THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Focus on the Positives: Self-Supervised Learning for Biodiversity Monitoring

Citation for published version:
Pantazis, O, Brostow, GJ, Jones, KE & Mac Aodha, O 2022, Focus on the Positives: Self-Supervised Learning for Biodiversity Monitoring. in *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision ICCV 2021 .* 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, pp. 10563-10572, International Conference on Computer Vision 2021, 11/10/21. https://doi.org/10.1109/ICCV48922.2021.01041

Digital Object Identifier (DOI):
10.1109/ICCV48922.2021.01041

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of 2021 IEEE/CVF International Conference on Computer Vision ICCV 2021

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 25. Apr. 2022

This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Focus on the Positives: Self-Supervised Learning for Biodiversity Monitoring

Omiros Pantazis[1] Gabriel J. Brostow[1,2] Kate E. Jones[1] Oisin Mac Aodha[3]
[1]University College London [2]Niantic [3]University of Edinburgh
www.github.com/omipan/camera_traps_self_supervised

Abstract

We address the problem of learning self-supervised representations from unlabeled image collections. Unlike existing approaches that attempt to learn useful features by maximizing similarity between augmented versions of each input image or by speculatively picking negative samples, we instead also make use of the natural variation that occurs in image collections that are captured using static monitoring cameras. To achieve this, we exploit readily available context data that encodes information such as the spatial and temporal relationships between the input images. We are able to learn representations that are surprisingly effective for downstream supervised classification, by first identifying high probability positive pairs at training time, i.e. those images that are likely to depict the same visual concept. For the critical task of global biodiversity monitoring, this results in image features that can be adapted to challenging visual species classification tasks with limited human supervision. We present results on four different camera trap image collections, across three different families of self-supervised learning methods, and show that careful image selection at training time results in superior performance compared to existing baselines such as conventional self-supervised training and transfer learning.

1. Introduction

Learning transferable representations of visual data without requiring explicit semantic supervision at training time is an important and open problem in computer vision. Recent progress on this front has been impressive, resulting in self-supervised methods that are capable of learning features that approach, and in some cases even surpass, their fully supervised counterparts across a range of downstream tasks [22, 17]. As a result of not having access to any semantic supervision (e.g. discrete category labels in the case of image classification), current best performing self-supervised methods typically use aggressive image augmentation strategies to generate different "views" of an input image during training [57, 26, 10, 23, 11]. The training time objective then consists of pulling these distinct "views" of the same image close to one another in feature space. The ability to generate plausible image variation using these manually designed image augmentation strategies would then appear to be a limiting factor in the further advancement of self-supervised learning (SSL).

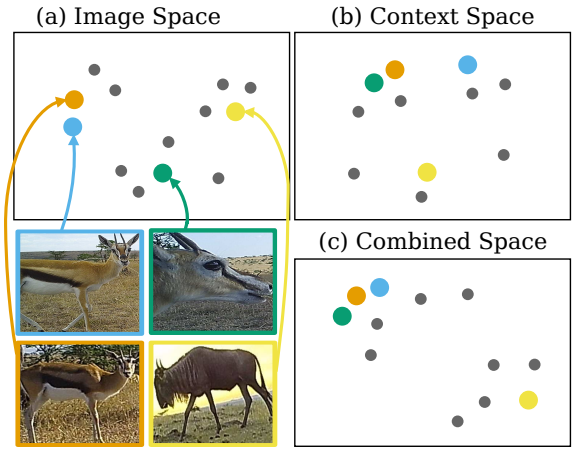

Figure 1. (a) Conventional self-supervised methods are capable of bringing visually similar examples closer in image embedding space. (b) There is often rich context information (e.g. encoding where and when an image was captured) that can also convey similarity. (c) By combing the complimentary nature of both signals, we can improve the quality of the final embedding space.

Current state-of-the-art self-supervised methods have predominantly been designed using image collections originally constructed for *supervised* learning e.g. [15, 62]. This has necessitated the exploration of different augmentation strategies to introduce appearance variation during training. However, a more natural signal to use is to exploit the fact that image observations that are made close in time and space are very likely to contain the same object instance. This form of natural variation has been used in self-supervised learning from video [43, 38, 39] or spatially distributed image collections [28, 3]. More generally, one can think of having access to not only a collection of images at training time, but also potentially rich context information pertaining to when and where each image was captured, in addition to other cues.

The central question that we address in this paper is how to make use of this context information during self-supervised learning to select more useful and varied image pairs at training time. The aim is to provide the self-supervised algorithm with "natural" positive image pairs and thus to establish connections in the latent feature space that were not possible with conventional augmentation-based approaches (Figure 1). We evaluate several different approaches and show, perhaps surprisingly, that the choice of images has more of an impact on performance than the underlying self-supervision algorithm. Our analysis is applicable to any self-supervised method that attempts to maximize the similarity between two "views" of the same visual concept, even those that do not require negative sample pairs [11].

We focus our evaluation on image collections that have been captured using camera traps - also known as "wildcams" or "trail cameras". These types of images are commonly collected for the purpose of biodiversity monitoring [58, 35, 6, 48, 7, 21]. Unlike more conventional image datasets typically used by the computer vision community, camera trap images exhibit some interesting properties that make them particularly well suited to evaluating self-supervised learning: (i) Camera trap images are not captured by humans directly, instead the cameras are automatically triggered based on the proximity of nearby animals. This overcomes the "iconic" object view bias that tends to be prevalent in datasets like ImageNet [15] or iNaturalist [52, 51]. They also contain other challenges such as partial depictions of objects due to occlusion, significant scene illumination changes, and strong object and location correlations [6]; (ii) The images are often captured in short temporal bursts and come with rich context data in the form of the time of day, the time of year, and location, and may also include additional information characterizing the local habitat. This information can provide useful clues about what animal species are potentially present at a given location; (iii) A conservative estimate is that tens of thousands of ever-active cameras are deployed globally [46], necessitating large amounts of tedious work on the part of ecologists and conservation biologists to manually annotate the incoming images or to correct errors made by automatically generated classifier predictions. Informative image features, derived from self-supervised learning, could significantly reduce this manual effort and have the potential to be an important tool in aiding the critical task of scalable global biodiversity monitoring.

We make the following three contributions:

1. We explore the benefits of self-supervised learning on four challenging camera trap datasets. We observe that self-supervised features are significantly more effective on average for downstream classification compared to widely adopted transfer learning baselines.
2. We show that how images from these datasets are selected during self-supervised training has much more impact on the quality of learned features compared to the choice of the underlying self-supervised training loss that is used.
3. While the role of *negative* image pairs in self-supervised learning has received significant attention, we show that current methods are surprisingly robust to incorrectly selected *positive* image pairs during training. This provides an important insight for the design of future self-supervised methods.

2. Related Work

2.1. Self-Supervised Learning

The goal of self-supervised learning (SSL) from visual data is to learn a function that can extract semantically meaningful image representations, without requiring any semantic annotations at training time. Until recently, the main focus of SSL research in vision centered on designing proxy tasks that, when solved at training time, would result in useful features for downstream transfer learning. Examples of approaches include visual puzzle solving [34], image colorization [60, 61], image in-painting [37], and image rotation prediction [20], to name a few.

There has been recent excitement for contrastive-based approaches [25, 24, 36]. Instead of requiring elaborate proxy tasks, contrastive methods learn features by pushing positive image pairs close to each other in feature space, while pushing negative pairs away from each other. Given semantic information [42, 30], positive and negative pairs can be defined based on object category labels, or some other form of semantic supervision. However, for SSL, this information is not available. As a result, the current most common solution is to artificially augment images with random transformations [16] to create positive and negative image pairs. Specifically, positive pairs are differently augmented versions of the same image, while negative pairs are augmented versions of any two *different* images. The assumption is that the selected augmentation space preserves semantic content, while introducing noise that the feature extractor will learn to tolerate through invariance.

Contrastive-based methods require negative image pairs, and various strategies have been proposed to ensure that useful negatives (i.e. hard negatives) are selected during training. Common strategies to ensure challenging negatives are selected include non-parametric memory banks [57, 63], momentum encoders [26], auxiliary modalities [50], local region rather than full image reasoning [36, 4], learning data prototypes online [9], or simply training with large batch sizes [10]. The body of work in this space is growing fast, see [41] for an overview.

One of the typical assumptions made by conventional contrastive-based approaches, when constructing negative

pairs, is that randomly sampled images from the same batch, or from a non-parametric memory, do not depict the same semantic content. In practice however this assumption is often violated, and thus results in a large number of false negative pairs during training. Recent research [14] has quantified the impact of this assumption by reporting performance gains under the presence of oracle (i.e. ground truth) negative labels and proposed an unsupervised negative sampling technique that partially improves performance by correcting for the introduced error. Another study [40] also observed the importance of hard negatives during SSL and introduced a user controllable hard negative sampling approach that improves over the baseline of uniform sampling. [29] showed that hard negatives are required for effective SSL and proposed an unsupervised feature space mixing method inspired by [59], to generate challenging training examples.

Given the aforementioned obstacles associated with effective negative sample selection, there have been recent attempts to forego the need for negative image pairs during SSL. One set of methods involve using additional positive prediction-based networks [23] or layers [11] during training. These approaches compete with contrastive methods, and have the added benefit of not requiring large batch sizes as they do not need to sample from a large pool of negatives. In light of this diminished reliance on negatives, we turn our attention to the role of the *positive* image pairs. We show that, perhaps unsurprisingly, there is a large performance gap between standard SSL methods and oracle baselines (i.e. given ground truth labels). By making use of context information, we can select positive pairs that exhibit more visual diversity compared to standard augmentation resulting in improved downstream performance.

2.2. Context-Aware Classification

While there are notable exceptions, for the most part, the computer vision community has predominately focused on using benchmark image datasets that lack additional data beyond the images themselves and their task-specific supervised annotations e.g. [15, 32]. Images acquired in the wild, as opposed to scraped from the web, often come with valuable metadata, encoding where and when the images were captured, along with other potentially valuable cues. We broadly refer to this free weak supervision as *context* data.

There are many instances where specific visual concepts can only be disambiguated by knowing where the images were captured. For example, in case of species identification, some categories may only be correctly classified when the image location is known at test time. Existing supervised works have explored the use of this geographical context via non-parametric density estimation to model the spatial distribution of object categories [8]. Others use deep networks with integrated geographical encoding layers [49, 13], or factorized models that reason about object spatial distributions in a latent feature space [33]. The previously mentioned approaches are all supervised, but geographical proximity has also been successfully used as a signal for SSL. In the case of representation learning from aerial images, recent analyses [3, 28] make the assumption that geographically close-by locations should be more similar in their latent features compared to more distant locations, and show that this results in superior image features.

Another form of readily accessible contextual information are temporal cues. For short temporal changes (e.g. nearby frames in a video), the observed raw pixel data might vary significantly from frame to frame, but it is often reasonable to assume that the underlying latent content being observed changes more slowly [55]. The assumption that nearby frames contain the same object instance has previously been successfully exploited in camera trap images by merging predictions from nearby time steps at test time [6]. Long range temporal reasoning is also a common component of supervised video analysis [56, 18], but without supervision it is more challenging to make use of this information. Combined spatial (in image space) and temporal reasoning was recently shown to be effective for the problem of object detection in static camera traps [7]. By exploiting the fact that objects (i.e. people or animals) may exhibit similar behavior over time at the same location, the proposed attention-based mechanism was able to use this signal to improve detection performance at test time, where no supervision is available. However, supervision in the form of bounding boxes and ground truth category labels are still required at training time.

Several approaches have been proposed for SSL in video, including posing the learning problem as a prediction task e.g. predicting future events [45], predicting motion and appearance statistics [53], predicting sequence order [19, 31], and also contrastive learning [39]. While the camera trap datasets that we use are related to video datasets, in that they also contain some temporal information, the frames are not uniformly sampled over time, but instead, the cameras are trigger-based events in the scene.

In this work, we evaluate different mechanisms for selecting positive image pairs during SSL. We also present a context-based image selection approach for choosing high probability positive pairs at training time to learn more effective self-supervised visual features. To achieve this, we make use of both spatial and temporal context in an unsupervised fashion.

3. Method

In self-supervised visual representation learning, we assume that we have a set of images $x_i \in \mathcal{X}$ at training time, but do not have access to any associated ground truth supervision e.g. object category labels. Our goal is to learn the parameters of a feature extractor f, so that when we apply

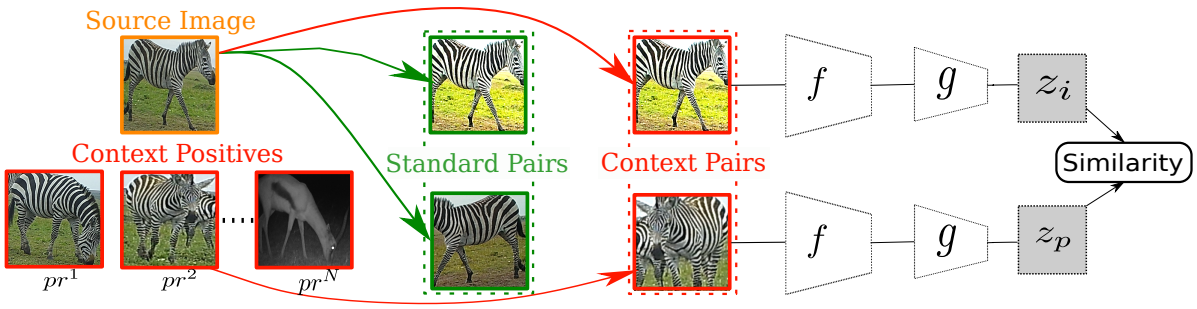

Figure 2. Overview of our approach. Conventional SSL typically generates augmented pairs of an input source image. Those are then passed to the model (right). Instead, we advocate that the careful selection of image pairs based on their context similarity (e.g. in space or time), generates more varied and useful image pairs that result in more informative visual representations.

the network on an image x_i, we obtain a feature vector that can be used for downstream tasks. Alternatively, we may also wish to fine-tune the feature extractor in an end-to-end fashion with additional supervision under the assumption that this will serve as a better initialization than randomly initialized parameters. In the case of images, the feature extractor f can be parameterized as a deep convolutional neural network, e.g. a deep Residual Network [27].

To overcome the lack of supervision at training time, many recent approaches rely on variants of augmentation-based SSL in an attempt to learn expressive features from the raw image data. Specifically, the standard pipeline involves taking a source image x_i and performing two sets of synthetic augmentations on the image (e.g. cropping, flipping, color jittering, *etc.*) to create two alternative "views" of the source image $(\tilde{x}_i, \tilde{x}_p)$. See [10] for representative augmentations. We refer to these two images as a positive augmented pair. The goal of the augmentations is to introduce visual diversity, while ideally preserving the underlying semantic information contained in the source image. During self-supervised training, the learning algorithm attempts to push the positive pair of images close to each other in feature space. The intuition is that we want to learn a feature extractor f that projects images containing the same visual concepts to similar regions in feature space. Later, we describe different approaches for selecting varied positive pairs. An overview of our pipeline is shown in Figure 2.

3.1. Self-Supervised Losses

Here we outline three broad methods for SSL that serve as the basis for our experimental evaluation. Common to each method is that they take the output of the feature extractor f and pass it through a projection network g, resulting in a lower-dimensional embedding vector $z_i = g(f(\tilde{x}_i))$. This projection network is commonly represented as a fully connected multi-layer neural network [10].

Triplet [54, 42]. The conventional triplet-based margin loss aims to push positive embedding pairs (z_i, z_p) close to each other, while also pushing negative pairs (z_i, z_n) further away from each other. Here, a "negative" pair, is a pair derived from two different images $(\tilde{x}_i, \tilde{x}_n)$, where $i \neq n$. In the language of triplet-based learning, we want the distance between the anchor z_i and the positive z_p to be smaller than the distance between the anchor and the negative z_n. The triplet loss is defined as

$$L_i^T = \max\left(D(z_i, z_p) - D(z_i, z_n) + m, 0\right), \quad (1)$$

where m is a scalar representing the desired margin of the loss, and $D(.)$ is an appropriate distance function, e.g. Euclidean distance or negative cosine similarity. The triplet loss requires positive pairs and, crucially, negative pairs. For SSL, we can generate positive pairs through image augmentation, and a common strategy for selecting negative pairs, in the absence of any other knowledge, is to simply randomly sample them from the entire dataset. While not commonly used for SSL, later we show that the triplet loss is surprisingly effective when compared to recent methods.

SimCLR [10]. Building on related approaches [44, 57], at the core of SimCLR is the use of a normalized temperature-based softmax cross-entropy loss. As with the triplet loss, the aim is to minimize the positive pair distance, while maximizing the negative pair distance. However, unlike the triplet loss, where we select one negative pair for each positive one, here we select a set of negatives. Unlike methods that maintain a memory bank containing embeddings for the entire training set to select negatives from [57], SimCLR simply uses large training batches and samples the negatives for each source instance from within the same batch. The loss used in [10] over a set of items $\mathcal{B}$ is defined as

$$L_i^S = -\log\left(\frac{\exp(-D(z_i, z_p)/\tau)}{\sum_{n\in\mathcal{B}} \mathbb{1}_{[n\neq i]} exp(-D(z_i, z_n)/\tau)}\right), \quad (2)$$

where τ is a temperature parameter used to control the scale of the distances and $\mathbb{1}_{[n\neq i]}$ ensures that an embedding vector is not compared with itself. Importantly, here the source-positive pair is also part of the sum in the denominator to ensure that the loss is normalized correctly i.e. $p \in \mathcal{B}$.

SimSiam [11]. SimSiam is a non-contrastive self-supervised method that innovates on the previous two losses

by removing the need for negative pairs during training. This is an important simplification because, as stated earlier, choosing appropriate negative samples for SSL is challenging. [11] introduced an additional fully connected network h, termed the prediction network. The goal of this network is to predict the embedding vector of the other image in the input pair. To avoid collapsing to a degenerate solution in the absence of negative examples, SimSiam applies a bottleneck structure in the prediction network and a stop-gradient operation [23] to prevent information flowing back through the projection and feature networks for a given image in the input pair. During training, the goal is to maximize agreement between the embedding vector z_i and the prediction $h(z_p)$ using the following symmetrized loss

$$L_i^M = \frac{1}{2}D(h(z_i), \texttt{sg}(z_p)) + \frac{1}{2}D(h(z_p), \texttt{sg}(z_i)). \quad (3)$$

Here, sg represents the application of the stop-gradient. Despite not requiring any negative image pairs, SimSiam still performs competitively compared to contrastive approaches, even when using more modest batch sizes [11].

3.2. Selecting Positive Image Pairs

One commonality shared by the three previous self-supervised methods is that they each require positive image pairs during training i.e. (x_i, x_p). Given only images, there are limited options for how to create these positive pairs. However, when additional information, such as time or location is available, this opens the door to alternative methods for selecting training pairs.

Image Augmentations. The standard approach used by many SSL methods that only train on unordered image collections is to create positive image pairs by stochastically augmenting the source images during training i.e. $(\tilde{x}_i, \tilde{x}_p)$, where $i = p$. Note that we also apply these augmentations to each of the following selection mechanisms, even if the image pairs are different, i.e. when $i \neq p$.

Oracle Positive Selection. If category labels were available, one could use them to select positive image pairs [30]. Here, we assume that our input image collection consists of pairs of (x_i, y_i), where y_i is the corresponding category label for image x_i. For each image x_i, x_p is selected from the set of images belonging to the same category i.e. $y_i = y_p$. This represents an idealized setting, and is not practical for actual SSL where this supervision is absent. However, it provides useful insight, and gives us an upper bound on performance. For methods that also require negative pairs, one could also select them using a similar procedure [14], i.e. $y_i \neq y_n$. Our focus is on the role of positive images, and as a result, we select negative pairs using random sampling in all cases. Later, we explore variants of this oracle positive selection mechanism where we introduce noise, so that with some probability λ, we intentionally select p to be from a different category, i.e. $y_i \neq y_p$.

Sequence Positives. Given sequential input data e.g. a video or burst of frames, one strategy is to select positive pairs based on their proximity in time. Here we choose images as pairs if they are from the same location and are within a specified number of frames or a unit of time from each other. We also ensure that the source image itself can be selected. This has the advantage of introducing more natural visual diversity compared to what is possible from conventional image-based augmentation. The disadvantage of this approach is that one must commit to a hard threshold on the number of frames, or unit of time, when deciding what constitutes a potential positive pair.

Context-Based Selection. In many real world datasets (e.g. camera traps) there can be extra information available in addition to the raw images. We can assume that each image x_i is associated with a K dimensional context vector c_i. This vector could include information such as an encoding of the geographical location of the image, date, time of day, the pixel coordinates of the image (i.e. it could be cropped from a larger image), *etc*. Instead of having to specify thresholds for each of these different dimensions, a simpler approach is to define a distance measure between a pair i and j,

$$D^c(c_i, c_j) = \sqrt{\sum_{k=1}^{K}(c_i^k - c_j^k)^2}. \quad (4)$$

When selecting a positive pair at training time for a source image i, we simply construct a distribution over all pairs

$$pr_i^j = \frac{\exp(-D^c(c_i, c_j)/\tau_c)}{\sum_{n=1}^{N}\exp(-D^c(c_i, c_n)/\tau_c)}. \quad (5)$$

Here, N is the number of images in the unlabeled training set, which includes i and j, and τ_c is another temperature hyperparameter. Then to select a positive item p for each i, we sample from a categorical distribution parameterized by the vector pr_i, i.e. $p \sim \text{Categorical}(pr_i)$. As a result, items that are close in context space will have a higher probability of being selected as a positive pair. Note that there is also a chance that we sample the same image as the source, i.e. $p = i$. As we perform random augmentations on each view of the source image, in cases where $p = i$, this results in the standard image augmentation-based selection mechanism.

4. Experiments

4.1. Datasets

We perform experiments on four representative camera trap datasets below and one satellite dataset [12] in the supplementary. The camera trap datasets exhibit complementary properties, e.g. collected in multiple countries versus one region, or collected over long (years) versus shorter (months) time periods. Camera trap images contain many difficulties that make automatic identification particularly

challenging [6]. While other benchmark datasets contain some useful context data related to when and where the images were taken e.g. [49, 52, 51], they do not explicitly contain sequences of images captured over time from the same location. We assume that the images are manually cropped around the objects of interest, since we are concerned with the problem of classification, not detection. High precision detectors are available, specifically tuned for camera trap images, in cases where ground truth bounding boxes are unavailable [5]. Image counts below refer to cropped images. The test splits for all datasets, except Snapshot Serengeti, were generated according to the protocol in [6]. This is designed to evaluate generalization behavior across novel camera trap sites. Each dataset consists of a set of locations (i.e. camera trap deployments), with multiple images from each location. Some locations are shared between the train and test sets and some are only present in the test set.

Caltech Camera Traps (CCT20) [6]. CCT20 contains images from 20 different sites in the Southwestern United States. In total, the dataset includes 15 species such as rabbits, raccoons, coyotes, bobcats, and others. There are 10 distinct camera trap locations in the test set and 10 that are shared between train and test, resulting in 12,617, 3,436, and 32,050 images in train, validation, and test.

Island Conservation Camera Traps (ICCT) [2]. ICCT consists of images taken across seven different islands, spanning six countries. Due to the global distribution of the dataset, it contains diverse ecosystems such as tropical, dry, and temperate. There are 116 locations, with 56 observed for training. We retain 12 species that had at least 100 appearances in the training locations, resulting in 11,378, 1,684, and 41,527 images in train, validation, and test.

Snapshot Serengeti (Serengeti) [47]. This contains cameras located around the Serengeti National Park in Tanzania, collected over multiple seasons. We use the subset that includes bounding box annotations. We use the author-provided train and test split, and retain at most 1,000 full images (i.e. not crops) for each species, keeping species that have at least 100 appearances. This results in 39 species, with 32,702 train and 8,492 test cropped images, from 179 train and 45 test locations. In contrast to the other datasets, the train and test locations do not overlap.

Maasai Mara Camera Traps (MMCT). Our final dataset is a fixed subset of our own ongoing collection efforts, which we will make available. MMCT contains images from 176 sites across the Maasai Mara in Kenya, which is primarily an open savanna and is known for its rich wildlife. It shares some ecosystem characteristics with Snapshot Serengeti, but it exhibits a larger class imbalance and more domesticated animals, given that the study was designed to study human impact on the environment. It contains 20 species of mammals, ranging from wildebeests, elephants, through to giraffes, and includes many threatened species. In total, there are 10,243 train images from 85 locations, along with 3,306 validation and 31,841 test images.

4.2. Implementation Details

For all experiments, we use a ResNet18 [27] as our feature extractor and add a two layer MLP for the projector g. For SimSiam, we use an additional MLP for the predictor h. As our input images are already cropped around the object of interest, we train with images of size 112×112. During self-supervised training, we augment the images with a set of transformations, similar to those used in SimCLR [10]. Unlike existing SSL methods that train from randomly initialized weights, we start from ImageNet pretrained weights and train for an additional 200 epochs with a batch size of 256. Results with random initialization are available in the supplementary material. A minor contribution is showing that SSL is still effective in this regime.

For each individual camera trap image, we construct a corresponding context vector c. While many different context variables are possible, for now we restrict ourselves to those encoding time and location. We encode date and time using wrapped coordinates as in [33]. For many camera trap datasets, the GPS coordinates of each camera are often kept secret as some of the animals present many be endangered. To overcome this, we use a one hot encoding to represent the deployment ID.

To evaluate the effectiveness of the different SSL approaches, we use the standard linear evaluation protocol [60, 36, 10]. Specifically, after self-supervised training we only keep the backbone feature extractor f, and use it to compute features for the training and test images. We then train a linear classifier using these features, and evaluate performance across the low to high data regimes i.e. by using 1%, 10%, or 100% of the available training labels. Unlike standard computer vision datasets e.g. [15], camera trap datasets tend to be heavily imbalanced, making this a particularly challenging task when there are few labels. Additional implementation details, hyperparameters, and experiments can be found in our supplementary material.

4.3. Results

Oracle Positive Selection. We first validate our hypothesis that SSL on camera traps data can benefit from seeing natural variation in the positive pairs when compared to standard image augmentations. We consider an idealized scenario where an oracle has access to the ground truth species labels during self-supervised training. We then sample positive pairs using the oracle selection strategy outlined in Section 3.1. In Figure 3, we observe a significant gap between standard SimCLR, which generates image pairs by augmenting the source image alone, and an oracle variant of SimCLR ("Oracle +ve") that selects true positives based on the ground truth species labels. More specifically, for MMCT, the top-one accuracy difference ranges from $\sim$9%

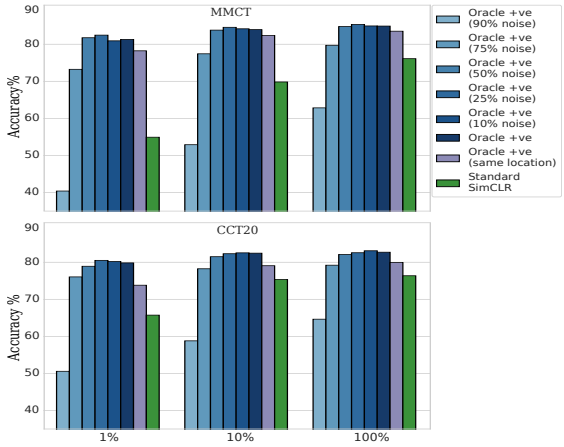

Figure 3. Oracle positive pair selection with SimCLR. We observe that SimCLR is surprisingly robust to label noise in the oracle positive selection setting. Performance only significantly deteriorates when the amount of noise is 90%. 1%, 10%, and 100% refer to proportions of supervised labeling used for evaluation after SSL.

and ∼28%, and between ∼6% and ∼15% for CCT20 across the different amounts of downstream supervision. It is worth noting that the oracle selection mechanism can select pairs from across the entire training set (i.e. from different locations and different points in times). We also include an oracle baseline that can only select positive pairs from the same camera trap deployment (i.e. the same physical location but the time could vary by as much as months). "Oracle +ve (same location)" is also significantly better than standard SimCLR. These results are encouraging as it provides evidence that more sophisticated methods for grouping images based on context from the same deployment (e.g. [7]) could result in further performance improvements.

Robustness to Oracle Noise. The oracle positive baseline in the previous section is an unreasonably strong and unrealistic baseline. We also evaluate the case where there is noise associated with the positive pair selection. Specifically, in Figure 3, "Oracle +ve (X% noise)" indicates that X% of the time, the oracle makes a mistake, pairing two images from different classes, i.e. $y_i \neq y_p$. Perhaps surprisingly, we observe across both datasets, that SimCLR is quite robust to this error, at least until the noise percentage gets very large e.g. 90%. Clearly, there is a benefit from being able to select images pairs that exhibit visual variety that would not be possible to create through image augmentation alone. This motivates our objective of selecting more varied positives by means of context-based mining.

Impact of Positive Selection. In Figure 5 we present an extensive set of experiments across four different camera trap datasets, three SSL methods, and three positive pair selection approaches. The first observation is that standard SSL results in superior representations for downstream linear classification compared to standard ImageNet initialized features in nearly all settings ("Standard" versus "ImageNet Init"). The second observation is that training with more "natural" positives is superior to standard SSL in almost all cases ("Sequence Positives" or "Context Positives" versus "Standard"). In many instances, this performance difference can be $> 5\%$ top one accuracy. Our "Context Positives" approach is consistently ranked first or second (excluding end-to-end supervised training which is provided for reference). We speculate that further gains may be achieved from using richer context information. Figure 4 shows examples of the nearest neighbors retrieved from the different models. We can see that "Context Positives" results in more visual diversity compared to standard SSL.

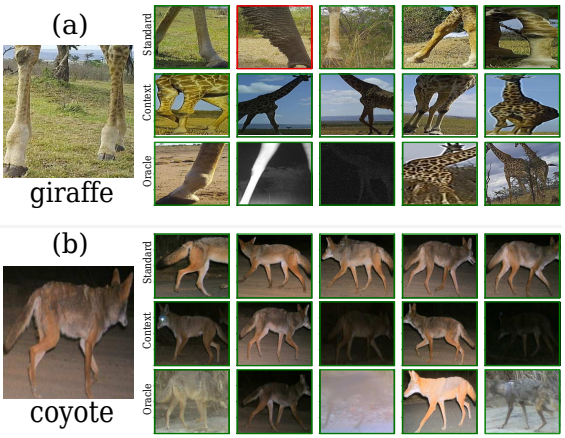

Figure 4. Here we show the top five nearest neighbors from the test set in the 128-dimensional embedding space (i.e. the output of the projector) from (a) MMCT and (b) CCT20, for three SimCLR variations. We see that the nearest neighbors for 'Standard' SimCLR display limited visual diversity. The unobtainable 'Oracle' model, which has been trained with ground truth labels, has the most variety. Our 'Context' approach is between the two extremes and shows non-trivial diversity which indicates that it contains more semantic information in its features compared to the standard augmentation-based SimCLR. Green and red outlines correspond to similar and different class neighbors respectively.

Impact of Self-Supervised Algorithm. One important observation from Figure 5 is that the choice of positive selection mechanism has more impact on downstream classification accuracy than the actual self-supervised method used. In fact, we show that even the conventional triplet loss when combined with any type of context information, results in better performance compared to recent state-of-the-art methods like SimCLR or SimSiam, when they only use standard image augmentation. This indicates that for "in the wild" datasets like camera traps, additional attention should be given to alternative approaches for positive image pair selection. This is further motivated by the large performance gap that still exists between the self-supervised methods and the oracle baselines in Figure 3. Moreover, we observe that between the three self-supervised approaches, given the proposed positive pair selection, SimCLR performs slightly better on average.

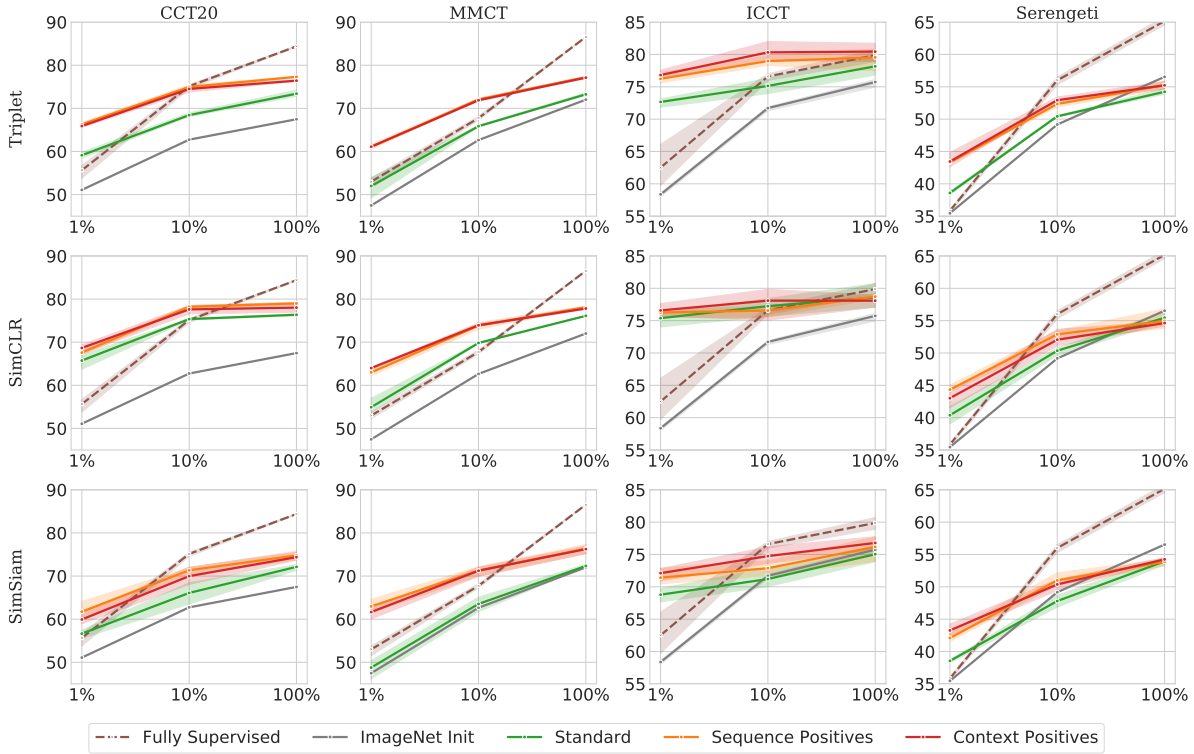

Figure 5. Top-one classification accuracy across four different camera trap datasets (columns) and three different SSL algorithms (rows). Each line represents a different mechanism for selecting positive image pairs during self-supervised training. Accuracy is computed by evaluating linear classifiers trained on different amounts of supervision (i.e. 1, 10, or 100%) from the respective training sets. The error bars represent the average accuracy and standard deviation over three repeats of SSL. For reference, we also include fully supervised baselines, trained end-to-end, with the proportion of labels indicated on each x axis. We see that SSL methods significantly outperform ImageNet derived features in nearly all cases. Furthermore, using natural positives (Sequence Positives or Context Positives) is always better than standard image augmentation alone (Standard), and has more impact than the choice of SSL method.

Discussion and Limitations. The datasets used for our evaluation have far fewer classes compared to datasets such as ImageNet [15] or iNaturalist [52, 51]. However, we argue that they are in fact more representative of the thousands of medium scale datasets that are generated by biodiversity researchers around the world every month. We also assumed that our images are cropped around the target of interest. While camera trap datasets can be very large, the majority of the images are often empty. Pretrained object detectors [5] can be used to filter the images to only these boxes of interest (i.e. animals). We present results with pretrained detectors in the supplementary material. Currently, our context-based sampling approach assumes that all context dimensions are weighted equally. Given how effective we have shown context information to be for self-supervised learning, a natural question is how can we improve performance further e.g. via learned context weights. Without supervision, this is a challenging problem which we leave for future work.

5. Conclusion

We have explored the problem of self-supervised learning in camera trap datasets. We showed that these types of image collections are a valuable testbed for benchmarking advances in self-supervised learning as they are complementary to the standard benchmarks commonly used in vision. In fact, conventional wisdom from well-explored standard benchmarks does not transfer directly to these types of images, as we observe that how images are selected during training can have a larger impact than the specific self-supervised algorithm. We posit that further exploration of this setting will lead to a greater understanding of the potential for self-supervised learning for "in the wild" collected datasets, and importantly, will directly benefit critical applications such as biodiversity monitoring.

Acknowledgements. The MMCT images are collected as part of the Biome Health Project [1] funded by WWF-UK. Thanks to Holly Pringle, Emily Madsen, Alex Rabeau, and Georgia Cronshaw for the annotation of the MMCT data.

References

[1] Biome Health Project. https://www.biomehealthproject.com/. 8

[2] Island Conservation Camera Traps. http://lila.science/datasets/island-conservation-camera-traps. 6

[3] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. *arXiv preprint arXiv:2011.09980*, 2020. 1, 3

[4] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019. 2

[5] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019. 6, 8

[6] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. 2, 3, 6

[7] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *CVPR*, 2020. 2, 3, 7

[8] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014. 3

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 4, 6

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 1, 2, 3, 4, 5

[12] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 5

[13] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *ICCV Workshops*, 2019. 3

[14] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NeurIPS*, 2020. 3, 5

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 3, 6, 8

[16] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NeurIPS*, 2014. 2

[17] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *CVPR*, 2021. 1

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 3

[19] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, 2017. 3

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv:1803.07728*, 2018. 2

[21] Paul Glover-Kapfer, Carolina A Soto-Navarro, and Oliver R Wearn. Camera-trapping version 3.0: current constraints and future priorities for development. *Remote Sensing in Ecology and Conservation*, 2019. 2

[22] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019. 1

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 1, 3, 5

[24] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 2

[25] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6

[28] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *AAAI*, 2019. 1, 3

[29] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, 2020. 3

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 2, 5

[31] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[33] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *ICCV*, 2019. 3, 6

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2

[35] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig

Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *PNAS*, 2018. 2

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2, 6

[37] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[38] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020. 1

[39] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 1, 3

[40] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021. 3

[41] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access*, 2021. 2

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2, 4

[43] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *ICRA*, 2018. 1

[44] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. 4

[45] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 3

[46] Robin Steenweg, Mark Hebblewhite, Roland Kays, Jorge Ahumada, Jason T Fisher, Cole Burton, Susan E Townsend, Chris Carbone, J Marcus Rowcliffe, Jesse Whittington, et al. Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, 2017. 2

[47] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1), 2015. 6

[48] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 2019. 2

[49] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *ICCV*, 2015. 3, 6

[50] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2

[51] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *CVPR*, 2021. 2, 6, 8

[52] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 2, 6, 8

[53] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, 2019. 3

[54] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009. 4

[55] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 2002. 3

[56] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019. 3

[57] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 1, 2, 4

[58] Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A Jansen, Tianjiang Wang, and Thomas Huang. Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 2013. 2

[59] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3

[60] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2, 6

[61] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 2

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 2017. 1

[63] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, 2019. 2