Delving into Robust Object Detection from Unmanned Aerial Vehicles: A Deep Nuisance Disentanglement Approach

Zhenyu Wu\textsuperscript{1}, Karthik Suresh\textsuperscript{2}, Priya Narayanan\textsuperscript{2}, Hongyu Xu\textsuperscript{3}, Heesung Kwon\textsuperscript{2}, Zhangyang Wang\textsuperscript{1}

\textsuperscript{1}Texas A&M University \quad \textsuperscript{2}U.S. Army Research Laboratory \quad \textsuperscript{3}University of Maryland

Abstract

Object detection from images captured by Unmanned Aerial Vehicles (UAVs) is becoming increasingly useful. Despite the great success of the generic object detection methods trained on ground-to-ground images, a huge performance drop is observed when they 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. Those nuisances constitute a large number of fine-grained domains, across which the detection model has to stay robust. Fortunately, UAVs will record meta-data that depict those varying attributes, which are either freely available along with the UAV images, or can be easily obtained. We propose to utilize those free meta-data in conjunction with associated UAV images to learn domain-robust features via an adversarial training framework dubbed Nuisance Disentangled Feature Transform (NDFT), for the specific challenging problem of object detection in UAV images, achieving a substantial gain in robustness to those nuisances. We demonstrate the effectiveness of our proposed algorithm, by showing state-of-the-art performance (single model) on two existing UAV-based object detection benchmarks. The code is available at https://github.com/TAMU-VITA/UAV-NDFT.

1. Introduction

Object detection has been extensively studied over the decades. While 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 machinery technology, Unmanned Aerial Vehicles (UAVs) equipped with cameras have been increasingly deployed in many industrial application, opening up a new frontier of computer vision applications in security surveillance, peacekeeping, agriculture, deliveries, aerial photography, disaster assistance [40, 25, 3, 14, 44], etc. One of the core features for the UAV-based applications is to detect objects of interest (e.g., pedestrians or vehicles).

Despite high demands, object detection from UAV is yet 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), such as but not limited to:

- **Variations in altitude and object scale:** The scales of objects captured in the image are closely affected by the flying altitude of UAVs. For example, the image captured by a DJI Inspire 2 series flying at 500 meters altitude [2] will contain very small objects, which are very challenging to detect and track. 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 shoots from different and free angles, in addition to the varying altitudes. For example, a UAV can look at one object from front view, to side view, to bird view, in a very short period of time. The diverse view angles cause arbitrary orientations and aspect ratios of the objects. Some view angles such as bird-view hardly 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 may fly under various weather and lighting conditions. The changes in illumination (daytime versus nighttime) and weathers (e.g. sunny, cloudy, foggy or rainy), will drastically affect the object visibility and appearance.

Most off-the-shelf detectors are trained with usually less varied, more restricted-view data. In comparison, the abundance of **UAV-specific nuisances** will cause the resulting
Figure 1: Examples showing the benefit of the proposed NDFT framework for object (vehicle) detection on the UAVDT dataset: starting from (a) Faster-RCNN [39] baseline, to gradually (b) disentangling the nuisances of altitude (A); (c) disentangling the nuisances of both altitude (A) and view angles (V); and (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 more nuisances (red rectangular boxes denote new correct detections beyond the baseline).

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 the 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 those massive number of fine-grained domains. Existing potential solutions include data augmentation [1, 13], domain adaption [34, 8], and ensemble of expert models [26]. However, neither of these approaches are easy to generalize to multiple and/or unseen domains [34, 8], and they could lead to over-parameterized models which is not suitable for UAV on-board deployments [1, 13, 26].

A (Almost) Free Lunch: Fine-Grained Nuisance Annotations. In view of the above, we cast UAV-based object detection problem as a cross-domain object detection problem with fine-grained domains. The object types of interest sustain across domains; such task-related features shall be preserved and extracted. The above UAV-specific nuisances constitute the domain-specific nuisances, that should be eliminated for transferable feature learning. For UAVs, major nuisance types are well recognized, e.g., altitude, angle and weather. More Importantly, in the specific case of UAVs, those nuisances annotations could be easily obtained or even freely available. For example, a UAV can record its flying altitudes as metadata by GPS, or more accurately, by a barometric sensor. For another example, weather information is easy to retrieve, since with each UAV flight’s time-stamp and spatial location (or path), one can straightforwardly obtain the weather of specific time/location.

Motivated by those observations, we propose to learn an object detection model that maintains its effectiveness in extracting task-related features while eliminating the recognized types of nuisances, across different domains (e.g., altitudes/angles/weathers). We take advantage of the free (or easy) access to the nuisance annotations. Based on them, we are the first to adopt an adversarial learning framework, to learn task-specific, domain-invariant features by explicitly disentangling task-specific and nuisance features in a supervised way. The framework, dubbed Nuisance Disentangled Feature Transform (NDFT), gives rise to highly robust UAV-based object detection models, that can be directly applicable to not only domains in training, but also more unseen domains, without needing any extra effort of domain adaptation or sampling/labeling. Experiments on two real UAV-based object detection benchmarks suggest the state-of-the-art effectiveness of NDFT.
2. Related Works

2.1. Object Detection: General and UAV-Specific

Object detection has progressed tremendously, partially thanks to established benchmarks (i.e., MS COCO [29] and PASCAL VOC [15]). There are primarily two main streams of approaches: two-stage detectors and single-stage detectors, based on whether the detectors have proposal-driven mechanism or not. Two stage detectors [18, 23, 17, 39, 10, 51, 52] contains region proposal network (RPN) to first generate region proposals, and then extract region-based features to predict the object categories and their corresponding locations. Single-stage detectors [36, 37, 38, 31] apply dense sampling windows over object locations and scales, and usually achieved higher speed than two-stage ones, although often at the cost of (marginal) accuracy decrease.

Aerial Image-based Object Detection A few aerial image datasets (i.e. DOTA [49], NWPU VHR-10 [9], and VEDAI [35]) were proposed recently. However, those above 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 more varied altitudes, poses and weather. Also, the common practice to detect objects from aerial images remains still to deploy off-the-shelf ground-based object detection models [21, 33].

Public benchmarks were unavailable for specifically UAV-based object detection until recently. Two datasets, UAVDT [12] and VisDrone2018 [54], were released to address this gap. UAVDT consists of 100 video sequences (about 80k frames) captured from UAVs under complex scenarios. Moreover, it also provides full annotations for weather conditions, flying altitudes, and camera views in addition to the ground truth bounding box of the target objects. VisDrone2018 [54] is a large-scale UAV-based object detection and tracking benchmark, composed of 10,209 static images and 179,264 frames from 263 video clips.

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. [7] proposed a super-resolution algorithm using coupled dictionary learning to transfer the target region into high resolution to “augment” its visual appearance. [47, 27, 30] proposed to internally super-resolve the feature maps of small objects to make them resemble similar characteristics as large objects. SNIP [42] 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 [43] further processed context regions around ground-truth instances at different appropriate scales to efficiently train the detector at multiple scales, improving the detection of tiny object detection more.

2.2. Handling Domain Variances

Domain Adaptation via Adversarial Training Adversarial domain adaptation [16] 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 fields too. [46] 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 adaptation on both image and instance levels. [5] introduced a Siamese-GAN to learn invariant feature representations for both labeled and unlabeled aerial images coming from two different domains. CyCADA [24] unified cycle-consistency with adversarial loss to learn domain-invariance. However, these domain adaptation methods typically assume one (ideal) source domain and one (non-ideal) target domain. The possibility of generalizing these methodologies to handling many 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 invariant features shared by many different “non-ideal” target domains (both seen and unseen), by disentangling domain-specific nuisances. The setting thus differs from typical domain adaptation and generalizes to task-specific feature extraction in unseen domains naturally.

Data Augmentation, and Model Ensemble Compared to the considerable amount of research in data augmentation for classification [16], less attention was paid on other tasks such as detection [1]. Classical data augmentation relies on a limited set of pre-known factors (such as scaling, rotation, flipping) that are easy to invoke, and adopt ad-hoc, minor perturbations that are unlikely to change labels, in order to gain robustness to those variations. However, UAV images will involve a much larger variety of nuisances, many of which are hard to “synthesize”, e.g., images from different angles. [13, 53] proposed learning-based approaches to synthesize new training samples for detection. But they focused on re-combining foreground objects and background contexts, rather than re-composing specific nuisance attributes. Also, the (much) larger augmented dataset adds to training burden and may cause over-parameterized models.

Another methodology was proposed in [26]. To capture the appearance variations caused by different shapes poses and viewing angles, it 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 has limitations as the model ensemble quickly becomes too expensive as more different domains are involved. It further cannot generalize to unknown or unseen domains.
Feature Disentanglement in Generative Models

Feature disentanglement [50] 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 [11, 41], in order to disentangle the factors of variation from the content in the latent feature space. In the image-to-image translation, a recent work [19] 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. Due to the different application scope from generative models, we do not add back the disentangled components to reconstruct the original input.

3. Our Approach

3.1. Formulation of NDFT

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

\[
\begin{align*}
\min_{f_O, f_T} & L_O(f_O(f_T(X)), Y_O) - \gamma L_N(f_N(f_T(X)), Y_N), \\
\min_{f_N} & L_N(f_N(f_T(X)), Y_N)
\end{align*}
\]

In (1), \( f_O \) denotes the model that performs the object detection task \( O \) on its input data. The label set \( Y_O \) is object bounding box coordinates and class labels provided on \( X \). \( L_O \) is a cost function defined to evaluate the object detection performance on \( O \). 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 \( N \). 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 (1), such that

- The object detection task performance \( L_O \) is minimally affected over \( f_T(X) \), compared to using \( X \).
- The nuisance prediction task performance \( L_N \) is maximally suppressed over \( f_T(X) \), compared to using \( X \).

In order to deal with the multiple nuisances case, we extend the (1) to multiple prediction tasks. Here we assume \( k \) nuisances prediction tasks associated with label sets \( Y^1_N, \ldots, Y^k_N \). \( \gamma_1, \ldots, \gamma_k \) are the respective weight coefficients.

The modified objective naturally becomes:

\[
\begin{align*}
\min_{f_O, f_T} & L_O(f_O(f_T(X)), Y_O) - \sum_{i=1}^{k} \gamma_i L_N(f^i_N(f_T(X)), Y^i_N), \\
\min_{f^1_N, \ldots, f^k_N} & L_N(f^i_N(f_T(X)), Y^i_N)
\end{align*}
\]

\( f_T, f_O \) and \( f^i_N \)s can all be implemented by deep networks.

**Interpretation as Three-Party Game**

NDFT can be derived from a three-competitor game optimization:

\[
\begin{align*}
\max_{f_N} \min_{f_O, f_T} & L_O(f_O(f_T(X)), Y_O) - \gamma L_N(f_N(f_T(X)), Y_N)
\end{align*}
\]

where \( f_T \) is an obfuscator, \( f_N \) as a attacker, and \( f_O \) as an utilizer (adopting ML security terms). In fact, the two sub-optimizations in (1) denote an iterative routine to solve this unified form (performing coordinate descent between \( \{ f_T, f_O \} \), and \( f_N \)). This form can easily capture many other settings or scenarios, e.g., privacy-preserving visual recognition [48, 45] where \( f_T \) encodes features to avoid peeps from \( f_N \) while preserving utility for \( f_O \).

3.2. Implementation and Training

**Architecture Overview: NDFT-Faster-RCNN**

As an instance of the general NDFT framework (2), Figure 2 displays an implementation example of NDFT using the Faster-RCNN backbone [39], while later we will demonstrate that NDFT can be plug-and-play with other 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 \( f_N \), shown by (2).

By jointly training \( f_T, f_O \) and \( f^i_N \)s in the above adversarial settings, 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. \( f_O \) 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 RoIAlign [22] layer, while it is fed to \( f_N \) after going through a spatial pyramid pooling layer [23].

**Choices of \( L_O \) and \( L_N \)**

\( L_O \) is the bounding box classification (e.g., softmax) and regression loss (e.g., smooth \( \ell_1 \)) as widely used in traditional two stage detectors. However, using \(- L_N \) as the adversarial loss in the first row of (2) 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 gradient reversal trick [16], we decide to follow [32] to choose the negative entropy func-
tion 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, close to uniform random guesses) on the nuisances.

Since we replace $L_N$ with $L_{ne}$ in the first objective in (2), it no longer needs $Y_N$. Meanwhile, the usage of $L_N$ and $Y_N$ remains unaffected in the second objective of (2). $L_N$ and $Y_N$ are used to pre-train $f_N^k$s at the initialization and keep $f_N^k$s as “sufficiently strong adversaries” throughout the adversarial training, in order to learn meaningful $f_T$ that can generalize better. Our final framework alternates between:

$$\min_{f_O, f_T} L_O(f_O(f_T(X)), Y_O) + \sum_{i=1}^{k} \gamma_i L_{ne}(f_N^i(f_T(X)))$$

$$\min_{f_N^1, \ldots, f_N^k} L_N(f_N^i(f_T(X)), Y_N^i)$$

### Training Strategy

Just like training GANs [20], our training is prone to collapse and/or bad local minima. We thus presented a carefully-designed training algorithm with the alternating update strategy. The training procedure is summarized in Algorithm 1 and explained below.

For each mini-batch, we first jointly optimize $f_T$ and $f_O$ weights (with $f_N^k$s frozen), by minimizing the first objective in (3) using the standard stochastic gradient descent (SGD). Meanwhile, we will keep “monitoring” $f_N^i$ branches: as $f_T$ is updated, 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 (3) 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 $f_N^1, \ldots, f_N^k$ to random initialization, and then re-train them on $f_T(X)$ (with $f_T$ fixed) to become strong nuisance predictors again, before we re-start the above alternative process of $f_T$, $f_O$ and $f_N^k$s. This re-starting trick is also found to benefit the generalization of learned $f_T$ [48], potentially due to helping get out of some bad local minima.

### 4. 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 general-ality of NDFT. First, we perform the main body of experiments on the UAVDT benchmark [12], 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 public benchmark, VisDrone2018. Originally, the nuisance annotations were not released on VisDrone2018. We manually annotate the nuisances on each image: those annotations will be released publicly, and hopefully will be contributed as a part of VisDrone. Learning NDFT gives a performance boost over the the best single model, and leads us to the (single model) state-of-the-art mean average precision (mAP)\(^1\) on VisDrone2018 validation set\(^2\).

In addition, we study a transfer learning setting from the NDFT learned on UAVDT, to VisDrone2018. The goal of exploring transfer is because UAVs often come across unseen scenarios, and a good transferability of learned features facilitates more general usability. When detecting the (shared) vehicles category, \(f_T\) shows strong transferability by outperforming the best single-model method currently reported on the VisDrone2018 leaderboard [4].

4.1. 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 by the authors in [12], 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 weather condition (daylight, 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) on UAVDT, to ensure a solid enough ground for NDFT. The authors reported a AP of \(\sim 20\) using a Faster-RCNN model with the VGG-16 backbone. We replace the backbone with ResNet-101, and fine-tune hyperparameters such as anchor scale \((16,32,64,128,256)\). We end up with an improved AP of 45.64 (using the same IoU threshold = 0.7 as the authors) as our baseline performance. We also communicated with the authors of [12] in person and they acknowledged this improved baseline. We then implement NDFT-Faster-RCNN using the architecture depicted in Figure 2, also with a ResNet-101 backbone. We denote \(\gamma_1\), \(\gamma_2\) and \(\gamma_3\) as the coefficients in (1), for the \(L_{ne}\) loss terms for altitude, view and weather nuisances, respectively.

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

Tables 1, 2, and 3 show the benefit of removing flying altitude (A), camera view (V) and weather condition (W) nuisances, individually. That could be viewed as learning NDFT-Faster-CNN (Figure 2) with only the corresponding one \(\gamma_i\) \((i = 1, 2, 3)\) to be nonzero. The baseline model without nuisance disentanglement has \(\gamma_i = 0, i = 1, 2, 3\).

As can be seen from Table 1, compared to the baseline \((\gamma_1 = 0)\), an overall AP gain is obtained at \(\gamma_1 = 0.03\), where we achieve a AP improvement of 0.28.

Table 2 shows the performance gain by removing the camera view (V) nuisance. At \(\gamma_2 = 0.01\), an overall AP improvement of 0.52 is obtained. Similar positive observations are found in Table 3 as well, when the weather (W) nuisance is removed.

---

\(^1\)mAP on the 10 categories of objects is the standard evaluation criterion on VisDrone2018.

\(^2\)The top-2 models on the UAVDT leaderboard are model ensembles. We compare with only single model solutions for fairness.

\(^3\)We discard another “foggy” class because of its too small size.
nuisance is removed: $\gamma_3 = 0.01$ results in an overall AP boost of 0.98 over the baseline, with the more challenging night class AP increased by 7.52.

Table 4 shows the full results by incrementally adding more adversarial losses 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 $\gamma_i = 0.01$ for both/all of them, as discovered to give the best single-nuisance results in Tables 1 - 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). In conclusion, removing nuisances using NDFT evidently contributes to addressing the tough problem of object detection on high-mobility UAV platforms. Furthermore, the final best-performer $A + V + W$ improves the class-wise APs noticeably on some most challenging nuisance classes, such as high-altitude, bird-view and nighttime. 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.

Table 4: UAVDT NDFT-Faster-RCNN with multiple attribute disentanglement.

<table>
<thead>
<tr>
<th>$\gamma_1$</th>
<th>A</th>
<th>V</th>
<th>Front</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>68.14</td>
<td>49.71</td>
<td>18.70</td>
<td>45.64</td>
</tr>
<tr>
<td>0.01</td>
<td>69.01</td>
<td>50.46</td>
<td>14.65</td>
<td>45.31</td>
</tr>
<tr>
<td>0.02</td>
<td>66.97</td>
<td>46.91</td>
<td>16.09</td>
<td>44.17</td>
</tr>
<tr>
<td>0.03</td>
<td>66.38</td>
<td>53.00</td>
<td>15.69</td>
<td>45.92</td>
</tr>
<tr>
<td>0.05</td>
<td>65.46</td>
<td>48.43</td>
<td>16.58</td>
<td>44.36</td>
</tr>
</tbody>
</table>

Adopting Stronger FPN Backbones We demonstrate that the performance gain by NDFT does not vanish as we adopt more sophisticated backbones, e.g. FPN [28]. Training FPN on UAVDT leads to the baseline performance improved from 45.64 to 49.05. By replacing Faster-RCNN with FPN in the NDFT training pipeline, the resulting model learns to simultaneously disentangle $A + V + W$ nuisances ($\gamma_i = 0.005, i = 1, 2, 3$). We are able to further increase the overall AP to 52.03, showing the general benefit of NDFT regardless of the backbone choices.

### Proof-of-Concepts for NDFT-based Tracking

With object detection as our main focus, we also evaluate NDFT on UAVDT tracking for proof-of-concept. We choose SORT [6] (a popular online and real-time tracker) and evaluate on the multi-object tracking (MOT) task defined on UAVDT. We follow the tracking-by-detection framework adopted in [12], and compare the tracking results based on the detection inputs from vanilla Faster-RCNN and NDFT-Faster-RCNN ($A + V + W$), respectively. All evaluation protocols are inherited from [12]. As in Table 5, NDFT-FRCNN largely outperforms the vanilla baseline in 10 out of the 11 metrics, showing its promise even beyond detection.

### Comparing NDFT with Multi-Task Learning

Another plausible option to utilize nuisance annotations is to jointly predict $Y_O$ and $Y_A$ as standard multi-task learning. To compare it with NDFT fairly, we switch the sign from $-\gamma$ to $+\gamma$ in (2) first row, through which the nuisance prediction tasks become three auxiliary losses (AL) in multi-task learning. We minimize this new optimization and carefully re-tune $\gamma$ for AL by performing grid search. As seen from Table 6, while AL is able to slightly improve over the baseline too (as expected), NDFT is evidently and consistently better thanks to its unique ability to encode invariances. The experiments objectively establish the role of adversarial losses versus standard auxiliary losses.

#### Table 5: NDFT versus vanilla baseline on MOT task.

#### Table 6: Comparing the baseline Faster-RCNN, adding auxiliary losses, and our proposed NDFT method.
4.2. VisDrone2018: Results and Analysis

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. We manually annotate the UAV-specific nuisances, with the same three categories as on UAVDT.

According to the leaderboard [4] and workshop report [55], the best-performing single model is DE-FPN, which utilized FPN (removing P6) with a ResNeXt-101 64-4d backbone. We implement DE-FPN by identically following their method description in [55], as our comparison subject.

Implementation Details Taking the DE-FPN backbone, NDFT is learned by simultaneously disentangling three nuisances (A+V+W). We create the DE-FPN model with NDFT, termed as NDFT-DE-FPN. The performance of DE-FPN and NDFT-DE-FPN are evaluated using the mAP over the 10 object categories on the VisDrone2018 validation set, since the testing set is not publicly accessible.

Table 7: mAP comparison on VisDrone2018 validation set.

<table>
<thead>
<tr>
<th>γ_i</th>
<th>DE-FPN</th>
<th>NDFT-DE-FPN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>mAP</td>
<td>48.41</td>
<td>49.75</td>
</tr>
</tbody>
</table>

Results and Analysis As in Table 7, NDFT-DE-FPN gives rise to a 4.36 mAP boost over DE-FPN, making it a new state-of-the-art single model on VisDrone2018. Figure 3 shows a visual comparison example.

4.3. Transfer from UAVDT to VisDrone2018

Problem Setting We use VisDrone2018 as a testbed to showcase the transferability of NDFT features learned from UAVDT. We choose DE-FPN as the comparison subject.

Implementation Details DE-FPN is trained on VisDrone 2018 training set and tested on the vehicle category of validation set. We then train the same DE-FPN backbone on UAVDT with three nuisances (A+V+W) disentangled (γ = γ_2 = γ_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 transferability using NDFT. Besides, we repeat the same above routine with γ_1 = γ_2 = γ_3 = 0, to create a transferred DE-FPN baseline 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 (since the testing set is not publicly accessible).

Results and Analysis The APs of DE-FPN, DE-FPN(r) and NDFT-DE-FPN(r) are 76.80, 75.27 and 79.50, respectively on the vehicle category. Directly transferring DE-FPN from UAVDT to VisDrone2018 (fine-tuned on the latter) does not give rise to competitive performance, showing a substantial domain mismatch between the two datasets. However, transferring the learned NDFT to VisDrone2018 leads to performance boosts, with a 4.23 AP margin over the transfer baseline without disentanglement, and 2.70 over DE-FPN. It demonstrates that NDFT could potentially contribute to a more generally transferable UAV object detector that handles more unseen scenes (domains). A visual comparison example on VisDrone2018 is presented in Figure 4.

5. Conclusion

This paper investigates object detection from UAV-mounted cameras, a vastly useful yet under-studied problem. The problem appears to be more challenging than standard object detection, due to many UAV-specific nuisances. We propose to gain robustness to those nuisances, by explicitly learning a Nuisance Disentangled Feature Transform (NDFT), utilizing the “free” metadata. Extensive results on real UAV imagery endorse its effectiveness.
References


[7] Liujian Cao, Rongrong Ji, Cheng Wang, and Jonathan Li. Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer. 2016. 3


[27] Jianan Li, Xiaodan Liang, Yun Chao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In IEEE CVPR, 2017. 3


Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. 3


