Presence-Only Geographical Priors for Fine-Grained Image Classification

Oisin Mac Aodha  Elijah Cole  Pietro Perona
Caltech
www.vision.caltech.edu/~macaodha/projects/geopriors

Abstract

Appearance information alone is often not sufficient to accurately differentiate between fine-grained visual categories. Human experts make use of additional cues such as where, and when, a given image was taken in order to inform their final decision. This contextual information is readily available in many online image collections but has been underutilized by existing image classifiers that focus solely on making predictions based on the image contents.

We propose an efficient spatio-temporal prior, that when conditioned on a geographical location and time, estimates the probability that a given object category occurs at that location. Our prior is trained from presence-only observation data and jointly models object categories, their spatio-temporal distributions, and photographer biases. Experiments performed on multiple challenging image classification datasets show that combining our prior with the predictions from image classifiers results in a large improvement in final classification performance.

1. Introduction

Correctly classifying objects into different fine-grained visual categories is a challenging problem. In contrast to generic object recognition, it can require knowledge of subtle features that are essential for differentiating between visually similar categories. However, without having access to additional information that may not be present in an image, many categories can be visually indistinguishable. For example, the two toad species in Fig. 1 are similar in appearance but tend to be found in very different locations in Europe. Knowing where a given image was taken can provide a strong prior for what objects it may contain.

Most images that are captured and shared online today also come with additional metadata in the form of where they were taken, when they were taken, and who captured them. This information not only offers the possibility of helping to resolve ambiguous cases for image classification, but can also enable us to generate predictions of where, and when, different objects are likely to be observed.

Existing work that uses location information to improve classification performance either discretizes the input data into spatio-temporal volumes [5], store the entire training set in memory at inference time [64], or jointly train deep images classifiers along with corresponding location information [55]. Methods that discretize or store the raw training data do not scale well in terms of memory, and jointly training image classifiers with location information necessitates that location information is present at test time - which may not always be the case. We take inspiration from species distribution modeling (SDM) [20], and instead model a separate geographical prior that can be combined with the predictions of any image classifier. However, unlike many approaches to SDM that assume they have access to presence and absence information at training time (e.g. [56]), we make a more general assumption that only presence information is available i.e. we know where the categories have been observed, but have no explicit data regarding where they are not found.

In this work we make the following contributions: (1) An efficient spatio-temporal prior that jointly models the relationship between location, time of year, photographer, and the presence of multiple different object categories. (2) A novel presence-only training loss to capture these relation-
ships. (3) Experiments that show that combining the probabilistic predictions of image classifiers with our prior significantly improves the test time performance on challenging fine-grained image datasets.

2. Related Work

Here we discuss work related to spatio-temporal models that encode the location of a set of discrete object categories. We do not address methods that explore other uses of location information such as inferring where an image was taken given only the raw pixels [29, 60], or methods that use location to disambiguate visually similar places for image localization [59, 68].

Fine-Grained Image Classification

Correctly determining which one of multiple possible fine-grained categories is present in an image requires understanding the relationship between subtle visual features and the corresponding image-level category label e.g. [61, 35, 66, 58]. Existing approaches have investigated the modeling of parts [41, 71, 9, 70, 32], higher order feature interactions [40, 23], attention mechanisms [65, 72, 62], noisy web data [37], novel training losses [14], and pairwise category information [17]. Orthogonal to those works, we propose a spatio-temporal prior that can be combined with the probabilistic predictions of any image classifier to improve the final classification performance.

Location and Classification

A small number of approaches have explored the use of location information to improve image classification at test time. Berg et al. [5] proposed a spatio-temporal prior that when combined with the output of an image classifier increased the accuracy of bird species classification. Their approach discretized location and time into spatio-temporal cubes and used an adaptive kernel density estimator to represent the distribution of each species independently. Also in the context of predicting the presence of different biological species, Wittich et al. [64] evaluated different nearest neighbor based lookup strategies for retrieving the most relevant instances from a training set of geo-tagged observations. These approaches are inefficient in terms their memory requirements as they necessitate storing either the entire training set or a discretized version of it. Existing repositories of citizen science data (e.g. [53, 2, 1]) can contain on the order of tens of millions of observations making them prohibitively large to store and retrieve on mobile devices. Choosing the correct discretization is challenging [48], and incorrect choices can significantly affect the final performance [38, 46]. A key benefit of our approach is that discretization is not required.

Tang et al. [55] explored different feature encodings for incorporating location information directly into deep neural networks at training time. This included raw location features (i.e. longitude and latitude), demographic information collected via a census, user provided hash-tags, and geographical map features (e.g. land use estimates). The disadvantage of their method is that it assumes that location information is present at test time and that all the required features can be computed for a given test location. Furthermore, they cannot use location information that does not have an associated image. They also need to retrain their entire model if new location data is collected. We instead propose an efficient spatio-temporal prior that jointly models the spatial distribution of multiple object categories that can be trained independently of the image classifier. Parallel to our work, [13] builds on [55] by exploring different ways to integrate location information into deep image classifiers.

Spatio-Temporal Distribution Modelling

Our goal is to estimate the spatio-temporal distribution of a set of object categories. Related to this, there is a rich literature exploring models for estimating the distribution of biological specimens across geographic space and time [31]. This is referred to as species distribution modelling or environmental niche modelling. Broadly, these methods can be divided into two groups, those that use presence-absence data and those that use presence-only data [28].

Making a presence-absence observation at a given location requires that every species from a predefined set of interest be confirmed as either present or absent for that sampling event. In practice, this kind of data is onerous to collect because it requires intense survey effort to confirm that a species is absent with a high degree of certainty [43]. However, once this data is collected it can be combined with standard supervised classification approaches such as logistic regression [28], probit regression [52], Gaussian processes [26], decision trees [20], and neural networks [67, 49, 45], among others [18, 47]. Presence-absence data is also compatible with traditional multi-label learning [34, 8, 69, 12, 63]. Recently deep models have been applied to this problem in order to jointly model the location preferences of different species [27, 11, 22, 56, 7] and human sampling biases [10].

In contrast, a presence-only (i.e. incidental) observation may be recorded wherever an object of interest is encountered - without requiring any absences to be verified. While presence-only data can be much easier to collect, the lack of absence information makes it more difficult to model. This limitation is typically dealt with in one of three different ways. The first approach is to generate ‘pseudo-negatives’ and then apply one of the presence-absence approaches from above. As no true negative information is available, these approaches randomly sample a set of locations and make the assumption that these locations are absences e.g. [19, 50, 3]. The second commonly used ap-
proach is to train a highly regularized model directly on the presence-only data e.g. by fitting a maximum entropy distribution [51] or a low-rank model [21], forcing the model to explain data where it has been observed and to be uncertain elsewhere. Finally, and most related to our work, there are approaches that use additional information such as the detectability of a given species and a photographer’s propensity to image them e.g. [44, 24].

Unlike many of the classic approaches for spatio-temporal distribution modelling, in this work we jointly learn a continuous spatio-temporal prior for each category of interest using a neural network to amortize the computation. In contrast to previous deep distribution models e.g. [27, 11, 56], we do not require presence-absence data or additional environmental features as input. We instead exploit the structure that exists in online image repositories, such as those collected by citizen scientists, to jointly model objects, their locations, and photographer biases.

### 3. Methods

Here we outline our spatio-temporal prior, which models the geographical and temporal distribution of a set of object categories and photographers. During training we assume that we have access to a set of tuples \(D = \{(I_i, x_i, y_i, p_i)\}_{i=1}^{N}\), where \(I_i\) is an image, \(y_i \in \{1, \ldots, C\}\) is the corresponding class label, \(x_i = [\text{lon}_i, \text{lat}_i, \text{time}_i]\) represents the location (longitude and latitude) and time the image was taken, and \(p_i\) is the individual, i.e. photographer, who captured the image. Note that the location does not need to be captured alongside the image. \(D\) can be assembled from unrelated image and location datasets as long as both contain the same categories.

At test time, given an image and where and when it was taken we aim to estimate which category it contains i.e. \(P(y|I, x)\). One approach is to model the joint distribution \(P(I, x)\) as in [55], but this necessitates that the location information is always available at test time. Instead, inspired by [5], we can incorporate location information as a Bayesian spatio-temporal prior. If we assume that \(I\) and \(x\) are conditionally independent given \(y\), then

\[
P(y|I, x) = \frac{P(I, x|y)P(y)}{P(I, x)} = \frac{P(I|y)p(y)P(y|x)}{P(I|y)p(y|x)} \propto P(y|I)p(y|x),
\]

where we assume a uniform prior \(P(y) = 1/C\) for \(y \in \{1, \ldots, C\}\). In reality an image may contain location information unrelated to the class label (e.g. the background), but we assume this factorization is valid. By factoring the distribution in this way we can represent the image classifier, \(P(y|I)\), and spatio-temporal prior, \(P(y|x)\), separately.

![Diagram](image.png)

**Figure 2: Inference time.** Our goal is to estimate if an object category \(y\) is present in an input image \(I\). At test time we make use of additional spatio-temporal information \(x\) in the form of where and when the image was taken.

That at test time we do not assume that we have any knowledge of the individual \(p\) who captured the image. In this work we focus our attention on representing \(P(y|x)\). For \(P(y|I)\) we can use any discriminative model that produces a probabilistic output e.g. a convolutional neural network.

**Presence-Absence Loss**

As we are modeling the spatio-temporal prior independently from the image classifier our training data is now of the form \(D = \{(x_i, y_i, p_i)\}_{i=1}^{N}\). In the ideal case we would have complete information consisting of where and when a given category has both been observed to be present and observed to be not present e.g. as in [11, 56]. Then instead of \(y_i \in \{1, \ldots, C\}\), each spatio-temporal location \(x_i\) would be associated with a binary multi-label vector \(y_i = [y_i^1, \ldots, y_i^C]\) where each entry \(y_i^c \in \{0, 1\}\) indicates whether or not category \(c\) has been observed as being present at \(x_i\). This formulation results in a standard multi-label learning problem, enabling us to estimate the parameters of the spatio-temporal model by solving

\[
\max_{\theta} \sum_{i=1}^{N} \sum_{c=1}^{C} y_i^c \log(\hat{y}_i^c) + (1 - y_i^c) \log(1 - \hat{y}_i^c),
\]

where we define \(\hat{y}_i^c = P(y_i^c|x_i)\) and \(P\) is parameterized by \(\theta\). However, as discussed previously, presence-absence information is both difficult and time consuming to acquire in real world settings.

**Presence-Only Loss**

In this work we explore the more challenging presence-only setting where each spatio-temporal location \(x_i\) is associated with a single label \(y_i \in \{1, \ldots, C\}\) indicating which category was observed. In essence, we have a label vector \(y_i\), where there is only one affirmative entry, i.e. \(y_i^c = 1\) for
some $c$, and the remaining entries are unknown. In this setting, Eqn. 4 can be written as

$$\max_{\theta} \sum_{i=1}^{N} \log(\hat{y}_i^c) + A_i,$$

where $A_i$ represents a proxy absence term for the $i^{th}$ training example and $c_i$ is the corresponding observed category. Now the question becomes how to choose $A_i$.

One common approach for representing $A_i$ is to generate ‘pseudo-negatives’ [3] by randomly sampling absence data from some parametric distribution. For instance, one might set

$$A_i = \log(1 - P(y_i|r_i)).$$

where $r_i$ is a randomly selected spatio-temporal location with $[\text{lon}(r_i), \text{lat}(r_i)] \sim \text{Unif}(S^2)$ and $\text{time}(r_i) \sim \text{Unif}([0, 1])$. The implicit assumption is that each category (whether man-made or naturally occurring) occurs in a relatively small subset of $S^2 \times [0, 1]$, so the probability of a category occurring at a randomly chosen location $r \in S^2 \times [0, 1]$ is small as well. To the extent that this assumption holds, these pseudo-negatives are likely to be valid.

An alternative approach is to instead sample absences over locations and times where the presence data for other categories occurs. In this case we would set $A_i$ according to Eqn. 6 but sample negative locations from the positive occurrence locations i.e. $r_i \sim \text{Unif}\{\{x_1, \ldots, x_N\}\}$. This biases the training towards regions that contain valid data.

### 3.1. Our Approach

In this section we outline how we model and train our spatio-temporal prior $P(y|x)$.

#### Location and Object Embedding

In many contexts, different objects do not occur independently at a given spatio-temporal location. Knowing that object A is present may provide information regarding the presence or absence of object B at the same place and time. Similarly, different spatio-temporal locations are not independent, and may share commonalities. We exploit this structure to encode low dimensional embeddings of objects and spatio-temporal locations.

Taking inspiration from [11], we model our spatio-temporal prior as $P(y|x) \propto s(f(x)|O)$. Here, $f: \mathbb{R}^3 \rightarrow \mathbb{R}^D$ is a multi-layered fully-connected neural network that maps a spatio-temporal location $x$ to a $D$ dimensional embedding vector. $O \in \mathbb{R}^{D \times C}$ represents an object embedding matrix, where each column is a different category. The product $f(x)O$ results in a $C$ dimensional vector, where each element represents the affinity that a spatio-temporal location $x$ has for category $y$. The intuition is that we are representing spatio-temporal locations and object categories in a shared embedding space where the inner product between the embedding of a location $x$ and an object $y$ is large if $y$ is likely to occur at location $x$. Finally, $s()$ is an entry-wise sigmoid operation to ensure that the resulting prediction are in the range $[0, 1]$.

#### Photographer Embedding

In online image collections we often have access to additional information at training time in the form of the photographer $p \in \mathcal{P}$ who captured the image. To see why this information is valuable, consider the following example. Suppose a photographer $p$ visits location $x$ and does not report object $y$. If $p$ has never taken an image of an object like $y$, then this non-report gives us little information. However, if $p$ has a history of reporting categories similar to object $y$, then this constitutes weak evidence that $y$ might actually be absent at that location. Thus, we can interpret the same presence-only information in different ways depending on the individual who provides it.

To capture photographer biases, we embed photographers into the same shared embedding space as the objects and locations. This is achieved by learning a photographer embedding matrix $P \in \mathbb{R}^{D \times |\mathcal{P}|}$ at training time. Like different object categories, photographers may have affinities for particular locations and times, and share similarities in their spatio-temporal patterns with other photographers. This enables us to represent both a photographer’s preference for a given location $P(p|x) \propto s(f(x)|P)$, and a photographer’s affinity for a given object category $P(y|p) \propto s(O^T_P)$. Once trained, the photographer embeddings $P$ are not required at test time, see Fig. 2.

#### Joint Embedding Loss

Our goal at training time is to estimate the set of parameters $\theta = [\theta_f, O, P]$, where $\theta_f$ denotes the weights of the location embedding network $f()$, $O$ is the category embedding matrix, and $P$ is the photographer embedding matrix.

We start with the constraint that our model should be conservative i.e. if a category $y$ has been observed at the spatio-temporal location $x$ in the training set, then $s(f(x|O_{:,y}))$ should be close to 1, otherwise it should be close to 0. Here, $O_{:,y}$ indicates the $y^{th}$ column of $O$. We rely on the location embedding function $f()$ to interpolate between presence locations. This is conservative in the sense that it assumes that an object is absent if it has not been observed. This is a very strong assumption, but it enables the spatio-temporal prior to be aggressive in downweighting incorrect predictions from the image classifier.

Our first loss encourages the model to predict the presence of objects where they have been observed in the training set and downweight their likelihood where they have not
been observed:

\[ \mathcal{L}_{\alpha,\text{loc}}(x, r, O, y) = \lambda \log(s(f(x)O_{:, y})) + \sum_{i=1}^{C} \log(1 - s(f(x)O_{:, i}))) \]

\[ + \sum_{i=1}^{C} \log(1 - s(f(r)O_{:, i}))) \]  

(7)

\[ \lambda \text{ is a hyperparameter used to weight the positive observations and } r \text{ is a uniformly random spatio-temporal data-point. Next, we want the affinity between a photographer } p \text{ and a location } x \text{ be high if } p \text{ was present at } x, \text{ and low otherwise:} \]

\[ \mathcal{L}_{p,\text{loc}}(x, r, P, p) = \log(s(f(x)P_{:, p})) + \log(1 - s(f(r)P_{:, p})). \]  

(8)

We assume that a photographer has a low affinity for a category unless they have previously observed it:

\[ \mathcal{L}_{p,o}(O, P, y, p) = \lambda \log(s(O^T_{:, y}P_{:, p}))) + \sum_{i=1}^{C} \log(1 - s(f(O^T_{:, i}P_{:, p}))). \]  

(9)

Finally, to estimate the parameters of our prior we maximize

\[ \mathcal{L} = \mathcal{L}_{\alpha,\text{loc}} + \mathcal{L}_{p,\text{loc}} + \mathcal{L}_{p,o}, \]  

by iterating over each of the datapoints in the training set.

4. Experiments

We evaluate the effectiveness of our spatio-temporal prior by performing experiments on several image classification datasets that have location and time information. We choose image classification because for other domains (e.g. species distribution modeling) it is challenging to obtain accurate ground truth information regarding the true spatio-temporal distributions of the categories of interest.

4.1. Datasets

While location metadata is readily available for online image collections, many popular image classification datasets do not contain this information e.g. [61, 57, 16, 39]. Some datasets exist with location information, but for only a subset of the images e.g. [25]. However, datasets containing images of different species of plants and animals are available with location, time, and photographer information. To this end, we perform experiments on the iNaturalist 2017 and 2018 (iNat2017 and iNat2018) species classification datasets which contain images collected and annotated by citizen scientists [58]. They have 5,089 and 8,142 categories respectively. While [5] evaluated their location prior on the BirdSnap dataset, the images and location metadata used are not provided by the authors. We recollect the images and location data from the web using the original image URLs. Despite the dataset consisting of images of species commonly found in North America, when we recollected the images and locations we found that the original images are from all over the world and 40% were missing location. Like [5], we also simulate location metadata for BirdSnap [5] and another fine-grained dataset of birds, NABirds [57], by associating each image with a species observation from eBird [53]. Our train locations and photographers are sampled from eBird 2015, and the test set is from 2016. BirdSnap and NABirds contain images from 500 and 555 different species of North America birds. Finally, we also perform experiments on YFCC100M-GEO100 [55] (YFCC). YFCC contains 100 everyday object categories with associated locations, but no date or photographer information is provided. The train and test split used in [55] is not available and so we created a new one. Unlike the other datasets, many of the object categories in YFCC are not geographically distinct e.g. ‘band’, ‘ford’, or ‘ipod’.

4.2. Implementation Details

Our location encoder \( f() \) is a fully-connected neural network consisting of an input layer, followed by multiple residual layers [30], and a final output embedding layer. We jointly train the location encoder, along with the photographer and object embeddings using Adam [36] for 30 epochs with a batch size of 1024, using dropout to prevent overfitting. The dimensionality of the shared embedding space is set to \( D = 256 \). When weighting the positive instances during training we set \( \lambda \) to the number of categories. To counteract the heavily imbalanced nature of many of the datasets, we limit the maximum number of datapoints for each category per epoch. We set the maximum number of datapoints to 100, and for each epoch we randomly select a different subset for each category. The only exception is for YFCC, where capping the data hurt performance. Details of our network architecture are in the supplementary material.

Except where noted, at test time, our model takes three inputs – longitude, latitude, and day of the year, specifying where and when the image of interest was captured. For these three input features \( x \) we explored different methods for "wrapping" the coordinates i.e. an observation taken on December 31st should result in a similar embedding to one captured on January 1st. Similarly, we want geographical coordinates to wrap around the earth. To achieve this, for each input dimension \( l \) of \( x \) we perform the mapping \( [\sin(\pi x^l), \cos(\pi x^l)] \), resulting in two numbers for each dimension. Here, we assume that each dimension of the input has been normalized to the range \( x^l \in [-1, 1] \).

For the image classifiers \( P(y|I) \) we fine-tune a separate InceptionV3 [54] network for each of the datasets beginning with ImageNet initialized weights [16] with an image
Table 1: **Classification accuracy**. Results after combining image classification predictions $P(y|x)$ with different spatio-temporal priors $P(y|x)$. All results are top 1 accuracy with classifier predictions extracted from an InceptionV3 [54] network fine-tuned on each of the respective datasets. $\dagger$ indicates that simulated locations, dates, and photographers from the eBird dataset [53] are used. The baseline algorithms do not use date information.

<table>
<thead>
<tr>
<th></th>
<th>YFCC</th>
<th>BirdSnap</th>
<th>BirdSnap$\dagger$</th>
<th>NABirds$\dagger$</th>
<th>iNat2017</th>
<th>iNat2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(y</td>
<td>x) -$ Prior Type</td>
<td>Test</td>
<td>Test</td>
<td>Test</td>
<td>Test</td>
<td>Val</td>
</tr>
<tr>
<td>No Prior (i.e. uniform)</td>
<td>50.15</td>
<td>70.07</td>
<td>70.07</td>
<td>76.08</td>
<td>63.27</td>
<td>64.16</td>
</tr>
<tr>
<td>Nearest Neighbor (num)</td>
<td>51.78</td>
<td>70.82</td>
<td>77.76</td>
<td>79.99</td>
<td>65.34</td>
<td>66.04</td>
</tr>
<tr>
<td>Nearest Neighbor (spatial)</td>
<td>51.21</td>
<td>71.57</td>
<td>77.98</td>
<td>80.79</td>
<td>65.85</td>
<td>67.02</td>
</tr>
<tr>
<td>Discretized Grid</td>
<td>51.06</td>
<td>71.09</td>
<td>77.19</td>
<td>79.58</td>
<td>65.49</td>
<td>66.62</td>
</tr>
<tr>
<td>Adaptive Kernel [5]</td>
<td>51.47</td>
<td>71.57</td>
<td>78.65</td>
<td>81.11</td>
<td>64.86</td>
<td>65.83</td>
</tr>
<tr>
<td>Tang et al. [55]</td>
<td>50.43</td>
<td>70.16</td>
<td>72.33</td>
<td>77.34</td>
<td>66.15</td>
<td>67.08</td>
</tr>
<tr>
<td>Ours no date</td>
<td>50.70</td>
<td>71.66</td>
<td>78.65</td>
<td>81.15</td>
<td>69.34</td>
<td>70.62</td>
</tr>
<tr>
<td>Ours full</td>
<td>-</td>
<td>71.84</td>
<td>79.58</td>
<td>81.50</td>
<td>69.60</td>
<td>70.83</td>
</tr>
</tbody>
</table>

Table 2: **Ablation**. Classification accuracy for different variants of our prior on the iNat2017 and iNat2018 [58] validation sets. In the case of iNat2018, we still observe improvements when combining our prior with a more powerful image classifier - see rows ‘InceptionV3 520 × 520’.

<table>
<thead>
<tr>
<th></th>
<th>Val $\times$ 299</th>
<th>Top1</th>
<th>Top3</th>
<th>Top5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>iNat2017 - InceptionV3 299 × 299</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Prior (i.e. uniform)</td>
<td>65.27</td>
<td>79.82</td>
<td>84.31</td>
<td></td>
</tr>
<tr>
<td>Ours no wrap encode</td>
<td>69.48</td>
<td>84.43</td>
<td>88.15</td>
<td></td>
</tr>
<tr>
<td>Ours no photographer</td>
<td>69.39</td>
<td>83.97</td>
<td>87.71</td>
<td></td>
</tr>
<tr>
<td>Ours no date</td>
<td>69.34</td>
<td>84.16</td>
<td>87.89</td>
<td></td>
</tr>
<tr>
<td>Ours full</td>
<td>69.60</td>
<td>84.41</td>
<td>88.07</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Val $\times$ 520</th>
<th>Top1</th>
<th>Top3</th>
<th>Top5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>iNat2018 - InceptionV3 520 × 520</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Prior (i.e. uniform)</td>
<td>60.20</td>
<td>77.90</td>
<td>83.29</td>
<td></td>
</tr>
<tr>
<td>Ours no wrap encode</td>
<td>72.12</td>
<td>87.00</td>
<td>90.52</td>
<td></td>
</tr>
<tr>
<td>Ours no photographer</td>
<td>72.84</td>
<td>87.30</td>
<td>90.75</td>
<td></td>
</tr>
<tr>
<td>Ours no date</td>
<td>72.41</td>
<td>87.19</td>
<td>90.60</td>
<td></td>
</tr>
<tr>
<td>Ours full</td>
<td>72.68</td>
<td>87.26</td>
<td>90.79</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Quantitative Evaluation

In Table 1 we evaluate how much our spatio-temporal prior improves image classification performance by comparing it to several baselines. We found that adding a uniform prior to the outputs of the nearest neighbor based baselines increases their performance. This adds robustness in cases where there are no objects from the training set present near the test locations. The lack of this uniform prior explains the poor results for nearest neighbor based approaches in [55]. For the comparison to Tang et al. [55], we jointly train a linear layer to embed the raw location information along with an output layer to combine the location embedding with the features from the last linear layer of the image classifier. The rest of the weights of the image classifier are not updated. For each of the baseline algorithms we select their hyperparameters (e.g. the number of neighbors) on a held out validation set for each dataset. When location information is not available at test time, we assume a uniform prior over the categories.

Our model performs on par, or better, than the baselines across all datasets. The advantage of our approach is that it is computationally efficient at test time and does not require features from the image classifier during training. Compared to nearest neighbor based methods, it only requires a forward pass through a compact fully-connected neural network. In addition, it also captures structural information such as object and photographer biases. One failure case that is worth noting are the results on YFCC [55]. We observe that all methods perform similar to using no location information (No Prior). This can be explained by the relative lack of spatio-temporal structure in the object categories present in the dataset. Again, this is consistent with the findings in [55], where the authors had to use additional features to increase the performance.

4.3.1 Ablation Study

In Table 2 we compare the performance of different variants of our model on iNat2017 and iNat2018 [58]. Again, across all metrics there is a large increase in performance compared to the baseline uniform prior. In some cases, we even observe that there is an additional boost in performance when we explicitly model photographer biases.

Training fine-grained image classifiers with larger input images can significantly increase classification performance [15]. We observe that the benefit of our spatio-temporal prior is still apparent even when we use a more powerful classifier that has been training for longer with larger images. This increase in accuracy is also present when we evaluate performance using more lenient evaluation metrics *i.e.* top 5 vs. top 1 accuracy. This is significant because it highlights that for some datasets the performance boost provided by the spatio-temporal prior is orthogonal to improvements in the underlying image classifier.
4.4. Qualitative Evaluation

Our model captures the relationship between objects, locations, and photographers. In Fig. 3 (a) we can see the resulting embeddings for each input location from our model trained on iNat2018 [58]. By applying the embedding function \( f(\cdot) \) to each location we can generate its \( D \) dimensional embedding vector. We then use ICA [33] to project the embedded features to a three dimensional space and mask out the ocean for visualization. Perhaps as expected, there is low frequency structure in the resulting image \( i.e. \) nearby locations tend to support similar objects. One advantage of our approach is that we are not restricted to a fixed discretization. As a result we can generate embeddings for any location and time. In Fig. 4 we visualize our learned object embedding \( O \). Objects that have similar spatio-temporal distributions tend to result in similar embedding vectors.

Distinct from other work, our prior also models the relationship between photographers and locations, and photographers and object categories. In Fig. 3 (b) we plot the estimated affinity for each input location across all photographers \( i.e. \sum_p s(f(x)P_{:,p}) \). We only show results for photographers who provided at least 100 observations in the iNat2018 [58] training set, resulting in 634 individuals. In Fig. 5 we display the estimated affinity for each object category for a set of photographers \( i.e. \) \( P(y|p) \propto s(O^T P) \). We observe that the embedding captures the similarity in object affinity held by different photographers.

Finally, in Fig. 6 we use our prior to generate spatio-temporal predictions for several different species from iNat2018 [58]. Each image is generated by querying every location on the surface of the earth, on a specified day of the year, to generate \( P(y = y^*|x) \) for the category of interest. In practice, we evaluate \( 1000 \times 2000 \) spatial locations for each time point \( (e.g. \) first day of the month). This step is very efficient as we can pre-compute \( f(x) \) for every location, independent of the category of interest. Again, for visualization we mask out the predictions over the ocean.

4.5. Limitations

We are limited by the quality of the provided location data \( e.g. \) it can be inaccurate or intentionally obfuscated. We also make strong assumptions about a photographer’s affinity for an individual object category. In reality, these interactions may be complex \( i.e. \) once a photographer captures an image of a particular category they may be less likely to take an image of the same object in the near future. There are also known spatial biases in the types of citizen science data we use [4, 10]. However, this may not be a major issue as we can assume that the distribution of test locations and dates is similarly biased. We currently only use location, time, and photographer ID during training. In practice, additional data such as environmental variables may be a valuable signal for specific object categories [6].

5. Conclusion

We introduce a spatio-temporal prior to help disambiguate fine-grained categories resulting in improved test time image classification performance. In addition to helping image classification, our model also naturally captures
Figure 5: **Photographer object affinity.** On the left we see a t-SNE [42] plot of the photographer embeddings $P$ for iNat2018 [58]. The three plots on the right depict the predicted affinities for three different photographers (A, B, and C) visualized on the category embedding from Fig. 4. Brighter colors indicate a higher affinity for a given category. We observe that individuals that are close in the photographer embedding space $P$ (e.g. A and B) have similar category affinities, compared to those that are far away (e.g. C).

Figure 6: **Spatio-temporal predictions.** Predicted distributions for several object categories for three different time points using our full model trained on iNat2018 [58]. Darker colors indicate locations where the categories are predicted to be found. In the first two rows we observe that our model captures seasonal migratory behaviors. On the bottom row, our model correctly predicts that the Western Honey Bee can be found on several different continents. It is worth noting that the results are affected by geographical sampling biases in the iNat2018 dataset.

the relationships between locations and objects, objects and objects, photographers and objects, and photographers and locations in an interpretable manner. Importantly, our prior is efficient at test time, both in terms of model size and inference speed, and scales to large numbers of categories.

**Acknowledgements** This work was supported by a GoogleFocused Research Award and an NSF Graduate Research Fellowship (Grant No. DGE1745301). We thank Grant Van Horn and Serge Belongie for helpful discussions, along with NVIDIA and AWS for their kind donations.


