# Recognition in Terra Incognita

# 在未知领域的识别

Sara Beery, Grant Van Horn, and Pietro Perona
Sara Beery, Grant Van Horn, 和 Pietro Perona
Caltech
加州理工学院
{ sbeery, gvanhorn, perona }@ caltech.edu

{ sbeery, gvanhorn, perona }@ caltech.edu

Abstract. It is desirable for detection and classification algorithms to generalize to unfamiliar environments, but suitable benchmarks for quantitatively studying this phenomenon are not yet available. We present a dataset designed to measure recognition generalization to novel environments. The images in our dataset are harvested from twenty camera traps deployed to monitor animal populations. Camera traps are fixed at one location, hence the background changes little across images; capture is triggered automatically, hence there is no human bias. The challenge is learning recognition in a handful of locations, and generalizing animal detection and classification to new locations where no training data is available. In our experiments state-of-the-art algorithms show excellent performance when tested at the same location where they were trained. However, we find that generalization to new locations is poor, especially for classification systems. [1]

摘要: 检测和分类算法能够在不熟悉的环境中进行泛化是理想的，但目前尚未有适合定量研究这一现象的基准。我们提出了一个数据集，旨在测量对新环境的识别泛化。我们数据集中的图像来自于部署在二十个相机陷阱中的动物种群监测。相机陷阱固定在一个位置，因此图像的背景变化很小；捕捉是自动触发的，因此没有人为偏差。挑战在于在少数地点学习识别，并将动物检测和分类泛化到没有训练数据的新地点。在我们的实验中，最先进的算法在测试时表现出色，尤其是在它们接受训练的同一位置。然而，我们发现对新地点的泛化效果较差，尤其是对于分类系统。[1]

Keywords: Recognition, transfer learning, domain adaptation, context, dataset, benchmark.

关键词: 识别，迁移学习，领域适应，背景，数据集，基准。

## 1 Introduction

## 1 引言

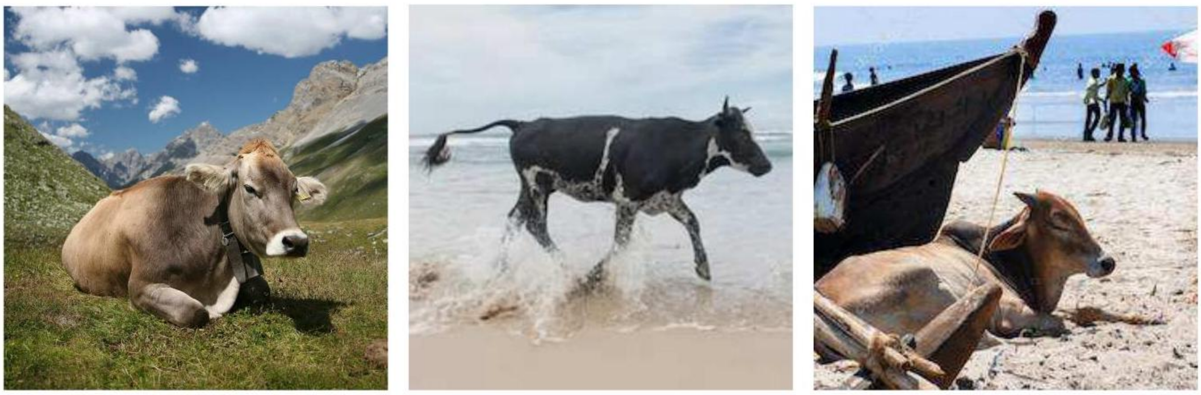
Automated visual recognition algorithms have recently achieved human expert performance at visual classification tasks in field biology [1 − 3] and medicine [4, 5] . Thanks to the combination of deep learning [6, 7] , Moore’ s law [8] and very large annotated datasets [9, 10] enormous progress has been made during the past 10 years. Indeed, 2017 may come to be remembered as the year when automated visual categorization surpassed human performance.

自动视觉识别算法最近在野外生物学 [1 − 3] 和医学 [4, 5] 的视觉分类任务中达到了人类专家的表现。得益于深度学习 [6, 7] 、摩尔定律 [8] 和非常大的标注数据集 [9, 10] ，在过去 10 年中取得了巨大的进展。事实上，2017 年可能会被铭记为自动视觉分类超越人类表现的一年。

However, it is known that current learning algorithms are dramatically less data-efficient than humans [11], transfer learning is difficult [12], and, anecdotally, vision algorithms do not generalize well across datasets [13, 14] (Fig. 1). These observations suggest that current algorithms rely mostly on rote pattern-matching, rather than abstracting from the training set ’visual concepts’ [15] that can generalize well to novel situations. In order to make progress we need datasets that support a careful analysis of generalization, dissecting the challenges in detection and classification: variation in lighting, viewpoint, shape, photographer’s choice and style, context/background. Here we focus on the latter: generalization to new environments, which includes background and overall lighting conditions.

然而，已知当前的学习算法在数据效率上远不及人类 [11]，迁移学习也很困难 [12]，并且根据经验，视觉算法在不同数据集之间的泛化能力较差 [13, 14] (图 1)。这些观察结果表明，当前的算法主要依赖于机械的模式匹配，而不是从训练集的“视觉概念”中进行抽象，这些概念可以很好地泛化到新情况 [15]。为了取得进展，我们需要支持对泛化进行仔细分析的数据集，剖析检测和分类中的挑战: 光照、视角、形状、摄影师的选择和风格、上下文/背景的变化。在这里，我们关注后者: 对新环境的泛化，包括背景和整体光照条件。

[1] The dataset is available at https://beerys.github.io/CaltechCameraTraps/



(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Fig. 1. Recognition algorithms generalize poorly to new environments. Cows in 'common' contexts (e.g. Alpine pastures) are detected and classified correctly (A), while cows in uncommon contexts (beach, waves and boat) are not detected (B) or classified poorly (C). Top five labels and confidence produced by ClarifAI.com shown.

图 1. 识别算法在新环境中的泛化能力较差。在“常见”背景 (例如阿尔卑斯草甸) 中，牛被正确检测和分类 (A)，而在不常见的背景 (海滩、波浪和船) 中，牛则未被检测到 (B) 或分类不佳 (C)。显示了由 ClarifAI.com 生成的前五个标签和置信度。

Applications where the ability to generalize visual recognition to new environments is crucial include surveillance, security, environmental monitoring, assisted living, home automation, automated exploration (e.g. sending rovers to other planets). Environmental monitoring by means of camera traps is a paradigmatic application. Camera traps are heat- or motion-activated cameras placed in the wild to monitor and investigate animal populations and behavior. Camera traps have become inexpensive, hence hundreds of them are often deployed for a given study, generating a deluge of images. Automated detection and classification of animals in images is a necessity. The challenge is training animal detectors and classifiers from data coming from a few pilot locations such that these detectors and classifiers will generalize to new locations. Camera trap data is controlled for environment including lighting (the cameras are static, and lighting changes systematically according to time and weather conditions), and eliminates photographer bias (the cameras are activated automatically).

在视觉识别能够推广到新环境的能力至关重要的应用中，包括监控、安全、环境监测、辅助生活、家庭自动化和自动探索 (例如，向其他行星发送探测器)。通过相机陷阱进行环境监测是一个典型的应用。相机陷阱是放置在野外的热感或运动激活相机，用于监测和研究动物种群及其行为。相机陷阱的成本已经降低，因此在特定研究中通常会部署数百个，从而生成大量图像。自动检测和分类图像中的动物是必要的。挑战在于从少数试点地点的数据中训练动物检测器和分类器，以使这些检测器和分类器能够推广到新地点。相机陷阱数据在环境上是受控的，包括光照 (相机是静态的，光照会根据时间和天气条件系统性变化)，并消除了摄影师偏见 (相机是自动激活的)。

Camera traps are not new to the computer vision community [16-27,2]. Our work is the first to identify camera traps as a unique opportunity to study generalization, and we offer the first study of generalization to new environments in this controlled setting. We make here three contributions: (a) a novel, well-annotated dataset to study visual generalization across locations, (b) a benchmark to measure algorithms' performance, and (c) baseline experiments establishing the state of the art. Our aim is to complement current datasets utilized by the vision community for detection and classification [9, 10, 28, 29] by introducing a new dataset and experimental protocol that can be used to systematically evaluate the generalization behavior of algorithms to novel environments. In this work we benchmark the current state-of-the-art detection and classification pipelines and find that there is much room for improvement.

相机陷阱对计算机视觉社区并不陌生 [16-27,2]。我们的工作首次将相机陷阱视为研究推广的独特机会，并在这一受控环境中提供了对新环境推广的首次研究。我们在此做出三项贡献:(a) 一个新颖且注释良好的数据集，用于研究跨地点的视觉推广；(b) 一个用于测量算法性能的基准；(c) 建立当前技术水平的基线实验。我们的目标是通过引入一个新的数据集和实验协议，补充视觉社区当前用于检测和分类的

[1] 数据集可在 https://beerys.github.io/CaltechCameraTraps/ 获取

数据集 [9, 10, 28, 29] ，以系统性地评估算法在新环境中的推广行为。在这项工作中，我们对当前最先进的检测和分类流程进行了基准测试，发现仍有很大的改进空间。

# 2 Related Work

# 2 相关工作

## 2.1 Datasets

## 2.1 数据集

The ImageNet [9], MS-COCO [10], PascalVOC [28], and Open Images [29] datasets are commonly used for benchmarking classification and detection algorithms. Images in these datasets were collected in different locations by different people, which enables algorithms to average over photographer style and irrelevant background clutter. However, as demonstrated in Fig. 1, the context can be strongly biased. Human photographers are biased towards well-lit, well-focused images where the subjects are centered in the frame [30, 31] . Furthermore, the number of images per class is balanced, unlike what happens in the real world [11].

ImageNet [9]、MS-COCO [10]、PascalVOC [28] 和 Open Images [29] 数据集通常用于基准分类和检测算法。这些数据集中的图像由不同的人在不同地点收集，这使得算法能够对摄影师的风格和无关的背景杂乱进行平均。然而，如图 1 所示，背景可能存在强烈的偏见。人类摄影师倾向于拍摄光线良好、对焦清晰且主体居中于画面的图像 [30, 31] 。此外，每个类别的图像数量是平衡的，这与现实世界中的情况不同 [11]。

Natural world datasets such as the iNaturalist dataset [1], CUB200 [32], Oxford Flowers [33], LeafSnap [34], and NABirds700 [35] are focused on fine-grained species classification and detection. Most images in these datasets are taken by humans under relatively good lighting conditions, though iNaturalist does contain human-selected camera trap images. Many of these datasets exhibit real-world long-tailed distributions, but in all cases there is a large amount of diversity in location and perspective.

自然世界数据集，如 iNaturalist 数据集 [1]、CUB200 [32]、牛津花卉 [33]、LeafSnap [34] 和 NABirds700 [35]，专注于细粒度物种分类和检测。这些数据集中的大多数图像是在相对良好的光照条件下由人类拍摄的，尽管 iNaturalist 确实包含人类选择的相机捕获图像。这些数据集中的许多展示了现实世界中的长尾分布，但在所有情况下，位置和视角的多样性都很大。

The Snapshot Serengeti dataset [21] is a large, multi-year camera trap dataset collected at 225 locations in a small region of the African savanna. It is the single largest-scale camera trap dataset ever collected, with over 3 million images. However, it is not yet suitable for controlled experiments. This dataset was collected from camera traps that fire in sequences of 3 for each motion trigger, and provides species annotation for groups of images based on a time threshold. This means that sometimes a single species annotation is provided for up to 10 frames, when in fact the animal was present in only a few of those frames (no bounding boxes are provided). Not all camera trap projects are structured in a similar way, and many cameras take shorter sequences or even single images on each trigger. In order to find a solution that works for new locations regardless of the camera trap parameters, it is important to have information about which images in the batch do or do not contain animals. In our dataset we provide annotations on a per-instance basis, with bounding boxes and associated classes for each animal in the frame.

Snapshot Serengeti 数据集 [21] 是一个大型的多年度相机陷阱数据集，收集于非洲草原小区域的 225 个地点。它是迄今为止收集的最大规模的相机陷阱数据集，包含超过 300 万张图像。然而，它尚不适合用于控制实验。该数据集是从每次运动触发时以 3 张图像的序列拍摄的相机陷阱中收集的，并根据时间阈值为图像组提供物种注释。这意味着有时会为多达 10 帧提供单一物种注释，而实际上动物只出现在其中的几帧中 (未提供边界框)。并非所有相机陷阱项目都以类似方式构建，许多相机在每次触发时拍摄的序列较短，甚至仅拍摄单张图像。为了找到适用于新地点的解决方案，无论相机陷阱参数如何，了解批次中哪些图像包含或不包含动物的信息是重要的。在我们的数据集中，我们提供逐实例的注释，为每个帧中的动物提供边界框和相关类别。

## 2.2 Detection

## 2.2 检测

Since camera traps are static, detecting animals in the images could be considered either a change detection or foreground detection problem. Detecting changes and/or foreground vs. background in video is a well studied problem [36], [37]. Many of these methods rely on constructing a good background model that updates regularly, and thus degrade rapidly at low frame rates. [38] and [39] consider low frame rate change detection in aerial images, but in these cases there are often very few examples per location.

由于相机陷阱是静态的，因此在图像中检测动物可以被视为变化检测或前景检测问题。在视频中检测变化和/或前景与背景是一个经过充分研究的问题 [36]，[37]。许多这些方法依赖于构建一个定期更新的良好背景模型，因此在低帧率下会迅速退化。[38] 和 [39] 考虑了航空图像中的低帧率变化检测，但在这些情况下，每个地点的示例通常非常少。

Some camera traps collect a short video when triggered instead of a sequence of frames. [20, 23, 22] show foreground detection results on camera trap video. Data that comes from most camera traps take sequences of frames at each trigger at a frame rate of ∼ 1 frame per second. This data can be considered ”video,” albeit with extremely low, variable frame rate. Statistical methods for background subtraction and foreground segmentation in camera trap image sequences have been previously considered. [16] demonstrates a graph-cut method that uses background modeling and foreground object saliency to segment foreground in camera trap sequences. [24] creates background models and perform a superpixel-based comparison to determine areas of motion. [25] uses a multilayer RPCA-based method applied to day and night sequences. [26] uses several statistical background-modeling approaches as additional signal to improve and speed up deep detection. These methods rely on a sequence of frames at each trigger to create appropriate background models, which are not always available. None of these methods demonstrate results on locations outside of their training set.

一些相机陷阱在被触发时收集短视频，而不是一系列帧。[20, 23, 22] 显示了相机陷阱视频中的前景检测结果。来自大多数相机陷阱的数据在每次触发时以 ∼ 1 帧每秒的帧率捕捉帧序列。这些数据可以被视为“视频”，尽管其帧率极低且可变。先前已考虑在相机陷阱图像序列中进行背景减除和前景分割的统计方法。[16] 展示了一种图切割方法，该方法利用背景建模和前景物体显著性在相机陷阱序列中分割前景。[24] 创建背景模型并执行基于超像素的比较以确定运动区域。[25] 使用了一种基于多层 RPCA 的方法，应用于昼夜序列。[26] 使用几种统计背景建模方法作为额外信号，以改善和加速深度检测。这些方法依赖于每次触发时的帧序列来创建适当的背景模型，但这些模型并不总是可用。这些方法均未在其训练集以外的地点展示结果。

## 2.3 Classification

## 2.3 分类

A few studies tackle classification of camera trap images. [18] showed results classifying squirrels vs. tortoises in the Mojave Desert. [17] showed classification results on data that provides image sequences of $\tilde{1}0$ frames. They do not consider the detection problem and instead manually crop the animal from the frame and balance the dataset, resulting in a total of 7,196 images across 18 species with at least 100 examples each. [19] were the first to take a deep network approach to camera trap classification, working with data from eMammal [40]. They first performed detection using the background subtraction method described in [16], then classified cropped detected regions, getting 38.31% top-1 accuracy on 20 common species. [27] show classification results on both Snapshot Serengeti and data from jungles in Panama, and saw a boost in classification performance from providing animal segmentations. [2] show 94.9% top-1 accuracy using an ensemble of models for classification on the Snapshot Serengeti dataset. None of the previous works show results on unseen test locations.

一些研究探讨了相机捕捉图像的分类。[18] 显示了在莫哈维沙漠中对松鼠与乌龟的分类结果。[17] 显示了对提供 $\tilde{1}0$ 帧图像序列的数据的分类结果。它们没有考虑检测问题，而是手动从帧中裁剪动物并平衡数据集，最终得到 18 个物种共 7,196 张图像，每个物种至少有 100 个样本。[19] 是首个采用深度网络方法进行相机捕捉分类的研究，使用了来自 eMammal [40] 的数据。他们首先使用 [16] 中描述的背景减法方法进行检测，然后对裁剪的检测区域进行分类，在 20 种常见物种上获得了 38.31% 的顶级准确率。[27] 显示了在 Snapshot Serengeti 和来自巴拿马丛林的数据上的分类结果，并通过提供动物分割提高了分类性能。[2] 显示了在 Snapshot Serengeti 数据集上使用模型集成进行分类的 94.9% 顶级准确率。之前的研

究均未在未见测试位置上展示结果。

## 2.4 Generalization and Domain Adaptation

## 2.4 泛化与领域适应

Generalizing to a new location is an instance of domain adaptation, where each location represents a domain with its own statistical properties such as types of flora and fauna, species frequency, man-made or other clutter, weather, camera type, and camera orientation. There have been many methods proposed for domain adaptation in classification [41]. [42] proposed a method for unsupervised domain adaptation by maximizing domain classification loss while minimizing loss for classifying the target classes. We generalized this method to multi-domain for our dataset, but did not see any improvement over the baseline. [43] demonstrated results of a similar method for fine-grained classification, using a multi-task setting where the adaptation was from clean web images to real-world images, and [44] investigated open-set domain adaptation.

泛化到新位置是领域适应的一个实例，其中每个位置代表一个具有自身统计特性的领域，如植物和动物类型、物种频率、人造或其他杂物、天气、相机类型和相机方向。已经提出了许多用于分类的领域适应方法 [41]。[42] 提出了一种通过最大化领域分类损失同时最小化目标类别分类损失的无监督领域适应方法。我们将该方法推广到多领域以适应我们的数据集，但未见对基线的改进。[43] 展示了类似方法在细粒度分类中的结果，采用了多任务设置，其中适应是从干净的网络图像到真实世界图像，[44] 则研究了开放集领域适应。

Few methods have been proposed for domain adaptation outside of classification. [45-47] investigate methods of domain adaptation for semantic segmentation, focusing mainly on cars and pedestrians and either adapting from synthetic to real data, from urban to suburban scenes, or from PASCAL to a camera onboard a car. [48-52] look at methods for adapting detectors from one data source to another, such as from synthetic to real data or from images to video. Raj, et. al., [53] demonstrated a subspace-based detection method for domain adaptation from PASCAL to COCO.

很少有方法被提出用于分类以外的领域适应。[45-47] 研究了语义分割的领域适应方法，主要集中在汽车和行人上，或者从合成数据适应到真实数据，从城市场景适应到郊区场景，或从 PASCAL 适应到车载摄像头。[48-52] 关注从一个数据源适应检测器到另一个数据源的方法，例如从合成数据到真实数据或从图像到视频。Raj 等人 [53] 展示了一种基于子空间的检测方法，用于从 PASCAL 到 COCO 的领域适应。

# 3 The Caltech Camera Traps Dataset

# 3 Caltech 摄像机陷阱数据集

The Caltech Camera Traps (CCT) dataset contains 243,187 images from 140 camera locations, curated from data provided by the USGS and NPS. Our goal in this paper is to specifically target the problem of generalization in detection and classification. To this end, we have randomly selected 20 camera locations from the American Southwest to study in detail. By limiting the geographic region, the flora and fauna seen across the locations remain consistent. The current task is not to deal with entirely new regions or species, but instead to be able to recognize the same species of animals in the same region with a different camera background. In the future we plan to extend this work to recognizing the same species in new regions, and to the open-set problem of recognizing never-before-seen species. Examples of data from different locations can be seen in Fig. 2.

Caltech 摄像机陷阱 (CCT) 数据集包含来自 140 个摄像机位置的 243,187 张图像，这些数据是从美国地质调查局 (USGS) 和国家公园管理局 (NPS) 提供的数据中整理而来。我们在本文中的目标是专门针对检测和分类中的泛化问题。为此，我们随机选择了来自美国西南部的 20 个摄像机位置进行详细研究。通过限制地理区域，各个位置所见的植物和动物保持一致。目前的任务不是处理全新的区域或物种，而是能够在相同区域内识别相同物种的动物，尽管背景摄像机不同。未来，我们计划将这项工作扩展到在新区域识别相同物种，以及开放集问题，即识别从未见过的物种。不同位置的数据示例可以在图 2 中看到。

Camera traps are motion- or heat-triggered cameras that are placed in locations of interest by biologists in order to monitor and study animal populations and behavior. When a camera is triggered, a sequence of images is taken at approximately one frame per second. Our dataset contains sequences of length 1 - 5 . The cameras are prone to false triggers caused by wind or heat rising from the ground, leading to empty frames. Empty frames can also occur if an animal moves out of the field of view of the camera while the sequence is firing. Once a month, biologists return to the cameras to replace the batteries and change out the memory card. After it has been collected, experts manually sort camera trap data to categorize species and remove empty frames. The time required to sort and label images by hand severely limits data scale and research productivity. We have acquired and further curated a portion of this data to analyze generalization behaviors of state-of-the-art classifiers and detectors.

相机陷阱是由生物学家放置在感兴趣地点的运动或热触发相机，用于监测和研究动物种群及其行为。当相机被触发时，约每秒拍摄一帧图像。我们的数据集包含长度为 1 到 5 的序列。这些相机容易受到风或地面升温引起的虚假触发，从而导致空帧。如果动物在序列拍摄期间移出相机的视野，也会出现空帧。生物学家每月返回相机更换电池和更换存储卡。数据收集后，专家手动整理相机陷阱数据，以对物种进行分类并删除空帧。手动排序和标记图像所需的时间严重限制了数据规模和研究生产力。我们已经获取并进一步整理了这部分数据，以分析最先进的分类器和检测器的泛化行为。

The dataset in this paper, which we call Caltech Camera Traps-20 (CCT-20), consists of 57,868 images across 20 locations, each labeled with one of 15 classes (or marked as empty). Classes are either single species (e.g. "Coyote" or groups of species, e.g. "Bird"). See Fig. 4 for the distribution of classes and images across locations. We do not filter the stream of images collected by the traps, rather this is the same data that a human biologist currently sifts through. Therefore the data is unbalanced in the number of images per location, distribution of species per location, and distribution of species overall (see Fig. 4).

本文中的数据集，我们称之为加州理工学院相机陷阱-20(CCT-20)，包含来自 20 个地点的 57,868 张图像，每张图像标记为 15 个类别之一 (或标记为空)。类别可以是单一物种 (例如“郊狼”) 或物种组 (例如“鸟类”)。请参见图 4 以了解各地点的类别和图像分布。我们不对陷阱收集的图像流进行过滤，而是使用人类生物学家当前筛选的相同数据。因此，数据在每个地点的图像数量、每个地点的物种分布以及整体物种分布上都是不平衡的 (见图 4)。

## 3.1 Detection and Labeling Challenges

## 3.1 检测和标记挑战

The animals in the images can be challenging to detect and classify, even for humans. We have determined that there are six main nuisance factors inherent to camera trap data, which can compound upon each other. Detailed analysis of these challenges can be seen in Fig. 3. When an image is too difficult to classify on its own, biologists will often refer to an easier image in the same sequence and then track motion by flipping between sequence frames in order to generate a label for each frame (e.g. is the animal still present or has it gone off the image plane?). We account for this in our experiments by reporting performance at the frame level and at the sequence level. Considering frame level performance allows us to investigate the limits of current models in exceptionally difficult cases.

图像中的动物可能难以检测和分类，即使对于人类也是如此。我们确定了相机捕捉数据中固有的六个主要干扰因素，这些因素可能相互叠加。这些挑战的详细分析可以在图 3 中看到。当一张图像本身难以分类时，生物学家通常会参考同一序列中较易分类的图像，然后通过在序列帧之间切换来跟踪运动，以便为每一帧生成标签 (例如，动物是否仍然存在，或者是否已经离开了图像平面？)。我们在实验中考虑了这一点，通过报告帧级别和序列级别的性能。考虑帧级别的性能使我们能够调查当前模型在极其困难情况下的极限。

## 3.2 Annotations

## 3.2 注释

We collected bounding box annotations on Amazon Mechanical Turk, procuring annotations from at least three and up to ten mturkers for each image for redundancy and accuracy. Workers were asked to draw boxes around all instances of a specific type of animal for each image, determined by what label was given to the sequence by the biologists. We used the crowdsourcing method by Branson et al. [54] to determine ground truth boxes from our collective annotations, and to iteratively collect additional annotations as necessary. We found that bounding box precisions varied based on annotator, and determined that for this data the PascalVOC metric of IoU $\geq 0.5$ is appropriate for the detection experiments (as opposed to the COCO IoU averaging metric).

我们在亚马逊机械土耳其上收集了边界框注释，为每张图像从至少三名到十名的 mturker 获取注释，以确保冗余和准确性。工作人员被要求在每张图像中围绕特定类型动物的所有实例绘制框，这些类型是

由生物学家根据序列所给的标签确定的。我们使用 Branson 等人提出的众包方法 [54] 来确定我们集体注释的真实框，并根据需要迭代收集额外的注释。我们发现边界框的精度因注释者而异，并确定对于这些数据，PascalVOC 的 IoU $\geq 0.5$ 指标适用于检测实验 (与 COCO IoU 平均指标相对)。

## 3.3 Data Split: Cis- and Trans-

## 3.3 数据划分: 同位和异位

Our goal is exploring generalization to new (i.e. untrained) locations. Thus, we compare the performance of detection and classification algorithms when they are tested at the same locations where they were trained, vs new locations. For brevity, we refer to locations seen during training as cis-locations and locations not seen during training as trans-locations.

我们的目标是探索对新 (即未训练) 地点的泛化。因此，我们比较检测和分类算法在训练时测试的相同地点与新地点的性能。为简洁起见，我们将训练期间看到的地点称为同位点，将训练期间未见过的地点称为异位点。

From our pool of 20 locations, we selected 9 locations at random to use as trans-location test data, and a single random location to use as trans-location validation data. From the remaining 10 locations, we use images taken on odd days as cis-location test data. From within the data taken on even days, we randomly select 5% to be used as cis-location validation data. The remaining data is used for training, with the constraint that training and validation sets do not share the same image sequences. This gives us 13,553 training images, 3,484 validation and 15,827 test images from cis-locations, and 1,725 val and 23,275 test images from trans-locations. The data split can be visualized in Fig. 4. We chose to interleave the cis training and test data by day because we found that using a single date to split the data results in additional generalization challenges due to changing vegetation and animal species distributions across seasons. By interleaving, we reduce noise and provide a clean experimental comparison of results on cis- and trans-locations.

在我们选择的 20 个地点中，我们随机选择了 9 个地点作为跨地点测试数据，并选择一个随机地点作为跨地点验证数据。在剩余的 10 个地点中，我们使用奇数天拍摄的图像作为同地点测试数据。在偶数天拍摄的数据中，我们随机选择 5% 用作同地点验证数据。剩余的数据用于训练，前提是训练集和验证集不共享相同的图像序列。这使我们从同地点获得了 13,553 张训练图像、3,484 张验证图像和 15,827 张测试图像，从跨地点获得了 1,725 张验证图像和 23,275 张测试图像。数据划分可以在图 4 中可视化。我们选择按天交错同地点的训练和测试数据，因为我们发现使用单一日期划分数据会导致由于季节性植被和动物种类分布变化而产生额外的泛化挑战。通过交错，我们减少了噪声，并提供了对同地点和跨地点结果的清晰实验比较。

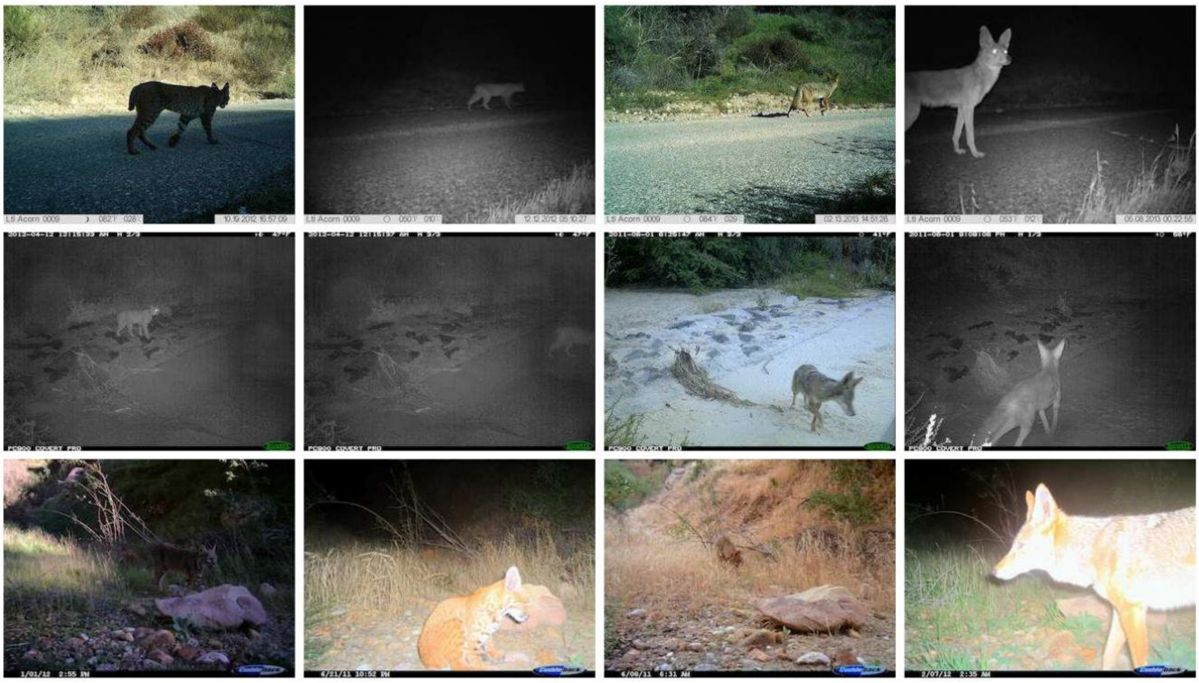


Fig. 2. Camera trap images from three different locations. Each row is a different location and a different camera type. The first two cameras use IR, while the third row used white flash. The first two columns are bobcats, the next two columns

图 2. 来自三个不同地点的相机捕捉图像。每一行代表一个不同的地点和不同的相机类型。前两台相机使用红外线，而第三行使用白色闪光灯。前两列是美洲狮，接下来的两列

are coyotes.

是郊狼。

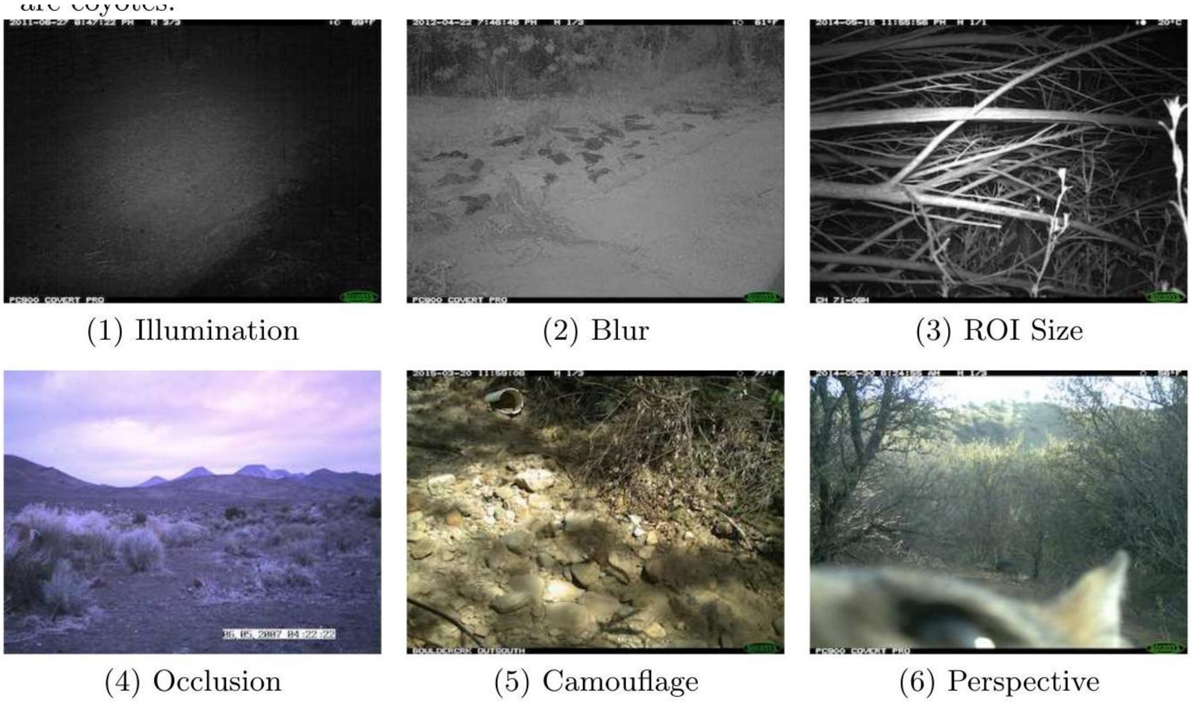

(1) Illumination
(2) Blur
(3) ROI Size
(4) Occlusion
(5) Camouflage
(6) Perspective

Fig. 3. Common data challenges: (1) Illumination: Animals are not always salient. (2) Motion blur: common with poor illumination at night. (3) Size of the region of interest (ROI): Animals can be small or far from the camera. (4) Occlusion: e.g. by bushes or rocks. (5) Camouflage: decreases saliency in animals' natural habitat. (6) Perspective: Animals can be close to the camera, resulting in partial views of the body.

图 3. 常见的数据挑战:(1) 照明: 动物并不总是显著可见。(2) 运动模糊: 在夜间光线不足时常见。(3) 感兴趣区域 (ROI) 的大小: 动物可能很小或离相机很远。(4) 遮挡: 例如，被灌木或岩石遮挡。(5) 伪装: 在动物的自然栖息地中降低显著性。(6) 视角: 动物可能离相机很近，导致身体部分视图。

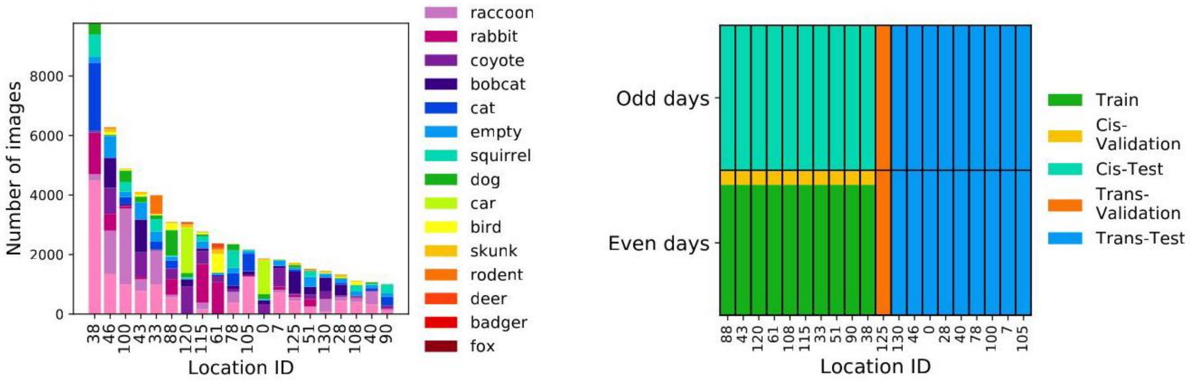


Fig. 4. (Left) Number of annotations for each location, over 16 classes. The ordering of the classes in the legend is from most to least examples overall. The distribution of the species is long-tailed at each

location, and each location has a different and peculiar distribution. (Right) Visualization of data splits. ”Cis” refers to images from locations seen during training, and ”trans” refers to new locations not seen during training.

图 4. (左) 每个位置的注释数量，涵盖 16 个类别。图例中类别的排序是从总体上样本最多到样本最少。每个位置的物种分布呈长尾分布，并且每个位置具有不同且独特的分布。(右) 数据划分的可视化。“Cis” 指的是在训练期间看到的位置的图像，而 “trans” 指的是在训练期间未见过的新位置。

# 4 Experiments

# 4 实验

Current state-of-the-art computer vision models for classification and detection are designed to work well on test data whose distribution matches the training distribution. However, in our experiments we are explicitly evaluating the models on a different test distribution. In this situation, it is common practice to employ early stopping [55] as a means of preventing overfitting to the train distribution. Therefore, for all classification and detection experiments we monitor performance on both the cis- and trans-location validation sets. In each experiment we save two models, one that we expect has the best performance on the trans-location test set (i.e. a model that generalizes), and one that we expect has the best performance on the cis-location test set (i.e. a model that performs well on the train distribution).

当前最先进的计算机视觉分类和检测模型旨在在测试数据的分布与训练分布匹配时表现良好。然而，在我们的实验中，我们明确地在不同的测试分布上评估模型。在这种情况下，通常的做法是采用早停法 [55] 来防止对训练分布的过拟合。因此，在所有分类和检测实验中，我们监控在 cis 和 trans 位置验证集上的性能。在每个实验中，我们保存两个模型，一个是我们预期在 trans 位置测试集上表现最佳的模型 (即一个能够泛化的模型)，另一个是我们预期在 cis 位置测试集上表现最佳的模型 (即一个在训练分布上表现良好的模型)。

## 4.1 Classification

## 4.1 分类

We explore the generalization of classifiers in 2 different settings: full images and cropped bounding boxes. For each setting we also explore the effects of using and ignoring sequence information. Sequence information is utilized in two different ways: (1) Most Confident we consider the sequence to be classified correctly if the most confident prediction from all frames grouped together is correct, or (2) Oracle we consider the sequence to be correctly classified if any frame is correctly classified. Note that (2) is a more optimistic usage of sequence information. For all classification experiments we use an Inception-v3 [56] model pretrained on ImageNet, with an initial learning rate of 0.0045, rmsprop with a momentum of 0.9, and a square input resolution of 299 . We employ random cropping (containing at least 65% of the region), horizontal flipping, and color distortion as data augmentation.

我们在两种不同的设置中探讨分类器的泛化: 完整图像和裁剪的边界框。对于每种设置，我们还探讨了使用和忽略序列信息的影响。序列信息以两种不同的方式被利用:(1) 最自信的情况下，如果所有帧组合在一起的最自信预测是正确的，我们认为该序列被正确分类；或者 (2) 在 Oracle 情况下，如果任何一帧被正确分类，我们认为该序列被正确分类。请注意，(2) 是对序列信息的更乐观的使用。在所有分类实验中，我们使用在 ImageNet 上预训练的 Inception-v3 [56] 模型，初始学习率为 0.0045，使用动量为 0.9 的 rmsprop，输入分辨率为 299 的平方。我们采用随机裁剪 (至少包含 65% 的区域)、水平翻转和颜色扭曲作为数据增强。

Table 1. Classification top-1 error across experiments. Empty images are removed for these experiments.

表 1. 实验中的分类 top-1 错误率。这些实验中移除了空图像。

| | Cis-Locations | | Trans-Locations | | Error Increase | |
|---|---|---|---|---|---|---|
| Sequence Information | Images | Bboxes | Images | Bboxes | Images | Bboxes |
| None | 19.06 | 8.14 | 41.04 | 19.56 | 115% | 140% |
| Most Confident | 17.7 | 7.06 | 34.53 | 15.77 | 95% | 123% |
| Oracle | 14.92 | 5.52 | 28.69 | 12.06 | 92% | 118% |

|  | 顺式位置 |  | 反式位置 |  | 错误增加 |  |
|---|---|---|---|---|---|---|
| 序列信息 | 图像 | 边界框 | 图像 | 边界框 | 图像 | 边界框 |
| 无 | 19.06 | 8.14 | 41.04 | 19.56 | 115% | 140% |
| 最有信心 | 17.7 | 7.06 | 34.53 | 15.77 | 95% | 123% |
| 预言者 | 14.92 | 5.52 | 28.69 | 12.06 | 92% | 118% |

Full Image. We train a classifier on the full images, considering all 15 classes as well as empty images (16 total classes). On the cis-location test set we achieve a top-1 error of 20.83% , and a top-1 error of 41.08% on the trans-location test set with a 97% cis-to-trans increase in error. To investigate if requiring the classifier to both detect and classify animals increased overfitting on the training location backgrounds, we removed the empty images and retrained the classifiers using just the 15 animal classes. Performance stayed at nearly the same levels, with a top-1 error of 19.06% and 41.04% for cis- and trans-locations respectively. Utilizing sequence information helped reduce overall error (achieving errors of 14.92% and 28.69% on cis- and trans-locations respectively), but even in the most optimistic oracle setting, there is still a 92% increase in error between evaluating on cis- and trans-locations. See Table 1 for the full results.

完整图像。我们在完整图像上训练分类器，考虑所有 15 个类别以及空图像 (总共 16 个类别)。在顺序位置测试集上，我们达到了 20.83% 的顶级错误率，而在逆序位置测试集上达到了 41.08% 的顶级错误率，错误率增加了 97% 。为了调查要求分类器同时检测和分类动物是否增加了对训练位置背景的过拟合，我们移除了空图像，并仅使用 15 个动物类别重新训练了分类器。性能几乎保持在相同水平，顺序和逆序位置的顶级错误率分别为 19.06% 和 41.04%。利用序列信息有助于减少整体错误 (顺序和逆序位置的错误率分别为 14.92% 和 28.69% )，但即使在最乐观的预言设置中，顺序和逆序位置之间的错误率仍然增加了 92% 。完整结果见表 1。

Bounding Boxes. We train a classifier on cropped bounding boxes, excluding all empty images (as there is no bounding box in those cases). Using no sequence information we achieve a cis-location top-1 error of 8.14% and a trans-location top-1 error of 19.56% . While the overall error has decreased compared to the image level classification, the error increase between cis- and trans-locations is still high at 140% . Sequence information further improved classification results (achieving errors of 5.52% and 12.06% on cis- and trans-locations respectively), and slightly reduced generalization error, bringing the error increase down to 118% in the most optimistic setting. See Table 1 for the full results. Additional experiments investigating the effect of number of images per location, number of training locations, and selection of validation location can be seen in the supplementary material.

边界框。我们在裁剪的边界框上训练分类器，排除了所有空图像 (因为在这些情况下没有边界框)。在不使用序列信息的情况下，我们在顺序位置的 top-1 错误率为 8.14% ，在横向位置的 top-1 错误率为 19.56% 。虽然与图像级分类相比，总体错误率有所降低，但顺序位置和横向位置之间的错误增加仍然很高，达到 140% 。序列信息进一步改善了分类结果 (在顺序位置和横向位置分别达到了 5.52% 和 12.06% 的错误率)，并略微降低了泛化误差，使得在最乐观的情况下错误增加降至 118% 。完整结果见表 1。关于每个位置图像数量、训练位置数量和验证位置选择的额外实验可以在补充材料中查看。

Analysis Fig. 5 provides a high level summary of our experimental findings. Namely, there is a generalization gap between cis- and trans-locations. Cropped boxes help to improve overall performance (shifting the blue lines vertically downward to the red lines), but the gap remains. In the best case scenario (red dashed lines: cropped boxes and optimistically utilizing sequences) we see a 92% increase in error between the cis- and trans-locations (with the same number of training examples), and 20x increase in training examples to have the same error rate. One might wonder whether this generalization gap is due to a large shift in class distributions between the two locations types. However, Fig. 7 shows that the overall distribution of classes between the locations is similar, and therefore probably does not account for the performance loss.

分析图 5 提供了我们实验发现的高级总结。即，顺序位置和横向位置之间存在一个泛化差距。裁剪框有助于提高整体性能 (将蓝线垂直向下移动到红线)，但差距依然存在。在最佳情况下 (红色虚线: 裁剪框并乐观地利用序列)，我们看到顺序位置和横向位置之间的错误增加为 92% (训练样本数量相同)，并且需要增加 20x 的训练样本才能达到相同的错误率。人们可能会想知道这个泛化差距是否是由于两个位置类型之间类分布的巨大变化。然而，图 7 显示两个位置之间的类总体分布是相似的，因此可能并不导致性能损失。

## 4.2 Detection

## 4.2 检测

We use the Faster-RCNN implementation found in the Tensorflow Object Detection code base [57] as our detection model. We study performance of the Faster-RCNN model using two different backbones, ResNet-101 [58] and Inception-ResNet-v2 with atrous convolution [57]. Similar to our classification experiments we analyze the effects of using sequence information using two methods: (1) Most Confident we consider a sequence to be labeled correctly if the most confident detection across all frames has an IoU $\geq 0.5$ with its matched ground truth box; (2) Oracle we consider a sequence to be labeled correctly if any frame' s most confident detection has IoU $\geq 0.5$ with its matched ground truth box. Note that method (2) is more optimistic than method (1).

我们使用在 Tensorflow 物体检测代码库 [57] 中找到的 Faster-RCNN 实现作为我们的检测模型。我们研究了使用两种不同骨干网络的 Faster-RCNN 模型的性能，分别是 ResNet-101 [58] 和带有空洞卷积的 Inception-ResNet-v2 [57]。与我们的分类实验类似，我们使用两种方法分析使用序列信息的效果:(1) 最自信，我们认为如果所有帧中最自信的检测与其匹配的真实框的 IoU $\geq 0.5$ ，则该序列标记为正确；(2) 甲骨文，我们认为如果任何帧的最自信检测与其匹配的真实框的 IoU $\geq 0.5$ ，则该序列标记为正确。注意，方法 (2) 比方法 (1) 更乐观。

Our detection models are pretrained on COCO [10], images are resized to have a max dimension of 1024 and a minimum dimension of 600 ; each experiment uses SGD with a momentum of 0.9 and a fixed learning rate schedule. Starting at 0.0003 we decay the learning rate by a factor of 10 at 90k steps and 120k steps. We use a batch size of 1, and employ horizontal flipping for data augmentation. For evaluation, we consider a detected box to be correct if its IoU $\geq 0.5$ with a ground truth box.

我们的检测模型在 COCO [10] 上进行了预训练，图像被调整为最大维度为 1024 和最小维度为 600；每个实验使用动量为 0.9 的 SGD 和固定的学习率计划。从 0.0003 开始，我们在 90k 步骤和 120k 步骤时将学习率衰减 10 倍。我们使用的批量大小为 1，并采用水平翻转进行数据增强。对于评估，我们认为如果检测框与真实框的 IoU $\geq 0.5$ 达到要求，则该检测框是正确的。

Results from our experiments can be seen in Table 2 and Fig 9. We find that both backbone architectures perform similarly. Without taking sequence information into account, the models achieve $\sim$ 77% mAP on cis-locations and $\sim$ 71% mAP on trans-locations. Adding sequence information using the most confident metric improves results, bringing performance on cis- and trans-locations to similar values at $\sim$ 85% . Finally, using the oracle metric brings mAP into the 90s for both locations. Precision-recall curves at the frame and sequence levels for both detectors can be seen in Fig. 9.

我们实验的结果可以在表 2 和图 9 中看到。我们发现两种骨干架构的表现相似。在不考虑序列信息的情况下，这些模型在顺式位置上达到 $\sim$ 77% mAP，在反式位置上达到 $\sim$ 71% mAP。使用最自信的度量添加序列信息可以改善结果，使顺式和反式位置的性能达到相似值 $\sim$ 85% 。最后，使用 oracle 度量使两个位置的 mAP 达到 90 以上。两个检测器在帧和序列级别的精确度-召回曲线可以在图 9 中看到。

Analysis There is a significantly lower generalization error in our detection experiments when not using sequences than what we observed in the classification experiments ($\sim$ 30% error increase for detections vs $\sim$ 115% error increase for classification). When using sequence information, the generalization error for detections is reduced to only $\sim$ 5% .

分析在我们的检测实验中，不使用序列时的泛化误差显著低于我们在分类实验中观察到的 (检测的 ($\sim$ 30% 误差增加与分类的 $\sim$ 115% 误差增加)。使用序列信息时，检测的泛化误差降低到仅 $\sim$ 5% 。

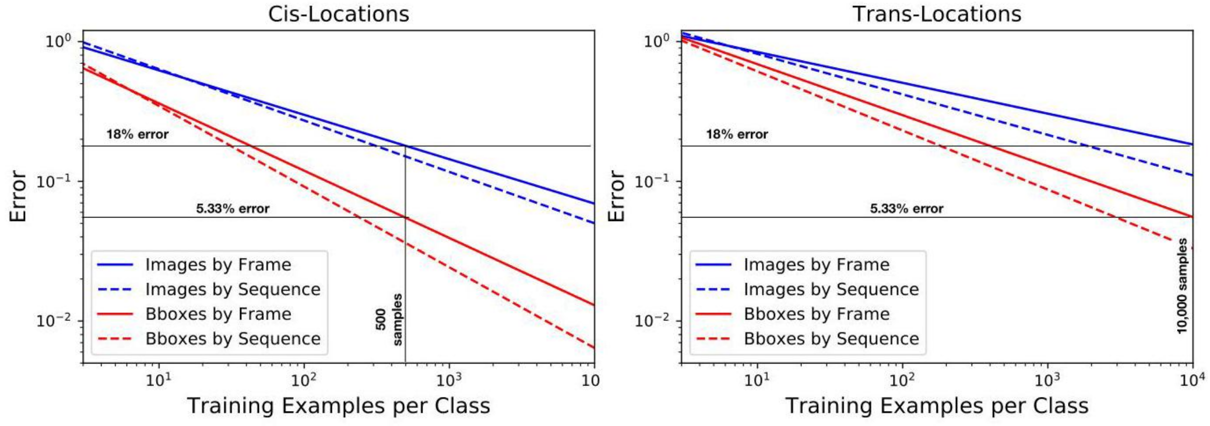


Fig. 5. Classification error vs. number of class-specific training examples. Error is calculated as 1 - AUC (area under the precision-recall curve). Best-fit lines through the error-vs-n.examples points for each class in each scenario (points omitted for clarity), with average $r^2 = 0.261$ . An example of line fit on top of data can be seen in Fig. 7. As expected, error decreases as a function of the number of training examples. This is true both for image classification (blue) and bounding-box classification (red) on both cis-locations and trans-locations. However, trans-locations show significantly higher error rates. To operate at an error rate of 5.33% on bounding boxes or 18% on images at the cis-locations we need 500 training examples, while we need 10,000 training examples to achieve the same error rate at the trans-locations, a 20x increase

图 5. 分类误差与类特定训练样本数量的关系。误差计算为 1 - AUC(精确度-召回曲线下的面积)。每个场景中每个类别的误差与样本数量的最佳拟合线 (为清晰起见省略点)，平均 $r^2 = 0.261$ 。数据上方的线性拟合示例可以在图 7 中看到。正如预期的那样，误差随着训练样本数量的增加而减少。这对于顺式位置和反式位置的图像分类 (蓝色) 和边界框分类 (红色) 都是如此。然而，反式位置显示出显著更高的误差率。为了在顺式位置的边界框上达到 5.33% 的误差率或在图像上达到 18% 的误差率，我们需要 500 个训练样本，而在反式位置达到相同的误差率则需要 10,000 个训练样本，增加了 20x 。

in data.

数据。

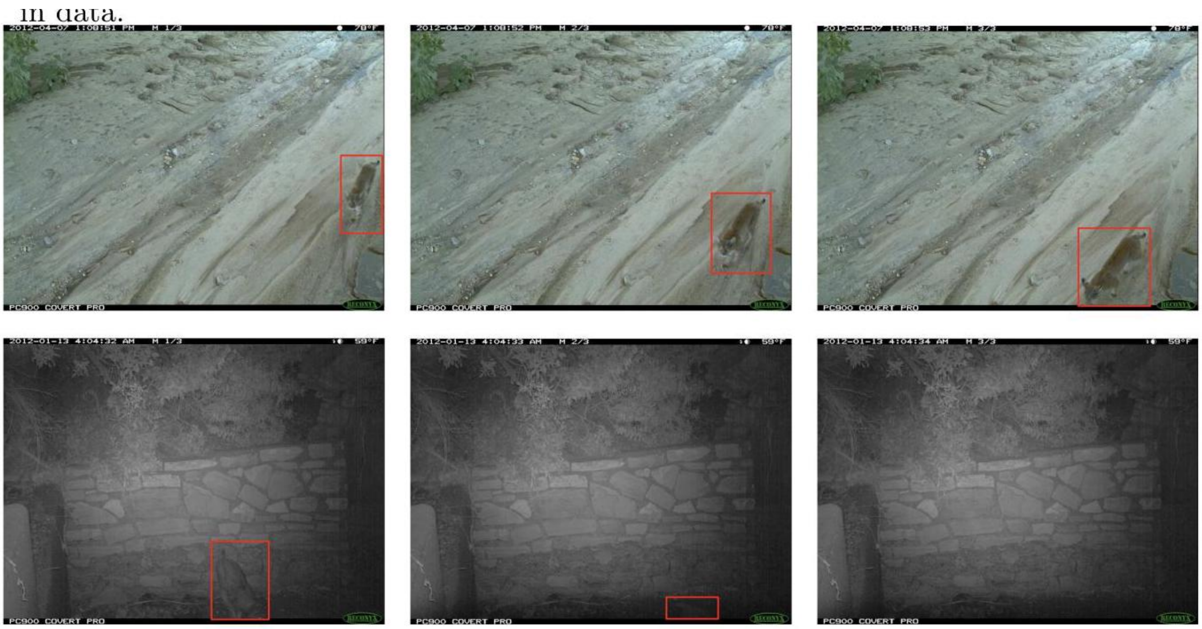


Fig. 6. Trans-classification failure cases at the sequence level: (Based on classification of bounding box crops) In the first sequence, the network struggles to distinguish between ’cat’ and ’bobcat’, incorrectly predicting ’cat’ in all three images with a mean confidence of 0.82 . In the second sequence, the network struggles to classify a bobcat at an unfamiliar pose in the first image and instead predicts ’raccoon’ with a confidence of 0.84 . Little additional sequence information is available in this case, as the next frame contains only a blurry tail, and the last frame is empty

图 6. 序列级别的跨分类失败案例:(基于边界框裁剪的分类) 在第一个序列中，网络在区分“猫”和“美洲狮”之间存在困难，在所有三张图像中错误地预测为“猫”，平均置信度为 0.82。在第二个序列中，网络在第一张图像中难以分类一个处于不熟悉姿势的美洲狮，而是以 0.84 的置信度预测为“浣熊”。在这种情况下，几乎没有额外的序列信息，因为下一帧仅包含一个模糊的尾巴，最后一帧是空的。

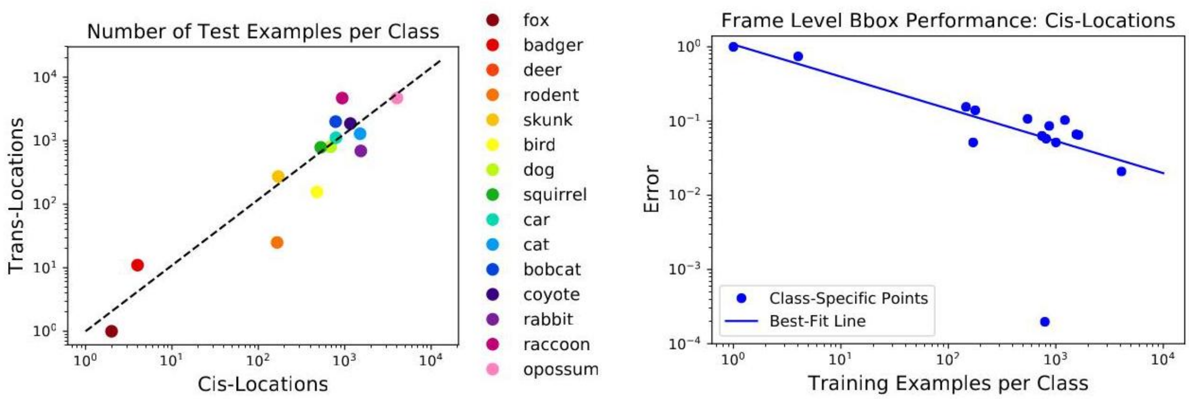


Fig. 7. (Left) Distribution of species across the two test sets. (Right) An example of line fit used to generate the plots in Fig. 5

图 7. (左) 两个测试集中的物种分布。(右) 用于生成图 5 中图表的线性拟合示例。

Qualitatively, we found the mistakes can often be attributed to nuisance factors that make frames difficult. We see examples of all 6 nuisance factors described in Fig. 3 causing detection failures. The errors remaining at the sequence level occur when these nuisance factors are present in all frames of a sequence, or when the sequence only contains a single, challenging frame containing an animal. Examples of sequence-level detection failures can be seen in Fig. 8. The generalization gap at the frame level implies that our models are better able to deal with nuisance factors at locations seen during training.

从定性上看，我们发现错误通常可以归因于使帧变得困难的干扰因素。我们在图 3 中描述的所有 6 种干扰因素导致检测失败的例子。序列级别上剩余的错误发生在这些干扰因素出现在序列的所有帧中，或者当序列仅包含一帧具有挑战性的动物图像时。序列级别检测失败的例子可以在图 8 中看到。帧级别的泛化差距表明，我们的模型更能处理在训练期间见过的位置的干扰因素。

Our experiments show that there is a small generalization gap when we use sequence information. However, overall performance has not saturated, and current state-of-the-art detectors are not achieving high precision at high recall values (1% precision at recall = 95%) . So while we are encouraged by the results, there is still room for improvement. When we consider frames independently, we see that the generalization gap reappears. Admittedly this is a difficult case as it is not clear what the performance of a human would be without sequence information. However, we know that there are objects that can be detected in these frames and this dataset will challenge the next generation of detection models to accurately localize these difficult cases.

我们的实验表明，当我们使用序列信息时，存在一个小的泛化差距。然而，整体性能尚未饱和，当前最先进的检测器在高召回值下并未达到高精度 (1% precision at recall = 95%) 。因此，尽管我们对结果感到鼓舞，但仍然有改进的空间。当我们独立考虑帧时，我们看到泛化差距再次出现。诚然，这是一个困难的案例，因为不清楚没有序列信息时人类的表现会如何。然而，我们知道在这些帧中有可以被检测到的物体，这个数据集将挑战下一代检测模型，以准确定位这些困难案例。

Table 2. Detection mAP at IoU = 0.5 across experiments.

表 2. 各实验中 IoU = 0.5 下的检测 mAP。

| | Cis-Locations | | Trans-Locations | | Error Increase | |
|---|---|---|---|---|---|---|
| Sequence Information | ResNet | Inception | ResNet | Inception | ResNet | Inception |
| None | 77.10 | 77.57 | 70.17 | 71.37 | 30% | 27.6% |
| Most Confident | 84.78 | 86.22 | 84.09 | 85.44 | 4.5% | 5.6% |
| Oracle | 94.95 | 95.04 | 92.13 | 93.09 | 55.8% | 39.3% |

| | 顺式位置 | | 反式位置 | | 错误增加 | |
|---|---|---|---|---|---|---|
| 序列信息 | 残差网络 | inception | 残差网络 | inception | 残差网络 | inception |
| 无 | 77.10 | 77.57 | 70.17 | 71.37 | 30% | 27.6% |
| 最自信 | 84.78 | 86.22 | 84.09 | 85.44 | 4.5% | 5.6% |
| 预言者 | 94.95 | 95.04 | 92.13 | 93.09 | 55.8% | 39.3% |

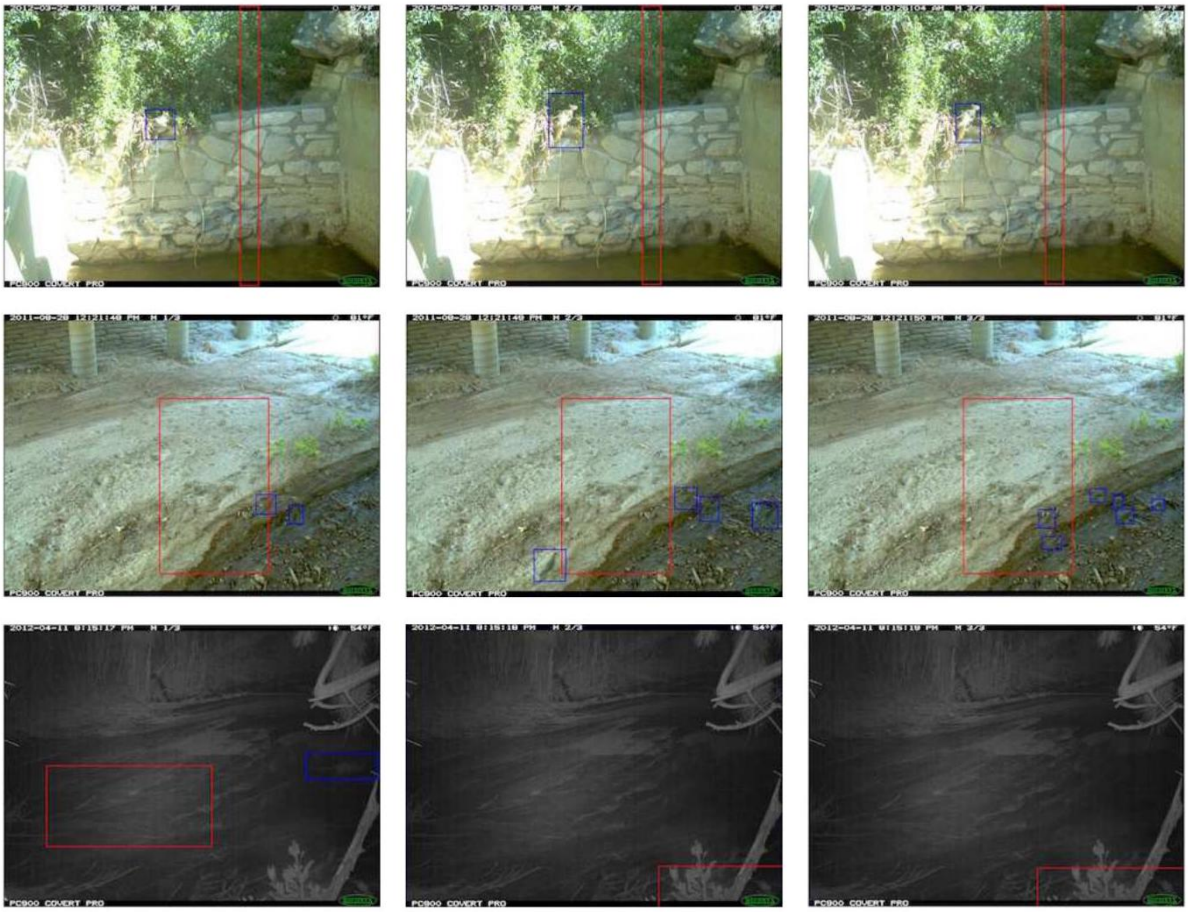


Fig. 8. Trans-detection failure cases at the sequence level: Highest-confidence detection in red, ground truth in blue. In all cases the confidence of the detection was lower than 0.2 . The first two sequences have small ROI, compounded with challenging lighting in the first and camouflaged birds in the second. In the third the opossum is poorly illuminated and only visible in the first frame.

图 8. 序列级别的跨检测失败案例: 最高置信度的检测用红色标出，真实值用蓝色标出。在所有案例中，检测的置信度均低于 0.2。前两个序列的 ROI 较小，第一序列的光照条件具有挑战性，第二序列中鸟类则伪装得很好。在第三个序列中，负鼠的照明条件较差，仅在第一帧中可见。

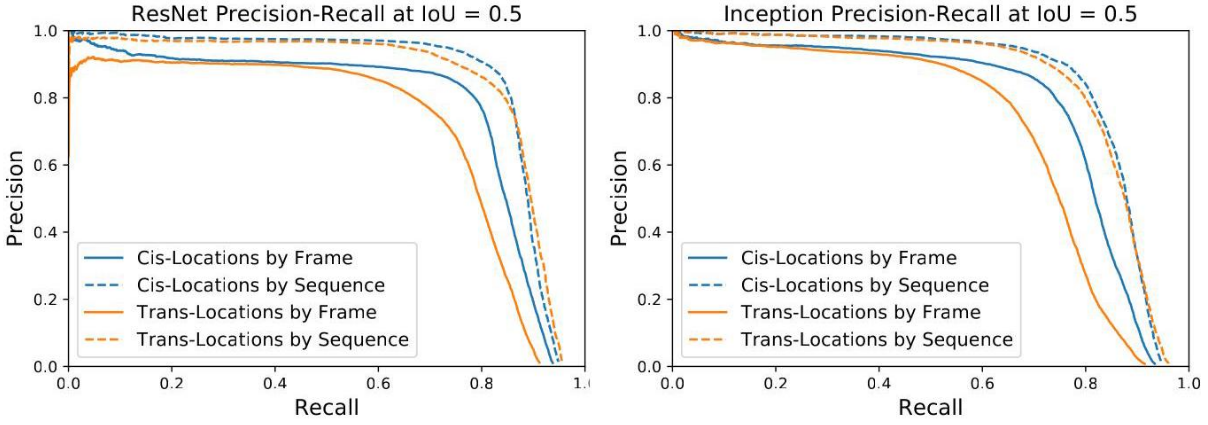


Fig. 9. Faster-RCNN precision-recall curves at an IoU of 0.5, by frame and by sequence, using a confidence-based approach to determine which frame should represent the sequence

图 9. 在 IoU 为 0.5 时，Faster-RCNN 的精度-召回曲线，按帧和按序列，使用基于置信度的方法来确定哪个帧应代表该序列。

# 5 Conclusions

# 5 结论

The question of generalization to novel image statistics is taking center stage in visual recognition. Many indicators point to the fact that current systems are data-inefficient and do not generalize well to new scenarios. Current systems are, in essence, glorified pattern-matching machines, rather than intelligent visual learners.

对新图像统计的泛化问题在视觉识别中正日益成为中心议题。许多指标表明，当前系统在数据效率上存在不足，并且在新场景中泛化效果不佳。当前系统本质上是被美化的模式匹配机器，而不是智能视觉学习者。

Many problem domains face a generalization challenge where the test conditions are potentially highly different than what has been seen during training. Self driving cars navigating new cities, rovers exploring new planets, security cameras installed in new buildings, and assistive technologies installed in new homes are all examples where good generalization is critical for a system to be useful. However, the most popular detection and classification benchmark datasets [9, 10, 28, 29] are evaluating models on test distributions that are the same as the train distributions. Clearly it is important for models to do well on data coming from the same distribution as the train set. However, we argue that it is important to characterize the generalization behavior of these models when the test distribution deviates from the train distribution. Current datasets do not allow researchers to quantify the generalization behavior of their models.

许多问题领域面临着一个泛化挑战，即测试条件可能与训练期间所见的条件有很大不同。自主驾驶汽车在新城市中导航、探测器在新行星上探索、安装在新建筑中的安全摄像头以及安装在新家中的辅助技术都是需要良好泛化能力的系统才能发挥作用的例子。然而，最流行的检测和分类基准数据集 [9, 10, 28, 29] 正在评估模型在与训练分布相同的测试分布上的表现。显然，模型在来自与训练集相同分布的数据上表现良好是重要的。然而，我们认为，当测试分布偏离训练分布时，表征这些模型的泛化行为是重要的。目前的数据集不允许研究人员量化他们模型的泛化行为。

We contribute a new dataset and evaluation protocol designed specifically to analyze the generalization behavior of classification and detection models. Our experiments reveal that there is room for significant improvement on the generalization of state-of-the-art classification models. Detection helps to improve overall classification accuracy, and we find that while detectors generalize better to new locations, there is room to improve their precision at high recall rates.

我们贡献了一个新的数据集和评估协议，专门设计用于分析分类和检测模型的泛化行为。我们的实验揭示了当前最先进的分类模型在泛化方面还有显著改进的空间。检测有助于提高整体分类准确性，我们发现尽管检测器在新位置的泛化能力更强，但在高召回率下提高其精度仍有改进的空间。

Camera traps provide a unique experimental setup that allow us to explore the generalization of models while controlling for many nuisance factors. Our current dataset is already revealing interesting behaviors of classification and detection models. There is still more information that we can learn by expanding our dataset in both data quantity and evaluation metrics. We plan to extend this dataset by adding additional locations, both from the American Southwest and from new regions. Drastic landscape and vegetation changes will allow us to investigate generalization in an even more challenging setting. Rare and novel events are frequently the most important and most challenging to detect and classify, and while our dataset already has these properties, we plan to define experimental protocols and data splits for benchmarking low-shot performance and the open-set problem of detecting and/or classifying species not seen during training.

相机陷阱提供了一个独特的实验设置，使我们能够在控制许多干扰因素的情况下探索模型的泛化能力。我们当前的数据集已经揭示了分类和检测模型的有趣行为。通过在数据量和评估指标上扩展我们的数据集，我们仍然可以学习到更多信息。我们计划通过增加来自美国西南部和新地区的额外位置来扩展这个数据集。剧烈的地形和植被变化将使我们能够在更具挑战性的环境中研究泛化。稀有和新颖事件通常是最重要且最具挑战性的检测和分类对象，尽管我们的数据集已经具备这些特性，但我们计划定义实验协议和数据划分，以基准低样本性能和检测和/或分类训练期间未见物种的开放集问题。

# 6 Acknowledgements

# 6 致谢

We would like to thank the USGS and NPS for providing data. This work was supported by NSFGRFP Grant No. 1745301, the views are those of the authors and do not necessarily reflect the views of the

NSF. Compute time was provided by an AWS Research Grant.

我们要感谢美国地质调查局和国家公园管理局提供的数据。本研究得到了国家科学基金会研究生研究奖学金 (NSFGRFP) 资助，资助编号为 1745301，观点仅代表作者个人，并不一定反映国家科学基金会的观点。计算时间由 AWS 研究资助提供。

## References

## 参考文献

1. Van Horn, G., Mac Aodha, O., Song, Y., Shepard, A., Adam, H., Perona, P., Be-longie, S.: The inaturalist challenge 2017 dataset. arXiv preprint arXiv:1707.06642 (2017)

2. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Packer, C., Clune, J.: Automatically identifying wild animals in camera trap images with deep learning. arXiv preprint arXiv:1703.05830 (2017)

3. van Horn, G., Barry, J., Belongie, S., Perona, P.: The Merlin Bird ID smartphone app (http://merlin.allaboutbirds.o

4. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639) (2017) 115

5. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R.: Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering (2018) 1

6. Fukushima, K., Miyake, S.: Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Springer (1982) 267-285

7. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11) (1998) 2278-2324

8. Schaller, R.R.: Moore's law: past, present and future. IEEE spectrum 34(6) (1997) 52–59

9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. (2009)

10. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740-755

11. Van Horn, G., Perona, P.: The devil is in the tails: Fine-grained classification in the wild. arXiv preprint arXiv:1709.01450 (2017)

12. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22** (10) (2010) 1345 − 1359

13. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1521-1528

14. Welinder, P., Welling, M., Perona, P.: A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3262-3269

15. Murphy, G.: The big book of concepts. MIT press (2004)

16. Ren, X., Han, T.X., He, Z.: Ensemble video object cut in highly dynamic scenes. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 1947-1954

17. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP Journal on Image and Video Processing **2013** (1) (2013) 52

18. Wilber, M.J., Scheirer, W.J., Leitner, P., Heflin, B., Zott, J., Reinke, D., Delaney, D.K., Boult, T.E.: Animal recognition in the mojave desert: Vision tools for field biologists. In: Applications of Computer Vision (WACV), 2013 IEEE Workshop on, IEEE (2013) 206-213

19. Chen, G., Han, T.X., He, Z., Kays, R., Forrester, T.: Deep convolutional neural network based species recognition for wild animal monitoring. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE (2014) 858-862

20. Lin, K.H., Khorrami, P., Wang, J., Hasegawa-Johnson, M., Huang, T.S.: Foreground object detection in highly dynamic scenes using saliency. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE (2014) 1125-1129

21. Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., Packer, C.: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. Scientific data **2** (2015) 150026

22. Zhang, Z., Han, T.X., He, Z.: Coupled ensemble graph cuts and object verification for animal segmentation from highly cluttered videos. In: Image Processing (ICIP), 2015 IEEE International Conference on, IEEE (2015) 2830-2834

23. Zhang, Z., He, Z., Cao, G., Cao, W.: Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. IEEE Transactions on Multimedia 18(10) (2016) 2079-2092

24. Miguel, A., Beery, S., Flores, E., Klemesrud, L., Bayrakcismith, R.: Finding areas of motion in camera trap images. In: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE (2016) 1334-1338

25. Giraldo-Zuluaga, J.H., Salazar, A., Gomez, A., Diaz-Pulido, A.: Camera-trap images segmentation using multi-layer robust principal component analysis. The Visual Computer (2017) 1-13

26. Yousif, H., Yuan, J., Kays, R., He, Z.: Fast human-animal detection from highly cluttered camera-trap images using joint background modeling and deep learning classification. In: Circuits and Systems (ISCAS), 2017 IEEE International Symposium on, IEEE (2017) 1-4

27. Villa, A.G., Salazar, A., Vargas, F.: Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. Ecological Informatics 41 (2017) 24-32

28. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision 88 (2) (2010) 303 – 338

29. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2017)

30. Ponce, J., Berg, T.L., Everingham, M., Forsyth, D.A., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B.C., Torralba, A., et al.: Dataset issues in object recognition. In: Toward category-level object recognition. Springer (2006) 29–48

31. Spain, M., Perona, P.: Some objects are more equal than others: Measuring and predicting importance. In: European Conference on Computer Vision (ECCV), Springer (2008) 523-536

32. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. (2011)

33. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2006) 1447-1454

34. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I., Soares, J.V.B.: Leafsnap: A computer vision system for automatic plant species identification. In: The 12th European Conference on Computer Vision (ECCV). (October 2012)

35. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 595-604

36. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: Subsense: A universal change detection method with local adaptive sensitivity. IEEE Transactions on Image Processing 24(1) (2015) 359–373

37. Babaee, M., Dinh, D.T., Rigoll, G.: A deep convolutional neural network for background subtraction. arXiv preprint arXiv:1702.01731 (2017)

38. Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., Qiu, X.: Change detection based on deep siamese convolutional network for optical aerial images. IEEE Geoscience and Remote Sensing Letters 14(10) (2017) 1845-1849

39. Benedek, C., Szirányi, T.: A mixed markov model for change detection in aerial photos with large time differences. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, IEEE (2008) 1-4

40. : emammal: a tool for collecting, archiving, and sharing camera trapping images and data. https://emammal.si.edu/ Accessed: 2018-03-13.

41. Csurka, G.: Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374 (2017)

42. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. (2015) 1180-1189

43. Gebru, T., Hoffman, J., Fei-Fei, L.: Fine-grained recognition in the wild: A multitask domain adaptation approach. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE (2017) 1358-1367

44. Busto, P.P., Gall, J.: Open set domain adaptation. In: The IEEE International Conference on Computer Vision (ICCV). Volume 1. (2017)

45. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)

46. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. arXiv preprint arXiv:1711.11556 (2017)

47. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: The IEEE International Conference on Computer Vision (ICCV). Volume 2. (2017) 6

48. Peng, X., Sun, B., Ali, K., Saenko, K.: Learning deep object detectors from 3d models. In: Computer Vision (ICCV), 2015 IEEE International Conference on, IEEE (2015) 1278-1286

49. Tang, K., Ramanathan, V., Fei-Fei, L., Koller, D.: Shifting weights: Adapting object detectors from image to video. In: Advances in Neural Information Processing Systems. (2012) 638-646

50. Sun, B., Saenko, K.: From virtual to reality: Fast adaptation of virtual object detectors to real domains. In: BMVC. Volume 1. (2014) 3

51. Hattori, H., Boddeti, V.N., Kitani, K., Kanade, T.: Learning scene-specific pedestrian detectors without real data. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 3819-3827

52. Xu, J., Ramos, S., Vázquez, D., López, A.M.: Domain adaptation of deformable part-based models. IEEE transactions on pattern analysis and machine intelligence 36(12) (2014) 2367-2380

53. Raj, A., Namboodiri, V.P., Tuytelaars, T.: Subspace alignment based domain adaptation for rcnn detector. arXiv preprint arXiv:1507.05578 (2015)

54. Grant Van Horn, Scott Laurie, S.B., Perona, P.: Lean multiclass crowdsourcing. Computer Vision and Pattern Recognition (2018)

55. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural networks: Tricks of the trade. Springer (2012) 437-478

56. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2818-2826

57. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE CVPR. (2017)

58. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770-778