BENCHMARKING STATE-OF-THE-ART FEW-SHOT LEARNING ALGORITHMS FOR VISUAL-BASED PILL DETECTION

A PREPRINT

Nhat Huy Phan
VinUni-Illinois Smart Health Center
VinUniversity
Hanoi, Vietnam
20huy.pn@vinuni.edu.vn

Huyen Thi-Ngoc Nguyen
VinUni-Illinois Smart Health Center
VinUniversity
Hanoi, Vietnam
huyen.ntn@vinuni.edu.vn

Phi Le Nguyen
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
lenp@soict.hust.edu.vn

Huy Hieu Pham
VinUni-Illinois Smart Health Center
VinUniversity
Hanoi, Vietnam
hieu.ph@vinuni.edu.vn

July 9, 2022

ABSTRACT

Identifying the various type of pills correctly from images is imperative ensuring patient safety and facilitate more effective patient care, especially in those populations that cannot identify medication on their own such as older peoples or children. Limited pill datasets are available, so few-shot learning is key to learn how to automatically recognize pill images. However, prior methods have been evaluated on publicly available datasets of high-quality, laboratory environment pill images, which prevents the utility of these models in real-world settings. To address this issue, we introduce an evaluation framework that benchmarks state-of-the-art few-shot learning algorithms for visual-based pill detection on a real-world dataset. First, we construct a dataset that contains pill images in a variety of conditions. We then re-evaluate several state-of-the-art few-shot methods for pill recognition tasks on our formulated evaluation framework. Finally, we provide a novel and strong baseline comprising of low-rank approximation techniques and semantic axes augmentation.

Keywords Pill detection · Few-shot object detection · Benchmark dataset

1 Introduction

Preventable medical errors, in particular medication error, are the third leading cause of death in the United States (1; 2; 3). The widespread growth of consuming medications has increased the need for applications that support medication reconciliation.

Typically, most pill identification approaches can be broken down into two main categories: manual entry based approaches and computer vision-based approaches. The first group allows the user to manually enter the characteristics of a pill to identify it. The second one automate this tasks can be considered superior than the first group. The second group can be broken down even further into deep learning and feature engineering based approaches. To the best of our knowledge, only the image classification model were available for the pill identification and no detection model a applied for this task until now.

Artificial intelligence (AI), powered by great advances in machine learning and deep neural networks, has made substantial progress across many areas of medicine in the past decade (4; 5; 6). Much of the AI work in healthcare is focused around disease prediction in clinical settings and the use of AI models to prevent medication error is

unexploited. Several studies show that AI-based applications can help correctly identifying prescription medication, which is currently a tedious and error-prone task. For instance, data-driven machine learning (ML) is currently a powerful approach for building accurate and robust pill identifier (7). However, ML algorithms will be prevented from reaching its full potential without access to sufficient data and the transition from research to clinical practice. In a real-life pill detection setting, the number of new classes is increasing such as new products, new manufactures, etc. Very few samples per new classes under noisy image conditions, taken by users. This raise the challenge of learning to detect pill types with very little examples.

Although the research of few-shot learning is developing rapidly, there is no existing work on the task of pill detection. The lack of a standard dataset and evaluation protocol has become an obstacle hindering fair comparison between few-shot algorithms. The above phenomena highlight the need for a common protocol for the evaluation of few-shot pill detection methods.

In this work, aiming at addressing the aforementioned challenge, we aim to perform a series of experiments that re-evaluate recent state-of-the-art few-shot image detectors powered by deep neural networks in the task of recognizing pills from images. The experimental results allow us to benchmark the SOTA models when considered for a real-world pill detection implementation. We show that most of SOTA approaches reported a high-level of accuracy in this setting.

1.1 Our contributions

To sum up, our contributions are as follows.

- Meanwhile all the existing pill identifier approaches are classification solution, we are the first evaluate detection task for pill identification.
- We introduce a new evaluation framework of few-shot pill detection and then re-evaluate in this paper state-of-the-art few-shot learning algorithms for pill detection from images. The benchmarking results serve as reference for measuring the effusiveness of current few-shot learning algorithms on a real-world setting. Thus our framework allows for more reliable comparison of few-shot pill detection methods.
- These benchmarks reflect the current state of the art few-shot learning for image-based pill detection task. This will serve as important baselines for future research. To support reproducing our results and benchmarking few-shot pill detection methods, we open-source our toolkit called FSPill that contains implementations of a number of state-of-the-art pill detection methods, data processing utilities, as well as our proposed evaluation framework.
- We introduce a new few-shot pill detection called LowRank TFA based on the low rank approximation of semantic layer in RoI module in standard faster-RCNN (8). This method shows its superiority in almost metrics and settings comparing with other state-of-the-art few-shot detection algorithm.

2 Related Works

2.1 Visual-based Pill Recognition

(9) was one of the first to study pill recognition based on visual characteristics. From there, various works continued to address the task of pill classification and detection based on a number of approaches. The early attempts carried out by (9) and temporary researchers focused primarily on the handcrafted features of pill images for pill classification. The most commonly addressed features were shape, color and imprints on pill surfaces. These features was used by (9) to construct descriptors that the author proposed to best complements to separate pill categories, and with a K-NN classifier, achieve 91% accuracy on more than 500 pill types. Other works put a strong emphasis on recognizing pills based on the pill imprints. (10) studied Modified Stroke Width Transform, Weighted Shape Context to better extract the imprint features, attaining accuracy of 92 for top 5 rank collected pills for 2000 pill types, respectively. The imprints are further stress in (11) works, as the authors used a two-step sampling distance sets to better cope with noise in the extracted imprints, obtaining accuracy of 97.16 for 2500 pill categories.

These early work addressed a data-set with limited number of samples, taken in typically constrains imaging conditions and not publicly available. They commonly incorporate a prior knowledge of pill domain and hand-crafted features for the recognition systems.

The more recent works in the field of visual-based pill recognition are constructed on more accessible public large data-set, such as NIH NLM (12), CURE(13) and incorporate more intricate techniques such as deep convolution neural networks. (14) addresses geometric and color distorting of pill images using QR-board, and construct a deep neural network to extract features of detected cropped pill image alongside with a baseline composing the hand crated features,

witnessing a boost in performance between handcrafted features and feature extracted using convolution networks. The authors achieved 95.35 top 1 return rate, but on a private data-set of 400 pill types . (15) used Google Le-Net Inception network on 1000 NIH pill types, obtaining a maximum of 22.86 top 1 accuracy. MobileDeepPill (16) utilized a multi-stream CNN structure of three streams color, gray and gradient images for pill classification. The model achieved 52.7 Top-1 accuracy in one-side and 73.7 Top-1 accuracy in two-side pill recognition scheme in NIH-NLM challenge.

ePill benchmark (17) approach the pill recognition as a fine-grained data-set. They run plain common convolution backbones along with multi-head metric learning classification and obtained the best performance mean average precision (MAP@1) of over 90 percent for two-side and over 70 percents for one-side input images.

2.2 Few-Shot Object Detection

Meta learning few-shot detection is further divided into single branch and dual branch few shot detection. Dual branch methods generally attracts more attention from community (18; 19; 20; 21; 22; 23; 24), whereas single branch methods are quite new with fewer deviations (25; 26; 27).

Single branch meta-learning simulates a simple, single-branch generic object detection with modified classification heads that are usually replaced with metric learning heads. RepMet (25) was one the first in this line of works, which defined the task as learning a distance metric. The extracted region proposals of a faster R-CNN is further used to learn an embedding space for all classes such given the embedding of a query, the similarity score of it to the representative vector of the class it belong is the largest. (26) utilize metric learning but with a cosine similarity for distance metric learning that is claimed to cater for the novel categories and improve their detection scores. Other works in the line of single branch meta learning tries to reduce learnable parameters, such MetaDet (28) that can learn to generate weights for a novel class.

Transfer learning is easily more simplistic than meta learning yet achieve a state-of-the-art performance and easily surpass some meta learning approaches for few-shot detection. TFA (29), one the first work in this line, proposed a simple two-phase training mechanism composed of training the base classes in the first phase and fine-tuning on the novel classes in the second phase. After the first phase, the weights of the networks are frozen so that only the weights from ROI head are open for training during the second stage. Like in meta learning, the common frameworks is the two-stage faster R-CNN. In FSCE (30), transfer learning is applied with a bit of modification to the classification head and the cosine similarity is used, which is claimed to be able to compensate for the mismatch between the categories of the base and novel classes. Many deviates the general transfer learning idea of TFA (29) by a modification to the loss function, such as (31), which utilized a object concentration loss for handling intra-class agreement, a background concentration loss for unlabelled instances, and a contrastive loss in the contrastive branch for obtaining the most representative embeddings. A mechanism to optimize the gradient flow is utilized in DeFRCN (32), which freeze and unfreeze specific layers at specific times in the training process.

2.3 Few-Shot Learning for Pill Recognition

Despite of the need of pill recognition system with few-shot learning ability, to the best of our knowledge, there is only one work about this ability of pill recognition algorithm: Multi-Stream deep network for pill classification (13). However, this work only focuses on the classification task and has complicated multi stages of training and components reducing the desired flexibility of any fast adaptive few-shot algorithm.

3 Evaluation Framework

3.1 Problem Formulation

Few-shot object detection aims at detecting novel objects with only few annotated instances. Formally, the training dataset $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{base}} \cup \mathcal{D}_{\text{novel}}$ is separated into two datasets $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{novel}}$ containing non-overlapping sets of base categories $\mathcal{C}_{\text{base}}$ and novel categories $\mathcal{C}_{\text{novel}}$, with $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \emptyset$. Each tuple $(I_i, \hat{y}_{o_1}, \ldots, \hat{y}_{o_M}) \in \mathcal{D}_{\text{train}}$ consists of an image $I_i = \{o_1, \ldots, o_M\}$ containing M objects $o_1, \ldots, o_M$ and their corresponding labels $\hat{y}_{o_i} = \{c_{o_i}, b_{o_i}\}$, including the category c_{o_i} and the bounding box $b_{o_i} = \{x_{o_i}, y_{o_i}, w_{o_i}, h_{o_i}\}$ with coordinates (x_{o_i}, y_{o_i}), width w_{o_i}, and height h_{o_i}. For the base categories $\mathcal{C}_{base}$ abundant training data are available in the base dataset $\mathcal{D}_{\text{base}}$. In contrast, the novel dataset $\mathcal{D}_{\text{novel}}$ contains only few annotated object instances for each novel category in $\mathcal{C}_{\text{novel}}$. For the task of K-shot object detection, there are exactly K annotated object instances available for each category in $\mathcal{C}_{\text{novel}}$. Therefore the number of annotated novel object instances $|\{o_j \in I_i \forall I_i \in \mathcal{D}_{\text{novel}}\}| = K \cdot |\mathcal{C}_{\text{novel}}|$ is relatively small. Note that the number of annotated object instances does not necessarily correspond to the number of images, as one image may

contain multiple instances. N-way object detection denotes a detector that is designed to detect object instances from N novel categories, where $N \leq |\mathcal{C}_{\text{novel}}|$. Few-shot object detection is therefore often referred to as N-way K-shot detection. In this track, **there are only two sets of data $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{novel}}$.**

3.2 Image Pill Dataset

Our benchmark will be based on current state VAIPE Pill Identification dataset (VAIPE-PI) from VAIPE project (33) The VAIPE Pill Identification was built to benefit the research on recognizing distinct types of medicines from mobile devices in order to ensure patient safety and promote more effective medical care. The original version of VAIPE-PI includes more than 10000 images, 60000 box-level pill annotations with variety of backgrounds and more than 400 pill categories, however in this few-shot pill detection benchmark, we will discard some poor-quality images and annotations as well as some pill categories that have too little samples for creating a credible test set.

Characteristic	Training set	Testing set	Total
Number of images	6461	833	7294
Number of pill categories	262	262	262
Instances per category	179.75	23.56	203.2
Image size (pixel x pixel, mean)	3311×3276	3276×3469	3300×3400
Instances per image	7.28	7.4	7.3
Number of boxes annotation	47097	6174	53271
Number of categories per image	5.18	5.76	5.32

Table 1: Common Statistics of VAIPE-P

3.3 Data Splits and Few-shots Construction

To create a training set that captures the imbalanced data between classes scenario of the original dataset and a credible, balanced test set as well few-shot samples, we used multi-label data stratification (34) implemented in iterative-stratification package (35) with 20% of samples for test set. Follow TFA (29), to create a balanced few-shot samples for pill detection, we will sample k boxes of each class (even for novel and base classes) from random images in the train set.

3.4 Formulation of Evaluation Framework

A fixed test set D_{test} will be used. In few-shot pill detection, firstly, a initial model will be trained on D_{base}, performance on test set $\mathcal{D}_{\text{test}} \setminus C_{\text{novel}}$ will be reported, where $\mathcal{D}_{\text{test}} \setminus C_{\text{novel}}$ can be comprehended as test set $\mathcal{D}_{\text{test}}$ discarding all annotations having same classes in C_{novel}. A random sampling algorithm will sample k (three different settings will be tested: $k \in \{5, 10, 20\}$) instances for each class (in total of $K * |C_{novel} \cup C_{base}|$ instances), this set called as $\mathcal{D}_{novel}$, after training on this novel set, three performance reports will be conducted on $\mathcal{D}_{\text{test}} \setminus C_{novel}$, $\mathcal{D}_{\text{test}} \setminus C_{base}$ and $\mathcal{D}_{\text{test}}$. Repeat above process 5 times and get the average performance. Follow (29) our performance metric will be COCO-style which include 4 main metrics: Average Precision (AP), AP Across Scales, Average Recall (AR), AR Across Scales. In this paper, when comparing different methods, we only report AP in different IoU thresholds: AP at IoU=0.05:0.95, AP at IoU=0.50 and AP at IoU=0.75 denoted as AP, AP50 and AP75.

4 Baseline Method

We first conducted seris of experiments for different modifications of TFA (29) including original TFA, unfreezing RoI TFA and unfreezing both RoI and last ResNet block TFA to discover which part in detection model plays key role in performance increment on few-shot VAIPE-P. The results are expressed in 2. This simple analysis motivates us to find a strong baseline that compromises between flexibility (ability to fitting to new data points) and catastrophic forgetting. Motivated by series of works on Adapters from Natural Language Processing (36; 37; 38), we propose a light weight adapting module composed by multiple rank-1 factors for Faster-RCNN to solve the aforementioned dilemma. We show mathematical intuition behind module design and superior results of the method over other State-of-The-Art algorithms in VAIPE-P benchmark.

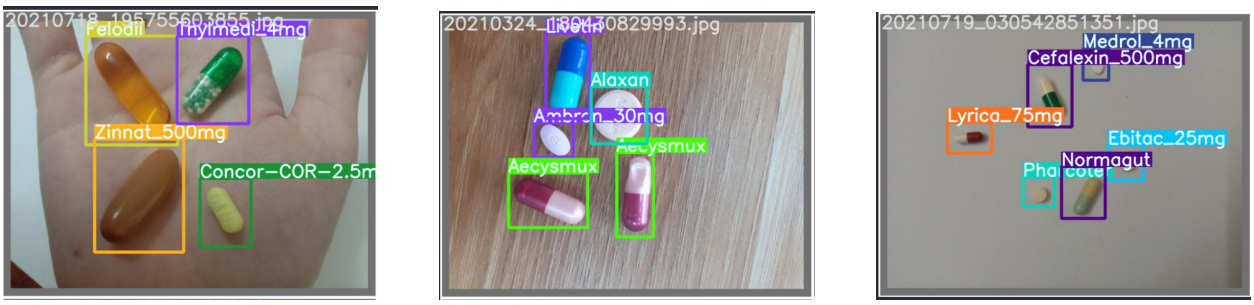

Figure 1: Examples of box-level labels

Algorithm 1: Few-Shot Evaluation Framework

Data: The dataset $\mathcal{D}_{\text{total}}$, number of random seeds S, base classes $\mathcal{C}_{\text{base}}$, novel classes $\mathcal{C}_{\text{novel}}$, all classes $\mathcal{C}_{\text{total}} = \mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}}$, Initial weight $\mathcal{W}_{\text{base}}$, number of base training epochs M, number of novel training epochs N

Result: test performance

Divide $\mathcal{D}_{\text{total}}$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$ according to data split construction strategy
$\mathcal{D}_{\text{base}} := \mathcal{D}_{\text{train}}$
$\mathcal{D}_{\text{novel}} := \{\emptyset\}$
for $c \in \mathcal{C}_{novel}$ **do**
  $\mathcal{C}_{\text{base}} = \text{Discard}(\mathcal{C}_{\text{base}}, c)$ `// Discarding all annotations contain class c in` $\mathcal{D}_{\text{base}}$
end
for *seed* $\in \{1, 2, ..., S\}$ **do**
  for $k \in \{5, 10, 20\}$ **do**
    for $c \in \mathcal{C}_{novel} \cup \mathcal{C}_{base}$ **do**
      Random sample k boxes from set of annotations that contains only class c from $\mathcal{D}_{\text{train}}$
      Extend $\mathcal{D}_{\text{novel}}$ with sampled boxes of class c
    end
  end
  for $i \in \{1, 2, ..., M\}$ **do**
    Train$\{\mathcal{W}_{\text{base}}\}$ with $\mathcal{D}_{\text{base}}$ `// Base training the model weight`
  end
  if *Enlarging Network* **then**
    $\mathcal{W}_{\text{novel}} := \text{Extend}(\mathcal{W}_{\text{base}})$ `// Adding more parameters for some methods`
  else
    $\mathcal{W}_{\text{novel}} := \mathcal{W}_{\text{base}}$
  end
  for $i \in \{1, 2, ..., N\}$ **do**
    Train$\{\mathcal{W}_{\text{base}}\}$ with $\mathcal{D}_{\text{base}}$ `// Novel training`
  end
  for $set \in \{base, novel, total\}$ **do**
    Report the performance on $\mathcal{D}_{\text{test}}$ that contain annotations from set only `// Evaluating Process`
  end
end
Average performance from all seeds for the final performance report.

4.1 Low-rank Update

By analyzing the RoI head weight matrices, we discover two last linear layers have low intrinsic ranks despite of their high dimension output space. This phenomenon is aligned with the general observation in big pre-trained language model (39). In other words, pre-training on large base dataset of pills encourages RoI head weights to collapse into small subset of principle semantic axes that linearizes the input features. Based on this observation, we hypothesize that the most transferable semantic transformation axes are learned in the base training phase, fine-tuning them on small dataset like novel set can destroy the transferable weights by gradient descent updating. Therefore, there's no need to update them in the novel phase of TFA. Meanwhile, in order to make RoI head weights to quickly adapt with new

Regressor and Classifier	RoI heads	ResNet block	bAP	nAP
✓			0.693	0.492
✓	✓		0.642 (**-0.05**)	0.583 (+0.09)
✓	✓	✓	0.623 (**-0.07**)	0.571 (+0.08)

Table 2: Performance of TFA methods with different modules to be unfrozen on 15 samples. The results is showing that adding more parameters in the fine-tuning phase of TFA, especially RoI heads weights, the detection model will excel on the novel classes however its performance on base classes will be dropped significantly due to catastrophic forgetting. Nevertheless, more parameters is not equivalent with the increase in novel classes shown by unfreezing last ResNet backbone and the upper bound of nAP is significant higher than the current one of original TFA.

novel input features, it's desired to have new free low-rank weights so that it can quickly adapt with novel features. To generate those desired semantic transformation axes, we use sparse SVD with chosen decomposed rank $r \ll n$ (the effect of variety of decomposed rank values is described in table). With i^{th} trained linear layer of RoI head represented as affine transformation of input features

$$y = \mathcal{W}_{\text{base},i}x + \mathcal{B}_{\text{base},i}, \tag{1}$$

with $\mathcal{W}_{\text{base},i} \in R^{m\times n}(m \geq n)$ and $\mathcal{B} \in R^n$

$$\mathcal{W}_{base,i} = U_{\text{base},i}S_{\text{base},i}V_{\text{base},i}^{\top} = L_{\text{base},i}R_{\text{base},i}^{\top}, \tag{2}$$

where $U_{\text{base},i} \in R$, $V_{\text{base},i}$ are two low-rank matrices with rank r and S_{base} is diagonal matrix with singular values (strength of each semantic axis). $L_{\text{base},i} = U_{\text{base},i}, R_{\text{base},i} = S_{\text{base},i}V_{\text{base},i}^{\top}$ are weight matrices with exactly rank r. To introduce new free weight factor, we use multiple r_{update} rank-1 factors (the effect of variety of update rank is described in table) update represented as dyads $u_{i,j}v_{i,j}^{\top}$ inspired by (40). The red parameters are frozen in second phase of TFA, while the blue is free to be trained.

$$y = \underbrace{L_{\text{base},i}R_{\text{base},i}^{\top}x}_{\text{base updating}} + \underbrace{\sum_{j=1}^{r_{\text{update}}} u_{i,j}v_{i,j}^{\top}x}_{\text{novel updating}} + \mathcal{B}_{\text{base},i} \tag{3}$$

4.2 Orthogonal Regularization

To allow positive knowledge transfer from pre-trained weights and efficient modelling of new novel factors, it's desired to have orthogonal constraints on semantic transformation axes. Firstly, if the new semantic transformation axes are free to be estimated by new data, there's a possibility that these new axes will have positive correlation with learned axes making learning new novel factors slower and data-inefficiency. The mutually orthogonal constraint has also been known as an effective method to reducing learning of new dataset interference with old learned weights (41).

$$u_{i,j} \perp u_{i,k},\ \forall j \neq k \quad \text{and} \quad u_{i,j} \perp U_{\text{base},i,k}\ \forall j,k \tag{4}$$

To ensure this orthogonal constraints, we adopt different orthogonal regularizations (42) on principle axes of i^{th} linear layer's weights. Here we adopt Mutual Coherence Regularization over Soft Orthogonality Regularization due to its emprical results (see table for more results on different regulariers).

$$L_{\text{orth}}\left(W_i\right) = \lambda \left\|W_i^T W_i - I\right\|_{\infty}, \tag{5}$$

where W is concatenation of novel factors and pre-trained factors $W = \text{concat}(u_{i,1}, \cdots, u_{i,r_{\text{update}}}, U_{\text{base},i})$ and $\lambda = 100$. Finally, the objective of training is.

$$\mathcal{W}_{\text{novel}} = \arg\max_{\mathcal{W}} L_{\text{cls}}(\mathcal{W}, D_{\text{novel}}) + L_{\text{reg}}(\mathcal{W}, D_{\text{novel}}) + \sum_i L_{\text{orth}}\left(W_i\right) \tag{6}$$

4.3 Novel Axes Augmentation

Due to data-scarcity scenarios, data augmentation is straight-foward method mitigate the overfitting. Here in low-rank method, we propose a novel augmentation method that augments latent code along new factor axes. Intuitively, if we

apply some noise in latent feature space, its inverse image is actually an augmented image. Specifically, in this method, we only apply Gaussian with dynamic class conditional variance to those features are composed by new novel factors.

$$y_{\text{aug}} = L_{\text{base},i} R_{\text{base},i}^{\top} x + \sum_{j=1}^{r_{\text{update}}} u_{i,j} x_{i,j}^{\text{aug}} + \mathcal{B}_{\text{base},i}, \tag{7}$$

where

$$x_{i,j}^{\text{aug}} \sim \mathcal{N}(v_{i,j}^{\top} x, \sigma_{i,j,c}^2) \text{ and } c = W_{\text{cls}}(x) \tag{8}$$

Notes that, $v_{i,j}^{\top} x$ is a scalar therefore, $\sigma_{i,j,c}^2$ should be a scalar as well. The variance is estimated by a cache of last m iterations.

$$\sigma_{i,j,c}^2 = \frac{1}{m} \sum_{k=1}^{m} \left(v_{i,j}^{\top} x_{k,c} - \frac{1}{m} \sum_{k=1}^{m} v_{i,j}^{\top} x_{k,c} \right)^2, \tag{9}$$

where $x_{k,c}$ is k^{th} feature in cache that has been classified as c class. Augmenting by estimated variance conditioned on specific class helps algorithm avoiding meaningless augmentation along axes that are not helpful in classify certain classes.

5 Experimental and Results

5.1 Experimental setup & Implementation details

Several codebases were used to re-implement State-of-The-Art methodologies due to the fundamental complexity of each few-shot object detection algorithm. However, to ensure the numerical stability as well as a fair comparison between detectors, we have upgraded CUDA implementation of FSCE and DeFRCN. For our baseline method, some hyperparameters are $r = 180, r_{\text{update}} = 15, \lambda = 100, m = 1000$.

5.2 State-of-the-Art Methods

Most of State-of-the-Art few-shot object detectors fall in two main different categories: transfer learning and meta-learning. Therefore, we will examine some algorithms that characterize well these differences and currently are state-of-the-art methods on well-established few-shot detection benchmark like MS-COCO (43) or Pascal VOC (44). They are namely TFA (29), Meta-DETR (45), DeFRCN (46),FSCE (30), FsDetView (47) and a baseline fine-tuning with Faster-RCNN (48) have been deployed for this task. TFA is a simple method involving in training a Faster-RCNN detector (8) with modified fine-tuning two-stage scheme. In the first stage, the detector will be trained on $\mathcal{D}_{\text{base}}$ then in second stage, k samples for each classes (even base classes) will be selected to create $\mathcal{D}_{\text{balance}}$ for slow learning rate fine-tuning the classifier and regressor of the detector. Being mentioned in section 2.2, GFSD, FSCE and DeFRCN are just variations of TFA but these changes in algorithm bring substantial increase in performance of the detector. For meta-learning direction, training a detector with a external re-weighting module in simulated episode scheme is the main idea among meta-learning few-shot object detection algorithms: Meta-DETR and FsDetView. All the networks were trained to localize pills using the stochastic gradient descent (SGD) optimizer, except for Meta-DETR which use Adam optimizer (49) since it involves in using Transformer (50). During the learning phase, the bounding box regression loss and region-level classification loss were usually jointly minimized, some meta-learning methods can add another meta-learning loss, commonly cross-entropy. To improve the generalization performance of the detectors while maintain the fairness evaluation between considering algorithms, a consistent set image augmentation operations for evaluated algorithms is used for augment the variations of the dataset.

5.3 Benchmarking results

In this section, we compare our proposed method Low-rank update for easy transferable few-shot learning, LowRank TFA, with the original method TFA (29) and other state-of-the-art few-shot object detection algorithms in different sets of categories: all classes, base classes and novel classes; different metrics: AP50:95, AP50, AP75 and different number of novel samples. By fine-tuning only small set of parameters, without doubting, TFA achieves almost the best results in base categories. However, FSCE (30), one of state-of-the-art methods, achieves substantial performance comparing with other methods in novel classes with 58.8% AP50:95 in setting of 15 shots but with the cost of low performance in base categories. Our TFA with unfreezing RoI's parameters got a decent balance point of performance between base and novel categories. LowRank TFA pushes this balance point further by achieving 59.7% AP50:95 in novel classes, the best among considering methods and 66.7% AP50:95, the closest to the original TFA. Surprisingly, the gap in novel categories performance between Low-rank TFA and others is more significant in lower shots setting (5 shots), LowRank

TFA got 45.6% AP50:95, 4.9% higher over the second highest method FSCE. With the smaller degradation in base knowledge, 1.4%, LowRank TFA achieves the best results in all categories even beating TFA. Section 5.4 suggests that this might be due to the improvements created by Axes Augmentation with Gaussian permutations. Notes that, with larger number of classes, the all categories performance mostly depends on the performance of base classes.

Table 3: Few-shot object detection performance of several methodologies, Fine-tuning with balanced set (TFA), Decoupled Faster R-CNN (DeFRCN), Faster-RCNN with Contrastive Proposal Encoding (FSCE), a fine-tuning baseline (Faster-RCNN+ft) and two meta-learning algorithms Meta-learning with DETR architecture (Meta-DETR), joint feature embedding module trained by episode training (FsDetView) for 5 and 15 samples per novel class. The best is in **bold** and *italic* respectively.

		All Categories			**Base Categories**			**Novel Categories**		
Shots	**Model**	**AP50:95**	**AP50**	**AP75**	**AP50:95**	**AP50**	**AP75**	**AP50:95**	**AP50**	**AP75**
5	TFA (29)	*0.471*	*0.745*	*0.601*	**0.497**	**0.764**	**0.636**	0.296	0.602	0.345
	TFA + RoI unfreezing **(ours)**	0.462	0.731	0.599	0.473	0.739	0.614	0.377	0.671	0.494
	DeFRCN (46)	0.371	0.653	0.48	0.374	0.65	0.484	0.354	0.672	0.445
	FSCE (30)	0.431	0.738	0.553	0.434	0.74	0.556	*0.407*	*0.72*	*0.528*
	Faster-RCNN + ft	0.305	0.526	0.412	0.307	0.523	0.416	0.289	0.549	0.382
	Meta-DETR (45)	0.355	0.529	0.436	0.376	0.552	0.449	0.322	0.494	0.399
	LowRank TFA **(ours)**	**0.478**	**0.756**	**0.608**	*0.483*	*0.756*	*0.625*	**0.456**	**0.752**	**0.546**
15	TFA (29)	**0.669**	**0.941**	**0.797**	**0.693**	**0.96**	**0.832**	0.492	0.798	0.541
	TFA + RoI unfreezing **(ours)**	0.635	0.917	0.768	0.642	0.921	0.78	0.583	0.888	0.688
	DeFRCN (46)	0.567	0.849	0.676	0.57	0.846	0.68	0.55	0.868	0.641
	FSCE (30)	0.627	0.934	0.749	0.624	0.936	0.752	*0.588*	*0.916*	*0.724*
	Faster-RCNN + ft	0.501	0.722	0.608	0.503	0.719	0.612	0.485	0.745	0.578
	Meta-DETR (45)	0.551	0.725	0.632	0.572	0.748	0.645	0.518	0.69	0.595
	LowRank TFA **(ours)**	*0.66*	*0.934*	*0.785*	*0.667*	*0.941*	*0.825*	**0.597**	**0.928**	**0.735**

5.4 Ablation Study

Module Analysis We study effectiveness of different modules in LowRank TFA by evaluating each module respectively on top original TFA. In table 4, ablation study shows that the Mutually Orthogonal Regularization plays a key role in mitigating the catastrophic forgetting effect by increase the bAP by 0.03. Meanwhile, both Low-rank Update and Axes Augmentation boost the adaption ability on novel set. Low-rank update increase 0.08 nAP over original TFA and Semantic Axes Augmentation further boost performance by 0.014 nAP. Notes that, original TFA only allow small set of parameters to be fine-tuned, therefore, the decrease in bAP is acceptable. Nevertheless, the Low-rank Update still has a strong effect in reducing catastrophic by increasing 0.025 bAP or 0.043 bAP (see table 3) over other much-parameters methodologies like TFA + RoI unfreezing or FSCE.

Low-rank Update ($r = 180$ and $r_{\text{update}} = 15$)	Orthogonal Regularization (Mutual Coherence)	Axes Augmentation (Gaussian)	bAP	nAP
			0.693	0.492
✓			0.665 **(-0.03)**	0.577 **(+0.08)**
✓	✓		0.649 **(-0.04)**	0.583 **(+0.09)**
✓	✓	✓	0.667 **(-0.03)**	0.597 **(+0.10)**
✓		✓	0.633 **(-0.06)**	0.584 **(+0.09)**

Table 4: Ablation study of different modules in LowRank TFA

Effect of different Decomposed and Update ranks We study empirically the effect of different values of decomposed ranks and update ranks on overall performance of LowRank TFA. In general, the value decomposed rank r is computed before training process by computing stable rank of weight matrices $r_{\text{stable}} = \frac{\|W_{\text{base},i}\|_F^2}{\|W_{\text{base},i}\|_\infty^2}$. The stable rank is a useful surrogate for the rank because it is largely unaffected by tiny singular values and cannot exceed the actual rank (51). By testing with different values of decomposed ranks and update ranks, we empirically show that LowRank-TFA is not sensitive to some extent with these hyper-parameters. A reasonable value for decomposed ranks can be chosen by stable rank analysis and update rank can be approximated by number of novel classes.

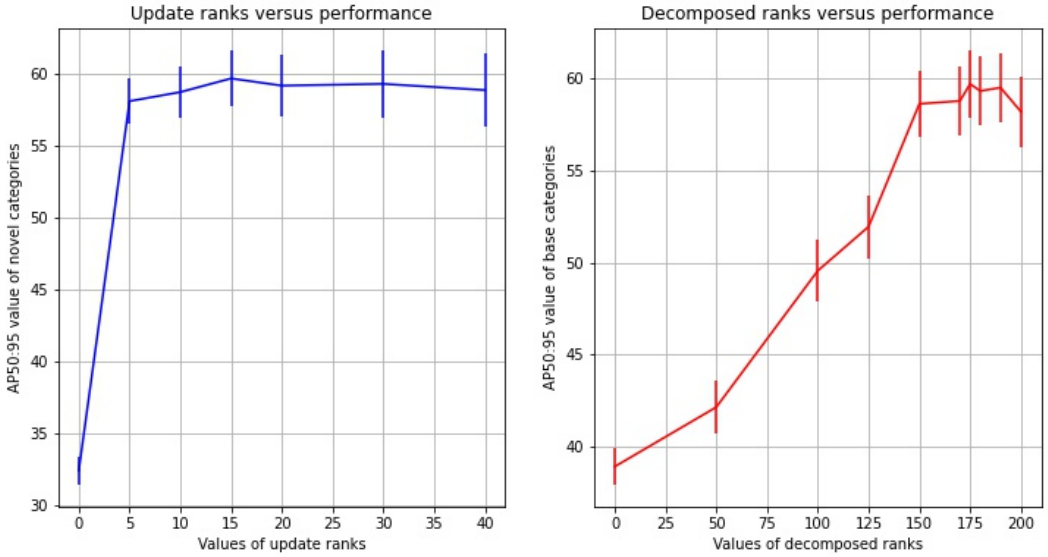

Figure 2: Sensitivity of different values of update ranks and decomposed ranks versus novel and base categories performance. The results are from 15 shots setting and the error is computed over 5 random seeds.

6 Conclusions

Acknowledgements This work was funded by Vingroup Joint Stock Company and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2021.DA00128. We thank XXX for valuable feedback on a draft of this paper.

References

[1] T. Bodenheimer and C. Sinsky, "From triple to quadruple aim: care of the patient requires care of the provider," *The Annals of Family Medicine*, vol. 12, no. 6, pp. 573–576, 2014.

[2] P. Aspden and P. Aspden, *Preventing medication errors*. National Acad. Press, 2007.

[3] M. A. Makary and M. Daniel, "Medical error—the third leading cause of death in the us," *Bmj*, vol. 353, 2016.

[4] A. L. Beam and I. S. Kohane, "Translating artificial intelligence into clinical care," *Jama*, vol. 316, no. 22, pp. 2368–2369, 2016.

[5] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[6] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 191, 2020.

[7] N. Larios Delgado, N. Usuyama, A. K. Hall, R. J. Hazen, M. Ma, S. Sahu, and J. Lundin, "Fast and accurate medication identification," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[8] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[9] J. J. Caban, A. Rosebrock, and T. S. Yoo, "Automatic identification of prescription drugs using shape distribution models," in *2012 19th IEEE International Conference on Image Processing*, 2012, pp. 1005–1008.

[10] Z. Chen and S. ichiro Kamata, "A new accurate pill recognition system using imprint information," in *Sixth International Conference on Machine Vision (ICMV 2013)*, B. Vuksanovic, J. Zhou, and A. Verikas, Eds., vol. 9067, International Society for Optics and Photonics. SPIE, 2013, pp. 199 – 203. [Online]. Available: https://doi.org/10.1117/12.2051168

[11] Z. Chen, J. Yu, S.-i. Kamata, and J. Yang, "Accurate system for automatic pill recognition using imprint information," *IET Image Processing*, vol. 9, 07 2015.

[12] Z. Yaniv, J. Faruque, S. Howe, K. Dunn, D. Sharlip, A. Bond, P. Perillan, O. Bodenreider, M. J. Ackerman, and T. S. Yoo, "The National Library of Medicine Pill Image Recognition Challenge: An Initial Report," *IEEE Appl Imag Pattern Recognit Workshop*, vol. 2016, Oct 2016.

[13] S. Ling, A. Pastor, J. Li, Z. Che, J. Wang, J. Kim, and P. Le Callet, "Few-shot pill recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9786–9795.

[14] Y. Wong, H. T. Ng, K. Leung, K. Chan, W. Chan, and C. C. Loy, "Development of fine-grained pill identification algorithm using deep convolutional network," *Journal of Biomedical Informatics*, vol. 74, 09 2017.

[15] Y. Wang, J. Ribera, C. Liu, S. Yarlagadda, and F. Zhu, "Pill recognition using minimal labeled data," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 2017, pp. 346–353.

[16] X. Zeng, K. Cao, and M. Zhang, "Mobiledeeppill: A small-footprint mobile deep learning system for recognizing unconstrained pill images," *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017.

[17] N. Usuyama, N. L. Delgado, A. K. Hall, and J. Lundin, "epillid dataset: A low-shot fine-grained benchmark for pill identification," *CoRR*, vol. abs/2005.14288, 2020. [Online]. Available: https://arxiv.org/abs/2005.14288

[18] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9576–9585, 2019.

[19] L. Zhang, S. Zhou, J. Guan, and J. Zhang, "Accurate few-shot object detection with support-query mutual guidance and hybrid loss," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 419–14 427.

[20] Y. Li, W. Feng, S. Lyu, Q. Zhao, and X. Li, "Mm-fsod: Meta and metric integrated few-shot object detection," *ArXiv*, vol. abs/2012.15159, 2020.

[21] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "One-shot object detection with co-attention and co-excitation," in *NeurIPS*, 2019.

[22] D.-J. Chen, H.-Y. Hsieh, and T.-L. Liu, "Adaptive image transformer for one-shot object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 242–12 251.

[23] G.-D. Zhang, Z. Luo, K. Cui, and S. Lu, "Meta-detr: Image-level few-shot object detection with inter-class correlation exploitation," 2021.

[24] J.-M. Pérez-Rúa, X. Zhu, T. M. Hospedales, and T. Xiang, "Incremental few-shot object detection," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 843–13 852, 2020.

[25] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, S. Pankanti, R. S. Feris, A. Kumar, R. Giryes, and A. M. Bronstein, "Repmet: Representative-based metric learning for classification and few-shot object detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5192–5201, 2019.

[26] Y. Yang, F. Wei, M. Shi, and G. Li, "Restoring negative information in few-shot object detection," *ArXiv*, vol. abs/2010.11714, 2020.

[27] S. Li, W. Song, S. Li, A. Hao, and H. Qin, "Meta-retinanet for few-shot object detection," in *BMVC*, 2020.

[28] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9924–9933.

[29] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," July 2020.

[30] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7348–7358, 2021.

[31] X. Chen, M. Jiang, and Q. Zhao, "Leveraging bottom-up and top-down attention for few-shot object detection," *ArXiv*, vol. abs/2007.12104, 2020.

[32] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8661–8670.

[33] "Two vinuniversity's funded projects by vinif in response to smart healthcare service and smart lighting technology," Feb 2022. [Online]. Available: https://vinuni.edu.vn/two-vinuniversitys-funded-projects-by-vinif-in-response-to-smart-healthcare-service-and-smart-lighting-technology/

[34] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases*, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 145–158.

[35] T. J. Bradberry, "Project title," https://github.com/trent-b/iterative-stratification, 2018.

[36] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," 2019. [Online]. Available: https://arxiv.org/abs/1902.00751

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[38] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," 2021. [Online]. Available: https://arxiv.org/abs/2110.04366

[39] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," 2020. [Online]. Available: https://arxiv.org/abs/2012.13255

[40] N. Mehta, K. J. Liang, V. K. Verma, and L. Carin, "Continual learning using a bayesian nonparametric dictionary of weight factors," 2020. [Online]. Available: https://arxiv.org/abs/2004.10098

[41] A. Chaudhry, N. Khan, P. K. Dokania, and P. H. S. Torr, "Continual learning in low-rank orthogonal subspaces," 2020. [Online]. Available: https://arxiv.org/abs/2010.11635

[42] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep cnns?" 2018. [Online]. Available: https://arxiv.org/abs/1810.09102

[43] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[44] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2009.

[45] G. Zhang, Z. Luo, K. Cui, and S. Lu, "Meta-detr: Few-shot object detection via unified image-level meta-learning," *CoRR*, vol. abs/2103.11731, 2021. [Online]. Available: https://arxiv.org/abs/2103.11731

[46] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8661–8670, 2021.

[47] Y. Xiao and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," in *ECCV*, 2020.

[48] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse r-cnn: End-to-end object detection with learnable proposals," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 449–14 458, 2021.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[50] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[51] N.Harvey, "Cpsc 536n lecture 15: Low-rank approximation of matrices," 2015. [Online]. Available: https://www.cs.ubc.ca/~nickhar/W12/Lecture15Notes.pdf