredes and Torr, 2015] proposes an easy but efficient method for zero-shot recognition with directly human-defined attributes, LatEM [Xian *et al.*, 2016] presents the first work that focuses on fine-grained object recognition in natural images with multiple dictionar-

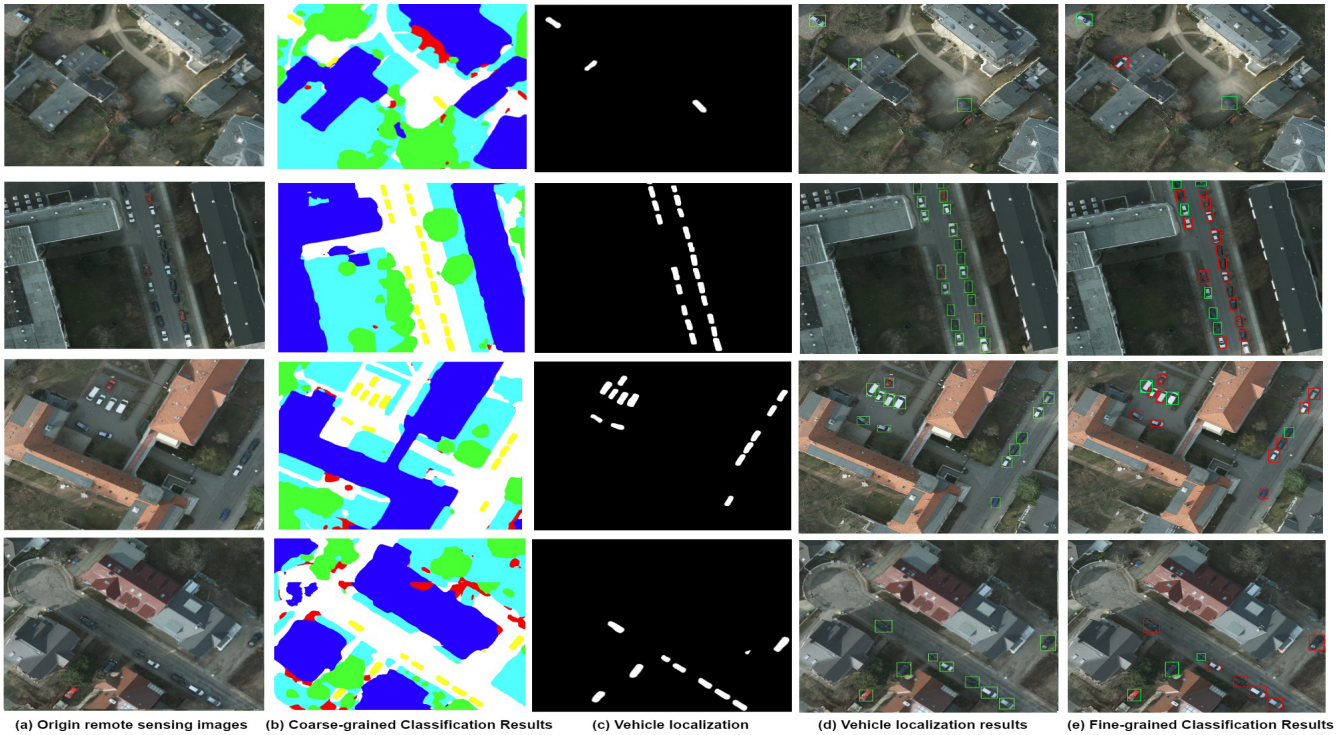

Figure 3: Generalized zero-shot detection results of the proposed framework. Each row corresponds to a group of results. Green rectangles and red rectangles in (e) represent localized vehicles with right and wrong label predictions, respectively.

Method	Backbone	Accuracy(%)
RITL_7	FCN-8s	88.4
KLab_3	-	86.4
DeepLab v3	VGG-16 backbone	89.2
Ours	VGG-16 backbone	90.1
Ours	Xception-65 backbone	92.3
Ours	Xception-71 backbone	**92.5**

Table 3: Comparison in the task of overall coarse-grained recognition in pixel-wise level.

Method	RITL_7	KLab_3	DeepLab v3	**Ours**
F1 score(%)	92.8	92.0	93.3	**94.1**

Table 4: Comparison in the task of coarse-grained vehicle recognition in pixel-wise level.

ies and DEM [Zhang *et al.*, 2017] delivers a work to learn an end-to-end model with human-defined attributes. The improvement of MCA among our framework, DEM and ESZSL demonstrates the success of learning latent attributes for generalized zero-shot fine-grained recognition.

5 Conclusion

In this paper, we propose and tackle a challenging problem of generalized zero-shot vehicle detection of remote sensing images. A coarse-to-fine framework that consists of the proposed hierarchical DeepLab v3 for recognizing and localizing vehicles in a coarse-grained manner following by generalized zero-shot vehicle classification with latent attributes learning for fine-grained vehicle classification is introduced to solve the problem. The experiment results on the new dataset constructed based on ISPRS Potsdam 2D Semantic Labeling Contest dataset demonstrates the effectiveness of the proposed framework. In the future, we would like to bring more generalized one-step deep networks for this task. We will also construct and test our framework on other remote sensing datasets and stick to improving the performance of generalized zero-shot vehicle detection.

Method	**Ours**+VGG-16	**Ours**+Xception-65	**Ours**+Xception-71
IoU	84.74	85.99	**86.13**

Table 5: IoU of coarse-grained recognition in pixel-wise level.

Acknowledgements

This work is supported by the Nature Science Foundation of China (No.61772443, No.U1705262, and No.61572410), National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Post Doctoral Innovative Talent Support Program under Grant BX201600094, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

[Ankan Bansal, 2018] Gaurav Sharma Rama Chellappa Ajay Divakaran Ankan Bansal, Karan Sikka. Zero-Shot Object Detection. *arXiv preprint arXiv:1804.04340*, 2018.

[Cao *et al.*, 2016] Liujuan Cao, Feng Luo, Li Chen, Yihan Sheng, Haibin Wang, Cheng Wang, and Rongrong Ji. Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning. *Pattern Recognition*, 2016.

[Chen *et al.*, 2017] Liang Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. 2017.

[Chen *et al.*, 2018a] 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*, 40(4):834–848, 2018.

[Chen *et al.*, 2018b] Liang Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. 2018.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[Demirel *et al.*, 2018] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. *arXiv preprint arXiv:1805.06157*, 2018.

[Girshick *et al.*, 2016] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. 2016.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[Jiang *et al.*, 2017] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2017.

[Kemker *et al.*, 2018] Ronald Kemker, Carl Salvaggio, and Christopher Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *Isprs Journal of Photogrammetry and Remote Sensing*, page S0924271618301229, 2018.

[Lampert *et al.*, 2014] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[Li *et al.*, 2017] Aoxue Li, Zhiwu Lu, Liwei Wang, Tao Xiang, and Ji Rong Wen. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):4157–4167, 2017.

[Liu *et al.*, 2017] Yansong Liu, Sankaranarayanan Piramanayagam, Sildomar T. Monteiro, and Eli Saber. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[Long *et al.*, 2014] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2014.

[Redmon and Farhadi, 2017] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Romera-Paredes and Torr, 2015] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Soleimani *et al.*, 2018] Amir Soleimani, Nasser M Nasrabadi, Elias Griffith, Jason Ralph, and Simon Maskell. Convolutional neural networks for aerial vehicle detection and recognition. 2018.

[Sumbul *et al.*, 2017] Gencer Sumbul, Ramazan Gokberk Cinbis, and Selim Aksoy. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–10, 2017.

[Xian *et al.*, 2016] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[Xian *et al.*, 2017] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. *arXiv preprint arXiv:1703.04394*, 2017.

[Yan *et al.*, 2018] Li Yan, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. 2018.

[Zhang *et al.*, 2017] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Zhao *et al.*, 2016] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. 2016.

[Zhu *et al.*, 2018] Pengkai Zhu, Hanxiao Wang, Tolga Bolukbasi, and Venkatesh Saligrama. Zero-shot detection. 2018.