Double Head Predictor based Few-Shot Object Detection for Aerial Imagery

Stefan Wolf[1,2] Jonas Meier[2] Lars Sommer[2,3] Jürgen Beyerer[2,1,3]

[1]Vision and Future Lab
Karlsruhe Institute of Technology
Karlsruhe, Germany

[2]Fraunhofer IOSB
Karlsruhe, Germany

[3]Fraunhofer Center
for Machine Learning
Munich, Germany

firstname.lastname@iosb.fraunhofer.de

Abstract

Many applications based on aerial imagery rely on accurate object detection, which requires a high number of annotated training data. However, the number of annotated training data is often limited. In this paper, we propose a novel few-shot detection method for aerial imagery that aims at detecting objects of unseen classes with only a few annotated examples. For this purpose, we extend the Two-Stage Fine-Tuning Approach (TFA), which achieves state-of-the-art results on common benchmark datasets. We propose a novel annotation sampling and pre-processing strategy to yield a better exploitation of base class annotations and a more stable training. We further apply a modified fine-tuning scheme to reduce the number of missed detections. To prevent loss of knowledge learned during the base training, we introduce a novel double head predictor, yielding the best trade-off in detection accuracy between the novel and base classes. Our proposed Double Head Few-Shot Detection (DH-FSDet) method outperforms state-of-the-art baselines on publicly available aerial imagery datasets. Finally, ablation experiments are performed in order to get better insight how few-shot detection in aerial imagery is affected by the selection of base and novel classes. We provide the source code at `https://github.com/Jonas-Meier/FrustratinglySimpleFsDet`.

1. Introduction

In recent years, object detection in aerial imagery often referred to as remote sensing imagery experienced significant advancements, which facilitated quantities of applications such as hazard detection, forecast of disasters, assistance in rescue operations, environmental monitoring and urban planning [5, 23]. Reason for these advancements is the usage of deep learning techniques, in particular convolutional neuronal networks, which led to powerful feature representations [31, 40, 42, 44]. Despite impressive results, deep learning based methods suffer from a common issue: the demand for large-scale datasets to train a deep neural network model. While the acquisition and annotation of additional training data is time-consuming and expensive or maybe not feasible, training a model with only a few samples may causes overfitting and thus, poor generalization abilities.

To circumvent this issue, few-shot learning concepts have been proposed, which aim at learning models from limited annotated training samples [16, 19, 28, 54, 52, 57]. In general, few-shot detection methods are initially trained on base classes with sufficient training data and the learned knowledge is then transferred to novel classes with limited training data, yielding detectors capable of localizing and classifying both the base and novel classes. Existing few-shot detection methods are typically designed on common benchmark detection datasets such as MS COCO [27] and PASCAL VOC [13]. While these datasets mostly comprise objects with unitary orientation, with small size variations and located in the image center, aerial imagery can contain objects with random orientation and clearly differing sizes, *e.g.* car and soccer ball field. Thus, few-shot detection methods designed on common benchmark datasets are not directly applicable on aerial imagery.

In this paper, we propose a novel few-shot detection method for aerial imagery based on Faster R-CNN [38] and the Two-Stage Fine-Tuning Approach (TFA) [47], which outperforms state-of-the-art methods with complex meta branch architectures on benchmark datasets by applying a straightforward fine-tuning scheme. To account for the often high number of object instances in aerial images, we propose a novel annotation sampling and pre-processing strategy, yielding a better exploitation of base class annotations and a more stable training. As the detector fails to generate region candidates for novel unseen classes due to the characteristics of aerial imagery and large inter-class variations, we propose an improved fine-tuning scheme by unfreezing the corresponding layers of the detector. Fi-

nally, we introduce a novel double head to prevent loss of knowledge learned during the base training, yielding the best trade-off in detection accuracy between the novel and base classes. We demonstrate the suitability of our proposed few-shot detection method for aerial imagery, clearly outperforming state-of-the-art baselines on publicly available aerial imagery datasets. In aerial imagery, splits of novel and base classes are typically randomly selected. In order to get better insight how few-shot detection in aerial imagery is affected by the number of base and novel classes, we further provide ablation experiments.

The remainder of this paper is organized as follows. In Section 2, we give an overview about deep learning based object detection and few-shot object detection in general and for aerial imagery. In Section 3, we discuss fundamental basics and introduce our proposed few-shot detection method. The experimental setup and results are given in Section 4. Finally, we conclude our paper in Section 5.

2. Related Work

In this chapter, we first give an overview about deep learning based object detection and approaches adopted for object detection in aerial imagery. Then, we present recent few-shot object detection methods in general and for aerial imagery.

2.1. Object Detection

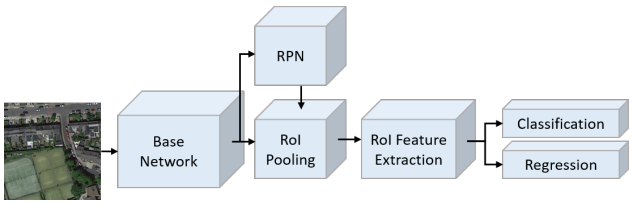
In recent years, a multitude of deep learning based object detectors has been proposed, achieving state-of-the-art results in numerous fields of application. These detectors are generally categorized into proposal-based and proposal-free methods. Proposal-based methods such as Faster R-CNN [38], R-FCN [8] and Cascade R-CNN [2] initially predict candidate regions termed proposals, which are classified in a subsequent stage, while proposal-free methods, *e.g*. SSD [30] and YOLO [35] and its variants [1, 36, 37], perform classification and detection at once. Large improvements in detection accuracy have been achieved by exploiting multiple feature maps within a feature pyramid network [1, 14, 25, 26, 37]. To circumvent the need for pre-defined anchor boxes used as reference for bounding box regression, anchor-free methods, *e.g*. FCOS [45], CenterNet [12] and FoveaBox [21], have been recently proposed, achieving comparable results on benchmark object detection datasets.

These deep learning based detection methods have been widely adapted for object detection in aerial imagery [9, 11, 10, 15, 17, 22, 32, 33, 34, 39, 40, 41, 42, 43, 44, 46, 55]. To account for the characteristics of aerial imagery, *e.g*. small-sized objects, adapting the feature map resolution and the anchor boxes has been proposed [39, 40, 41, 42]. Exploitation of multiple feature maps [9, 11, 15, 34, 46], integration of semantic context [32] and modified loss functions [55] have been applied to further improve aerial object detection. In recent years, the emerge of large-scale datasets with rotated ground truth, *e.g*. DOTA [51], facilitates oriented object detection [10, 22, 43].

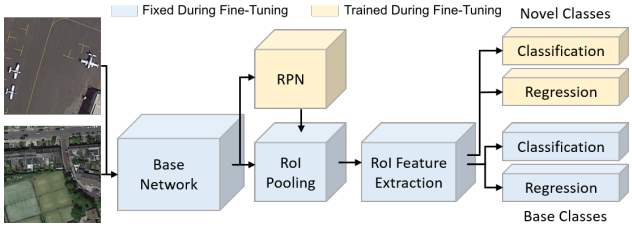
2.2. Few-Shot Object Detection

Since the available training data are often extremely rare, few-shot learning – learning from only a few training samples – has gained great interest. In the following, the literature under review is restricted to few-shot object detection methods. Feature reweighting methods, *e.g*. MetaYOLO [19], Meta R-CNN [54], FSDetView [52] and AFD-Net [28] typically comprise two branches: a main branch to extract features from a query image and a separate support branch to extract per-class feature vectors from support images, which are used to re-weight the features from the main branch. Instead of feature reweighting, PNSD [57] and Meta Faster R-CNN [16] make use of a distance metric to compute the similarity between query features and different support features, which are extracted from different branches. Instead of directly computing features per support image, RepMet [20] and FSOD^{up} [50] extract support features to generate representative features. To obtain a detector for novel classes, MetaDet [48] and GenDet [29] estimate detection parameters for novel classes, using only a few support images. Recent approaches use attention mechanisms to improve the detection performance [56, 4, 3]. Wang *et al*. [47] propose an alternative strategy based on a two-staged fine-tuning scheme, avoiding an auxiliary meta branch.

While most few-shot detection methods are examined on MS COCO, only few approaches are developed for aerial imagery. Li et al. [24] proposed a few-shot object detector for aerial imagery termed FSODM, whose functional principle is similar to MetaYOLO. YOLOv3 is used as feature extractor in the meta branch and a lightweight CNN generates feature vectors used for reweighting. The re-weighted feature maps are then fed into three separate prediction layers to produce bounding box coordinates, objectness scores and class scores. P-CNN [7] extracts query features and support features similar to Meta R-CNN, which are then multiplied channel-wise before being processed by the detector head. To address the issue that objects in remote sensing images are arbitrary oriented, while a small number of samples leads to a sparse orientation space, the authors introduce a prototype learning network and replace the original RPN by a prototype-guided RPN. Xiao et al. [53] integrates a Self-Adaptive Attention Network into Faster R-CNN, which takes features from support images as input and updates a hidden relation graph in order to improve the classification of novel classes by memorizing similar objects.



(a) Base training



(b) Few-shot fine-tuning

Figure 1: Overview of the architecture of our proposed DH-FSDet with the double-head configuration for fine-tuning.

3. Methodology

In this chapter, we describe Faster R-CNN [38] and the Two-Stage Fine-Tuning Approach (TFA) [47] as basis of our work. Afterwards, we present the deficiencies of the existing approach and our novel solution to these problems.

3.1. Base Method

Our proposed method termed Double Head Few-Shot Detection (DH-FSDet) is based on Faster R-CNN and the TFA. Faster R-CNN is a two-stage object detection method generating region proposals likely containing objects in the first stage. The second stage is responsible for classifying and refining the proposed object regions and dropping those regions that only contain image background.

The first stage is implemented by a base network and a Region Proposal Network (RPN). The base network extracts semantically strong feature maps from the input image. We apply a ResNet-50 [18] network as base network and extract the features after every stage. To use the semantically high-level information from late stages combined with the high-resolution feature maps from early stages, we apply a Feature Pyramid Network (FPN) [25] which aggregates the output of each stage with the output of the previous stage by addition. Afterwards, the RPN is applied on each output of the FPN which decides for each pixel of the feature map whether an object is present in that region and estimates a bounding box for the object. For each proposed region, a RoI Align operation is applied that extracts a segment from the feature map representative for the object.

In the second stage, for each region proposal two fully-connected (FC) layers are applied to refine the extracted features. Afterwards, a softmax-based classifier and an anchor-based class-specific regressor is applied to generate the final detection predictions.

The TFA is targeted towards few-shot object detection. In the few-shot object detection setting, the total number of classes in the dataset is split into N_b base classes and N_n novel classes. While for the base classes all annotations from the dataset are available, for novel classes only a subset of K annotations is available. In the first stage, the TFA applies a regular Faster R-CNN training for the base classes. The target of this base training is to learn features that are general enough to be reused with novel classes. In the second stage called fine-tuning stage, the novel classes are added to the classification stage of the network with randomly initialized weights. Thus, the classification head's prediction layer outputs $N_b + N_n$ dimensions instead of N_b. Afterwards, the softmax is applied over all $N_b + N_n$ dimensions. For each class, base and novel, K shots are sampled to create a balanced training set. During training only the last layers, i.e. the final classification layer and the final regression layer, are adjusted. All other weights are fixed to prevent overfitting. Additionally, the learning rate is reduced by a factor of 20 compared to the first stage.

3.2. Extensions

To improve precision and recall for base and novel classes, we apply multiple improvements. These extensions are described in this section and are necessary because of the new challenges imposed by iSAID compared to MS COCO like a higher number of objects per image and a more difficult distinction between foreground and background.

3.2.1 Annotation Sampling Strategy

In the few-shot object detection setting, each class, base and novel, only uses K shots in the fine-tuning stage. However, this is neither sensible to represent a practical application since more samples for base classes are available nor is this technically sensible since it reduces the variance of the data. Thus, we introduce a base shot multiplier M_{BSM}. Instead of only K shots, we use $M_{BSM} \cdot K$ shots for the base classes. However, this introduces an imbalance in the training since objects of base classes appear more often. Thus, we introduce a novel-class oversampling factor M_{NOF} which duplicates the images containing objects from novel classes until each image is present M_{NOF} times in the fine-tuning dataset. Due to preliminary experiments, we set both parameters to 5.

3.2.2 Data Pre-Processing

In the TFA, a single annotation is used for each image. If multiple annotations are present on an image, the image is duplicated for each annotation. However, this leads to slow convergence and unstable training since the network is not

learned with a consistent mapping of an input image to object predictions. Thus, we always use all available annotations of a single class for an image during training.

3.2.3 Unfreezing the Region Proposal Network

In the base training of TFA, the RPN is trained to filter out predictions not containing objects from base classes. This implies that the RPN is trained to predict novel classes as background. In the fine-tuning stage, the RPN is fixed and no weight adjustments are made to generate proposals for novel classes. This is an adequate approach for datasets like MS COCO which tend to have large objects in the foreground that are clearly separable from the background and most of these foreground objects are annotated. Thus, the RPN will likely learn to generate a region proposal for every foreground object which results in a class-agnostic RPN. In contrast, aerial imagery has a much noisier background since many objects are not annotated. For example, in iSAID, trees and buildings are not annotated while they are clearly separable from the background. Thus, the RPN is learned to be more class-specific than in the case of MS COCO.

Thoroughly analyzing the results has supported this assumption and has shown that the fixed RPN leads to a low recall since missed proposals by the RPN can not be recovered in later stages. To increase the recall, we unfreeze the RPN in the fine-tuning stage. This adjustment enables the RPN to learn generating proposals for novel classes.

3.2.4 Double Head Predictor

Since the base classes are trained only with K shots in the fine-tuning stage, they are subject to a phenomena called catastrophic forgetting. Thus, the impact of the large dataset used for the base training is diminishing over time and only the generalization of the K shots remains. Another problem of TFA is the impact of the randomly initialized novel classes on the base classes due to the softmax-classifier. If the classifier erroneously predicts a high score for a novel class because of insufficient training data, the score of all base classes will be low.

To solve this problem, we propose a novel double head predictor design as shown in Figure 1. In contrast to TFA, the head for predicting the base classes is not extended during fine-tuning but a second head is introduced for predicting the novel classes. For the model of the base training, the two FC layers after RoI Align are called FC_1 and FC_2 while the predictor layer is called FC_p. In our double head predictor design used for fine-tuning, FC_1 is shared for both heads. The weights of FC_2 are duplicated with the two new layers being called $\mathrm{FC}_{2,\mathrm{b}}$ and $\mathrm{FC}_{2,\mathrm{n}}$ for predicting base classes and novel classes, respectively. While the weights of $\mathrm{FC}_{2,\mathrm{b}}$ are fixed during training, the weights of $\mathrm{FC}_{2,\mathrm{n}}$ are trained to enable adjustments towards the novel classes. The old predictor layer FC_p is now called $\mathrm{FC}_{\mathrm{p,b}}$ for the prediction of the base classes and its weights are fixed to prevent base class degradation. To predict the novel classes, a newly initialized $\mathrm{FC}_{\mathrm{p,n}}$ is introduced. The softmax activation is applied separately on the results of $\mathrm{FC}_{\mathrm{p,b}}$ and $\mathrm{FC}_{\mathrm{p,n}}$. This decouples the predictions of base and novel classes and prevents a negative impact of the fine-tuning on the prediction of the base classes.

4. Experimental Results

In the following section, we first introduce the experimental settings. Then, we compare our proposed few-shot detection method to state-of-the-art on the iSAID dataset in quantitative and qualitative manner followed by an ablation study. Furthermore, the impact of the number of base and novel classes on the detection performance is examined. Finally, we present experimental results on a differing dataset.

4.1. Experimental Settings

For our experiments, we use the iSAID dataset [49], which comprises 2806 aerial images. The dataset provides ground truth annotations for 655,451 object instances, which are divided into 15 categories (see Figure 2). On average, 3.27 classes co-exist per image. The large variation in Ground Sampling Distance (GSD), *i.e*. 1.3e-6 to 4.5 meters per pixel, make the detection task, especially few-shot detection, more difficult. Another difficulty is posed by the unclear separation of foreground and background whereas annotated objects in MS COCO are typically centered in front of a distinct background like a wall. Moreover, objects in iSAID can be arbitrarily rotated in contrast to MS COCO. Following the official data preprocessing protocol[1], each image is cropped into tiles of size 800×800 pixels, whereby adjacent tiles exhibit an overlap of 25%.

We consider three different class splits to perform few-shot object detection. Table 1 gives an overview of the novel classes per split, which are unseen in the base training. The remaining classes are used as base classes. In the first split, the novel classes only comprise different vehicle classes, *i.e*. Helicopter, Ship, Plane and Large Vehicle. Mainly small-sized objects with a large variation in appearance are typical for this split. Note that the class Small Vehicle is not considered due to its high occurrence, which would clearly restrict the number of suitable images for the fine-tuning stage. The second split consists of three novel classes, *i.e*. Baseball Diamond, Soccer Ball Field and Roundabout. In contrast to the first split, the object dimensions are mostly large and the variation in appearance is generally small. Compared to the first and second split, the third split comprises more

[1] https://github.com/CAPTAIN-WHU/iSAID_Devkit

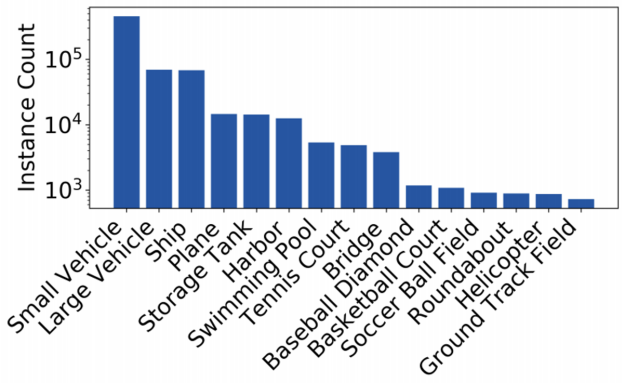


Figure 2: Histogram showing the number of instances per class. Image taken from [49].

Split	Novel Classes
1	Helicopter, Ship, Plane, Large Vehicle
2	Baseball Diamond, Soccer Ball Field, Roundabout
3	Ground Track Field, Helicopter, Baseball Diamond Roundabout, Soccer Ball Field, Basketball Court

Table 1: Overview of novel classes per split.

novel classes, *i.e*. six. For this, we considered the classes with the lowest occurrence.

To evaluate the performance of our proposed few-shot detection method, we compute the mean Average Precision (mAP) for the novel classes, the base classes and all classes together. Detections with an Intersection-over-Union (IoU) to a ground truth annotation above 0.5 are considered as correct. In novel fine-tuning phase, we set the number of annotated bounding boxes per class to 10, 50, and 100, respectively. The number of annotated bounding boxes are selected in order to determine an appropriate number of required samples for the few-shot detection task.

4.2. Results on iSAID

The few-shot object detection performance of our proposed DH-FSDet on the iSAID dataset is given in Table 2. For comparison, we consider two differing few-shot detection methods that achieve state-of-the-art results on the MS COCO benchmark dataset. The first method termed few-shot object detection and viewpoint estimation (FSDetView) [52] is a feature reweighting approach based on Meta R-CNN. To account for the large variation in object dimensions, we attach a FPN onto the base network. The second method referred to as two-stage fine-tuning approach (TFA) [47] is a straightforward approach, which is used as basis for our approach. As described in Section 3.1, TFA initially trains the entire object detector, *i.e*. Faster R-CNN with FPN, on the base classes followed by fine-tuning both base and novel classes on a small balanced training set. Note that only the last layers of the detector head are fine-tuned, while all other parameters of the model are kept fixed. For fair comparison, the same base network, *i.e*. ResNet-50, is used for all methods and weights pre-trained on ImageNet are used for initialization. As the results strongly depend on a small number of samples, three separate fine-tunings are performed per model and the mAP is averaged over all runs.

Our proposed method clearly outperforms the baseline methods on both the novel and base classes. While FSDetView achieves better detection accuracies on the novel classes than TFA, the detection accuracies on the base classes drop by a large margin compared to the results achieved after the base training, indicating the loss of knowledge gained during base training. TFA circumvents this large drop on the base classes by keeping most parameters fixed during fine-tuning. However, only fine-tuning the last layers of the detector head is not sufficient to accurately localize and classify the novel classes as will be discussed in Section 4.3. Our proposed method facilitates the learning of novel classes, while the results on the bases classes are in contrast to the baselines almost similar to the results after the base training.

As expected, the mAP values increase with more shots for all methods. In particular, using only 10 shots is not sufficient to adequately learn the large variation in appearance and size of occurring objects in the iSAID dataset. Comparing the results of the first and second split confirms this assumption, as clearly higher mAP values are achieved for the novel classes in the second split, which exhibit comparatively less variation in appearance. Though the third split mainly comprises novel classes with small variation in appearance, the mAP values for the novel classes are worse compared to the second split. This indicates that considering more novel classes impede the few-shot learning task.

Qualitative experiments given in Figure 3 confirm the improved detection accuracy for both novel and base classes compared to the baseline approaches.

4.3. Ablation Experiments

In the following, ablation experiments are performed to analyze the impact of our proposed extensions in more detail. Results are exemplarily given for a single run with 100 shots on the second split in Table 3. We first analyze the impact of parameters that affect the data preparation during fine-tuning, *i.e*. the base shot multiplier M_{BSM} and the novel-class oversampling factor M_{NOF}. As described in Section 3.2.1, M_{BSM} is introduced to increase the number of samples for the base classes, while M_{NOF} adjusts the duplication of novel class annotations to balance the class distribution. Note that both values are set to 1 for the baseline method. Increasing the number of samples for the base class by a factor of 5 yields an improved mAP for the

Method	Shot	mAP (in %) - Split 1		mAP (in %) - Split 2		mAP (in %) - Split 3	
		Novel	Base	Novel	Base	Novel	Base
FSDetView [52]	10	1.3 ± 0.3	33.8 ± 0.5	8.7 ± 2.1	29.8 ± 1.6	4.6 ± 1.2	32.9 ± 3.4
	50	7.2 ± 2.3	35.3 ± 0.5	26.8 ± 2.8	30.0 ± 1.1	17.1 ± 1.1	34.6 ± 1.1
	100	10.2 ± 1.2	36.4 ± 0.6	32.8 ± 2.0	30.4 ± 0.4	24.1 ± 1.1	34.5 ± 1.3
TFA [47]	10	3.3 ± 0.8	58.6 ± 0.3	9.0 ± 2.6	56.5 ± 0.8	3.8 ± 1.1	59.0 ± 1.5
	50	4.7 ± 0.0	60.7 ± 0.5	12.1 ± 1.9	58.5 ± 0.8	5.6 ± 1.4	60.9 ± 0.3
	100	5.0 ± 0.3	61.4 ± 0.3	14.4 ± 1.5	59.2 ± 0.2	5.4 ± 1.1	61.6 ± 0.4
DH-FSDet (Ours)	10	5.2 ± 0.8	65.0 ± 0.2	14.5 ± 1.7	64.5 ± 0.1	9.7 ± 2.2	67.8 ± 0.1
	50	12.8 ± 0.8	65.1 ± 0.1	28.9 ± 3.4	64.7 ± 0.1	19.6 ± 2.4	68.0 ± 0.1
	100	16.7 ± 1.7	65.2 ± 0.1	36.0 ± 1.7	64.8 ± 0.1	23.1 ± 0.9	68.1 ± 0.1

Table 2: Few-shot object detection evaluation on iSAID. We report the mAP for 3 different splits for 10, 50 and 100 shots. Note that the results are averaged over three runs. Our proposed method outperforms two differing state-of-the-art methods by a large margin for both novel and base classes.

Figure 3: Qualitative results for FSDetView (top row), TFA (middle row) and our proposed method for 100 shots. Our proposed method exhibits clearly better recall rates for novel classes in the first split (left three columns) and in the second split (right three columns) as well as better results for the base classes.

base classes, as more diverse samples are considered during fine-tuning. However, the detection accuracy for the novel classes decreases, which indicates that an unbalanced class distribution in the fine-tuning step yields worse results for the underrepresented classes. Setting both values to 5 yields improved mAP values for both novel and base classes.

By default, only a single annotation is considered per image. In case of multiple objects per image, duplicate images are generated for each object and only the corresponding annotation is considered for the respective duplicate image. As this procedure may impede the classification accuracy, we use all annotations instead. While the detection accuracy slightly increases for the novel classes, the detection accuracy for the base classes is improved by 2.2% in mAP.

To analyze the still low mAP values for the novel classes in more detail, we examine the precision-recall curves (PRCs). Figure 4 shows precision-recall curves (PRCs) exemplarily for class Roundabout, which exhibits the lowest AP values compared to Baseball Diamond and Soccer Ball Field. The red PRC clearly indicates that one reason for the low AP is the poor recall rate. We assume that the high number of missed detections is due to the fixed RPN. As the

Data Preparation		Number of	Unfixed	Unfixed	Double	mAP (in %)		
M_{BSM}	M_{NOF}	Annotations	RPN	Last FC	Head	Novel	Base	All
1	1	one	-	-	-	14.8	57.8	49.2
5	1	one	-	-	-	13.0	59.0	49.8
5	5	one	-	-	-	16.0	58.9	50.3
5	5	all	-	-	-	16.4	61.1	52.2
5	5	all	✓	-	-	25.8	62.5	55.2
5	5	all	✓	✓	-	**35.5**	59.2	54.4
5	5	all	✓	✓*	✓	34.0	**64.7**	**58.6**

Table 3: Ablation results showing the impact of modifying the data preparation during fine-tuning, *i.e.* increasing the base-shot multiplier M_{BSM} and the novel-class oversampling factor M_{NOF}, the number of annotations per fine-tuning sample, unfreezing of layers and our proposed double head. The results are exemplarily reported for a single run on the second split with 100 shots. * indicates that the last FC is only unfixed in the novel head.

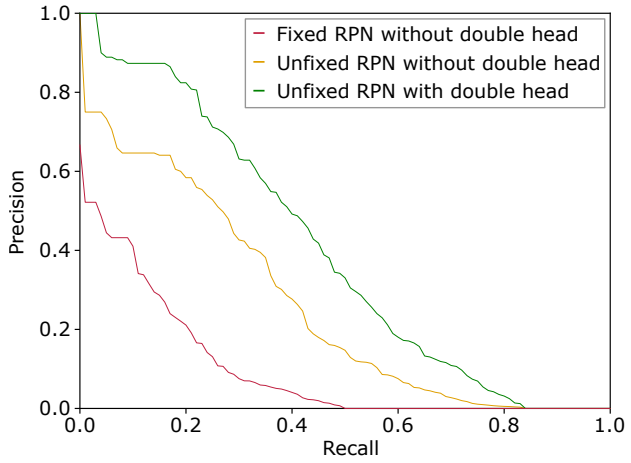


Figure 4: Precision-recall curves for our proposed approach with various configurations exemplarily for class Roundabout to demonstrate the impact of the proposed extensions. Unfreezing the RPN results in higher recall rates, while introducing the double head yields higher precision values.

generation of region candidates is only learned during base training, region candidates for novel classes that clearly differ from the base classes may not be generated and thus, cannot be detected.

Hence, we unfreeze the RPN during fine-tuning in order to explicitly learn to generate region candidates for the novel classes, yielding an improved detection accuracy by 9.4% in mAP for the novel classes. As shown in Figure 4 (yellow curve), the recall rate considerably increases by unfreezing the RPN. This shows that the generalization ability of the RPN is limited in case of unseen classes, whose appearance clearly differs from the classes seen during training. However, the achieved precision is still low, as the model is not able to confidently distinguish between the novel class and the background class.

To improve the precision, we analyze the impact of unfreezing more parameters. While by default only the parameters in the prediction layers are fine-tuned, we further unfreeze the parameters of the last fully connected layer, which is prior to the prediction layers. Note that no further layers are unfixed to avoid overfitting. The detection accuracy considerably improves for the novel classes, while the mAP for the base classes drops. This indicates that learning more parameters is essential to achieve good precision for novel classes, but knowledge about the base classes learned during the base training gets lost.

Thus, we apply our proposed double head so that the parameters can be kept fixed for the base classes and learned for the novel classes, which yields the best trade-of in detection accuracy between the novel and base classes. As exemplarily shown for class Roundabout in Figure 4 (green curve), the precision is clearly improved by introducing the proposed double head, which facilitates the learning of more parameters in case of the novel classes.

4.4. Impact of the Number of Base / Novel Classes

Analyzing the per class mAP for the novel classes in the second and third split, indicate that more novel classes yield worse mAP values and that less base classes result in worse mAP values for the novel classes, respectively. Hence, we analyze the impact of the number of base and novel classes in more detail. The results are averaged over three seeds.

First, we vary the number of classes in the base training (see Table 4). For each experiment, the novel classes of the third split are used as novel classes. As base classes we consider the three, six and nine classes with the highest instance count (see Figure 2), respectively. Using more base classes results in higher mAP values for the novel classes. A reason for the higher mAP is the higher diversity of the classes and images during base training, yielding less class-specific features that are kept fixed during fine-tuning, which seems beneficial to learn novel classes with few examples.

We further vary the number of novel classes in the fine-tuning stage (see Table 5). Note that the same base training

# Base Classes	mAP (in %) - Novel
3	19.1 ± 0.55
6	20.4 ± 1.67
9	23.1 ± 0.91

Table 4: Impact of different number of base classes during base training. The novel classes of the third split are used as novel classes.

# Novel Classes	mAP (in %)					
	RA	BD	HC	GTF	SBF	BC
1	**40.4**	41.1	**24.6**	**17.7**	**16.8**	**15.0**
3	40.1	**42.3**	24.4	13.0	14.7	10.7
6	39.9	38.8	21.6	14.8	14.3	9.2

Table 5: Impact of different number of novel classes during fine-tuning. Classes divided by a vertical line are trained separately. Note that the same base training is used for all fine-tunings. Abbreviations: RA - Roundabout, BD - Baseball Diamond, HC – Helicopter, GTF - Ground Track Field, SBF - Soccer Ball Field, BC - Basketball Court.

is used for all fine-tunings. For the base training, we use the base classes of the third split. We consider three different numbers of novel classes, *i.e*. 1, 3 and 6. In case of three novel classes, we define two sets of classes based on their mAP values for fine-tuning with 6 novel classes. The first set contains the classes exhibiting the highest values and the second set contains the classes exhibiting the lowest values. Training with only one novel class leads to the highest mAP values for all classes but Baseball Diamond which is most accurately detected when trained with three novel classes. Increasing the number of novel classes from three to six only improves the mAP of Ground Track Field. This indicates that considering more classes that are novel impairs the learning of the single novel classes.

As the number of base and novel classes clearly affects the few-shot detection performance, the selection of base and novel classes is an important setting for specific few-shot detection tasks. So far, the impact of the selection of base and novel classes is not examined in detail, which has to be addressed in future work.

4.5. Results on NWPU VHR-10

For comparison with state-of-the-art approaches, we evaluate our method on the NWPU VHR-10 [6] dataset. It contains 800 high resolution remote sensing images with 650 of them including objects of the annotated classes. Of the total of 10 classes, 7 classes (Ship, Storage Tank, Basketball Court, Ground Track Field, Harbor, Bridge, Vehicle) are base classes and 3 classes (Airplane, Baseball Diamond, Tennis Court) are novel classes. We compare to the results of [24]. Thus, we follow the evaluation protocol of [24] and use the identical samples for fine-tuning.

Method	Shot	mAP (in %)	
		Novel	Base
Faster R-CNN (ResNet-101)	20	33.7	70.0
YOLOv3	20	27.7	76.6
Yolo-Low-Shot	3	12.0	76.1
FSODM	3	32.3	77.9
DH-FSDet (Ours)	3	**35.6**	**93.2**

Table 6: Comparison of our proposed few-shot object detector to different state-of-the-art detectors on the NWPU VHR-10 dataset.

The results are shown in Table 6. For novel classes, we have a significantly higher mAP than dedicated few-shots methods for aerial imagery as well as conventional object detectors like Faster R-CNN even though the conventional detectors have been trained with 20 shots instead of 3. The better accuracy is due to the advanced fine-tuning strategy with careful unfreezing of certain layers. Looking at the base classes, our proposed method has an even higher advantage since the use of a FPN and a fixed-scale second stage induces a higher scale-invariance which supports the detection in aerial imagery containing objects of highly different scales. Note that the base class performance of our method is evaluated after fine-tuning while the base classes of the other models are evaluated after base training. Evaluating after fine-tuning is a harder task since only a small number of samples is available for the base classes during fine-tuning, which makes a model prone to catastrophic forgetting.

5. Conclusion

In this paper, we proposed a novel few-shot detection method for aerial imagery based on TFA. To account for the often high number of object instances in aerial images, we applied a novel annotation sampling and pre-processing strategy, yielding a better exploitation of base class annotations and a more stable training. We further proposed a modified fine-tuning scheme to reduce the number of missed detections. To prevent loss of knowledge learned during the base training, we introduce a novel double head predictor, exhibiting the best trade-off in detection accuracy between the novel and base classes. Our proposed method outperforms state-of-the-art baselines on publicly available aerial imagery datasets. Furthermore, we demonstrated how the selection of novel and base classes affects the detection performance. In future work, we will analyze in more detail how the selection of base and novel classes, *e.g*. similarity of base and novel classes, affects the detection performance. As the size, appearance and orientation of objects strongly vary in aerial imagery, we will evaluate the impact of different data augmentation techniques for few-shot detection.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[3] Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, and Winston H Hsu. Should i look at the head or the tail? dual-awareness attention for few-shot object detection. *arXiv preprint arXiv:2102.12152*, 2021.

[4] Xianyu Chen, Ming Jiang, and Qi Zhao. Leveraging bottom-up and top-down attention for few-shot object detection. *arXiv preprint arXiv:2007.12104*, 2020.

[5] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11–28, 2016.

[6] Gong Cheng, Junwei Han, Peicheng Zhou, and Lei Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:119–132, 2014.

[7] Gong Cheng, Bowei Yan, Peizhen Shi, Ke Li, Xiwen Yao, Lei Guo, and Junwei Han. Prototype-cnn for few-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[9] Zhipeng Deng, Hao Sun, Shilin Zhou, Juanping Zhao, Lin Lei, and Huanxin Zou. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, 145:3–22, 2018.

[10] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019.

[11] Peng Ding, Ye Zhang, Wei-Jian Deng, Ping Jia, and Arjan Kuijper. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 141:208–218, 2018.

[12] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[14] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[15] Wei Guo, Wen Yang, Haijian Zhang, and Guang Hua. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. *Remote Sensing*, 10(1):131, 2018.

[16] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. *arXiv preprint arXiv:2104.07719*, 2021.

[17] Xiaobing Han, Yanfei Zhong, and Liangpei Zhang. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. *Remote Sensing*, 9(7):666, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

[20] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2019.

[21] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.

[22] Ke Li, Gong Cheng, Shuhui Bu, and Xiong You. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2337–2348, 2017.

[23] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

[24] Xiang Li, Jingyu Deng, and Yi Fang. Few-shot object detection on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–14, 2021.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Longyao Liu, Bo Ma, Yulin Zhang, Xin Yi, and Haozhi Li. Afd-net: Adaptive fully-dual network for few-shot object detection. *arXiv preprint arXiv:2011.14667*, 2020.

[29] Liyang Liu, Bochao Wang, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Gendet: Meta learning to generate detectors from few shots. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[31] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

[32] Kun Nie, Lars Sommer, Arne Schumann, and Jurgen Beyerer. Semantic labeling based vehicle detection in aerial imagery. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 626–634. IEEE, 2018.

[33] Jiangmiao Pang, Cong Li, Jianping Shi, Zhihai Xu, and Huajun Feng. R2-cnn: Fast tiny object detection in large-scale remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5512–5524, 2019.

[34] Xiaoliang Qian, Sheng Lin, Gong Cheng, Xiwen Yao, Hangli Ren, and Wei Wang. Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion. *Remote Sensing*, 12(1):143, 2020.

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[36] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[39] Yun Ren, Changren Zhu, and Shunping Xiao. Small object detection in optical remote sensing images via modified faster r-cnn. *Applied Sciences*, 8(5):813, 2018.

[40] Wesam Sakla, Goran Konjevod, and T Nathan Mundhenk. Deep multi-modal vehicle detection in aerial isr imagery. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 916–923. IEEE, 2017.

[41] Lars Sommer, Tobias Schuchert, and Jürgen Beyerer. Comprehensive analysis of deep learning-based vehicle detection in aerial images. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2733–2747, 2018.

[42] Lars Wilko Sommer, Tobias Schuchert, and Jürgen Beyerer. Fast deep vehicle detection in aerial images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 311–319. IEEE, 2017.

[43] Tianyu Tang, Shilin Zhou, Zhipeng Deng, Lin Lei, and Huanxin Zou. Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks. *Remote Sensing*, 9(11):1170, 2017.

[44] Hilal Tayara and Kil To Chong. Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network. *Sensors*, 18(10):3341, 2018.

[45] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9627–9636, 2019.

[46] Chen Wang, Xiao Bai, Shuai Wang, Jun Zhou, and Peng Ren. Multiscale visual attention networks for object detection in vhr remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 16(2):310–314, 2018.

[47] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

[48] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9925–9934, 2019.

[49] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.

[50] Aming Wu, Yahong Han, Linchao Zhu, Yi Yang, and Cheng Deng. Universal-prototype augmentation for few-shot object detection. *arXiv preprint arXiv:2103.01077*, 2021.

[51] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[52] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European Conference on Computer Vision*, pages 192–210. Springer, 2020.

[53] Zixuan Xiao, Jiahao Qi, Wei Xue, and Ping Zhong. Few-shot object detection with self-adaptive attention network for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[54] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9577–9586, 2019.

[55] Michael Ying Yang, Wentong Liao, Xinbo Li, Yanpeng Cao, and Bodo Rosenhahn. Vehicle detection in aerial images. *Photogrammetric Engineering & Remote Sensing*, 85(4):297–304, 2019.

[56] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: tackling object confusion for few-shot detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12653–12660, 2020.

[57] Shan Zhang, Dawei Luo, Lei Wang, and Piotr Koniusz. Few-shot object detection by second-order pooling. In *Proceedings of the Asian Conference on Computer Vision*, 2020.