DEEP NUISANCE DISENTANGLEMENT FOR ROBUST OBJECT DETECTION FROM UNMANNED AERIAL VEHICLES

A Thesis

by

KARTHIK SURESH

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Zhangyang Wang
Committee Members,	Ulisses Braga-Neto
	Yoonsuck Choe
Head of Department,	Dilma Da Silva

May 2019

Major Subject: Computer Engineering

ABSTRACT*

Object detection from images captured by Unmanned Aerial Vehicles (UAVs) is becoming dramatically useful. Despite the great success of the generic object detection methods trained on ground-to-ground images, a huge performance drop is observed when these methods are directly applied to images captured by UAVs. The unsatisfactory performance is owing to many UAV-specific nuisances, such as varying flying altitudes, adverse weather conditions, dynamically changing viewing angles, etc., constituting a large number of fine-grained domains across which the detection model has to stay robust. Fortunately, UAVs record meta-data corresponding to the same varying attributes, which can either be freely available along with the UAV images, or easily obtained. We propose to utilize the free meta-data in conjunction with the associated UAV images to learn domain-robust features via an adversarial training framework. This model is dubbed Nuisance Disentangled Feature Transforms (NDFT), for the specific challenging problem of object detection in UAV images. It achieves a substantial gain in robustness to these nuisances. This work demonstrates the effectiveness of our proposed algorithm by showing both quantitative improvements on two existing UAV-based object detection benchmarks, as well as qualitative improvements on self-collected UAV imagery.

Reprinted with permission from the Abstract section of *Deep Nuisance Disentanglement for Robust Object Detection from Unmanned Aerial Vehicles* by Zhenyu Wu†, Karthik Suresh†, Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang, 2019, International Conference on Computer Vision (ICCV 2019) Proceedings (Under Review).

† indicates equal contribution

DEDICATION

To my parents, all my teachers, almighty God and everybody who has helped me reach where I am today.

ACKNOWLEDGMENTS

First and foremost, I would like to sincerely thank Dr. Zhangyang Wang for patiently guiding me throughout the thesis. I have grown as a student and a researcher under his tutelage.

I would like to express gratitude to my parents as they have always supported me. I would also like to thank my colleague Zhenyu Wu for his invaluable guidance and collaboration through which this work was possible. I would like to thank my committee members, Dr. Braga Neto and Dr. Yoonsuck Choe and for their guidance and support throughout my thesis work and coursework.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Zhangyang Wang [advisor] and Dr. Yoonsuck Choe of the Department of Computer Science and Engineering at Texas A&M University and Dr. Ulisses Braga Neto of the Electrical and Computer Engineering Department at Texas A&M University. This work is a result of my collaboration with Zhenyu Wu, my peer at VITA lab , Priya Narayanan and Heesung Kwon at the Army Research Lab, and Hongyu Xu, a PhD student at University of Maryland. I have made contributions such as developing the best model on the Visdrone dataset, cleaning the UAVDT dataset, meta-data labelling of the Visdrone dataset, transfer learning experiments from UAVDT to Visdrone dataset, etc. which has made me the joint first author on the resulting paper along with Zhenyu Wu.

Funding Sources

Graduate study was not supported via any funding sources.

NOMENCLATURE

CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
FPN	Feature Pyramid Network
UAV	Unmanned Aerial Vehicles
NDFT	Nuisance Disentangled Feature Transform
AP	Average Precision
mAP	Mean Average Precision

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION*

Object detection has been studied extensively over decades. Most of the promising detectors are able to detect objects of interest in clear images, such images are usually captured from ground-based cameras. With the rapid development of technology, Unmanned Aerial Vehicles (UAVs) equipped with cameras have been increasingly deployed in many industrial applications, opening up a new frontier of computer vision applications in security surveillance, peacekeeping, agriculture, deliveries, aerial photography, disaster assistance [1, 2, 3, 4], etc. One of the core features of the UAV-based applications is to detect object of interests (pedestrians or vehicles). While it is in high demand, object detection from UAVs is insufficiently investigated. In the meantime, the large mobility of UAV-mounted cameras bring in greater challenges than traditional object detection (using surveillance or other ground-based cameras). Some of the UAV-specific nuisances are enumerated below.

- **Variations in altitude and object scale**: The scales of objects captured in the image are closely affected by the flying altitude of the UAVs. For example, an image captured by a DJI Inspire 2 series flying at 500 meters altitude [5] will contain very small objects, which are very challenging to detect and track as compared to an image captured at an altitude of say 50 meters. In addition, a UAV can be operated in a variety of altitudes while capturing images. When shooting in lower altitudes, its camera can capture more details of objects of interest. When it flies to higher altitudes, the camera can inspect a larger area and more objects will be captured in the image. As a consequence, the same object can vary a lot in terms of scale throughout the captured video, with different flying altitudes during a single flight.

- **Variations in view angle**: The mobility of UAVs leads to video shots from different and free

Reprinted with permission from the Abstract section of *Deep Nuisance Disentanglement for Robust Object Detection from Unmanned Aerial Vehicles* by Zhenyu Wu[†], Karthik Suresh[†], Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang, 2019, International Conference on Computer Vision (ICCV 2019) Proceedings (Under Review).

[†] indicates equal contribution

angles, in addition to the varying altitudes. For example, a UAV can look at one object from the front view, side view, bird view, or a combination of the above views in a single video. The diverse view angles cause arbitrary orientations and aspect ratios of the objects. Some view angles such as bird view seldom occur in traditional ground-based object detection. As a result, the UAV-based detection model has to deal with more different visual appearances of the same object. Note that more view angles can be presented when altitudes grow higher. Also, wider view angles often lead to denser objects in the view.

- **Variations in weather and illumination**: A UAV operated in uncontrolled outdoor environments is very likely to fly under various weather and lighting conditions. The changes in illumination (e.g. daytime and nighttime), as well as weather changes (e.g. sunny, cloudy, foggy or rainy), will drastically affect the object visibility and appearance. Videos shot in daylight introduce interference of shadows and videos captured in the night containing dim street lamps are prone to missing important feature information. In the meantime, frames captured during foggy weather lack sharp details resulting in the contours of the objects vanishing in the background.

Most off-the-shelf detectors are trained with usually less varied, more restricted-view data. In comparison, the abundance of **UAV-specific nuisances** enumerated above will cause the resulting UAV-based detection model to operate in a large number of different **fine-grained domains**. Here a domain could be interpreted as a specific combination of nuisances, for example, the images taken at low-altitude and daytime, and those taken at high-altitude and nighttime domain, constitute two different domains. Therefore, our goal is to train a *cross-domain object detection* model that stays robust to the massive number of fine-grained domains. Existing potential solutions include data augmentation [6], domain adaption [7, 8], and ensemble of expert models [9]. However, none of these approaches are easy to generalize to multiple and/or unseen domains [7, 8], and they could lead to over-parameterized models which are not suitable for the UAV on-board deployments [6, 9].

A (Almost) Free Lunch: Fine-Grained Nuisance Annotations: Motivated by the previous discussion, we cast the UAV-based object detection problem as a cross-domain object detection problem with multiple fine-grained domains. The above UAV-specific nuisances constitute the domain-specific nuisances that should be eliminated for transferable feature learning. However, the features of the objects of interest in the image must be preserved across all transformations and nuisance elimination for robust detection. For UAVs, the major nuisance types are known to be altitude, angle and weather. More importantly in the specific case of UAVs, the nuisance annotations can easily be obtained or can even be freely available. For example, a UAV can record its flying altitudes as metadata by GPS, or more accurately by a barometric sensor. Taking weather as another example, since it is easy to retrieve each UAV flight's time-stamp and spatial location (or path), one can easily obtain the weather information of the specific time/location.

Motivated by the above facts, we propose to learn an object detection model that maintains its effectiveness in extracting task-related features while eliminating the known types of nuisances across different domains (altitudes/angles/weather). We take advantage of the easy (or free) access to the nuisance annotations. We are the first to adopt an adversarial learning framework to learn task-specific, domain-robust features by explicitly disentangling task-specific features from nuisance features in a supervised way. The framework, dubbed *Nuisance Disentangled Feature Transforms* (**NDFT**) gives rise to UAV-based object detection models that can be directly applicable not only to domains seen in the training data, but also to unseen domains without needing any extra effort of domain adaptation or sampling/labeling. Experiments on real UAV datasets demonstrate its effectiveness and robustness.

2. RELATED WORK*

2.1 Object Detection

Object detection has progressed tremendously thanks to the extensive study by the academia and the emergence of benchmarks (*i.e.* MS COCO [10] and PASCAL VOC [11]). There are primarily two main-stream approaches: two-stage detectors and single-stage detectors, based on whether the detectors have a proposal-driven mechanism or not. Two stage detectors [12, 13, 14, 15, 16, 17], which contain a region proposal network (RPN), first generate regions within an image which is likely to contain an object. The detector then localizes and classifies each of these proposed regions into different categories. On the other hand, single-stage detectors [18, 19, 20, 21] apply dense sampling windows over object locations and scales and have no region proposal stage. The single-stage detectors usually achieved high speed by directly exploiting multiple layers in a deep CNN networks while the detection accuracy is compromised compared to two-stage detectors.

Aerial Image-based Object Detection: Despite a few aerial image datasets (*i.e.* DOTA [22], NWPU VHR-10 [23], and VEDAI [24]) being proposed, the common practice to detect objects from aerial images is only to deploy off-the-shelf ground-based object detection models [25]. Moreover, these datasets only contain geo-spatial images (e.g., satellite) with bird-view small objects, which are not as diverse as UAV-captured images with greatly varied altitudes, poses and weather conditions. Publicly available benchmark datasets were not available for UAV-based object detection until recently. Two datasets, UAVDT [26] and VisDrone2018 [27], were released to address this issue. Specifically, UAVDT consists of 100 video sequences (about 80k frames) captured from UAVs under complex scenarios. Moreover, it also provides full annotations for weather

*Reprinted with permission from the submitted paper, *Deep Nuisance Disentanglement for Robust Object Detection from Unmanned Aerial Vehicles* by Zhenyu Wu†, Karthik Suresh†, Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang, 2019, International Conference on Computer Vision (ICCV 2019) Proceedings (Under Review).

† indicates equal contribution

conditions, flying altitudes, and camera views in addition to the ground truth bounding box of the target objects. VisDrone2018 [27] is a large-scale UAV-based object detection and tracking benchmark, which composed of 10,209 static images and 179,264 frames from 263 video clips. In this work, these two benchmark datasets for UAV-based detection will be adopted as our test beds.

Detecting Tiny Objects: A typical ad hoc approach to detect tiny objects is through learning representations of all the objects at multiple scales. This approach is however highly inefficient with limited performance gains. [28] proposed a super-resolution algorithm using coupled dictionary learning to transfer the target region into a high resolution to "augment" its visual appearance. [29] proposed to internally super-resolve the feature maps of small objects to make them resemble similar characteristics as large objects. SNIP [30] showed that CNNs were not naturally robust to the variations in object scales. It proposed to train and test detectors on the same scales of an image pyramid, and selectively back-propagate the gradients of object instances of different sizes as a function of the image scale during the training stage. SNIPER [31] proposed to process context regions around ground-truth instances (chips) at different appropriate scales and further to efficiently train the detector at multi-scales, which improved the detection performance on tiny objects.

2.2 Handling Domain Variances

Domain Adaptation via Adversarial Training: Adversarial domain adaptation [32] was proposed to reduce the domain gap by learning with only labeled data from a source domain plus massive unlabeled data from a target domain. This approach has recently gained increased attention in the detection field too. [33] learned robust detection models to occlusion and deformations, through hard positive examples generated by an adversarial network. [8] improved the cross-domain robustness of object detection by enforcing adversarial domain adaption on both image and instance levels. [34] introduced a Siamese-GAN to learn invariant feature representations for both labeled and unlabeled aerial images coming from two different domains. CyCADA [35] unified cycle-consistency with adversarial loss to learn domain invariance. However, these domain adaption methods typically assume one (ideal) source domain and one (non-ideal) target domain. The possibility of generalizing these methodologies to handling multiple fine-grained domains is questionable. Once a new unseen domain emerges, domain adaptation needs explicit re-training. In comparison, our proposed framework does not assume any ideal reference (source) domain, but rather tries to extract robust features shared by many different "non-ideal" target domains. The setting thus differs from typical domain adaptation and generalizes naturally to task-specific feature extraction in unseen domains.

Data Augmentation, and Model Ensemble: Compared to the considerable amount of research in data augmentation for classification [32], much less attention has been paid on other tasks such as detection. Classical data augmentation relies on a limited set of known factors (such as scaling, rotation, flipping) that are easy to invoke, and adopt ad hoc. The idea is to gain robustness to minor variations such as scaling, rotating, or flipping which does not change the label of an input. However, UAV images involve a much larger variety of nuisances, many of which are hard to "synthesize". A recent work [6] proposed novel learning-based approaches to synthesize new training samples for detection. But it focused on recombining foreground objects and background contexts, rather than recomposing specific nuisance attributes. Also, the (much) larger augmented dataset adds to training burden and may cause over-parameterized models.

Another methodology was proposed in [9] to better capture the appearance variations caused by different shapes, poses, and viewing angles. The authors proposed a Multi-Expert R-CNN consisting of three experts, each responsible for objects with a particular shape: horizontally elongated, square-like, and vertically elongated. This approach is apparently limited, since the model ensemble quickly becomes computationally intractable as many different domains are involved. It further cannot generalize to unknown or unseen domains.

Feature Disentanglement: Feature disentanglement [36] leads to non-overlapped groups of factorized latent representations, each of which would properly describe corresponding information to particular attributes of interest. It has mostly been applied to generative models [37, 38], and lately to reinforcement learning [39]. In the image-to-image translation literature, a recent work [40] also disentangled image representations into shared parts for both domains and exclusive parts for either domain. NDFT extends the idea of feature disentanglement to learning cross-domain robust discriminative models.

3. FORMULATION OF NDFT*

Our proposed UAV-based cross-domain object detector can be characterized as an adversarial training framework. Assume our training data X is associated with an **O**bject detection task $\mathcal{O}$, and a UAV-specific **N**uisance prediction task $\mathcal{N}$. We mathematically express the goal of cross-domain object detection as alternatively optimizing two objectives as follows (γ is a weight coefficient):

$$\begin{aligned} &\min_{f_O, f_T} L_O(f_O(f_T(X)), Y_O) - \gamma L_N(f_N(f_T(X)), Y_N), \\ &\min_{f_N} L_N(f_N(f_T(X)), Y_N) \end{aligned} \tag{3.1}$$

In (3.1), f_O denotes the model that performs the object detection task $\mathcal{O}$ on its input data. The label set Y_O are object bounding box coordinates and class labels provided on X. L_O is a cost function defined to evaluate the object detection performance on $\mathcal{O}$. On the other hand, the labels of the UAV-specific nuisances Y_N come from metadata along with X (e.g., flying altitude, camera view or weather condition), and a standard cost function L_N (e.g., softmax) is defined to evaluate the task performance on $\mathcal{N}$. Here we formulate nuisance robustness as the suppression of the nuisance prediction accuracy from the learned features.

We seek a *Nuisance Disentangled Feature Transform* (**NDFT**) f_T by solving (3.1), such that

- The object detection task performance L_O is minimally affected when $f_T(X)$ is used instead of X.
- The nuisance prediction task performance L_N is maximally suppressed over $f_T(X)$, compared to that of X.

In order to deal with the multiple nuisances case, we extend the (3.1) to multiple prediction tasks.

Reprinted with permission from the submitted paper, *Deep Nuisance Disentanglement for Robust Object Detection from Unmanned Aerial Vehicles* by Zhenyu Wu†, Karthik Suresh†, Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang, 2019, International Conference on Computer Vision (ICCV 2019) Proceedings (Under Review).

† indicates equal contribution

Here we assume k nuisance prediction tasks associated with label sets $Y_N^1, ..., Y_N^k$. $\gamma_1, ..., \gamma_k$ are the respective weight coefficients. The modified objective naturally becomes:

$$\begin{aligned} &\min_{f_O, f_T} L_O(f_O(f_T(X)), Y_O) - \sum_{i=1}^{k} \gamma_i L_N(f_N^i(f_T(X)), Y_N^i), \\ &\min_{f_N^1, ..., f_N^k} L_N(f_N^i(f_T(X)), Y_N^i) \end{aligned} \tag{3.2}$$

All modules, f_T, f_O and f_N^is, participate in training and can be implemented by neural networks. The object detection models (f_O) used in this work are briefly described below.

- **Faster-RCNN**: Faster-RCNN [15] is a two-stage detector which is a standard benchmark in the object detection community. It is a descendant of the RCNN family of object detectors which have traditionally used a region proposal algorithm (such as Selective Search) to computer probable object locations in an image. These probable object locations called *regions of interest* are then passed on to a CNN to extract features, localize objects and classify the objects in the image to different classes. Faster-RCNN came up with a novel *Region Proposal Network*(RPN) which is nothing but a small object detector in itself to propose regions of interest by classifying certain predefined boxes of varying sizes and aspect ratios into two classes: object or no object. This RPN can be trained by backpropagation similar to the main object detector using the classification and the localization losses. The ROI pooling layer converts various different-sized regions of interest detected by the RPN into a fixed size so that it can be further fed into the fully connected layers.

- **Feature Pyramid Network**: Feature Pyramid Networks [41] (FPN) is a feature extractor designed to take advantage of the pyramidal shape of the convolutional feature maps to extract better features and predict objects from each level of the feature pyramid. FPNs have a bottom-up pathway which is similar to any feature extractor used in detection like VGG-16 or Resnet. The top-down pathway uses upsampling and element-wise addition to combine two adjacent feature maps. The features extracted this way combine high resolution feature maps in the beginning which has valuable low level information about small objects and the low resolution feature maps which have a very high semantic content. This drastically increases the performance of the detector on small objects [41].

4. IMPLEMENTATION AND TRAINING*

4.1 Architecture Overview of NDFT-Faster-RCNN

As an instance of the general NDFT framework (4.1), Figure 4.1 displays an implementation example of NDFT using the Faster-RCNN backbone [15], while later we will demonstrate that NDFT can be plug-and-play with more sophisticated object detection networks (e.g., FPN).
During training, the input data X first goes through the NDFT module f_T, and its output $f_T(X)$ is passed through two subsequent branches simultaneously. The upper object detection branch f_O, uses $f_T(X)$ to detect objects, while the lower nuisance prediction model f_N predicts nuisance labels from the same $f_T(X)$. Finally, the network minimizes the prediction penalty (error rate) for f_T, while maximizing the prediction penalty for f_N, as shown in (4.1).
By jointly training T, f_T and f_N^is in the above adversarial setting, the NDFT module will find the optimal transform that preserves the object detection related features while removing the UAV-specific nuisances prediction related features, fulfilling the goal of cross-domain object detection that is robust to the UAV-specific nuisances.
Choices of f_T, f_O and f_N: In this NDFT-Faster-RCNN example, f_T includes the conv1_x, conv2_x, conv3_x and conv4_x of the ResNet101 part of Faster-RCNN (convk_x here indicates the feature map output produced by the k$^{\text{th}}$ convolutional layer of the backbone network). f_O includes the conv5_x layer, attached with a classification and regression loss for detection. We further implement f_N using the same architecture as f_O (except the number of classes for prediction). The output of f_T is fed to f_O after going through the RoIAlign [42] layer, whereas it is fed to f_N after going through a spatial pyramid pooling layer [13].
Choices of L_O and L_N: In object detection, L_O is the bounding box classification (e.g., softmax)

Reprinted with permission from the submitted paper, *Deep Nuisance Disentanglement for Robust Object Detection from Unmanned Aerial Vehicles* by Zhenyu Wu†, Karthik Suresh†, Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang, 2019, International Conference on Computer Vision (ICCV 2019) Proceedings (Under Review).

$\dagger$ indicates equal contribution

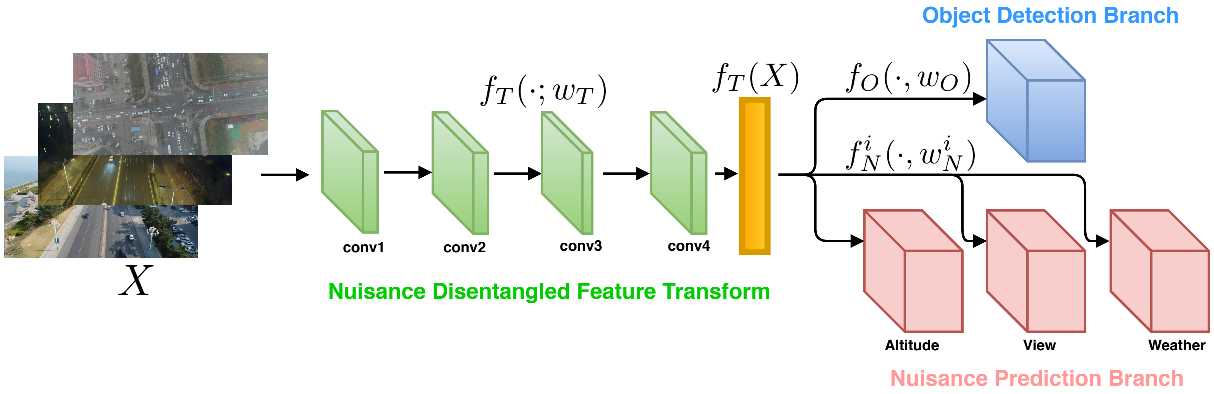


Figure 4.1: Our proposed NDFT-Faster-RCNN network.

and regression loss (e.g., smooth ℓ_1) as widely used in traditional two stage detectors. However, using $-L_N$ as the adversarial loss in the first row of (4.1) is not straightforward. If we choose L_N as some typical classification loss such as the softmax, then maximizing it directly is prone to gradient explosion. After experimenting with several solutions such as the standard gradient reversal trick [32], we decide to follow the suggestion of [43] to choose the negative entropy function of the predicted class vector as the adversarial loss, denoted as L_{ne}. Minimizing L_{ne} will encourage the model to make "uncertain" predictions (equivalently, as good as a uniform random guess) on the nuisances.

Since we replace L_N with L_{ne} in the first objective in (4.1), it no longer needs Y_N. Meanwhile, the usage of L_N and Y_N remains unaffected in the second objective of (4.1). L_N and Y_N are used to keep f_N^is as "sufficiently strong adversaries" throughout training, in order to learn meaningful NDFT f_T that can generalize better. The final NDFT framework alternatively optimizes the following two objectives:

$$\begin{aligned}&\min_{f_O,f_T} L_O(f_O(f_T(X)),Y_O)+\sum_{i=1}^{k}\gamma_i L_{ne}(f_N^i(f_T(X))),\\&\min_{f_N^1,\dots,f_N^k} L_N(f_N^i(f_T(X)),Y_N^i)\end{aligned}\tag{4.1}$$

We also use Y_N to pre-train sufficiently strong f_N^is at the initialization.

4.2 Training Strategy

Just like training GANs [44], our training is prone to collapse and/or reach bad local minima. We thus presented a carefully-designed training algorithm with three-module alternating update strategy , which could be interpreted as a three-party game. The training procedure is summarized in Algorithm 1, and is explained below.

Algorithm 1 Learning Nuisance Disentangled Feature Transforms in UAV-based Object Detection via Adversarial Training

```
Given pre-trained NDFT module $f_T$, object detection task module $f_O$, and nuisances prediction module $f_N^i$s
for number of training iterations do
    Sample a mini-batch of n examples $\{X^1, \cdots, X^n\}$
    Update NDFT module $f_T$ (weights $w_T$) and object detection module $f_O$ (weights $w_O$) with stochastic gradients:
    $\nabla_{w_T \cup w_O} \frac{1}{n} \sum_{j=1}^{n} \Big[ L_O(f_O(f_T(X^j)), Y_O^j) + \sum_{i=1}^{k} \gamma_i L_{ne}(f_N^i(f_T(X^j))) \Big]$
    while at least one nuisance prediction task has training accuracy $\leq 0.9$ do
                                        ▷ Avoid $f_N^i$s becoming too weak competitors.
        Update nuisance prediction modules $f_N^i, \dots, f_N^k$ (weights $w_N^1, \dots, w_N^k$) with stochastic gradients:
        $\nabla_{w_N^i} \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} L_N(f_N^i(f_T(X^j)), Y_N^j)$
    Restart $f_N^i, \dots, f_N^k$ every 1000 iterations, and repeat Algorithm 1 from the beginning. ▷ An empirical trick to alleviate overfitting.
```

For each mini-batch, we first jointly optimize f_T and f_O weights (with f_N^is frozen), by minimizing the first objective in (4.1) using the standard stochastic gradient descent (SGD). Meanwhile, we keep "monitering" f_N^i branches as f_T is updated, such that if at least one of the f_N^i becomes too weak (i.e., showing poor predicting accuracy on the same mini-batch), another update will be triggered by minimizing the second objective in (4.1) using SGD. The goal is to "strengthen" the nuisance prediction competitors. Besides, we also discover an empirical trick, by periodically re-setting the current weights of $f_N^1, ..., f_N^k$ to random initialization, and then retraining them on

$f_T(X)$ (with f_T fixed) to become strong nuisance predictors again. We again restart the above alternative training process of f_T, f_O and f_N^is. This restarting trick is also found to benefit the generalization of the learned f_T, potentially due to helping the learning process get out of some trivial local minima.

5. EXPERIMENTAL RESULTS*

Since public UAV-based object detection datasets (in particular those with nuisance annotations) are currently of very limited availability, we design **three sets of experiments** to validate the effectiveness, robustness, and generality of NDFT. First, we perform the main body of experiments on the **UAVDT** benchmark [26], which provides all three UAV-specific nuisance annotations (altitude, weather, and view angle). We demonstrate the clear observation that the more variations are disentangled via NDFT, the larger AP improvement we will gain on UAVDT; and eventually we achieve the state-of-the-art performance on UAVDT.

We then move to the other publicly available UAV-based detection benchmark, **VisDrone2018**. Since nuisance annotations are not available on it, we utilize it as a *transfer learning* testbed. f_T shows strong transferability, and actually outperforms the best single-model method currently reported on the VisDrone2018 leaderboard.

Lastly, we present the detection examples where the NDFT model directly learned on UAVDT, is adapted to our self-collected UAV imagery using a DJI Phantom 4 Pro, showing its remarkable generalization ability to real-world UAV videos.

5.1 Improving UAV-based Object Detection on UAVDT: Results and Ablation Study

Problem Setting: The image object detection track on UAVDT consists of around 41k frames with 840k bounding boxes. It has three categories: car, truck and bus, but the class distribution is highly imbalanced (the latter two occupy less than 5% of bounding boxes). Hence, following the convention of the authors in [26], we combine the three into one *vehicle* class and report AP based on that. All frames are also annotated with three categories of UAV-specific nuisances: flying altitude (*low, medium and high*), camera views (*front-view, side-view and bird-view*), and

Reprinted with permission from the submitted paper, *Deep Nuisance Disentanglement for Robust Object Detection from Unmanned Aerial Vehicles* by Zhenyu Wu†, Karthik Suresh†, Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang, 2019, International Conference on Computer Vision (ICCV 2019) Proceedings (Under Review).

† indicates equal contribution

weather condition (*daylight*[1], *night*). We will denote the three nuisances as **A**, **V**, and **W** for short, respectively.

Implementation Details: We first did our best due diligence to improve the baseline (without considering nuisance handling) performance on UAVDT, to ensure a solid enough ground for NDFT. The authors reported a AP of 20 using a Faster-RCNN model with the VGG-16 backbone. We replace the backbone into ResNet-101, and fine-tune more hyperparameters such as anchor scale (16,32,64,128,256). We end up with an improved AP of 44.04 (using the same IoU threshold = 0.7 as the authors) as our baseline performance. We also communicated with the authors of [26] in person and they acknowledged this improved baseline. We then implement NDFT-Faster-RCNN using the architecture depicted in Figure 4.1. We denote γ_1, γ_2 and γ_3 as the coefficients in (5.1), for the L_{ne} loss terms for altitude, view and weather nuisances, respectively.

Table 5.1: Learning NDFT-Faster-RCNN on altitude nuisance only, with different γ_1 values on the UAVDT dataset.

A γ_1	Low	Med	High	Overall
0.0	63.04	48.62	13.88	44.04
0.01	**69.01**	50.46	14.63	**45.31**
0.02	66.97	46.91	16.69	44.17
0.03	66.38	**53.00**	15.69	45.92
0.05	61.14	49.71	**18.70**	44.64

Results and Analysis: We unfold our full ablation study on the UAVDT dataset in a progressive way. We first study the impact of removing each individual nuisance type (A, V, and W and then gradually proceed to removing two and three nuisance types simultaneously, and show the resulting consistent performance gains.

Tables 5.1, 5.2, and 5.3 show the the benefit of removing flying altitude (A), camera view (V) and weather condition (W) nuisances one at a time. That could be viewed as learning NDFT-Faster-

[1] We discard another "foggy" class because its overly small number of samples; and those sample are mostly taken in nighttime.

Table 5.2: Learning NDFT-Faster-RCNN on view angle nuisance only, with different γ_2 values on the UAVDT dataset.

V γ_2	Front	Side	Bird	Overall
0.0	44.36	66.94	25.03	44.04
0.01	57.45	67.61	25.60	**46.16**
0.02	61.49	66.85	24.93	45.73
0.03	54.55	**68.22**	23.07	45.42
0.04	**65.82**	65.23	**26.77**	45.14

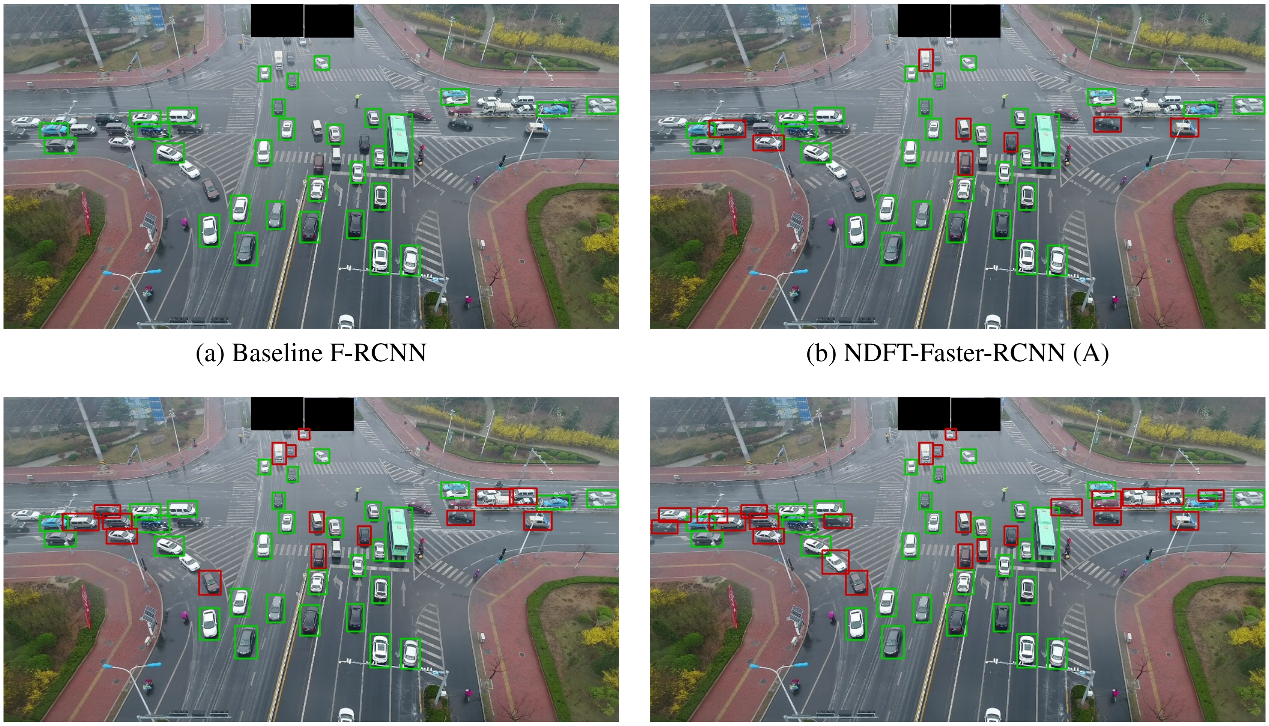
Table 5.3: Learning NDFT-Faster-RCNN on weather nuisance only, with different γ_3 values

W γ_3	Day	Night	Overall
0.0	43.56	46.39	44.04
0.01	45.18	**59.66**	**46.62**
0.025	43.72	57.41	44.43
0.05	**45.61**	56.14	46.03
0.1	44.28	48.78	43.60

CNN (Figure 4.1) with only the corresponding one γ_i (i = 1, 2, 3) to be nonzero. The baseline model without nuisance disentanglement could be viewed as γ_i = 0, i = 1, 2, 3.

As can be seen from Table 5.1, compared to the baseline (γ_1 = 0), increasing γ_1 gradually enables and enforces the disentanglement of A, which leads to a consistent superiority over the baseline. As Table 5.1 shows, under all $\gamma_1 > 0$ values tested, the overall AP always improves, and so are most single-class APs. The peak AP gain is obtained at $\gamma_1 = 0.01$, where we achieve a AP improvement of 1.88. The low, medium and high-altitude class APs increase by 3.34, 4.38 and 1.81 over the baseline case respectively.

Table 5.2 shows the performance gain by removing the camera view (V) nuisance. At $\gamma_2 = 0.01$, a large overall AP improvement of 2.12 is obtained. On the front-view frames, the AP increases by 8.09. Similar positive observations are found in Table 5.3 as well: $\gamma_3 = 0.01$ results in an overall AP boost of 2.58 over the baseline, with daylight and (the more challenging) night class



(a) Baseline F-RCNN

(b) NDFT-Faster-RCNN (A)

(c) NDFT-Faster-RCNN (A+V)

(d) NDFT-Faster-RCNN(A+V+W)

Figure 5.1: Example showing the benefit of the proposed NDFT framework for object detection on the UAVDT dataset

APs increased by 1.62 and 12.74, respectively.

Table 5.4 shows the full results by incrementally adding more adversarial losses simultaneously into training. For example, $A + V + W$ stands for simultaneously disentangling flying altitude, camera view and weather nuisances. When using two or three losses, unless otherwise stated, we apply γ_i = 0.01 for both/all of them, as discovered to give the best single-nuisance results in Tables 5.1 - 5.3.

As a consistent observation throughout the table, the more nuisances removed through NDFT, the better AP values we obtain (e.g., $A+V$ outperforms any of the three single models, and $A+V+W$ further achieves the best AP among all). To conclude on UAVDT, removing nuisances using NDFT evidently addresses the tough problem of object detection on high-mobility UAV platforms. Furthermore, taking a closer look the final best-performer $A+V+W$, we are encouraged to discover it improves the class-wise APs with noticeable margins on some most challenging nuisance classes, such as high-altitude, bird-view and nighttime. These observations can be clearly observed looking at Figure 5.1. Going from Figure 5.1(a), the result for the Faster-RCNN [15] baseline, to gradually (b) disentangling the nuisances of altitude (A), to (c) disentangling the nuisances of both altitude (A) and view angles (V), to (d) disentangling all the nuisances of altitude (A), view angles (V), and weather (W), the detection performance gradually improves from (a) to (d) with disentanglement on all the nuisances (red rectangular boxes denote new correct detections beyond the baseline).Improving object detection in those cases can be significant for deploying camera-mounted UAVs to uncontrolled, potentially adverse visual environments with better reliability and robustness.

Adapting Stronger Backbones-FPN: Besides, we observe the performance gain by NDFT does not vanish as we adopt more sophisticated backbones, e.g. FPN [41]. Training FPN on UAVDT leads to the baseline performance improved from 44.04 to 49.05. By plugging FPN into the proposed NDFT-Faster-RCNN training pipeline, the resulting model learns to simultaneously disentangle $A + V + W$ nuisances (γ_i = 0.005, i = 1,2,3). We are able to further increase the overall AP to 52.03, showing the general benefit of NDFT regardless of the backbone choices.

Table 5.4: UAVDT NDFT-Faster-RCNN with multiple attribute disentanglement.

	Baseline	A	V	W	A+V	A+V+W
	Flying Altitude					
Low	63.04	66.38	71.09	**75.32**	66.05	70.33
Med	48.62	53.00	52.29	51.59	54.07	**54.10**
High	13.88	15.69	16.62	16.08	18.60	**18.87**
	Camera View					
Front	49.36	53.90	57.45	**62.36**	61.23	56.88
Side	65.29	67.41	67.61	68.47	**68.82**	68.18
Bird	24.03	24.56	25.60	23.97	24.43	**28.80**
	Weather Condition					
Day	44.37	**47.32**	45.30	45.18	46.26	45.94
Night	46.92	45.82	56.70	59.66	59.16	**60.29**
Overall	44.04	45.92	46.16	46.62	46.88	**47.07**

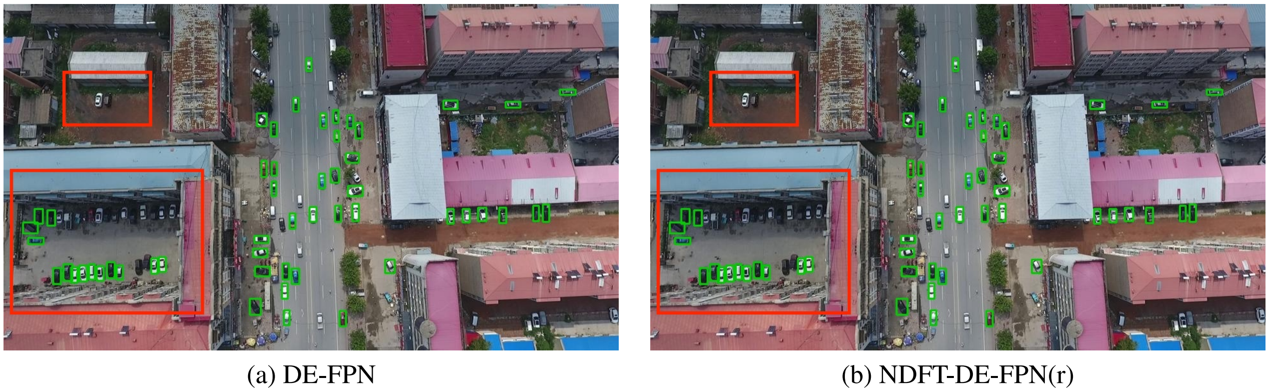
5.2 Transfer Learning to VisDrone2018

Problem Setting: The image object detection track on VisDrone2018 provides a dataset of 10,209 images, with 10 categories of pedestrians, vehicles and other traffic objects annotated. Unfortunately, it does not provide nuisance annotations, so directly training NDFT here is not feasible. However, we use it as a testbed to showcase the superior transferablity of NDFT features. According to the leaderboard [45] and the workshop technical report [46], excluding top-performer ensemble models, the best-performing single-model approach is DE-FPN, which utilized FPN (removing P6). We hence choose DE-FPN as the comparison subject here and implemented a few more tricks to ensure that our model is as good as the best single model baseline.

Implementation Details: We implement DE-FPN by identically following their method description in [46]. It is trained on VisDrone 2018 training set and tested on the vehicle category of validation set (since the testing set is not publicly accessible). We then train the same DE-FPN backbone on UAVDT with three nuisances (A+V+W) disentangled, with $\gamma_1 = \gamma_2 = \gamma_3 = 0.005$. The learned f_T is then transferred to VisDrone2018, by only re-training the classification/regression layer while keep other featured extraction layers all fixed. In that way, we focus on assessing the learned feature transferablity using NDFT. Besides, we repeat the same above routine with $\gamma_1 = \gamma_2 = \gamma_3 = 0$, to create a transferred baseline model without nuisance disentanglement. We denote the two transferred models as NDFT-DE-FPN(r) and DE-FPN(r), respectively. Since vehicle is the only shared category between UAVDT and VisDrone2018, we compare average precision on the vehicle class only to ensure a fair transfer setting. The performance of DE-FPN, NDFT-DE-FPN(r) and DE-FPN(r) are compared on the VisDrone 2018 validation set.

Table 5.5: Comparison on the VisDrone2018 validation set.

DE-FPN	DE-FPN(r)	NDFT-DE-FPN(r)
76.80	75.27	79.50


(a) DE-FPN

(b) NDFT-DE-FPN(r)

Figure 5.2: Benefit of the NDFT approach on the VisDrone2018 dataset.

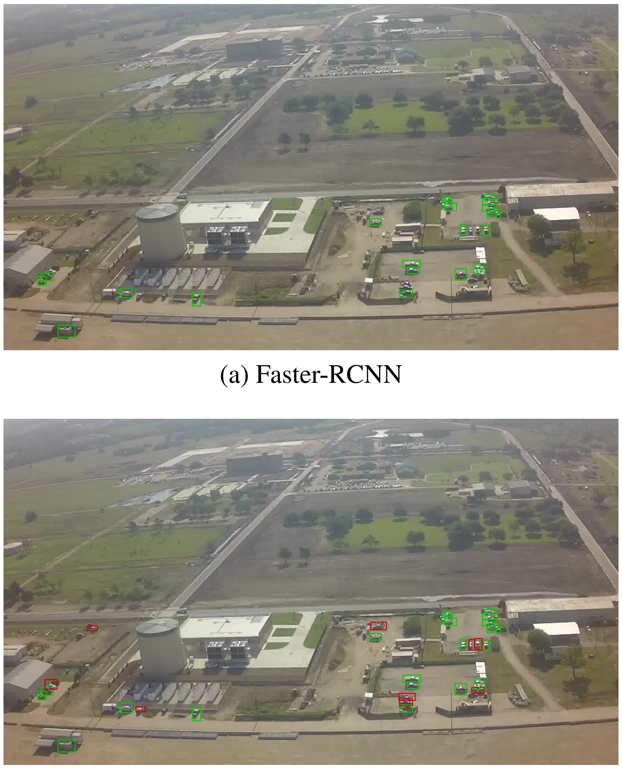
The qualitative results for the performance of NDFT on Visdrone dataset can been seen in Figure 5.2(a) and 5.2(b).

Green boxes are the correct boxes predicted by both models. Red boxes highlight the local regions where NDFT-DE-FPN(r) is able to detect substantially more vehicles than DE-FPN (the state-of-the-art single-model method on VisDrone2018).

Results and Analysis As observed in Table 5.5. directly transferring DE-FPN from UAVDT to VisDrone2018 (with fine-tuning on the latter) does not give rise to competitive performance, showing a substantial domain mismatch between the two datasets (even we only compare on their shared object category). However, transferring the learned NDFT to VisDrone2018 leads to a much boosted result, with a 4.23 AP margin over the transfer baseline without nuisance disentanglement, and a 2.70 margin over DE-FPN. It demonstrates that NDFT accounts for eliminating domain nuisances that potentially hurt transfer, and provides a powerful tool for cross-domain object detection.

5.3 Qualitative Results on In-House UAV Videos

We finally test our algorithm on real-world UAV video captured by our own platform (a DJI Phantom 4 Pro), in uncontrolled outdoor environments. Those flights were for agriculture planning proposes, and vehicle is one of the main categories of interest. We directly apply the NDFT-Faster-RCNN trained on UAVDT, with default $\gamma_1 = \gamma_2 = \gamma_3 = 0.1$, without any tuning or do-



(a) Faster-RCNN

(b) NDFT-Faster-RCNN

Figure 5.3: Examples showing the benefit of learning NDFT for object detection on our self-collected UAV video.

main adaption. We also compare it with the baseline UAVDT model without disentanglement ($\gamma_1 = \gamma_2 = \gamma_3 = 0$). Figure 5.3 clearly manifests that NDFT-Faster-RCNN picks up more challenging objects, e.g., small vehicles from non-front view angles.

6. CONCLUSION AND IMPACT*

This work investigates object detection from UAV-mounted cameras, a very useful and understudied problem. The problem appears to be more challenging than standard object detection, due to many UAV-specific nuisances, such as varying flying altitudes, adverse weather conditions, and dynamically changing viewing angles. We propose to gain in robustness to those nuisances, by explicitly learning a Nuisance Disentangled Feature Transforms (NDFT), utilizing the "free" meta-data as auxiliary attributes. Extensive results on real UAV imagery endorse its effectiveness. Our future interests will be devoted to generalizing NDFT to semi-supervised and weakly supervised training, to alleviate the dependence on annotated datasets.
Our approach has multiple highlights:

- Our proposed approach has high practical impact since we are the first to fully utilize the freely-acquirable attributes (meta-data).
- NDFT is the first approach aiming to learn robust feature w.r.t domain variation in object detection on UAV collected images, and it is extensible to multiple UAV-specific nuisances.
- The proposed alternative training strategy among the three modules stabilizes the adversarial training.
- We have tried multiple forms of the adversarial loss, including negative cross entropy between prediction and true label, KL divergence between prediction and uniform label and negative entropy. We have empirically found the negative entropy to work the best.

Reprinted with permission from the submitted paper, *Deep Nuisance Disentanglement for Robust Object Detection from Unmanned Aerial Vehicles* by Zhenyu Wu†, Karthik Suresh†, Priya Narayanan, Hongyu Xu, Heesung Kwon, Zhangyang Wang, 2019, International Conference on Computer Vision (ICCV 2019) Proceedings (Under Review).

† indicates equal contribution

REFERENCES

[1] E. Semsch, M. Jakob, D. Pavlicek, and M. Pechoucek, “Autonomous uav surveillance in complex urban environments,” in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 02*, pp. 82–85, IEEE Computer Society, 2009.

[2] E. Honkavaara, H. Saari, J. Kaivosoja, I. Pölönen, T. Hakala, P. Litkey, J. Mäkynen, and L. Pesonen, “Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight uav spectral camera for precision agriculture,” *Remote Sensing*, vol. 5, no. 10, pp. 5006–5039, 2013.

[3] “Drones for deliveries.” `https://scet.berkeley.edu/wp-content/uploads/ConnCarProjectReport-1.pdf`, 2018.

[4] M. Erdelj and E. Natalizio, “Uav-assisted disaster management: Applications and open issues,” in *Computing, Networking and Communications (ICNC), 2016 International Conference on*, pp. 1–5, IEEE, 2016.

[5] “Dji inspire 2 specs.” `https://www.dji.com/inspire-2/info#specs`, 2018.

[6] N. Dvornik, J. Mairal, and C. Schmid, “Modeling visual context is key to augmenting object detection datasets,” in *European Conference on Computer Vision*, pp. 375–391, Springer, 2018.

[7] A. Raj, V. P. Namboodiri, and T. Tuytelaars, “Subspace alignment based domain adaptation for rcnn detector,” *arXiv preprint arXiv:1507.05578*, 2015.

[8] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, “Domain adaptive faster r-cnn for object detection in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3339–3348, 2018.

[9] H. Lee, S. Eum, and H. Kwon, “Me r-cnn: Multi-expert r-cnn for object detection,” *arXiv preprint arXiv:1704.01069*, 2017.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in *European conference on computer vision*, pp. 346–361, Springer, 2014.

[14] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, pp. 91–99, 2015.

[16] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in *Advances in neural information processing systems*, pp. 379–387, 2016.

[17] H. Xu, X. Lv, X. Wang, Z. Ren, N. Bodla, and R. Chellappa, “Deep regionlets for object detection,” in *European Conference on Computer Vision*, pp. 827–844, 2018.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[19] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,”

[20] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[22] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,”

[23] G. Cheng, J. Han, P. Zhou, and L. Guo, “Multi-class geospatial object detection and geographic image classification based on collection of part detectors,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.

[24] S. Razakarivony and F. Jurie, “Vehicle detection in aerial imagery: A small target detection benchmark,” *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016.

[25] S. Han, W. Shen, and Z. Liu, “Deep drone: object detection and tracking for smart drones on embedded system,” 2016.

[26] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, “The unmanned aerial vehicle benchmark: Object detection and tracking,” *arXiv preprint arXiv:1804.00518*, 2018.

[27] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, “Vision meets drones: A challenge,” *arXiv preprint arXiv:1804.07437*, 2018.

[28] L. Cao, R. Ji, C. Wang, and J. Li, “Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer.,” 2016.

[29] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, “Perceptual generative adversarial networks for small object detection,” in *IEEE CVPR*, 2017.

[30] B. Singh and L. S. Davis, “An analysis of scale invariance in object detection–snip,”

[31] B. Singh, M. Najibi, and L. S. Davis, “Sniper: Efficient multi-scale training,” *arXiv preprint arXiv:1805.09300*, 2018.

[32] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” *arXiv preprint arXiv:1409.7495*, 2014.

[33] X. Wang, A. Shrivastava, and A. Gupta, “A-fast-rcnn: Hard positive generation via adversary for object detection,”

[34] L. Bashmal, Y. Bazi, H. AlHichri, M. M. AlRahhal, N. Ammour, and N. Alajlan, “Siamese-gan: Learning invariant representations for aerial vehicle image categorization,” *Remote Sensing*, vol. 10, no. 2, p. 351, 2018.

[35] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, “Cycada: Cycle-consistent adversarial domain adaptation,” *arXiv preprint arXiv:1711.03213*, 2017.

[36] X. Xiang and T. D. Tran, “Linear disentangled representation learning for facial actions,” *arXiv preprint arXiv:1701.03102*, 2017.

[37] G. Desjardins, A. Courville, and Y. Bengio, “Disentangling factors of variation via generative entangling,” *arXiv preprint arXiv:1210.5474*, 2012.

[38] N. Siddharth, B. Paige, A. Desmaison, J.-W. van de Meent, F. Wood, N. D. Goodman, P. Kohli, and P. H. Torr, “Learning disentangled representations in deep generative models,” 2016.

[39] I. Higgins, A. Pal, A. A. Rusu, L. Matthey, C. P. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner, "Darla: Improving zero-shot transfer in reinforcement learning," *arXiv preprint arXiv:1707.08475*, 2017.

[40] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," *arXiv preprint arXiv:1805.09730*, 2018.

[41] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection.,"

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988, IEEE, 2017.

[43] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Multi-task adversarial network for disentangled feature learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[45] "Visdrone2018 object detection in images leaderboard." `http://aiskyeye.com/views/getInfo?loc=13`, 2018.

[46] P. Zhu, L. Wen, D. Du, X. Bian, *et al.*, "Visdrone-det 2018: The vision meets drone object detection in image challenge results," *ECCV Vision Meets Drone Workshop*, 2018.

[47] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong, "Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis," in *European Conference on Computer Vision*, pp. 151–166, Springer, 2014.

[48] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "End-to-end united video dehazing and detection," *arXiv preprint arXiv:1709.03919*, 2017.