Unsupervised Learning of Interpretable Models for Sparse, Smooth Data

Arnold Kalmbach

Master of Computer Science (Thesis)

Department of Computer Science

McGill University

Montreal, Quebec

December 8, 2018

A thesis submitted to McGill University in partial fulfillment of requirements for the degree of Master of Computer Science

DEDICATION

This thesis is dedicated to my teachers throughout the years, inside and outside classrooms, foremost to my parents Faye and Arnold, and to my sister Ilana.

ACKNOWLEDGEMENTS

Studying at McGill has been an incredible journey and a great priviledge, and I would like to thank everyone who made my time here as memorable as it was illuminating. Foremost, I would like to thank my supervisor Gregory Dudek for his guidance, encouragement, and infectious enthusiasm for interesting ideas. I also would like to express my gratitude for labmate, mentor and friend, Yogesh Girdhar, whose work is the starting point of this thesis, and whose own adventures led me to an extremely rewarding semester abroad at Woods Hole Oceangraphic Institution. I'd also like to thank all of the members of the Mobile Robotics Lab; Florian, Juan, Sandeep, Travis, Jimmy, Nikhil, Andrew, Johanna, Lucas, Wei-Di, Karim, Monica, Ed, Scott, Anqi, Malika, Isabelle, Yiannis, Yogi, Dave, and Greg, for building a vibrant, energizing community, and for fruitful collaborations and valuable feedback throughout my studies. I'd like to express extra gratitude to Florian and Juan, for being extra dependable friends and sources of good advice, as well as to Wei-Di, for all his help in our active vision experiments over the past year. Thanks go to my co-authors outside of McGill, especially to Maia Hoeberechts at ONC and Heidi Sosik at WHOI. Thanks also go to my friends outside the lab, Steve, Brian, Matt, and Ben, who help keep my work connected to the world at large, and who enrich my perspective and my life. Finally, I'd like to thank my family: my partner, Katie (who helps keep me sane every single day), my mother, Faye, my father, Arnold, and my sister Ilana; for their unconditional support and caring, no matter how many times I tried out unfinished, incomprehensible explanations of my work on them.

ABSTRACT

This thesis considers learning unsupervised representations that facilitate understanding how complex data varies over space and time – in other words learning interpretable semantic spatio-temporal maps. We focus on representations that are *sparse*, meaning most locations are described by few types, and *smooth*, meaning most locations are described by the same types as their neighbors. We investigate how a spatio-temporal topic model – a descendent of Latent Dirichlet Allocation (LDA) and other probabilistic latent variable models – can be used to learn such representations for problems outside the typical text domain associated with LDA. We apply spatio-temporal topic models to a broad range of domains including modelling ambient audio, phytoplankton populations, and images observed by a robotic camera. In developing this variety of applications, we introduce novel interpretations of the learned topic model; beyond simple clustering or unsupervised classification we also consider using the topic model to make predictions in feature space and in the space of raw observations. In our experiments we demonstrate that sparse and smooth map representations are more interpretable than alternative unsupervised representations. We evaluate this interpretability in terms of both alignment with natural human-centric representations as well as the ease of learning later supervised models.

ABRÉGÉ

Cette thèse considère l'apprentissage des représentations non supervisées qui facilitent la compréhension de la variation des données complexes dans l'espace et dans le temps, c'est-á-dire l'apprentissage des cartes spatio-temporelles sémantiques interprétables. Nous nous concentrons sur les représentations qui sont *clairsemées*, ce qui signifie que la plupart des endroits sont décrits par quelques types, et *régulières*, ce qui signifie que la plupart des endroits sont décrits par les mêmes types que leurs voisins. Nous étudions comment un modèle thèmatique spatio-temporel ('spatio-temporal topic model') - un descendant de 'Latent Dirichlet Allocation' (LDA) et d'autres modèles de variables latentes probabilistes - peut être utilisé pour apprendre de telles représentations pour des problèmes extérieurs du domaine de texte associé à LDA. Nous appliquons des modèles thèmatiques spatio-temporels à une large gamme de domaines, y compris la modélisation de l'audio ambiant, des populations de phytoplancton et des images observées par une caméra robotisée. En développant cette variété d'applications, nous introduisons de nouvelles interprétations du modèle thèmatique appris; plus que du simple regroupement ou de la classification non supervisée, nous envisageons également d'utiliser le modèle thèmatique pour faire des prédictions dans l'espace objet et dans l'espace des observations propres. Dans nos essais, nous démontrons que les représentations de carte clairsemées et lisses sont plus interprétables que les représentations standards non supervisées. Nous évaluons cette interprétabilité en relation avec l'alignement avec les représentations naturelles centrées sur l'homme et aussi la facilité d'apprentissage des modèles supervisés basés sur nos représentations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure page

LIST OF SYMBOLS

x	Location, spatial or spatio-temporal
w	Word, an instance of a discrete-valued feature ($\mathbf{w}$, all words)
z	Topic assignment ($\mathbf{z}$, all topic assignments)
$g(x)$	The neighborhood of location x
g	Parameter controlling neighborhood size
θ, $\theta_{g(x)}$	Topic prior for one neighborhood (Θ, all topic priors)
ϕ, ϕ_k	Word distribution for one topic (Φ, all topics)
α	Parameter of symmetric Dirichlet prior on θ
β	Parameter of symmetric Dirichlet prior on ϕ
K	Number of topics
V	Vocabulary size, number of distinct values of w
$N^k_{g(x)}$	Number of times topic k was used in neighborhood $g(x)$
N^v_k	Number of times topic k was used for word v
$N^{\cdot}_{\cdot,-i}$	Count as above, but excluding the i-th datapoint
σ	The softmax function where the i-th dimension $\sigma(x)_i = e^{x_i} / \sum_j e^{x_j}$
$H[X]$	Entropy of the random variable X:
	$\mathbb{E}[-\log P(x)]$
$D_{KL}[P, Q]$	KL-Divergence (information gain) of PMFs P, Q:
	$\mathbb{E}_X\left[\log P(x) - \log Q(x)\right]$
$MI[X, Y]$	Mutual information between random variables X, Y:
	$\mathbb{E}_X\left[\mathbb{E}_Y\left[P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)\right]\right]$

CHAPTER 1
Introduction

1.1 Context

A fundamental goal of machine learning is to generate useful representations of rich, high-dimensional datasets, simplifying human understanding by learning to transform data. In many cases the goal of these representations is to imitate a training dataset with both inputs and labels, often generated by substantial manual annotation effort. In many other cases, however, a dataset of these labels is unavailable. For example, when a robot explores a new type of environment for the first time, when a scientist is encountered with a new type of data, or when one algorithm is presented with a novel learned representation produced by another algorithm (all examples discussed in detail within this thesis).

Unsupervised learning is an area of machine learning that aims to circumvent the need for manual annotation, learning useful representations despite the absence of labeled data. Classic methods include Principal Component Analysis (PCA) and k-means clustering, and are both efficient to train and relatively straightforward to interpret. Unfortunately, these approaches are too simplistic to capture natural categories in data, such as images, text documents or audio that are most prevalent and interesting in human-oriented applications. PCA simply learns a linear projection based on the eigenvectors of the training data, i.e. a basis for the directions of greatest variance, while k-means just clusters data-points by approximately minimizing the distance to cluster centers. Although these approaches are very widely applicable, more elaborate methods are needed to capture detail in very large, highly varying datasets.

In contrast, more recent unsupervised learning techniques employ deep neural networks (DNNs) to define more expressive unsupervised models; restricted Boltzmann machines (RBMs), autoencoders, and generative adversarial networks (GANs) are model families of particular note for impressive performance on large-scale unsupervised learning tasks with real data (Bengio et al., 2013; Goodfellow et al., 2014). In general, these approaches use a sequence of non-linear activations trained by stochastic gradient descent to find a low dimensional representation of the data which preserves as much information about it as possible. Such 'representation learning' techniques have been shown to be powerful as early computation stages for later supervised learning tasks. Typically, however, these supervised tasks are solved with complex models (often another DNN) rather than simple ones, and may require extensive training data. In addition, although these methods achieve impressive performance in an absolute sense, the representations themselves are opaque and resist human interpretation. Even when regularization steps are added, the significance of a feature dimension with respect to a 'semantic dimension' is rarely natural.

A final group of unsupervised learning techniques are those that are deliberately constructed as Bayesian probabilistic graphical models (PGMs). These models take the data $\mathbf{x}$ and hypothesize related hidden variables $\mathbf{z}$ used as the learned representation, along with a joint distribution that factors into a conditional likelihood and a prior, $P(\mathbf{x}, \mathbf{z}) = P(\mathbf{x}|\mathbf{z})P(\mathbf{z})$. The representation for a data point is produced by posterior inference, for example choosing $z = \operatorname{argmax}_z P(z|x)$. This group is not mutually exclusive with the previous two: PCA can be interpreted as a PGM with a single latent multivariate normal variable (Bishop, 1999), k-means can be interpreted as PGM by its equivalence to a special case of the well-known Gaussian Mixture Model (Bishop, 2006a), and autoencoders with a particular form can be trained for approximate posterior inference with particular choices of likelihood and prior (Kingma and Welling, 2014). The essential feature of the Bayesian graphical model formalism is that it provides a mechanism to control the types of representations we would like to obtain without hard constraints, allowing models to take advantage of both weak domain knowledge and very general priors on the kinds of representations that can be easily interpreted.

1.2 Scope

In this thesis we consider Bayesian unsupervised learning for spatial, temporal, or spatio-temporal data, in other words, for a dataset lacking associated labels but with associated locations. We hypothesize that by choosing an appropriate prior we can learn to represent data in ways that are accurate without sacrificing interpretability. By interpretable representations, we mean those that are useful with little extra learning; for instance representations that can be understood intuitively by a person, or those that reduce the data required for a complex supervised learning task. Despite the fact that we are constrained to unsupervised techniques, we aim to find representations that can be easily aligned with concrete, human-centric explanations, for instance through a one-to-one matching process, or through a small linear model.

In particular, our approach centers on Realtime Online Spatio-temporal Topic models (ROST) (Girdhar, 2014), a descendant of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) whose priors favor representations that are simple and spatio-temporally smooth. In Ch. 2 we provide a comprehensive review of the spatio-temporal topic model. We hypothesize that given an appropriate feature representation this model family favors interpretable representations in a wide variety of domains. Further, we hypothesize that these representations can be interpreted in a variety of ways – as classifications, as features for other learning problems, and as mechanisms to simplify source data without losing spatio-temporal variety.

1.3 Motivation

The introduction of cheap commercial GPS receivers and improvements in inertial and visual state estimation has resulted in an ever expanding set of data sources which come with associated spatial data. In the context of robotics, in addition to classic obstacle based landmark and occupancy maps, it is advantageous for a robot to have a qualitative sense of what kind of place it is in. For instance it would be prudent for a flying robot to recognize whether it is over park, industrial, or residential land and modify its behavior for the appropriate balance of caution and efficiency. Nevertheless, these types of land may appear vastly different in different regions and with different cameras (or completely distinct sensor modalities), therefore a robot should be able to learn to differentiate between them autonomously, and it should do so with as little manual human labeling as possible. We note that when people manually draw such label maps, the categories form spatially coherent regions; a central theme of our work is to exploit this simplification of the space of relevant representations (see Fig. 1–1).

In addition to robot mapping problems, we are interested in scientific mapping and forecasting problems, and in particular focus on issues in the marine sciences. These applications often involve novel or otherwise unusual sensor data. Further, an intuitive simplification of the data may not be known ahead of time to use to produce a training set, or else producing may require extremely specialized expertise, and therefore be impractical in the required size. We aim to address these two issues by using unsupervised learning, and constraining our representations to those with spatio-temporally coherent regions. Rather than replacing more traditional statistical methods involved in the marine sciences, we envision this unsupervised system as a tool to generate intuitions and suggest hypotheses that domain researchers can choose to test in more detail.

More generally, there is often a gap between measurements we can take (eg. images, sound, raw sensor data, etc.) and the regions that we want to map. Data requirements to learn a fully-supervised mapping from these measurements to interpretable regions are often impractical, therefore we seek to let unsupervised methods do as much of the work as possible, simplifying the supervised learning task. Although we focus mainly on robotics and scientific applications, we note the opportunity to understand semantic regions in many aspects of public and personal life

Figure 1–1: Aerial image of McMurdo Antarctic Research Station, illustrating what we mean by 'spatially-coherent regions' when we describe our hypothesized latent structure for spatial classification. Consider manually categorizing the content of a sliding window across this image. Although the set of categories which would best describe the entire image is not obvious, or even necessarily unique, their spatial structure is evident. Namely, whatever set of labels you choose, you would prefer to label the image by drawing the outlines of the areas where each label applies, rather than considering every pixel totally independently. These regions need not be entirely disjoint, meaning a single-label model is not sufficient, but at the same time most locations are best described as belonging to one or a few classes only, and are likely to be similar to their neighbors. (Photo cred: USGS [Public domain])

through the ubiquity of location as a field in data produced by smart-phones and other mobile devices.

In addition to finding abstract high-level representations of data, we are interested in finding spatially coherent representations of datasets in their original domain. These applications arise when input data are in terms of a meaningful yet complex feature space, for example a classifier with a large number of potential outputs. To make this more concrete, consider a toy scenario: You are driving down the highway and can observe the buildings near each exit as you pass it. As you do so, if you build a model of the types of buildings that are often found at the same exit you will find groupings corresponding to residential areas, commercial areas, industrial parks etc. Even if you do not have any previous experience about what these categories mean, as you are driving you can use this information to make predictions about the buildings you will find. For instance, if you see a book store and a restaurant before the exit, you can guess you will be likely to find a movie theater nearby as well, and could choose to take the exit based on this prediction before actually seeing your destination. In addition, building a model of associations between each building type and other observations you can make would require extensive experience, but if you already know which buildings are associated with one another, this requirement is lessened. In this scenario, even if the high-level categories are not aligned with any human semantics, they are still useful for another model to use to make concrete predictions.

1.4 Contributions

Our main contribution is a demonstration of the breadth of problems for which spatio-temporal topic models can provide interpretable representations. We do this by making original contributions in terms of the application of topic models in several different domains. We apply spatio-temporal topic models to a variety of unsupervised learning problems in distinct domains and demonstrate that their representations outperform more direct unsupervised baselines. We demonstrate that our approach can be used to generate a simplified map (or timeseries) representation of the data, with naturally occurring mostly-disjoint regions. The applications presented in this thesis each explore the hypothesis that such a map contains an ideal economy of information, preserving detail while also being congruent with human explanations and compatible with simple downstream supervised learning problems.

Choosing a feature representation is one of the key challenges of working with topic models. The feature is the lowest level of observation present in the probabilistic structure of the PGM, and as a result common cross-validation metrics like the log probability of held-out data are not meaningful across different feature representations. Nevertheless, the feature representation is an important mechanism to ensure the learned model reflects the desired semantic representation of the data. In this thesis we extensively discuss how to choose a feature function for various domains and applications as well as exploring in detail how to learn an appropriate feature function.

Finally, our work demonstrates how topic model outputs can be used more broadly than previous work. Because we apply topic models in a wide variety of domains, we find that different applications require distinct interpretations of the learned parameters. In addition to the aforementioned semantic-region maps, we enable applications where it is advantageous to use the topic model to produce predictions in the original feature space. We demonstrate approaches which make these predictions using auxiliary data, fill missing feature data with these predictions, and predict features for locations which we have not yet observed. Further, we develop a method to learn features that are both good for topic modeling and invertible back to their original domain. This invertible feature representation allows us to directly view and evaluate what our model has learned as images, a much more intuitive representation than a bag-of-words in a vision context.

1.5 Thesis Outline

This thesis is organized according to incrementally more complex topic model applications. First, in Ch. 2, we review the spatio-temporal topic model, formally outlining its generative model and assumptions, and why they lead to interpretable models. Then, in Ch. 3 we give two example applications demonstrating that careful choice of a feature function combined with the model's assumptions can be used to produce topic assignments that align with independently generated hand labeling (Kalmbach et al., 2016, 2013). The work in this chapter is most closely related to previous applications of the spatio-temporal topic model and topic models in general.

After we have developed tools to directly interpret the learned representations, we explore applications where predictions in a more concrete domain are necessary. In Ch. 4 we explore how spatio-temporal topic models can be used to gain insight into a novel scientific dataset (Kalmbach, Girdhar, Sosik, and Dudek, 2017; Kalmbach, Sosik, Dudek, and Girdhar, 2017). In contrast to the preceding examples, rather than gaining insights in the abstract topic domain, we prefer them in the feature domain. We develop tools to predict feature distributions from topic priors, and to approximate topic priors from external data. A central theme in this chapter is that because the topic priors are encouraged to be interpretable, they are easier to predict from auxiliary data than feature distributions.

Finally, in Ch. 5 we consider using spatio-temporal topic models make spatial predictions in the image domain, rather than the image feature domain. We develop a convolutional autoencoder architecture such that its latent space gives a suitable image feature distribution function for topic modeling We extend the feature prediction methods of Ch. 4 and demonstrate how a spatio-temporal topic model can be used to interpolate between camera perspectives using our topic model's predicted features and our autoencoder's decoder stage. By using this method, we are able to evaluate the topic model in terms of the quality of its image predictions. Further, this approach allows us to create a unique, intuitive interface to explore what the topic model has learned and search for images based on a mixture of topics represented as images.

CHAPTER 2
The Spatio-Temporal Topic Model

2.1 Topic Models and Latent Dirichlet Allocation

Topic models are a family of probabilistic latent variable model, developed as a tool for document and word classification according to semantic clusters called 'topics'. Topic models aim to find word clusters derived solely from word co-occurrence, and are thus suitable for large, unlabeled datasets. For instance, in a corpus of of news articles, a successful topic model might learn that the words 'film', 'show', 'audience', and 'actor' form one cluster, and 'budget', 'market', 'plan', and 'spending' form another. These natural groupings of words give topic models their name.

Compared to other unsupervised learning methods, a document is treated directly as a collection of many of discrete-valued variables (words), rather than as a single high-dimensional vector. As a result, topic models provide a much more direct way to reason about and place priors on individual observations, ultimately leading to powerful, interpretable representations of document collection datasets.

Topic models posit a generative model for text documents: K topics are defined, each formulated as a probability distribution over words from a finite vocabulary of size V. Each document is defined by a distribution of topics and a total number of words. The process for generating a document is modeled such that, for each word, first a topic z is drawn from the topic distribution for its document (the document-topic prior θ_d). Then, a word w is drawn from the corresponding topic (the topic-word distribution ϕ_z). The computational goal of topic modeling algorithms is to invert this process, observing the words from a set of documents and inferring the posterior over topic assignments $P(\boldsymbol{z}|\boldsymbol{w})$. In addition to the maximum likelihood topic assignments for individual words, this posterior also gives maximum likelihood estimates (MLEs) for the document-topic prior matrix Θ (who's rows are formed by each θ_d) and global topic-word distribution matrix Φ (who's rows are formed by each ϕ_z).

Topic models were first proposed as an alternative to Latent Class Indexing (LCI, also known as Latent Class Analysis). In LCI, the latent space is defined by the singular value decomposition (SVD) of the document-word count matrix for a corpus (Deerwester et al., 1990). This model is similar to how one might naively apply PCA or other multidimensional scaling techniques to a matrix of word counts. The generative model described above, and the name for the factors, 'topic', were introduced as probabilistic LCI (pLCI) (Hofmann, 1999). In this work, Hoffman et al. showed that LCI is equivalent to a probabilistic generative model for document-word counts with additive Gaussian noise (due to the SVD's relationship to the L_2 norm), and proposed the topic model as a more reasonable noise model. Topic models have been widely successful in text modeling, and throughout the 2000s many more variants of the generative model were proposed and demonstrated, as reviewed by (Blei et al., 2010).

Although pLCI is extremely general, with no further assumptions topic models are unlikely to find intuitive topics and also often generalize poorly as they are prone to overfitting (Blei et al., 2003). Latent Dirichlet Allocation improves this situation, while minimally decreasing the expressive power of the model in practice by making a further assumption, namely, placing independent Dirichlet priors on each individual document topic prior $\theta_d \sim Dir(\alpha)$ and topic word distribution $\phi_z \sim Dir(\beta)$. Recall that the Dirichlet distribution $Dir(\theta; \boldsymbol{\alpha})$ is the multivariate generalization of the Beta distribution, i.e. a distribution over the $K-1$ dimensional simplex with parameter $\boldsymbol{\alpha} \in \mathbb{R}^K$. In other words, it is a distribution over the K dimensional discrete-valued probability mass functions (PMFs). To be exact the, Dirichlet distribution is a (conjugate) prior for categorical and multinomial distributions, with probability density function (PDF):

$$P(\theta|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \boldsymbol{\alpha}_i)}{\prod_{i=1}^{K} \Gamma(\boldsymbol{\alpha}_i)} \prod_{i=1}^{K} \theta_i^{\boldsymbol{\alpha}_i - 1} \tag{2.1}$$

Where Γ denotes the gamma function.

It has mean $\boldsymbol{\alpha}_i / \sum_i \boldsymbol{\alpha}_i$ and when $\sum_i \boldsymbol{\alpha}_i < 1$ is said to be ‘sparse’, meaning the PDF has most of its mass on the corners of the simplex, and typical samples have few non-zero entries. In the case of LDA, a symmetric Dirichlet prior is chosen, where $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j \;\; \forall i, j$, and the concentration, $\alpha = \sum_i \boldsymbol{\alpha}_i$, is considered as a hyperparameter. Fig. 2–1 illustrates the effect of varying α. Choosing a symmetric, sparse Dirichlet prior for each topic-word distribution ϕ_z encodes the assumption that few words are strongly associated with a topic, some words are weakly associated with a topic, and

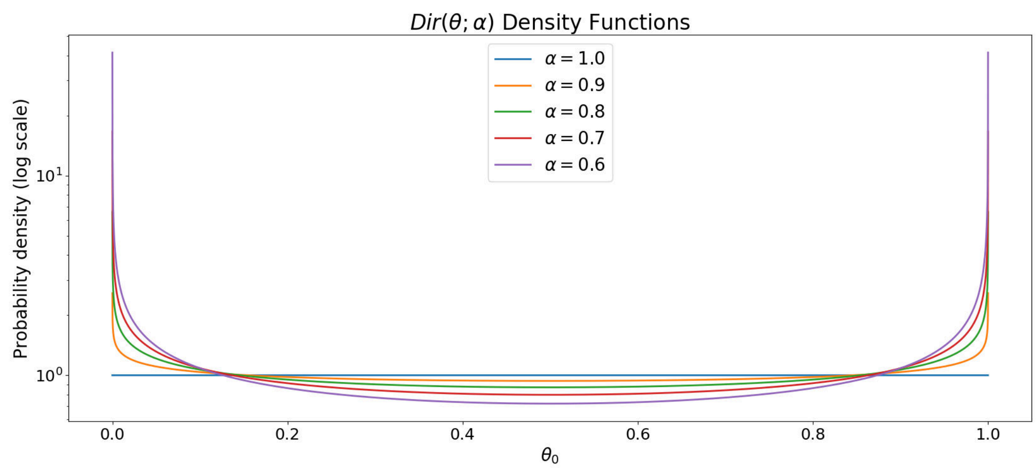


Figure 2–1: PDFs of Dirichlet distributions with 2D symmetric parameters α. These can be used as a prior over the 2D PMFs. Note that this means there is $K - 1 = 1$ degree of freedom, i.e. this is a prior for a Bernoulli random variable and this case is identical to a Beta(α, α) distribution. It is helpful to imagine each of these curves as giving a distribution of probabilities for coins from an unfair coin factory, where the average over all coins looks fair, but (for $\alpha \approx 0$) most individual coins are very unfair. For a K-dimensional Dirichlet distribution, the corresponding analogy is a prior for unfair K-sided dice.

most words are not at all associated with a topic without requiring any external information about which individual words are associated with which topics.

The posterior for θ given observations $z_i \sim Cat(\theta),\ i = 1, \ldots n$ can be easily derived using Bayes' rule to be $P(\theta|\boldsymbol{N}) = Dir(\alpha + \boldsymbol{N})$, where $\boldsymbol{N}$ is the vector of counts for each of the K possible values of z, and consequently the predictive distribution for z_{n+1} is given by (Bishop, 2006b):

$$\begin{aligned} P(z_{n+1} = k|\alpha, \boldsymbol{z}) &= \int P(z_{n+1} = k|\theta) P(\theta|\alpha, \boldsymbol{z}) d\theta \\ &= \int \theta_k Dir(\alpha + \boldsymbol{N}) d\theta \\ &= \frac{N_k + \alpha}{\sum_{j=1}^{K} N_j + \alpha} \end{aligned} \tag{2.2}$$

Although exact inference of the full LDA posterior $P(\boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$ is intractable, the posterior for a single topic assignment z_i can be estimated with a collapsed Gibbs sampler based on a the product of the predictive distributions due to θ and ϕ (Griffiths and Steyvers, 2004). Given a set of observed words $\mathbf{w}$, and topic assignments for all words except the current word, w_i, $\boldsymbol{z_{-i}}$ with counts N_k^v being the number of times word v was assigned topic k and N_d^k the number of times topic assignment k has been used in document d

$$P(z_i = k|\boldsymbol{z_{-i}}, \boldsymbol{w}) \propto \left(\frac{N_d^k + \alpha}{\sum_{j=1}^{K} N_d^j + \alpha} \right) \left(\frac{N_k^{w_i} + \beta}{\sum_{v=1}^{V} N_k^v + \beta} \right) \tag{2.3}$$

Various other methods for approximate posterior inference for LDA have been explored, particularly variational methods and more recently stochastic gradient approximations to variational methods (For a comprehensive review, see (Geigle, 2016)). Despite the efficiency of these methods, they lack the appealing anytime, (in the sense of Zilbertstine, (Zilberstein, 1996)) online aspect of a sampling based approach, and are much less flexible to modification of the model. For these reasons in this work we rely only on the Gibbs sampling based approach.

2.2 Realtime Online Spatio-Temporal Topic Models

Although LDA has been developed and used primarily in the context of modeling collections of text documents, its fundamental assumptions are compatible with an extremely wide range of applications. Virtually any feature function may be quantized or otherwise tweaked to produce discrete-valued pseudo 'words' and collected into 'documents'. Sparse, probabilistic 'topics' are an attractive, interpretable form of dimensionality reduction for wide variety of domains. Non-textual descendants of LDA have most extensively been used in computer vision applications such as learning natural scene categories (Fei-Fei and Perona, 2005), taxonomies of images (Bart et al., 2011), human action categories from video (Niebles et al., 2008), and image-caption pairing (Blei and Jordan, 2003). Further works illustrate how LDA can be used in bioinformatics, such as for learning genealogies (taxonomies of species) based on their collections of gene sequences (Pritchard et al., 2000).

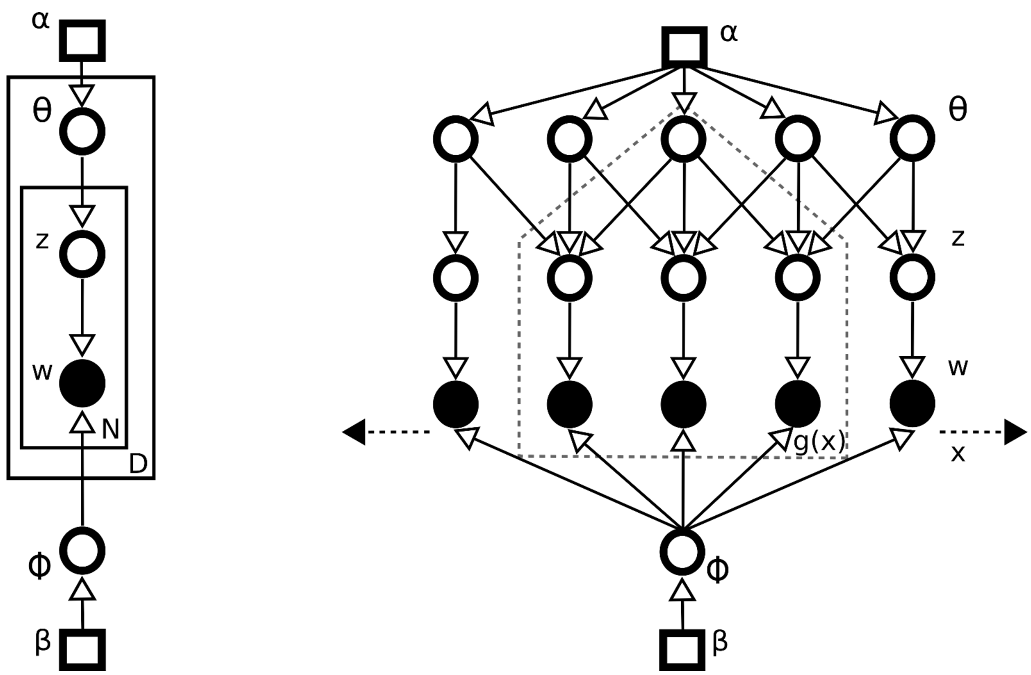
This body of literature hints that the LDA family of models is very general, and could be considered for a much broader set of applications where unsupervised representation learning is desired. One of the aims of this thesis is to demonstrate the breadth of this set. To clearly enumerate the assumptions which must be met by an LDA application: (1) The data must be comprised of collections of discrete observations, (2) the observations within a collection must be conditionally independent given a set of topics and a prior over these topics[1] , (3) the observations in separate collections must be conditionally independent given the topics and a symmetric Dirichlet prior, and (4) the topics must be conditionally independent given another symmetric Dirichlet prior.

Nevertheless, real-world data rarely meet these criteria exactly. Specifically, observations associated with a particular place or time cannot usually be considered independent from nearby observations (Assumption 3). While we could accommodate this by simply increasing the size of our collections of observations, this choice would come at the cost of decreasing the spatio-temporal resolution of the learned document-topic priors. What we would like is to enforce smoothness rather than equality between nearby topic priors.

The Realtime Online Spatio-temporal Topic model, ROST, implements exactly such a variant of Latent Dirichlet Allocation (Girdhar, 2014). With ROST, fully discrete *documents* are replaced by *neighborhoods*, each comprised of a topic prior and a collection of observations. We denote the neighborhood of a spatial, temporal or spatio-temporal location x by $g(x)$. Crucially, neighborhoods may be overlapping, and consequently the priors for nearby neighborhoods are not independent – instead they must describe an intersecting set of observations (see Fig. 2–2).

---

[1] Note, this 'bag of words' assumption implies that word order within a document does not matter, an important shortcoming of LDA in the text modeling literature, which is not necessarily as important in other domains.



(a) (b)

Figure 2–2: *Graphical models for topic models:* (a) Graphical model for Latent Dirichlet Allocation. Words within a document are conditionally independent given their topic assignments z and the topics ϕ, which in turn are conditionally independent given the topic prior θ for the associated document. (b) The graphical model for spatio-temporal topic models relaxes these independence assumptions, as the topic prior is shared by all topic assignments within overlapping spatio-temporal neighborhoods (represented by the dashed pentagon).

Formally, ROST includes the same 3 hyperparameters as LDA, namely β, the Dirichlet prior concentration on each ϕ_z, α, the Dirichlet prior concentration on $\theta_{g(x)}$ (with the neighborhood at x replacing the document index d), K, the number of topics, as well as an additional hyperparameter, g, the spatio-temporal extent of a neighborhood. Increasing g yields larger neighborhoods, where individual topic priors must account for more varied observations, are more overlapping with their neighbors, and therefore learning gives smoother document-topic distributions. Conversely, as g decreases to 0, ROST becomes identical to LDA and the local topic priors have no tendency towards spatio-temporal smoothness. To implement posterior inference with the collapsed Gibbs sampler, the only change required to Eqn. 2.3 is to replace the document d with the neighborhood $g(x)$ if word w_i was observed at location x.

The authors have shown that the spatio-temporal smoothness assumption leads to more semantically relevant topics in real-time computer vision problems as well as other domains (Girdhar and Dudek, 2012; Girdhar et al., 2014; Girdhar, 2016). The combination of smoothness and sparsity assumptions consistently leads the map of θ_x to contain regions with high probability of a single topic, with somewhat less sparse boundaries between them. Because these assumptions are implemented as priors, precise knowledge of 'how smooth' and 'how sparse' is not required, in contrast to a harder or non-Bayesian approach. As a result, ROST produces remarkably natural, interpretable topic maps for a wide variety of datasets without an onerous hyperparameter tuning procedure.

2.3 Topic Models in Practice

Although topic models are a relatively robust modeling technique, recovering semantically meaningful topics still requires careful problem setup. This part of the process is often downplayed in the literature as less-than-generalizable, nevertheless it is important to obtaining the desired data representation. The applications outlined in this thesis at least partially aim at exploring some of these issues, and how they can be systematized for varied applications and application domains.

The first issue encountered by an aspiring topic-modeller is how to obtain features which meet the assumptions of the model. Even in the text-modeling literature, where discrete-valued observations (words) and collections (documents) are readily available, significant pre-processing is common, including removing stop words and rare words, as well as stemming. These steps ensure a bag-of-words model is appropriate, ensuring that all the data is relevant to a semantic interpretation in addition to minimizing the importance of structure not captured by a histogram of words. These measures also have significant beneficial side-effects: stop words often comprise a significant portion of the dataset, so removing them significantly increases the variability between document-word distributions while greatly decreasing the total number of topic assignments to learn. In addition, stemming can greatly reduce the vocabulary, reducing the dimensionality of each word-topic distribution. In the context of natural language processing, these measures are acceptable – the set of stop words in English is relatively small and can be easily listed, and reasonably performant stemming models are available.

Nevertheless, often the most difficult step in setting up a topic model for a non-text application is formulating a feature function with similar properties. In offline contexts, for a given feature function, filtering the most common and least common words is often a reasonable heuristic approximating removing stop words to enhance the variability between documents. In contrast, approximating stemming requires domain knowledge. Despite the challenge this entails, as we will describe in Sec. 3.1, it also presents an opportunity to insert a weak form of supervision beyond using purely pre-defined feature functions. By hand-coding or learning a relationship between raw features and data relevant to a semantic interpretation, the algorithm designer can refine the way in which the topic model will describe the data without resorting to the manual classifications that would be required for fully supervised techniques.

Fortunately, although topic models are often used in situations where labeled classification data is impractical to collect, weak supervisory signals through cross-validation related to later, simpler, supervised learning problems are often enough to choose a model that captures the relevant insights. In the context of text-documents, this approach has been explored by Wallach et al. through a method to measure the semantic coherence of topics (Wallach et al., 2009). We take similar inspiration for non-text applications. Although this approach is expensive, requiring en exhaustive search of hyperparameter space, compared to DNN models, cross-validation is relatively cheap. This is firstly due to the hyperparameter space being much smaller, just 4 real-valued hyperparameters as opposed to the combinatorial space of DNN architectures. And is secondly due to the fact that topic model learning is much faster than fitting the most popular DNN architectures, on the order of just an hour on a consumer laptop for 1 million - 10 million word datasets.

The success of cross-validation through downstream tasks depends heavily on how we choose to interpret the topic model. Throughout the rest of this thesis each application will explore different ways of embedding the maximum-likelihood topic assignments, topic-word distributions, and document-topic distributions in tasks where providing ground truth is relatively easy.

CHAPTER 3
Soft, Unsupervised Classification with Spatio-Temporal Topic Models

Our first two example applications explore the most direct connection to the existing topic modeling literature, performing unsupervised classification by defining a feature function, fitting a topic model, and grouping observations by their neighborhood's topic distribution. In Sec. 3.1 we detail our work on deep-sea substrate classification from a challenging video dataset (Kalmbach et al., 2016). This application highlights the impressive closeness between the representation learned by our completely unsupervised technique and labels produced by a human expert which can be achieved by carefully designing a feature-function. Then in Sec. 3.2 we review our work on recognizing locations based on topic models of ambient audio (Kalmbach et al., 2013). This application highlights the suitability of spatio-temporal topic models for problems outside of computer vision, as well as the importance of spatio-temporal smoothness to representations learned in the absence of training targets.

3.1 Substrate Classification from Repurposed Dive Videos

Substrate classification, the task of creating a spatial description of the nature of the seabed, is a fundamental factor in many aspects of ocean research. Domain research in marine biology, physical oceanography and geology – including classifying benthic habitat (i.e. sea-floor environments), modeling deep-sea circulation and analyzing tectonic motion depends on accurate classification of substrate.

In the context of the deep sea, basic questions remain unanswered about what terrain can be found, particularly in geologically diverse mid-ocean ridge environments. High exploration cost and difficulty of sampling drive the need for remote sensing options for data acquisition, including visual and acoustic surveys, which in turn generate large volumes of data requiring analysis. Manual analysis of substrate type is time consuming and requires geological expertise. Subjective factors, such as the choice of salient environmental features, make manual analysis for multidisciplinary use subject to observer bias.

Further, single-label classifications of substrate regions do not always adequately describe the complexity of types encountered In contrast, simple continuous valued substrate descriptors are sometimes used, such as the Udden-Wentworth Scale which proposes to describe the substrate by $-log_2(d)$ where d is average grain diameter in mm (Krumbein and Aberdeen, 1937). Nevertheless, such descriptors are often impractical to evaluate with remote sensing, fail to capture the shape of the terrain, and do not account for the fact that terrain types, and therefore grain diameters, are rarely uniform within even a small region. Adding further dimensions to a substrate descriptor could account for some of these issues, at the cost of losing interpretability, and complicating the work of further domain research relying on this data. In this work we explore using soft classifications to add generality to the single-label paradigm, and seek to maintain interpretability by imposing a prior towards sparse label distributions. This approach can be seen as trying to maintain the best parts of both single label and higher dimensional substrate descriptors.

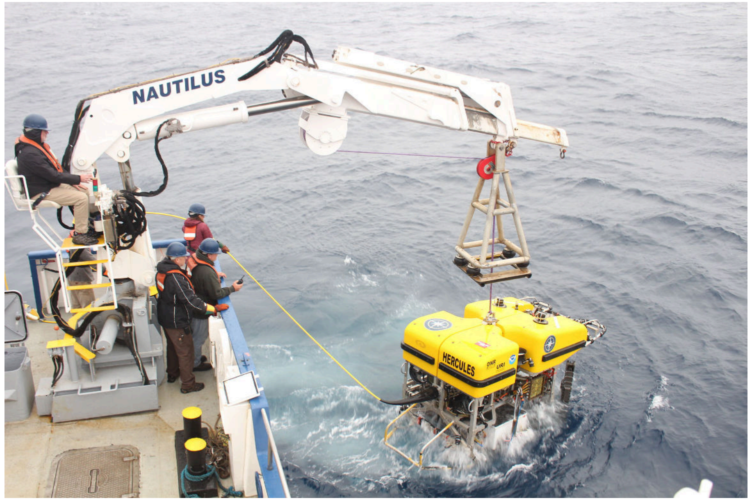


Figure 3–1: ROV Hercules returning to surface after a deep-sea cable route survey near Endeavour Ridge. (Photo cred. Ocean Networks Canada, Creative Commons NonCommercial-ShareAlike 1.0])

Organizations such as Ocean Networks Canada (ONC), who operate cabled remote sub-sea observatories, regularly conduct manual visual cable route surveys in the deep-sea using Remotely Operated underwater Vehicles (ROVs). Collecting purpose-built deep-sea substrate measurements at high-spatial resolution is prohibitively expensive. In their absence, the historic video logs of these surveys contain extensive close-up video of the substrate, presenting the opportunity to train a substrate classification model.

Unfortunately, deep-sea ROV video contains significant artifacts due to a variety of factors. For instance, hard, directional lighting is typical because sunlight does not reach the depths which interest us. Further, regions with the most interesting substrate such as hydro-geothermal regions, canyons, and sea-mounts all also feature challenging terrain for the ROV pilot. This leads to a wide variety of perspectives on the substrate. Finally, sediment and marine snow cloud the water column in the deep-sea, reducing contrast when the substrate is not very close to the camera, and filling images with particulate noise which confuses unmodified terrestrial computer vision techniques. Developing techniques to 'see through' these artifacts to the substrate is necessary component of any system aiming to perform substrate classification using this data.

The spatio-temporal topic model's prior on neighborhood distributions is a good fit for this problem. It explicitly aims to capture the distribution of labels at each spatial location, rather than applying a single label like many techniques. Further, the sparse Dirichlet prior allows the modeler to ensure that the output distributions are more interpretable compared to a uniform or Gaussian prior when the meaning of the feature function is not well defined or understood with respect to the categories of interest. Finally, the smoothness component of the prior implemented by overlapping neighborhoods ensures that the topics reflect our intuition that labels at adjacent locations should be similar to one-another. By encouraging our model to capture our intuitions about the labels, we make the most of large quantities of available unlabeled data so that with a highly restricted amount of labeled data we can learn to predict expert annotations.

3.1.1 Using domain knowledge to define a feature function

As discussed in Ch. 2.3, one of the chief challenges of applying spatio-temporal topic models to a new domain is defining a feature function. In this work, we chose between standard, off the shelf features, combined with careful pre-processing steps to ensure that they encode the relevant information for substrate classification. Specifically we consider two feature functions: SIFT codebook and histogram of Improved Local Binary Patterns (ILBPs).

Based on the literature of topic models in computer vision, the most obvious choice of feature function is a codebook of Scale-Invariant Feature transform (SIFT) descriptors. SIFT is an extremely popular keypoint-based image feature, introduced by Lowe et. al as a scale and rotation invariant descriptor and detector for distinctive (i.e. easy to re-identify) points within an image (Lowe, 2004). SIFT has been remarkably successful in object recognition, image stitching, visual mapping, and many other computer vision tasks. We take SIFT as a reasonable surrogate for a variety of similar keypoint descriptors such as SURF and ORB. While SIFT features are 128-dimensional vectors, our topic models require discrete-valued features (words). Before running our algorithm, we produce a codebook of SIFT features. That is, using ROV logs from other dives, we compute the keypoints and descriptors for each frame. Then, choosing a vocabulary size (i.e. $V = 3000$), we compute the V k-means centroids of the descriptors. At training time for our topic model, we then extract keypoints for each new frame and compute the descriptors for each, replacing the descriptor with the index of the nearest neighbor centroid.

In contrast, Local Binary Patterns represent textures by encoding the local brightness variations in 9-pixel squares as 8-bit codes representing the relationship between each edge pixel and the center. Typically, these codes are computed for every pixel in an image, and grouped together into 256-bin histograms for non-overlapping windows. These histograms are then used as a descriptor for the textures in their windows. ILBP, originally introduced under the name MultiBlock LBP in (Liao et al., 2007), extends LBP, firstly by accommodating scaled pixel areas, and secondly by adding information about the center pixel. In ILBP, each pixel, including the center pixel, is compared to the mean value over an $N \times N$ pixel region. Liao et al. have shown that this improves the robustness of LBP as well as the ability to encode larger image structures. We use the multi-channel, multi-scale, windowed ILBP implementation described in (Paris et al., 2012). In contrast to descriptors such as SIFT, ILBPs are usually collected directly as histograms, and therefore no codebook is necessary. We concatenate the histograms for each window, scale, and color channel of an image, and treat each bin of the resulting histogram as a discrete feature when running our model.

In order to ensure that our model learns to distinguish substrate types, and not some other aspect of the dataset, it is important to carefully choose what data we will model. Although it is impractical to collect a large dataset of substrate labels, some knowledge of how substrate types will appear in the video data can be used to help the learning process. In this example, we focus on removing features which are created by factors other than the substrate.

A typical cable route survey creates 6-12 hours of continuous dive video, of which at least a few hours are usually irrelevant. For instance, surveys often contain extensive video of the ROV descending or ascending, sections with minutes to hours of video where the ROV is parked on the sea-floor, and yet more sections where the substrate is not in view due to challenging navigational conditions and sediments disturbed by the ROV itself. Further, it is atypical for surveys to be conducted at constant speed, as the video is designed for live, manual analysis. To counter these effects, we first train a substrate detector, and then select a set of training images that are uniformly distributed in space.

We implement a substrate detector with a support vector machine (SVM) on the spatial envelopes of relevant and irrelevant frames. By spatial envelope, we refer to the holistic representation of the shape of a scene implemented as the GIST descriptor described in (Oliva and Torralba, 2001). GIST descriptors have been shown to encode high-level perceptual dimensions of the spatial envelope such as naturalness, openness, roughness etc. This descriptor is computed using a method based on Gabor filter responses in multiple orientations and scales, and in a grid of image windows across the image. GIST descriptors are appropriate as the local appearance of images with substrate varies dramatically, and the spatial envelope is more important that any local features.

We manually labeled 2000 randomly selected frames as either relevant or irrelevant (i.e. with or without substrate) for our subsequent classification problem. We trained an SVM using 1000 frames randomly selected from these 2000 labeled examples, and evaluated its performance on the other 1000. Table 3–2 shows the precision and recall computed over all frames in the test set.

Dive logs contain navigation data collected by an ultra-short baseline (USBL) acoustic positioning device, in the form of Latitude, Longitude, and Depth measurements in addition to video. We convert Lat., Lon. into Eastings and Northings (meters East or North from start point), and time-synchronized the navigation and

	Substrate	No Substrate
Train Set (True)	564	436
Test Set (True)	589	411
Full Dataset (Estimated)	6859	3141

Table 3–1: Numbers of relevant and irrelevant frames in our example video.

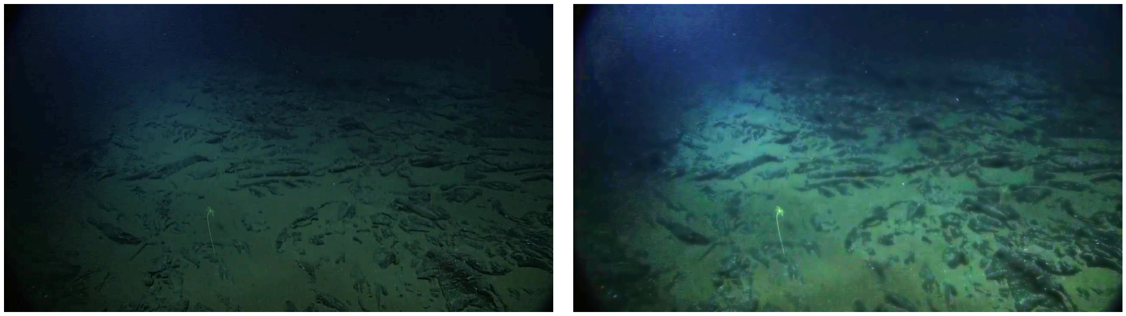
	Precision	Recall	F1-Score
Train Set	0.7942	0.9645	0.8711
Test Set	**0.8117**	**0.9440**	**0.8729**

Table 3–2: Performance of GIST SVM substrate detector.

Table 3–3: GIST SVM performance

video logs. Next we run our substrate detector on each frame from the video, discarding frames which do not contain substrate. Finally, we construct a substrate imagery dataset by passing through the data in temporal order, and only keeping video frames which are more than a fixed radius away from all other points in the training set.

In addition to ensuring that the training data comes only from potentially relevant locations, we must also ensure that the features encode the relevant aspects of the images. Most images in the survey videos contain strong brightness variation due to hard directional lighting and not the shape or texture of the substrate. Typical frames contain a well-lit foreground in the bottom middle and darker regions elsewhere. Due to marine snow and sediment, many images also contain particulate noise. To reduce the visual impact of the irrelevant, but ubiquitous spotlight effects manifested as low-frequency image content and particulate matter manifested



(a) (b)

Figure 3–2: *Challenging ROV video data:* (a) An example image collected by the ROV, before bandpass filtering and (b) after.

as high-frequency image content, we use a bandpass filter, with a 2D Gaussian kernel:

$$K = \left((1 - e^{-(x^2+y^2)/2f_{low}^2}\right)\left(e^{-(x^2+y^2)/2f_{high}^2}\right) \tag{3.1}$$

where x, y refer to the pixel offset, and where we set the frequency cutoff parameters f_{low} and f_{high} to be 4 and 100 cycles/image respectively based on the approximate minimum and maximum size of relevant substrate features. Typical effects of this filter are illustrated in Fig 3–2. While particulate noise was a detrimental factor, we found that the SIFT keypoints lost by smoothing over image details were more detrimental. To compensate, before extracting SIFT features we applied the high-pass portion of the filter only. Finally, even with these preprocessing steps, not all features contain enough variation to help distinguish frames, and therefore only contribute noise to the modeling process. The concatenation of all the LBP histograms for an image results in a 10,752 dimensional histogram. We found that keeping only the $V = 3000$ bins with highest variance greatly reduced inference time for our topic model while slightly improving classification performance.

3.1.2 Evaluation

We validate our method using data recorded at the Endeavour Segment of the Juan de Fuca Ridge, a midocean ridge environment 300 km West of Vancouver Island, British Columbia, at approximately 2.2 km depth (Department of Fisheries and Oceans Canada, 2009). Specifically, our data is from the East Flank of Endeavour, a gently sloping area featuring a variety of substrate types. Mission objectives on this dive were to perform a visual survey for suitability of a scientific instrument installation – much of the video shows substrate, but distance, angle, and speed are inconsistent throughout the recording.

The data consist of 30 fps HD video and 10 Hz USBL acoustic positioning. The ROV moved at an average rate of 0.5 knots (approx. 0.26 m/s), through two partially overlapping lawnmower patterns, one up and down a gentle slope, and one across the flat area at the base of the slope, with a total distance traveled of just over 5.6 km. We ran the substrate detector on every third frame (synchronized with the USBL data), and applied spatial subsampling with a minimum inter-frame distance of 1.5 m. This process resulted in a dataset of just over 3000 images.

Of these, we randomly selected 500 images, and with guidance from an expert in deep-sea geology, we defined seven categories that represent the types of substrate seen in the sample. These categories were Sedimented (SED), Interrupted Lava Flow (INT), Pillow Lava Flow (PIL), Cliff or Wall (CLIFF), Other Rock (O.RCK), Turbid Water (TURB), and Substrate out of Range (DARK). We labeled each image in this sample with proportions of these seven types, using a minimum increment of 0.25 for each category. Images exemplifying each category can be seen in Fig. 3–3a.

We generated the SIFT and LBP bag-of-words representations for each image in the full dataset, using the noise suppression techniques described above. We used scale factor 2 for the ILBP region with a uniform 6x7 grid of non-overlapping windows. Initial experimentation showed that these values produced good results, and that additional scales did not cause significant improvement. We then fit the spatio-temporal topic model on each feature representation, choosing a neighborhood size of $g = 15$m and $K = 7$ topics based on the estimated variation of the terrain and the number of true categories. We repeated this process for each pair of sparsity hyperparameters $\alpha, \beta \in \{0.01, 0.1, 0.2, 0.4, 0.8, 0.9, 0.99\}$. We chose the model with maximum log-probability on the held-out 500 image labeled set: $\alpha = 0.1, \beta = 0.1$ for SIFT and $\alpha = 0.1, \beta = 0.8$ for LBP. For these 500 data points, we infer topic assignments without updating the model. Although the most extreme values of α and β tended to perform poorly, intermediate values had similar log-probabilities, and the exact hyperparameter choice did not seem to have a strong effect on the learned topics.

To evaluate the degree to which each ground-truth category was represented by some topic generated by our algorithm, we produced a one-to-one pairing between categories and topics. First, we calculated the Pearson ρ correlation between each ground-truth category and each topic produced by our algorithm, using all 500 of the labeled frames. Then, we defined the cost of associating category i with topic j as $Cost(i, j) = 1 - \rho_{i,j}$, and used the well known Kuhn-Munkres ('Hungarian') Algorithm (Kuhn, 1955), to find a minimal total cost assignment of topics to categories. In other words, we paired categories and topics so as to maximize the sum of correlations in the set of pairings. As the pairing is between topics and classes (a 7×7 matching problem), the cost of this process is negligible

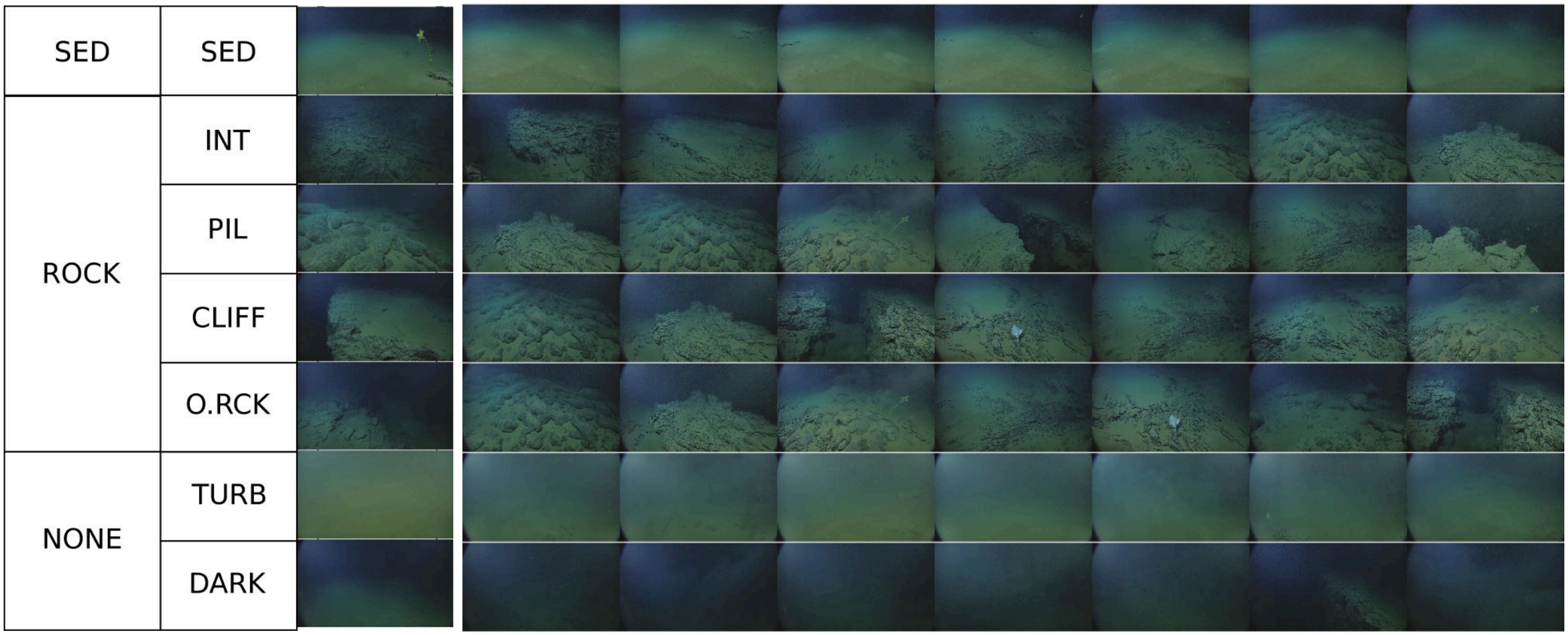
We also computed 7 k-means centers for each image based on the distribution of each feature; SIFT, LBP, and top-3000-dimensions LBP. Then, we computed the cluster labels for test images, and applied the same optimal pairing algorithm for each. The results of these approaches offer insight into how well a method which does not take spatial smoothness into account might perform.

Fig. 3–3b shows the most representative frames for each topic. These images show the 7 frames that have the highest proportion of each topic throughout the dataset. Note that some topics were much more prevalent than others, so an image may be among the best examples of a particular topic in the dataset without that topic being the top label for the image. For instance, many of the strongest examples of interrupted lava flows contain more sediment than interrupted lava flow. Comparison with the examples in Fig. 3–3a suggests that the topics assigned to sediment, turbid water, and substrate out of range accurately recover their categories, and that the other assignments are somewhat less accurate.

Tab. 3–4 reports the correlations between topics (or cluster labels) and ground-truth labels in the optimal pairing. Note that the topic model obtained the highest correlation of any method for every category, improving the strength of association by a mean of 3.58 times over the best k-means approach (strength of association is defined as the absolute value of correlation). Strong positive correlations (i.e. near 1) indicate a linear relationship between variables, in this case indicating that the proportion of an image composed of a certain label can be predicted by a linear function of the weight of a single topic much more confidently than from its k-means cluster assignment. Also note that the LBP topic model outperformed the SIFT topic model for all classes except pillow lava, while at the same time the LBP k-means models also outperformed the SIFT k-means model for all classes except pillow lava. This suggests that the performance of a simplistic k-means model may be a reasonable simplistic surrogate for the full topic model when choosing which feature function to use for a topic model.

Note that for the LBP topic model, p-values were $\lll 0.001$, whereas for the other approaches, p-values varied, sometimes taking significantly higher values. These p-values represent a strong rejection of the hypothesis that the topics were not correlated with the categories for our method, and a weaker one for the baseline strategies.

In absolute terms, the values of the correlations reflect the intuition given by the best examples of each topic: The categories sediment, turbid water, and no substrate are each strongly correlated with their assigned topic, but the other categories are only weakly correlated with their assignment. Therefore, we additionally analyze our system's performance in classifying the high-level categories Sediment, Rocky, and No Substrate. These categories were constructed by combining the original categories, as seen in the leftmost column of Fig. 3–3. We combined the groundtruth labels, topics, and labels for the three baseline methods by summing over the groups to be combined into each of the three categories. The performance of the resulting models is presented in Table 3–5, showing that the LBP topics show strong predictive ability for the high-level class labels.



(a) (b)

Figure 3–3: *Example images of substrate types, and paired topics:* (a) Representative examples of the categories (top to bottom) 'Sedimented', 'Interrupted, 'Lava Flow' 'Pillow Lava Flow', 'Cliff or Wall', 'Other Rock', 'Turbid Water', and 'Substrate out of Range'. (b) Top 7 images paired with each category from (a), based on the max-likelihood topic distribution, and the max-correlation pairing algorithm.

Table 3–4: *Detailed substrate classification performance:* Pearson $\rho\,(498)$ for best-match topic with 7 ground truth categories (see 3–3a). For ROST Filt. LBP p-value was $\ll 0.001$. Our method had the highest correlation for every category except PIL

	K-M SIFT	K-M LBP	K-M Filt. LBP	ROST SIFT	**ROST Filt. LBP**
SED	0.2898	0.3526	0.3661	0.5196	**0.7489**
INT	0.0718	0.0092	0.1116	0.3517	**0.3899**
PIL	0.4162	-0.0029	0.0849	**0.7024**	0.3884
CLIFF	0.1721	0.3474	0.2916	0.3601	**0.4152**
O.RCK	-0.0137	0.0508	-0.0685	0.1901	**0.2580**
TURB	-0.0271	0.6385	0.7509	0.1723	**0.8153**
DARK	0.3630	0.7509	0.3466	0.5454	**0.5664**

Table 3–5: *High-level substrate classification performance:* Pearson $\rho\,(498)$ for best-match topic with 3 high-level categories. Our method had correlation coefficient above 0.7 for every category, with p-value $\ll 0.001$.

	K-M SIFT	K-M LBP	K-M Filt. LBP	ROST SIFT	**ROST Filt. LBP**
SED	0.4413	0.5059	0.5350	0.5196	**0.7489**
ROCK	0.3492	0.3541	0.4930	0.6852	**0.7156**
NONE	0.0459	0.4930	0.4601	0.3116	**0.8015**

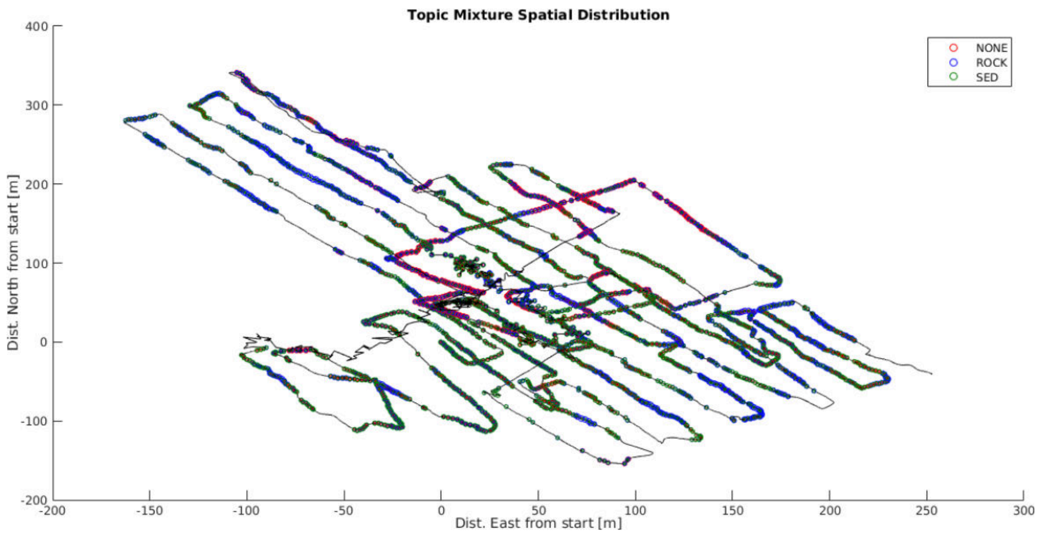


Figure 3–4: *Estimated substrate mixture map for Endeavour Ridge, BC:* ROV track with topic mixtures at each sample point. The line represents the path of the ROV, and each point is the location of a substrate sample image. At each point, there are circles for each of the three topics, with their sizes representing the mixture of the topics in that sample.

Finally, in Fig. 3–4 we present the map of high-level substrate types produced by the LBP topic model. Maps of this type provide an interface between our method and biological or geological research which seeks to use information about substrate type. For instance, this map could be used in conjunction with a map of observations of a certain species to help answer questions about how habitat suitability is related to substrate type. Note that most points are described mainly by a single type, and that the substrate type assignments are spatially consistent. This outcome is crucial to an intuitive understanding of the map; it allows scientists to think naturally about regions of the substrate while still admitting that transitional areas which cannot be labeled as a single type exist. By choosing a spatial topic model to capture the visual data, we explicitly require our method to recover such an interpretable model. The automatic substrate classifications of our work are included in the GIS database for the Endeavour marine protected area, and are publicly available to researchers interested in the area (Douglas et al., 2017) [1] .

3.2 Same-Place Recognition from Ambient Sound

Our second example application of spatio-temporal topic models highlights the versatility of the model, operating on features from ambient audio. To the best of our knowledge, this is the first application of topic models to environmental audio that takes advantage of spatially coherent semantic regions. Our ability to consider

[1] An interactive version of this database is available online at `http://www.oceannetworks.ca/endeavour-hydrothermal-vents-marine-protected-area`

temporal dependencies is crucial for this application as sound is fundamentally linked to temporal variation.

In this work we consider how to identify similar world regions by modeling the statistics of ambient sounds. We hypothesize that transitions in acoustic space correlate with transitions in other characteristic properties of the environment. Therefore, we aim to use the ambient sound to identify particular places and detect when we have returned to places we have already visited. In the context of robotics, we envision this system as one of many inputs to a filter-based probabilistic localization system. Common sensors such as range-finders and cameras often produce precise location hypotheses, but they are also susceptible to ambiguities which give rise to multi-modal beliefs. In contrast, an ambient-sound based localization system would give relatively imprecise hypotheses, only identifying the high-level region rather than an exact pose. Ambiguities in the soundscape of an environment, however, are potentially very different from keypoint or corner ambiguities which confuse traditional sensing modalities. As a result, even such 'weak' loop closures from sound could help improve the performance of localization systems.

Similar to the substrate classification example, in this work, an agent traverses the environment, makes observations along the way, and seeks to produce an intuitive description of these observations. As with substrate classification, an intuitive account is one that describes the world in terms of a few types, where neighboring locations should be assigned to the same type most of the time. Our spatio-temporal topic modeling framework aims to exploit this intuition to learn an appropriate representation without using any ground-truth labeling

As we are interested in an input to a localization system, in this example the topic model features temporal smoothness rather than spatio-temporal, and observations are not accompanied by their position in the world. In addition, this system is designed to be fully unsupervised. That is, rather than matching topics with labels from a small training set, in this work we allow the topics to describe the data in whatever way represents it most accurately, and aim to extract information by comparing topic priors for different temporal neighborhoods to one-another directly.

3.2.1 Acoustic Features

In the previous example we have seen the importance of carefully designing a feature function. For this example, we must take a short window of ambient audio and turn it into a discrete-valued feature. We extract such a feature using a k-means codebook of Mel Frequency Cepstral Coefficients (MFCC).

MFCCs are a compact representation of the timbre of an acoustic signal over a short time window. They have been commonly used to detect the source of a signal, especially in speaker and instrument recognition problems. Successive coefficients of a cepstrum represent the amount of the signal due to increasingly complex sinusoids. These are calculated using the inverse DFT of the log of squared DFT magnitude of a signal. MFCCs in particular use a re-weighting of the magnitude spectrum, so that lower frequencies are treated as more important than higher ones in order to achieve a more perceptually relevant cepstrum.

To be exact, MFCCs are computed using a filter bank with L overlapping bands with triangular magnitude responses $\hat{h}_l[k]$ for $l = 1, 2, ..., L$. The center frequency of band l is given by $f_c(l) = 700(e^{l/1127} - 1)$. The overlap percentage is constant and

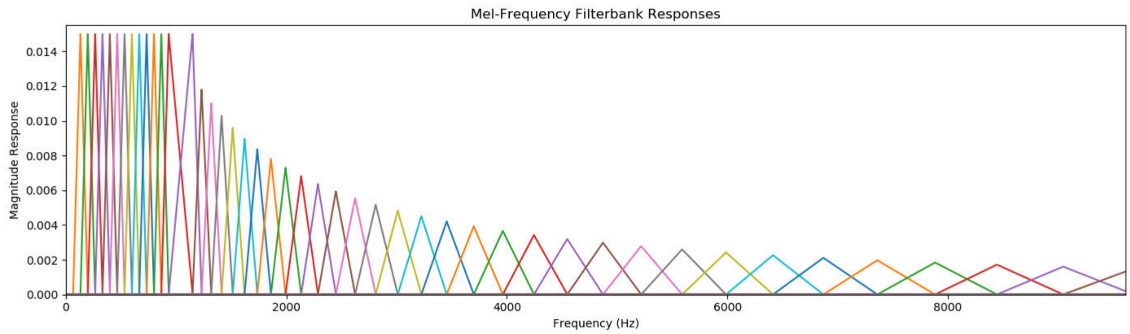


Figure 3–5: Mel- filterbank h_l frequency responses for 48-bands and 50% overlap. This filterbank implements perceptual reweighting of an audio signal.

the centers are spaced further apart as frequency increases (see Fig. 3–5), resulting in higher precision and weight on the lower frequency part of the signal. Given an input window x the Mel-Frequency Cepstrum is given by

$$cc(x) = \left| \mathcal{F}^{-1} \left(\log \left(\left| \sum_{l=0}^{L} \mathcal{F}\left(x\right) \hat{h}_l \right|^2 \right) \right) \right|^2 \tag{3.2}$$

The first M coefficients can be computed efficiently using a DCT formulation where the filterbank is applied directly to the time-domain signal. In this domain, an interpretation where the m-th coefficient represents the part of the signal that can be described by the m-th cosine basis function arises.

We consider the truncated cepstra, i.e. the first M coefficients for short, overlapping windows of audio, as is typical for instrument and speaker identification problems which employ cepstral techniques. More detail on the construction of the filter bank and on choosing appropriate parameters for the window size and number of coefficients to calculate can be found in (Sturm et al., 2010).

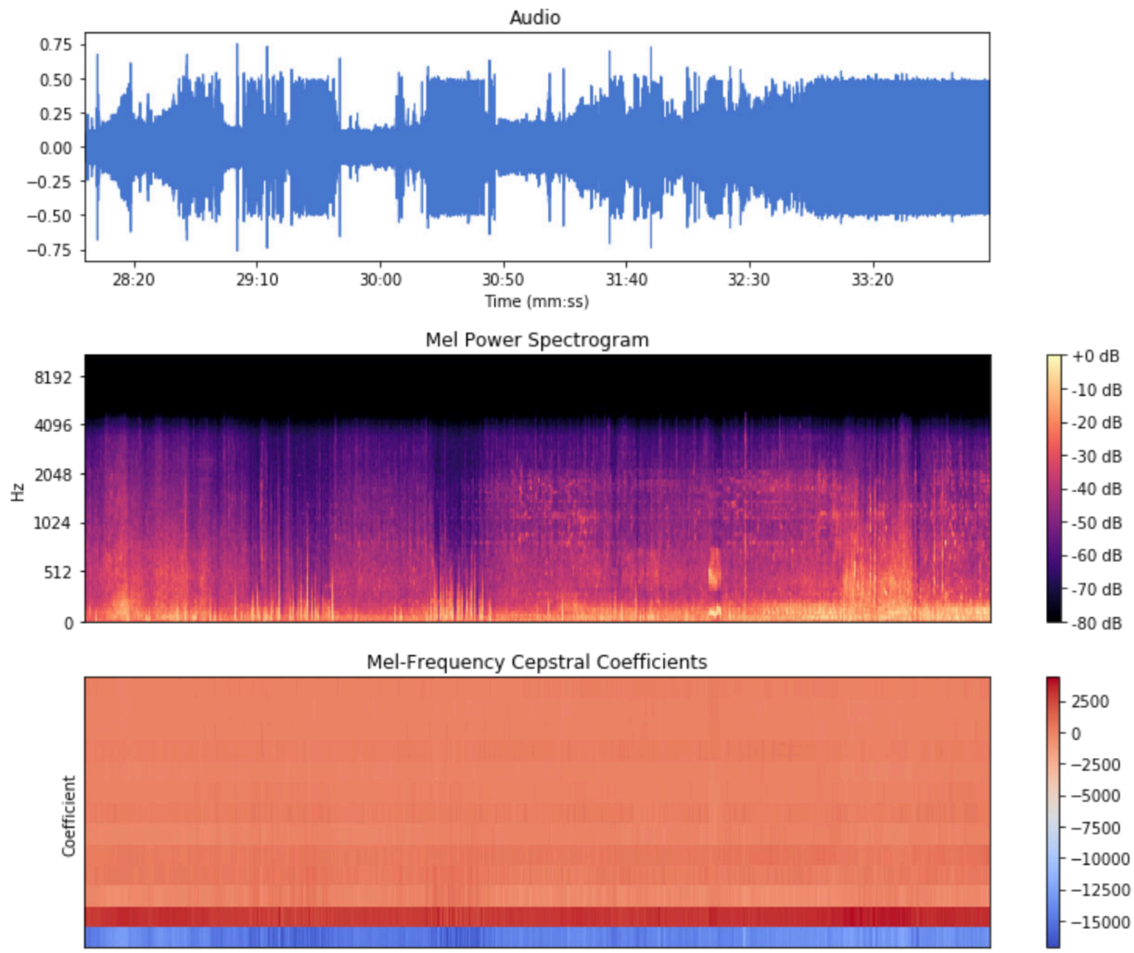


Figure 3–6: Segment of our audio processing pipeline corresponding to approximately 6 minutes from the 'Campus' dataset detailed in 3–7a (region labels 4,5,6,7). Compared to more traditional problems involving speech or music, obvious changes in the RMS amplitude and power spectrum are relatively subtle and infrequent.

By only considering the first M coefficients, we discard the most complex parts of the signal, and preserve only the 'spectral envelope', i.e. the general shape of the sound. The higher order coefficients encode more ephemeral aspects of the sound, which in the case of matching ambiances would be detrimental to robustness. Although the decomposition of an ambient sound's spectrum into envelope and noise (which coefficient is the appropriate boundary for a fundamentally noisy signal?) is much more obscure than that of an instrument or voice, we hypothesize that consistent noise sources such as air conditioners, passing vehicles, and distant vocal chatter are sufficiently distinct to be captured (for example, see Fig. 3–6). The final stage of our feature extractor is quantization via a k-means codebook, similar to the quantized SIFT features in the previous example. With reasonable window size and overlap parameters, this method produces dozens of feature observations per second.

3.2.2 Evaluation

We recorded audio datasets from two trajectories through the McGill campus and surrounding downtown area of Montreal, 51 and 43 minutes long respectively. The audio was recorded in stereo from a commodity hand-held video camera at a 44.1 kHZ sample rate while walking at approximately constant speed, and later combined into a single channel. The trajectories were chosen to contain varied sound environments, and contain both indoor and outdoor sounds, as well as sounds from busy and quiet environments. The maps of these trajectories are shown in Figs. 3–7a and 3–8a. The dotted segments correspond to indoor environments. The *Four-loops* dataset consists of four loops through the trajectory shown in the map, while *Figure-8* is a trajectory which is topologically equivalent to figure '8', and is looped twice.

We graphically represent when the trajectory returns to the same place with similarity matrices, where element (i, j) is colored (non-black) if the agent was labeled to be in the same region at time i and j, i.e., region labels $r_i = r_j$. For each trajectory, we hand-labeled the region for each temporal window according to the segments shown in Figs. 3–7a and 3–8a, giving the similarity matrices seen in Fig. 3–7b and 3–8b. Thus, colored blocks in the ground truth similarity matrices correspond to sets of locations that belong to a single spatial region with a consistent acoustic profile. For example, the red squares in Fig. 3–7b correspond to sets of pairs of points along the trajectory that all have acoustic profiles produced at the roadway marked as region 1 in Fig. 3–7a.

The regions were produced by identifying points on the map where environment transitions occur, and were chosen to be geographically relevant rather than acoustically relevant in order to test the hypothesis that these two types of transitions often align. Doorways for entering and exiting buildings as well as the edges of campus were the main landmarks. Some gradual environment transitions occur in the datasets, for instance going from quiet outdoors parts of campus to busy ones. We do not try to capture these gradual transitions in the ground-truth, and instead just pick a single point where this transition occurs, as is in the transition from region 2 to 3 in Fig. 3–7a.

We first generated two vocabularies of size 1500 by clustering MFCC features from the two datasets, and then used the vocabulary from the first dataset to generate MFCC words for the second, and vice versa. Each MFCC word corresponds to a 92 millisecond window of the sound, with a 50% overlap with the previous window. We

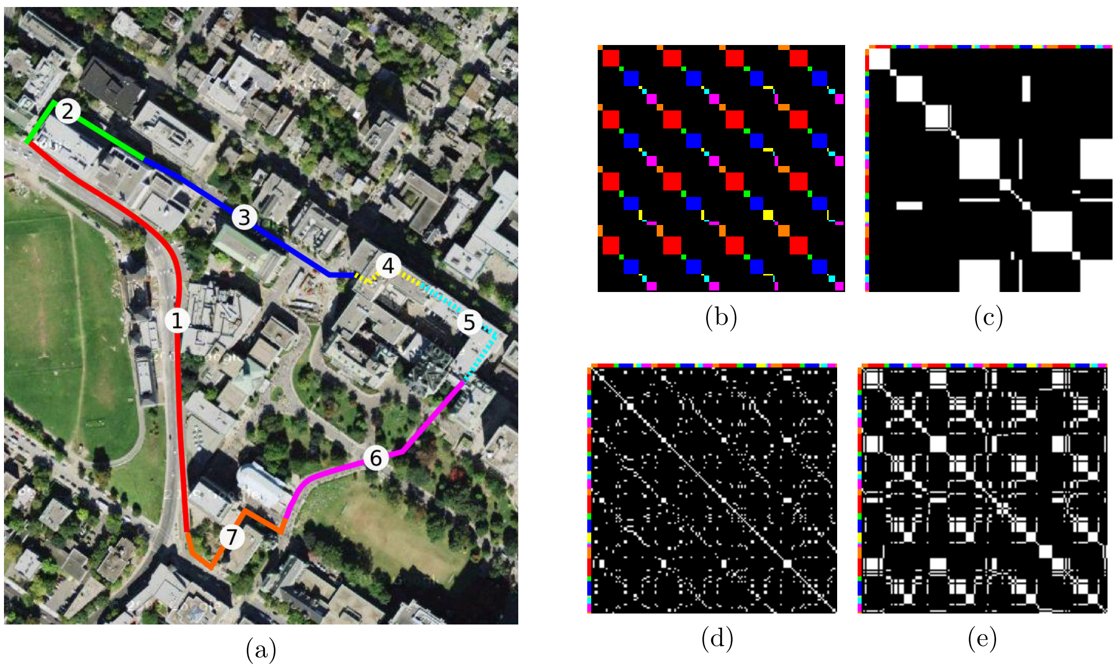
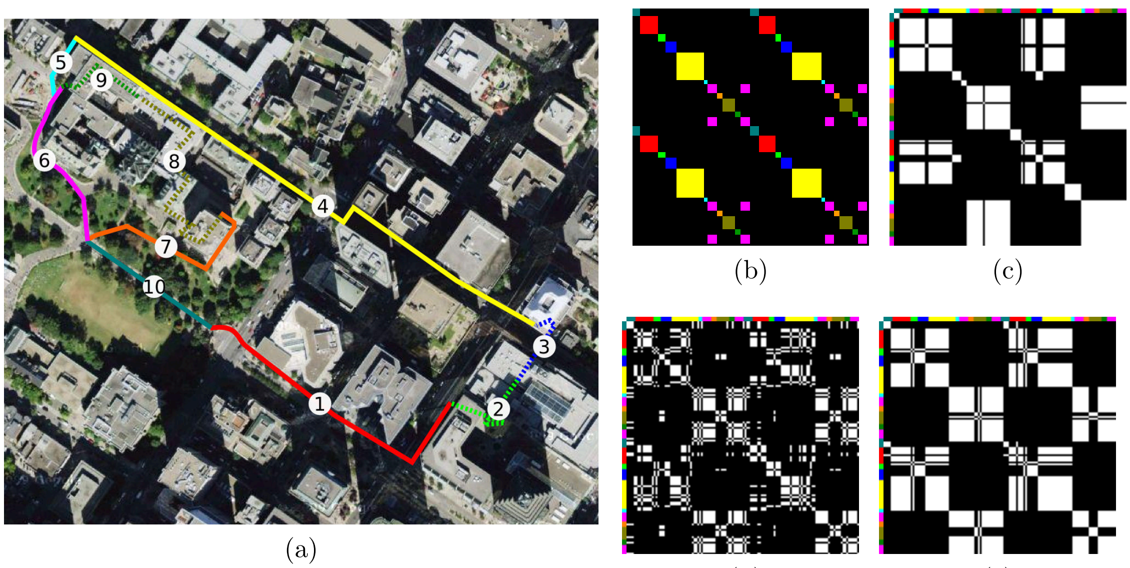


Figure 3–7: *Four-loops dataset:* (a) Map showing the path traversed while recording the dataset. (b) Ground truth similarity matrix. (c) Similarity matrix for feature-based region labeling ($g = 12$). (d) Similarity matrix for LDA-topics based region labeling ($g = 0$). (e) Similarity matrix for temporally smoothed LDA region labeling ($g = 4$).



(a) (b) (c) (d) (e)

Figure 3–8: *Figure-8 dataset:* (a) Map showing the path traversed while recording the dataset. (b) Ground truth similarity matrix. (c) Similarity matrix for feature-based region labeling ($g = 14$). (d) Similarity matrix for LDA-topics based region labeling ($g = 0$). (e) Similarity matrix for temporally smoothed LDA region labeling ($g = 2$).

then grouped these words into windows each representing 20 seconds of sound, with no overlap. The *Four-loops* dataset has 151 such windows, and the *Figure-8* dataset has 128 windows. We ran the temporally smoothed LDA on these datasets, varying the neighborhood size g, in number of 20-second windows.

For each window, we then compute the region label r_t by counting the most popular topic label in that window. Then, for each pair of windows, we compare the corresponding region labels, and mark the corresponding times to belong to the same region if the region labels match.

We experimented with neighborhoods of $g = 0 \ldots 10$ windows, and computed the true positive rates (TPR) and false positive rates (FPR) resulting from comparison with the ground truth matrix. TPR refers to the fraction of true positive similarity matches out of all positive results returned by the algorithm. Similarly, FPR refers to the fraction of false positive similarity matches out of all negative matches returned by the algorithm. An ideal algorithm has TPR of 1.0 and FPR of 0.0. The resulting plots of TPR vs FPR (known as a Receiver Operating Characteristic, or ROC, curve) are shown in Fig. 3–9. Figs. 3–7e and 3–8e show the similarity matrices with the best performance, chosen by their distance from the baseline performance on the ROC curve. Fig. 3–7d, 3–8d show the similarity matrix for neighborhood size $g = 0$, which is equivalent to a standard LDA topic model, where windows are assumed to be independent. This case is the leftmost point on the “topics” ROC curve.

In addition to computing the region assignment through our topic model, we also computed generated region labels directly from the window feature distributions. This model assigns the region label to be the index of the most common feature in

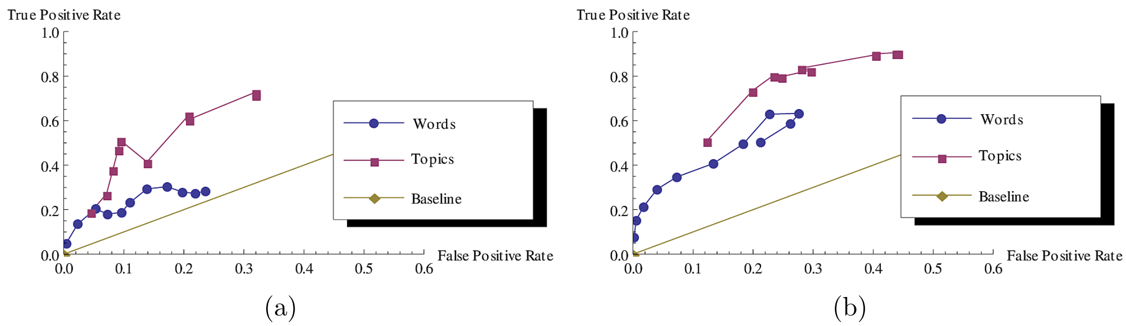


Figure 3–9: *ROC curves for same-place recognition based on ambient sound:* (a) Four-loops dataset. (b) Figure 8 dataset.

that window's neighborhood. We varied neighborhood sizes for the direct region labeling approach, using the same set of neighborhood sizes as we did for the topic model. The similarity matrices for the best performing neighborhood size setting are shown in Figs. 3–7c and 3–8c.

Our experiments indicate that the temporal topic model finds a representation where a simple region assignment is much more meaningful than directly considering the features or using a topic model without any temporal considerations. Fixing the false positive rate to be less than 0.25, Table 3–6 shows the best detection accuracy for each algorithm for both the datasets. It should be noted that the reported false positive rate is probably higher than the actual false positive rate because similar sounding regions at physically different locations were marked as distinct in our ground truth. For example, regions (1) and (4) from the *Figure-8* dataset are both recorded from busy streets, and sound the same, and as result are detected to be the same region by the proposed algorithm (Fig. 3–8e). From the similarity matrices

Table 3–6: Best accuracy of unsupervised same place recognition from sound.

Dataset	Best Detection Accuracy (False positive rate < 0.25)					
	MFCC-words		LDA		Temp. Smooth. LDA	
	TPR	FPR	TPR	FPR	TPR	FPR
Four-loops	0.29	0.15	0.19	0.04	0.61	0.21
Figure-8	0.63	0.23	0.50	0.12	0.80	0.23

produced by the system for the two datasets (Fig. 3–7e, 3–8e), we can see that the algorithm is successfully able to detect loop closure at many, but not all locations.

In both datasets, the temporally smoothed topic-based region labels outperformed the top-word-based region labels. Further, the topic model with temporal independence of windows (the leftmost point in the topics line) does not significantly outperform the naive feature-based approach. This indicates how critical it is to explicitly model the dependencies between adjacent observations. To reiterate, we do not require topics to explicitly correspond to particular real-world regions. Yet in order to find a self-consistent region labeling where returning to the same location means our model will produce the same label we still must consider the temporal smoothness of region labels.

CHAPTER 4
Combining topics and domain expertise to develop insight into plankton ecology

As we have seen, the spatio-temporal topic model is a powerful technique for unsupervised classification of location specific data. By carefully defining a feature function, we can achieve intuitive classifications that match human semantic annotations. Our interpretation of the model's outputs, however, has thus far been somewhat straightforward. In this chapter, we delve deeper into how topic models can be used for tasks beyond classification by interpreting neighborhood-topic distributions as a feature that can be used to predict word-distributions.

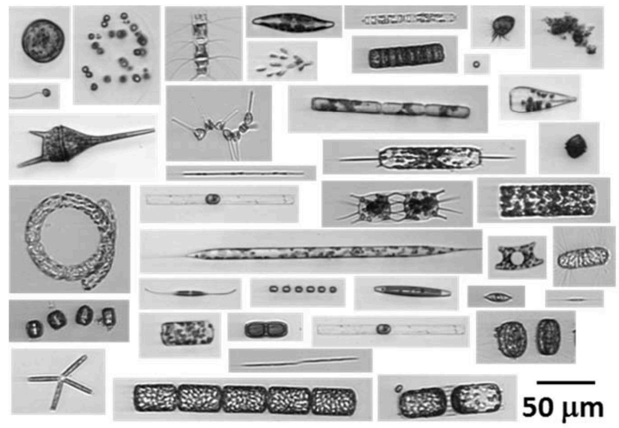
More specifically, we consider a dataset of population counts for a large number of species recorded throughout a spatio-temporal interval. Ecologists are interested in taking this type of dataset and producing a model that can predict the population of a species at another time and place, perhaps given some environmental factors such as climate conditions or other observed species. However, applying this approach directly does not scale well with the number of species in the dataset. As the number of recorded species populations increases, the data for a given time and place becomes more detailed and the interactions between species more complex. Capturing these details requires more powerful models, consequently increasing the required number of training observations, as well as reducing the intuitive value of the model even if it is well fit.

We address this scaling problem by applying a spatio-temporal topic model, capturing inter-species interactions by unsupervised means before considering other factors. We observe that given a neighborhood-topic prior the topic model gives a natural way to predict the distribution of observations that can be expected in that neighborhood. This means that if we can predict these priors, we can make predictions that can be interpreted in terms of real observations, not just abstract topic representations that we hope to align to real categories. Further, neighborhood-topic priors are relatively low-dimensional, and are designed to vary smoothly. In this chapter we demonstrate that the result of these properties is a representation that is relatively straightforward to model. By predicting the simplified topic representation we achieve more accurate species distribution predictions than other representations, and do so in a way that simultaneously suggests intuitive explanations for the data. Although we discuss this approach in the context of population prediction it can be applied much more generally as a method to simplify discrete feature representations.

4.1 Phytoplankton community model

In this chapter, we specifically consider population distribution data for a variety of phytoplankton. Our work represents novel contributions to both the machine learning and phytoplankton ecology literatures. Phytoplankton are microscopic organisms that form the base of marine food webs. They produce chlorophyll and other pigments to harvest sunlight and fuel photosynthesis, so they can utilize CO_2 and other nutrients to produce O_2 and new organic matter. As such, they play critical roles in global biogeochemical cycles and in structuring marine ecosystems. Marine

(a)



(b)

Figure 4–1: (a) Imaging FlowCyto Bot (IFCB) being deployed at Martha's Vineyard, MA (Photo cred: T. Crockford). (b) Example phytoplankton detections from the Martha's Vineyard IFCB dataset (Sosik et al., 2014) (Photo cred: WHOI-Plankton IFCB wiki [Public domain]). `https://ifcb-data.whoi.edu/about` offers an informal introduction to phytoplankton ecology and IFCB.

scientists have long used techniques to measure the amount of chlorophyll in a water sample as a proxy for phytoplankton biomass (Lorenzen, 1966). These methods are coarse and give only bulk indices, with no information about which species of phytoplankton are present. Phytoplankton are extremely diverse, however, and their community structure plays a major role in shaping ecosystems and their functions. As an extreme example, particular species are known to cause toxic blooms that can threaten wildlife as well as human health.

To meet the gap in observational capability that includes taxonomic resolution, Sosik and Olson have developed the automated, submersible Imaging FlowCytobot (IFCB, Fig. 4–1a) (Olson and Sosik, 2007). IFCB acts as an autonomous underwater microscope, collecting water samples and imaging small droplets at high resolution

($\sim$1 µm). The coupled analysis system (Sosik and Olson, 2007) is able to detect and crop phytoplankton within these images (See Fig. 4–1b). Many of these contain sufficient detail to be automatically classified to the genus or species level by a random-forest based system (Sosik et al., 2016). IFCB is designed to serially sample ocean water and perform detection and classification for long periods of time (weeks to years) with minimal manual intervention.

Sosik, Olson, and other IFCB users have collected impressively detailed datasets of plankton taxon (i.e. species or genus, pl taxa) abundance, both with long-term fixed location IFCB deployments, and over large areas, by deploying IFCB underway on a ship.[1] Nevertheless, this level of detail comes at the price of intuitive, comprehensive models using classic statistical techniques. To our knowledge, no other work has attempted to derive understanding about the interactions between multiple taxa in this dataset, nonetheless all of them considered together.

We hypothesize that by taking sparsity and spatio-temporal smoothness into account, a spatio-temporal topic model can capture the interactions between taxa, simplifying understanding this dataset both formally through further modeling and intuitively for domain researchers. Specifically, we construct a spatio-temporal topic model where the discrete-valued features (words) are the count of occurrences of each taxon. The neighborhoods are the temporal or spatio-temporal windows in which the observations were made by IFCB. We interpret the resulting word-topic distributions as *communities* – probabilistic groupings of plankton taxa that tend to

[1] Datasets available online at `http://ifcb-data.whoi.edu/mvco/`

co-occur. Under this interpretation, the topic-neighborhood distributions describe the mixture of communities found in a given location.

As we have seen in previous examples, the strength of the topic model is that it finds sparse, spatially smooth neighborhood-topic distributions. In this application, this is crucial for maintaining interpretability. Sparsity means that a researcher can look at the resulting map or timeseries of community distributions, and form a conceptual summary of what was observed there. In addition, spatial smoothness means that the neighborhood-topic distributions are simple to predict based on their locations or the distributions of their neighbors.

Nevertheless, there are few proposals for 'real' communities within the phytoplankton ecology literature. As a result, we cannot evaluate our model by attempting to match predicted communities with real labels. Instead, we consider the distribution of taxa implied by a neighborhood-topic distribution. Recall that the posterior probability of topic assignments $P(\boldsymbol{z}|\boldsymbol{w})$ can be estimated via the collapsed Gibbs sampler update (in Eqn. 2.3). Given a set of topic assignments, the posterior distributions for the neighborhood-topic priors Θ and the topic-word priors Φ are given by $Dir(\alpha + \boldsymbol{N}_{g(x)}^{1:K})$ and $Dir(\beta + \boldsymbol{N}_{k}^{1:V})$ respectively, where $\boldsymbol{N}_{\cdot}^{\cdot}$ represents the appropriate count of topic assignments. From these posteriors the maximum likelihood estimates, denoted $\hat{\Theta}$ and $\hat{\Phi}$ are trivial to recover:

$$\hat{\Theta}_{g(x),k} = \frac{N_{g(x)}^{k} + \alpha}{\sum_{j=1}^{K} N_{g(x)}^{j} + \alpha}, \quad \hat{\Phi}_{k,w_i} = \frac{N_{k}^{w_i} + \beta}{\sum_{v=1}^{V} N_{k}^{v} + \beta}, \tag{4.1}$$

Now consider fixing the values of the neighborhood-topic and topic-word priors, θ and Φ. In this case, the distribution of words for a neighborhood is completely determined by the values of θ and Φ. In the following two applications, we demonstrate methods which train a simplistic supervised model to predict the MLE neighborhood-topic distributions at a location x, in other words to predict some $\check{\theta}_x \approx \hat{\theta}_x$. Combining these two pieces, we predict the word distribution for a new location by

$$P(w = v|\theta, \Phi) = \sum_{k=0}^{K} \theta_{x,k}\Phi_{k,v} \approx \sum_{k=0}^{K} \check{\theta}_{x,k}\hat{\Phi}_{k,v} \tag{4.2}$$

Then, we compare this predicted taxon distribution to those that are actually observed in that location. This source of weak-supervision allows us to better guide our representation learning process, as well as to interpret the final learned representation in terms of concrete evidence compatible with the knowledge of domain researchers.

4.1.1 Learning the number of topics online

A final issue in the formulation of our phytoplankton community model is that because these datasets are so unique, to our knowledge no domain research has attempted to define phytoplankton communities, and therefore we do not have strong information about how to set the number of topics, K. For this reason, instead of fixing K ahead of time, we replace the Dirichlet prior on the topic priors with a Dirichlet Process (DP) prior, and learn the number of topics from data, following the approach of Bayesian Non-parametric ROST (BNP-ROST) (Girdhar, 2016). Using a DP instead of the plain Dirichlet prior for θ means that our model can account for observations that are better described by creating a new community rather than assigning it to one of the existing ones. At the same time it still allows enough control over how often this occurs to ensure the final set of learned communities is small enough to remain useful. With this approach, the model can be started with a small number of communities and be allowed to grow as the data demands.

More formally, the DP prior is defined with respect to a distribution G that has positive support over some probability space A. The defining property of a DP is that for any partition[2] of A, the probability of observing an event from a particular element of the partition is jointly Dirichlet distributed. For all partitions composed of the subsets $\{A_0, \ldots, A_r\}$

$$\begin{gathered} \text{if } H \sim DP(\eta, G) \\ H\left(A_0, \ldots, A_r\right) \sim Dir(\eta G(A_0), \ldots, \eta G(A_r)) \end{gathered} \tag{4.3}$$

where η is a positive real-valued parameter that controls the sparsity of the DP, analogous to α in a finite, symmetric Dirichlet distribution (Teh, 2010).

In our case, we are interested in a prior over topic assignments. Consider the probability space A to be the (infinite) set of non-negative integers, each indexing a topic. To construct a DP over this space, we consider the partition $\{z = 0, z = 1, \ldots, z = K-1, z \geq K\}$. If there are K topics currently in use, this partition defines the events of choosing any of the existing topics, or else creating a new topic. Because we have chosen a DP, the topic assignment prior is still (finite) Dirichlet distributed at any given Gibbs sampler iteration. We can view this in terms of the 'Chinese

[2] i.e. any finite, disjoint, and covering set of subsets

Restaurant Process' construction, reparameterizing G and η in terms of our original symmetric Dirichlet parameter α and the probability γ of incrementing the number of topics (Jordan, 2005). This interpretation leads to a predictive distribution which is only a minor modification from the Dirichlet predictive distribution (Eqn. 2.3):

$$P(z_i = k|\boldsymbol{z_{-i}}) = \begin{cases} \frac{(N^k_{g(x)}+\alpha)}{\gamma+\sum_{j=1}^K N^j_{g(x)}+\alpha} & \text{if } k < K \\ \frac{\gamma}{\gamma+\sum_{j=1}^K N^j_{g(x)}+\alpha} & \text{if } k = K \end{cases} \tag{4.4}$$

where K is incremented whenever some $z = K$ is drawn from the resulting distribution.

4.2 Learning communities that are explained by the environment

Our first application involving IFCB plankton data aims at learning an interpretable community model for a fixed-location deployment at Martha's Vineyard Coastal Observatory (MVCO), one mile off the south-shore of Martha's Vineyard, MA (Kalmbach et al., 2017). In addition to an IFCB, MVCO has a suite of other sensors, enabling inquiry into the relationships between phytoplankton abundance and environment variables including oceanographic and meteorological factors such as dissolved 0_2 content, salinity, air and water temperature, and rainfall. While the combination of these two data sources could be used as the basis to learn intricate models of the precise relationship between each taxon and the environment, in this work we show that by first accounting for the associations between taxa, our method can capture most of the same information while requiring minimal domain knowledge.

We propose to use a simple linear-ridge regressor to predict the distribution of phytoplankton taxa from environment variables. We first train the community model and compute the MLE $\hat{\Theta}$ and $\hat{\Phi}$. Because of the prior imposed by our model, $\hat{\Theta}$ is relatively low-dimensional, sparse, and temporally smooth in comparison to the full phytoplankton distributions. Therefore, we hypothesize that estimating the communities and then predicting the phytoplankton taxa by Eqn. 4.2 will produce more reliable predictions than a direct regression technique.

In addition, this regression task provides complementary feedback to the topic model, ensuring that the final community model is both interpretable and accurate. For this application, instead of using the probability of held-out data as the cross-validation metric, we propose to use the agreement (which we define as $1/D_{KL}$) between the predicted taxon distributions based on environmental factors and the ground-truth taxon distribution. By doing so, we ensure that the maximum likelihood taxon distributions given the topics closely resemble the true taxon distributions, but also that the topic distributions themselves can be explained simply in terms of the chosen environment variables. For this reason, rather than choosing a complex regression model, which may have better overall performance on the final predictions, we take an unsophisticated linear ridge approach to regression.

4.2.1 Martha's Vineyard multi-year timeseries experiment

We demonstrate our method on a dataset recorded continuously from Jan. 2009 to Jul. 2016 at MVCO. IFCB was configured to automatically sample from 5 ml of surface seawater approximately every 20 min. The classification system generated an average of over 1100 observations per day, distributed over 47 taxa (Fig. 4–2).

We aggregated the observations to produce the taxon distribution for each day during the 7.5 year period, and used this as input to our method. For the regression model, we chose a suite of 18 environment variables from the MVCO ocean data and meteorological data summaries (Fig. 4–4). In addition to being naively chosen, the environment data features significant gaps and systematic noise due to the practical challenges of long-term ocean sensor deployments. Much of the systematic noise was suppressed by rejecting outliers based on the median absolute deviation of each variable, yet the regression task remains extremely challenging.

We trained our community model on the taxon distributions over the entire period multiple times for different hyperparameter settings, varying the topic concentration α, the topic-prior concentration β, the DP prior parameter γ, and the neighborhood size g. For each community model corresponding to a different combination of choices of these hyperparameters, we trained the regression model 8 times, once using each year as the test set and the other 6.5 years for training. Within each training set, we chose the regularization parameter for ridge regression using hold-one-out cross validation. Finally, we used the resulting regressors to predict the respective held-out community distributions for each year, and used the respective community models to predict the taxon distributions.

We demonstrate the utility of our method in comparison to two more standard regression techniques. The first is to predict the taxon distributions directly with a similar ridge regression model and training procedure. The second is to first take a PCA decomposition of the taxon count data, using the first K principle components, where K is the same as the number of communities used by our model, and then use

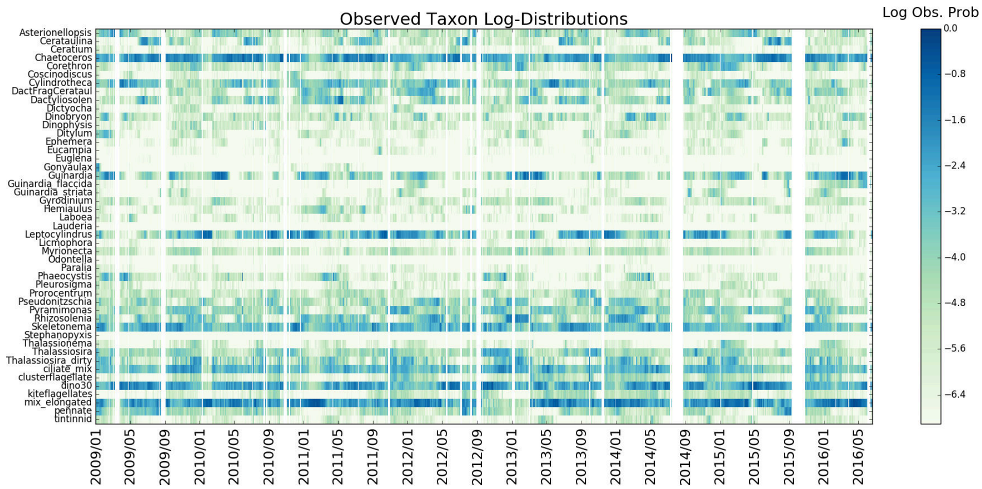


Figure 4–2: Observed daily taxon log-distributions at Martha's Vineyard Coastal Observatory over 7.5 years.

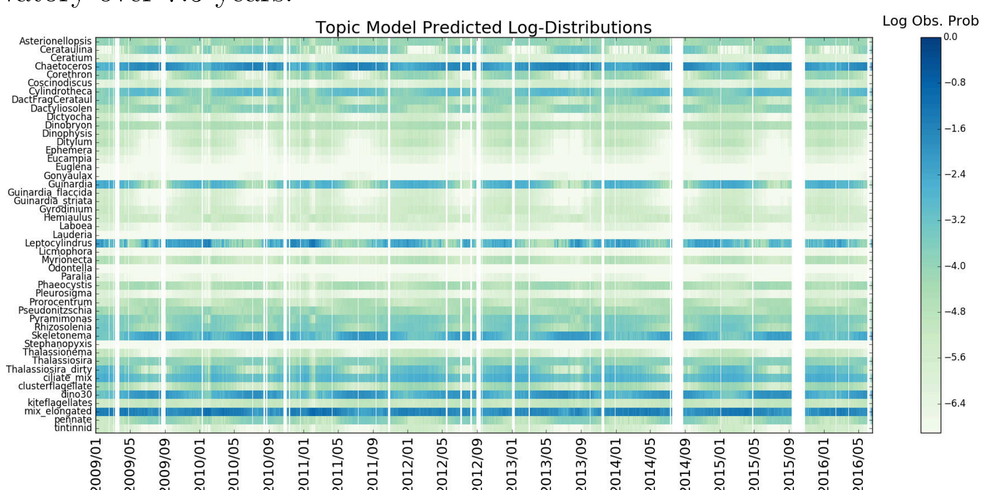


Figure 4–3: Predicted taxon log-distributions using our community model and regression system.

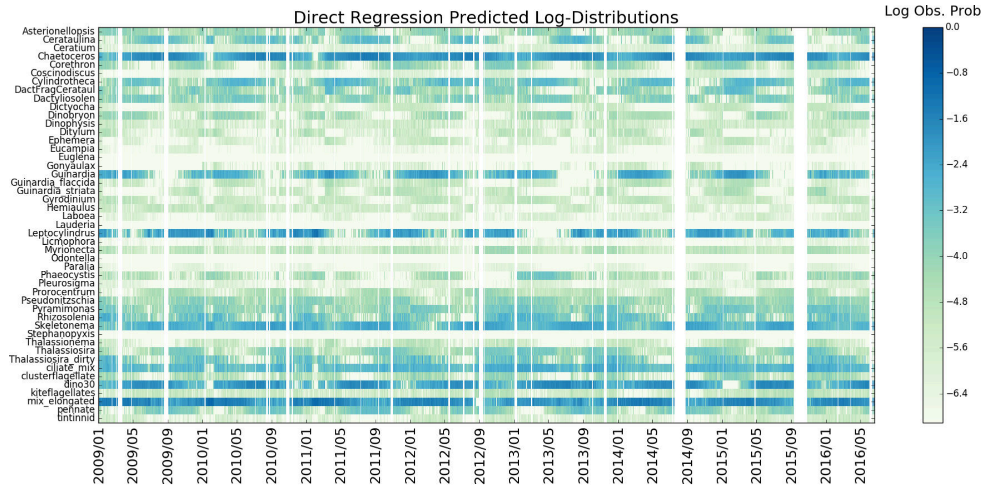


Figure 4–3: Predicted taxon log-distributions using direct regression on taxon distributions.

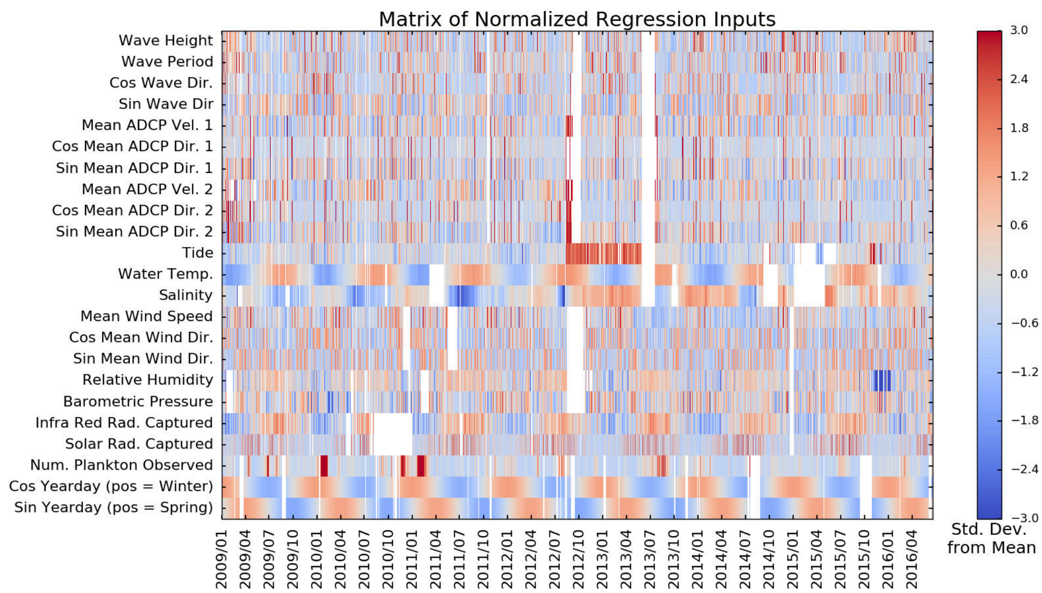


Figure 4–4: *Environment variables used to predict phytoplankton distributions*: Oceanographic and meteorological factors potentially related to phytoplankton lifecycles, centered and scaled to a normal distribution as our regression system receives them. White spaces indicate gaps in the data or where outlier data was removed.

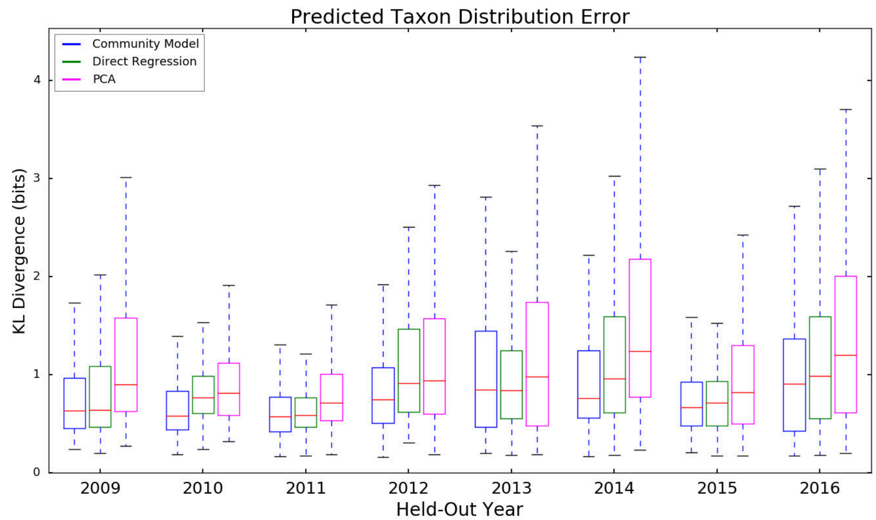


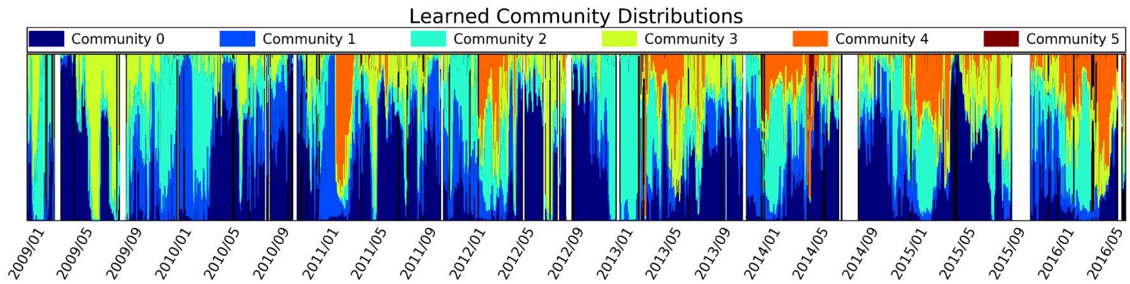
Figure 4–5: Comparison of daily taxon distribution prediction divergences for each of the three regression methods and each of the years in the dataset. For each year, the other 6.5 years were used as training data. The community based regression method (ours, left) shows the lowest median KL-Divergence for all years.

the same ridge regression model and training procedure to predict the PCA weights, and finally project the weights back to predict the taxon distributions.

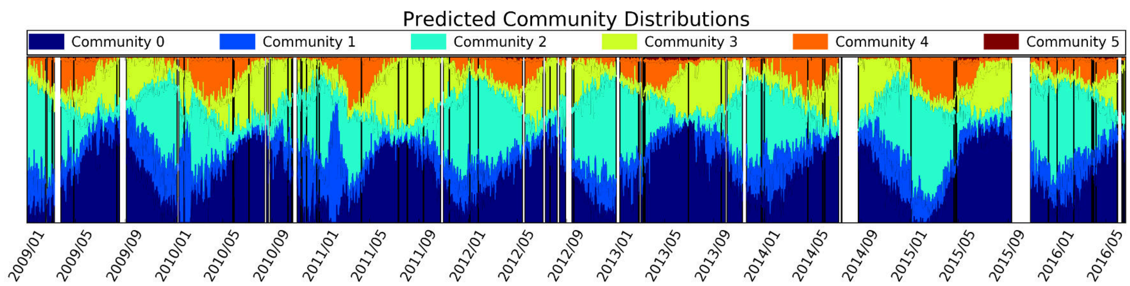
We evaluated the resulting regression systems for all hyperparameter settings as well as the two baseline methods by comparing the predicted taxon distributions to the true distributions on each day in the dataset. Our error measure is the KL-Divergence between the predictions and the held-out distributions We chose the community model hyperparameters settings with the lowest average KL-Divergence over all the days in the dataset, ultimately picking a model with 6 active communities. The predicted taxon distributions for this model are shown in Fig. 4–3.

Fig. 4–5 shows the taxon distribution prediction errors for this community model and our two baseline models, broken out by year. The boxes represent the distribution of prediction errors for nearly 365 days in 2009 through 2015, and 172 days in 2016 (nearly 365 because of some small gaps in the taxon count data). Our community model (leftmost for each year) achieved the lowest median error on every year in the dataset. We found that by optimizing the hyperparameters for the regression task, we were able to choose an interpretable representation of the community structure. In contrast, PCA does not feature any prior for temporal smoothness. As a result although its prediction error is on average only a little less accurate than our model's, the sequences of predictions it makes are sometimes implausible, featuring taxon distributions that change much more rapidly than the observed data. We found that both baselines were extremely susceptible to noise in the environment data, on average performing better than expected, but occasionally making extremely poor predictions (Fig. 4–7). With our model, the regression problem is of a lower dimensionality than for direct regression, and therefore less susceptible to overfitting. For this reason when both models are presented with the same small amount of training data, our model is more able to avoid large errors for new inputs unlike the training data.

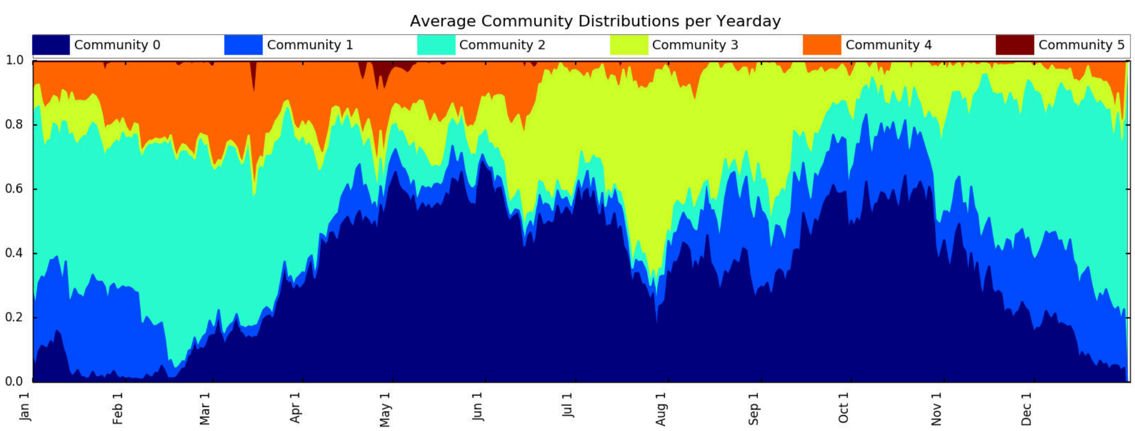
An intriguing result of our community regression model is that nearly all of the magnitude in the weights of the learned regression parameters is either on the day of the year, the water temperature, or the number of plankton classified for a given day (Fig. 4–8b). From our regression matrix, we can see that 5 out of 6 communities has a seasonal niche with which most strongly predicts its presence. We found that the best



(a)



(b)



(c)

Figure 4–6: The learned and predicted community distributions. The horizontal axis represents time, and each community is represented by a color. The fraction of observations on a day belonging to a particular community are shown by the size of the colored area (totaling 1 for each day). Our model finds strong seasonal structure in the data without having such an assumption built-in. (a) Daily community distributions over 7.5 years for the best performing community model on the regression task. (b) Daily community distributions predicted from environment data. (c) Average community distribution for each day of the year over the entire dataset.

performing community models showed strong seasonal structure, rather than relying on other variables. This is made evident by the average community distribution for each day of the year over the entire dataset before regression (Fig. 4–6c). Note that there are many possible community decompositions, and although our model makes weak assumptions about the temporal smoothness of the communities, it does not have any prior knowledge of the seasonal nature of the data. By performing hyperparameter optimization over the downstream regression task, we were able to select a model with just the right level of sparsity and temporal smoothness to emphasize this seasonal aspect and describe the data in an interpretable way.

Another outcome of our experiment is the communities themselves learned by our model (Fig. 4–8a). We found that across different hyperparameter choices, the top few most active communities were relatively similar to those presented here. Some associations based on our model have ready explanations. For instance community 1 is dominated by the taxa "mix_elongated", representing miscellaneous centric diatom chains, and "leptocylindrus", which both exhibit elongated morphologies and easily confuse the IFCB's vision-based classification system. As a more exciting example, communities 2 and 4 are the only communities with significant probability of observing the taxon "Guinardia", and are predicted by warm water temperatures, while *Guinardia delicatula* populations have been noted to be negatively associated with parasites that do not survive during cold winters (Peacock et al., 2014).

Our community model takes a complex, high-dimensional population dataset, and offers immediate questions to pursue related to the ecology of specific taxa near MVCO. For instance, in Fig. 4–6a we see that community 4 begins to appear

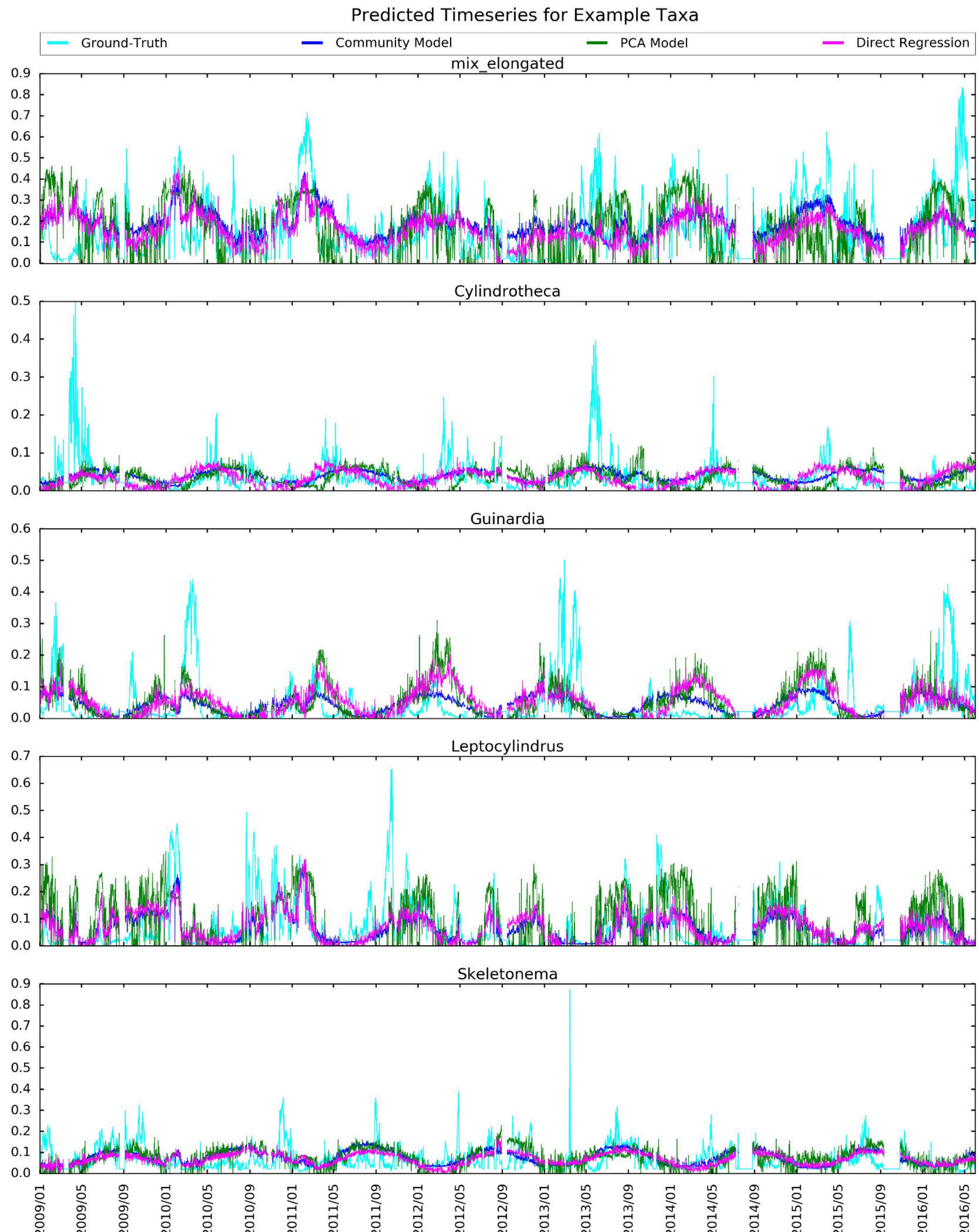
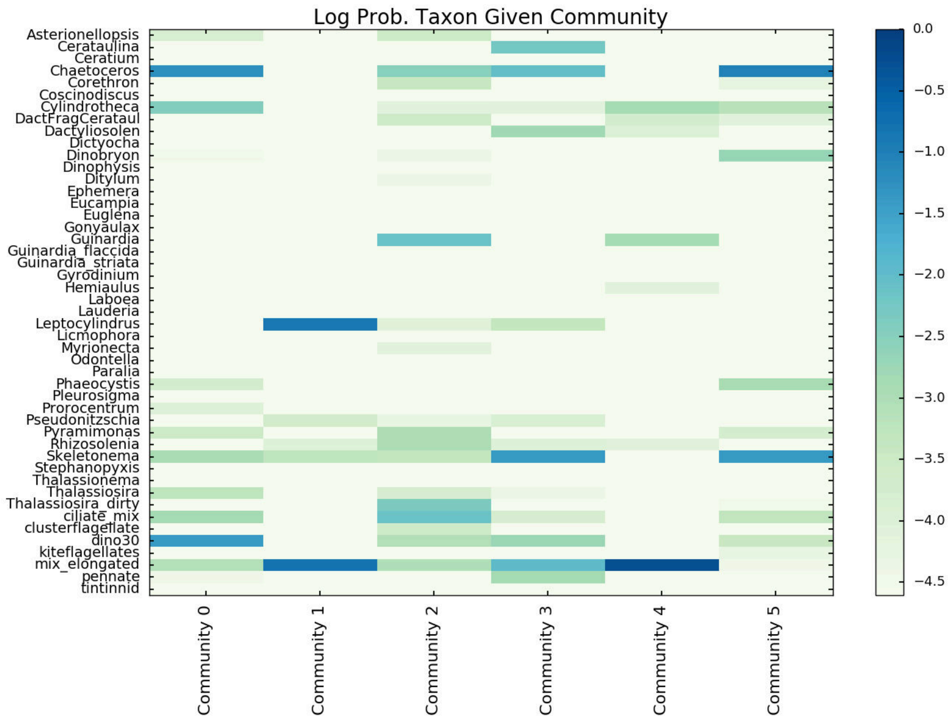
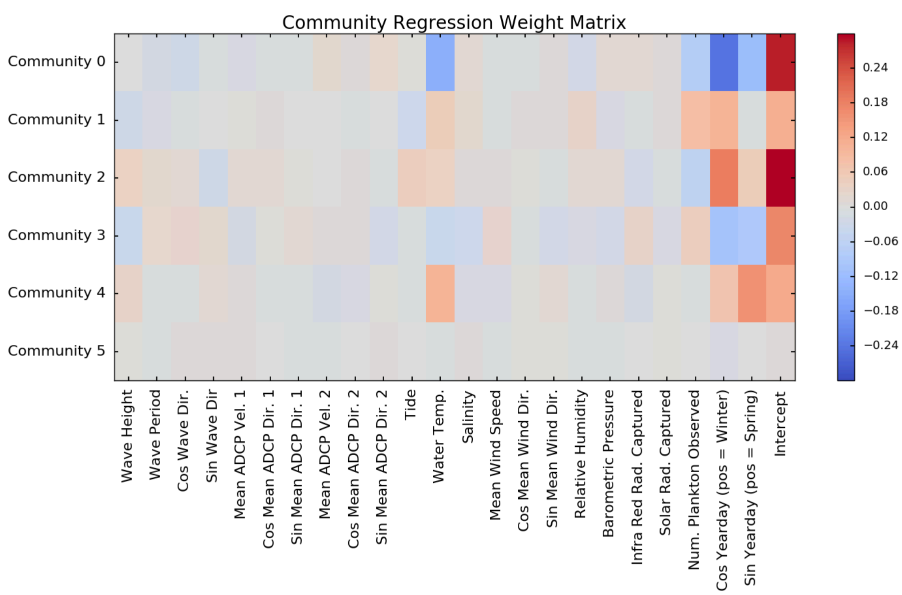


Figure 4–7: Timeseries representation of individual probability of observing 5 common taxa on each day in the dataset. Note that the community model's predictions are less susceptible to noise that the other two strategies.



(a)



(b)

Figure 4–8: (a) Log-probability of each taxon for each community. (b) Linear regression weight matrix for top-performing community model.

a little earlier each year from 2011, at first appearing only in the spring, and by 2015 persisting throughout the Winter. If we were to only look at the timeseries for the individual taxa, “mix_elongated”, “Guinardia”, and “Cylindrotheca” which dominate community 4, this pattern is not obvious at all (See Ground-Truth, Fig. 4–7). Our model, however, highlights that these taxa often co-occur, and that the pattern of co-occurrence shifts over the years.

4.3 Predicting presence of missing phytoplankton by their associates

In our second study of IFCB plankton data, we ground our community model by its ability to predict the presence of an artificially held-out plankton taxon from the others, rather than predicting the full distribution from auxiliary data (Kalmbach et al., 2017). Further, in this application we explore spatial phytoplankton classification data, demonstrating our model’s ability to capture both the temporal and spatial smoothness aspects of community structure.

This work is motivated by a scenario where a plankton ecologist is searching for a particular plankton taxon. If the ecologist deploys IFCB on a ship and periodically collects samples, but never observes the taxon of interest, she requires further information to decide whether to move on or keep collecting samples in the same location. Because IFCB samples are a small volume of water, yet ecologists would like to describe large volumes of ocean that have complex ecosystems, it is very likely that samples will sometimes be deficient in taxa that are, in fact, found near the sample location. Therefore, we propose to model the associations between taxa in such a way that if we disregard the target taxon we can still estimate the community mixture for a location accurately, and consequently predict the presence or absence of the target from its known associates

We would like to estimate the probability of observing a falsely-missing target taxon $v^\star$, based on the observed distribution of the $V-1$ other taxa. We assume that the taxon is not systematically missing, in other words that we can collect a training set at some other locations comprised of all V taxa. Then given this training set, we learn the topic assignments and MLE priors $\hat{\Theta}$ and $\hat{\Phi}$. At test-time we disregard the target taxon, assuming it is falsely-missing. To address this we consider the topics excluding $v^\star$ as the maximum likelihood of the Dirichlet posterior over the other $V-1$ taxa. In other words, at test time we use topics which ignore the missing taxon:

$$P(w = v|z = k, \boldsymbol{w}) \approx \hat{\Phi}_k^{-v^\star} \triangleq \frac{N_k^{w_i} + \beta}{\sum_{u \neq v^\star} N_k^u + \beta}, \tag{4.5}$$

Next, we obtain topic assignments $\boldsymbol{z_{-v^\star}}$ by substituting $\hat{\Phi}_k^{-v^\star}$ for Φ in the predictive distribution Eqn. 2.3. In the iterative process of sampling topic assignments, we fix $\hat{\Phi}_k^{-v^\star}$, and only update topic assignments for *new* observations based using the 'target deficient' topic assignment counts. These define the approximate MLE topic priors $\check{\Theta}_{g(x)}$. Finally, we estimate the proportion of the data which would have been made of the target class using the original MLE topic matrix

$$P(w_x = v^\star|\mathbf{w}_{-v^\star}) \approx \sum_{k=0}^{K} \check{\Theta}_{g(x),k} \hat{\Phi}_{k,v^\star} \tag{4.6}$$

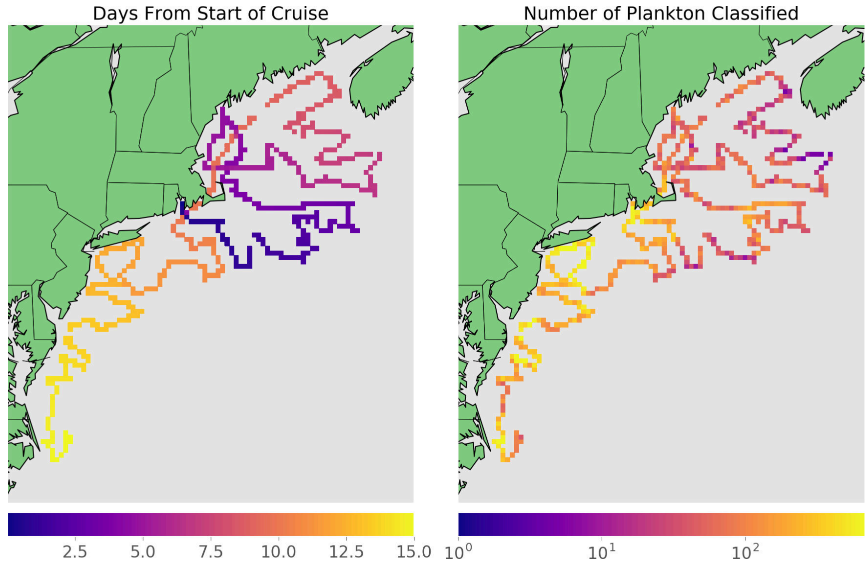


Figure 4–9: Summary of data recorded during the Pisces 14-05 cruise. Left, color-scale shows progress in time. Right, color-scale shows the number of plankton observed at each sample location (log scaled).

4.3.1 US Atlantic coast hotspot prediction experiment

We demonstrate this approach with IFCB classification results from NOAA's Fall 2014 EcoMon Survey aboard the Research Vessel Pisces (Cruise PC 14-05). The IFCB was configured to automatically sample from underway flowing surface seawater during the period 4-19 November 2014. The classification system generated over 140,000 individual phytoplankton observations from these water samples. Classification results comprise a dataset with 47 taxa at 852 locations spanning the US Atlantic coast from North Carolina to Maine (See Fig. 4–9).

We divide the sample locations into equal-sized parts, representing the training and test phases of the simulated mission. The counts of all 47 taxa were kept in the training set and used to learn the topic model. For the test set, we held out each of the 8 most-frequently observed phytoplankton taxa one at a time. We define the hotspots of a taxon to be the top 50 sample locations in the test data, where the relative abundance of the taxon to all other taxa was highest. The 8 tested taxa make up just over 81% of all the observations in the dataset. The most common taxa are miscellaneous centric diatom chains ("mix_elongated"), mixed species of pennate diatoms, *Thalassiosira* spp., *Guinardia delicatula, Guinardia striata, Dictyocha* spp., *Ephemera* spp., and *Phaeocystis* spp. While our method accounts for the sparsity of taxon distributions, this dataset also features sparsity in terms of the locations of observations. To address this separate issue, we resort to a 2D spatial median filter. Finally, we apply a threshold to identify the hotspot locations.

We compared our method to an exhaustive search strategy and a k-means search strategy. For each sample in the test set, exhaustive search estimates the probability of observing $v^\star$ by looking up the sample in the training set with the most similar distribution to the observed data. This represents the strategy which makes the most use of all the data available for every test sample, at the cost of a linear computational complexity in the number of sample locations in the dataset. In the k-means strategy, we fix a constant test-time complexity by reducing the search space to the K centroids returned by a standard k-means clustering implementation. These centroids are defined such that if each class distribution in the training set were replaced by the nearest of the K centroids, the sum of squared error is approximately minimized, however it does not take into account the sparsity or spatial smoothness of the underlying distributions.

We carried out experiments for two different train/test regimes. First, we used every second sample location for training (see Fig. 4–10, column 1). This regime simulates a mission where the classifier frequently fails to identify examples of a class, for instance because its classifier was poorly tuned. Because nearby sample locations tend to have similar distributions, this regime tests the ability of a model to interpolate over small distances. Second, we used the first half of the sample locations as training (Fig. 4–11, column 1), and the second half for testing. This latter case simulates a mission where the capabilities of the classifier have changed from the first half to the second half. It tests the ability of a model to predict in a new location that is not likely to have any spatially linked correlation with the training data.

We ran our model for a range of choices of the hyperparameters $\alpha, \beta \in \{0.001, 0.01, 0.1, 0.5, 1\}$ and $\gamma \in \{10^{-6}, 10^{-5}, 10^{-4}\}$ with each of the top 8 taxa held out of the testing data and for both training regimes. We also ran the exhaustive search and k-means strategies for each. The strategies each produce an estimate for $P(w = v^{\star}|g(x))$, which we then smooth with a median filter with size parameter σ. For a scalar threshold τ, we predict that location x is a hotspot if $\Pi_{\sigma}(P(w = v^{\star}|g(x))) > \tau$, where Π_{σ} is the median function over a square region with side length σ.

To evaluate our results we compare the held-out locations in the test set (Fig. 4–10 and 4–11, column 2) to predictions from each of the proposed strategies (Fig. 4–10 and 4–11, our model, column 3; exhaustive nearest neighbor search, column 4; and k-means search, column 5). The input to the models is illustrated with the observed values of the held-out class at the training locations (Fig. 4–10 and 4–11, column 1).

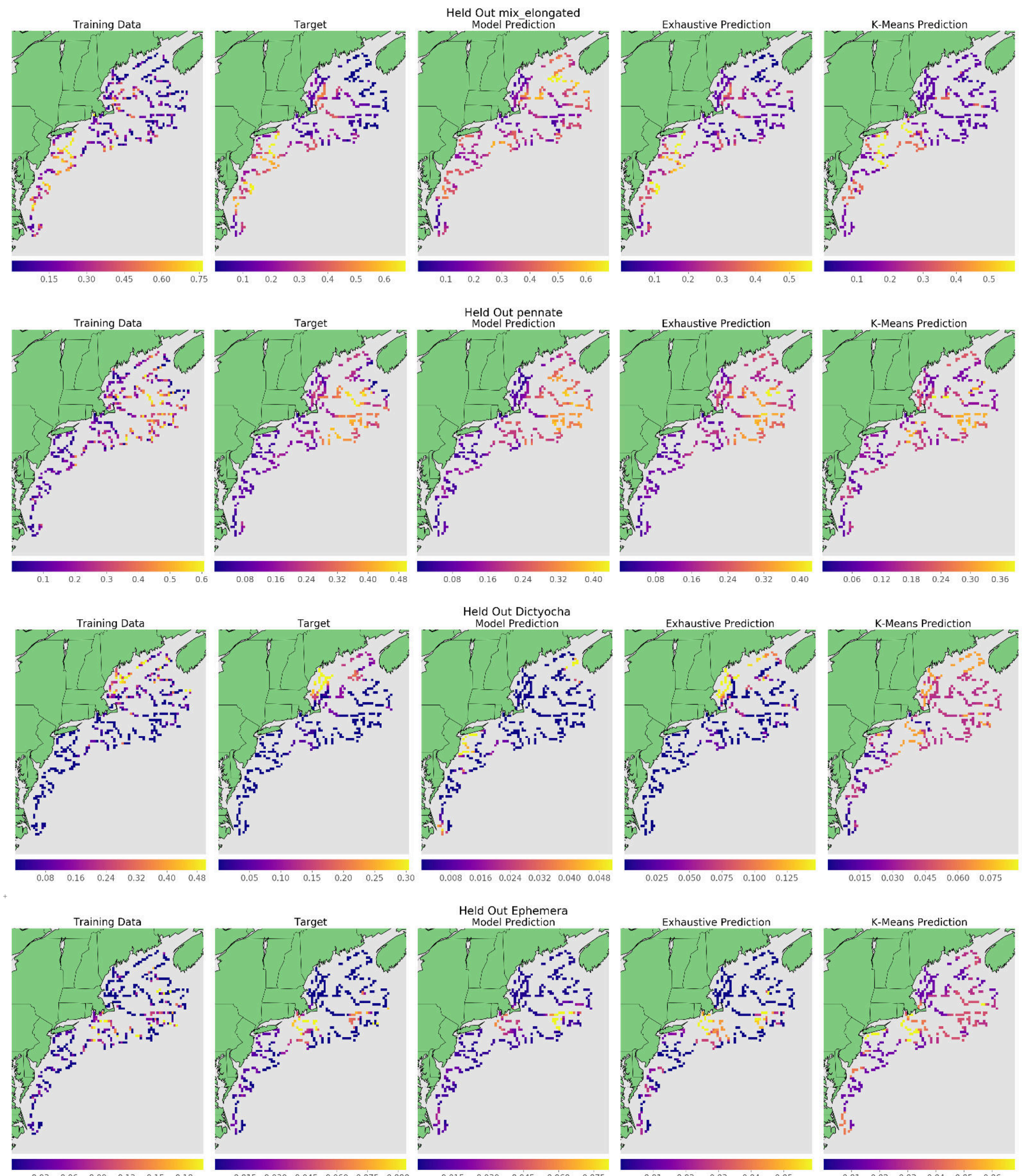


Figure 4–10: Spatial distribution for four target classes (rows) in interleaved training/testing samples. The columns correspond to training data (col. 1), held-out target locations (col. 2), and the three models under evaluation (col. 3-5). We find close correspondence between the proposed model and the target data, but exhaustive nearest-neighbor approach has the most similar distribution to held-out target locations. This is because the distribution of plankton is correlated with its spatial neighbors, and hence simple interpolation of the training data is likely to give an accurate plankton distribution at the held-out locations.

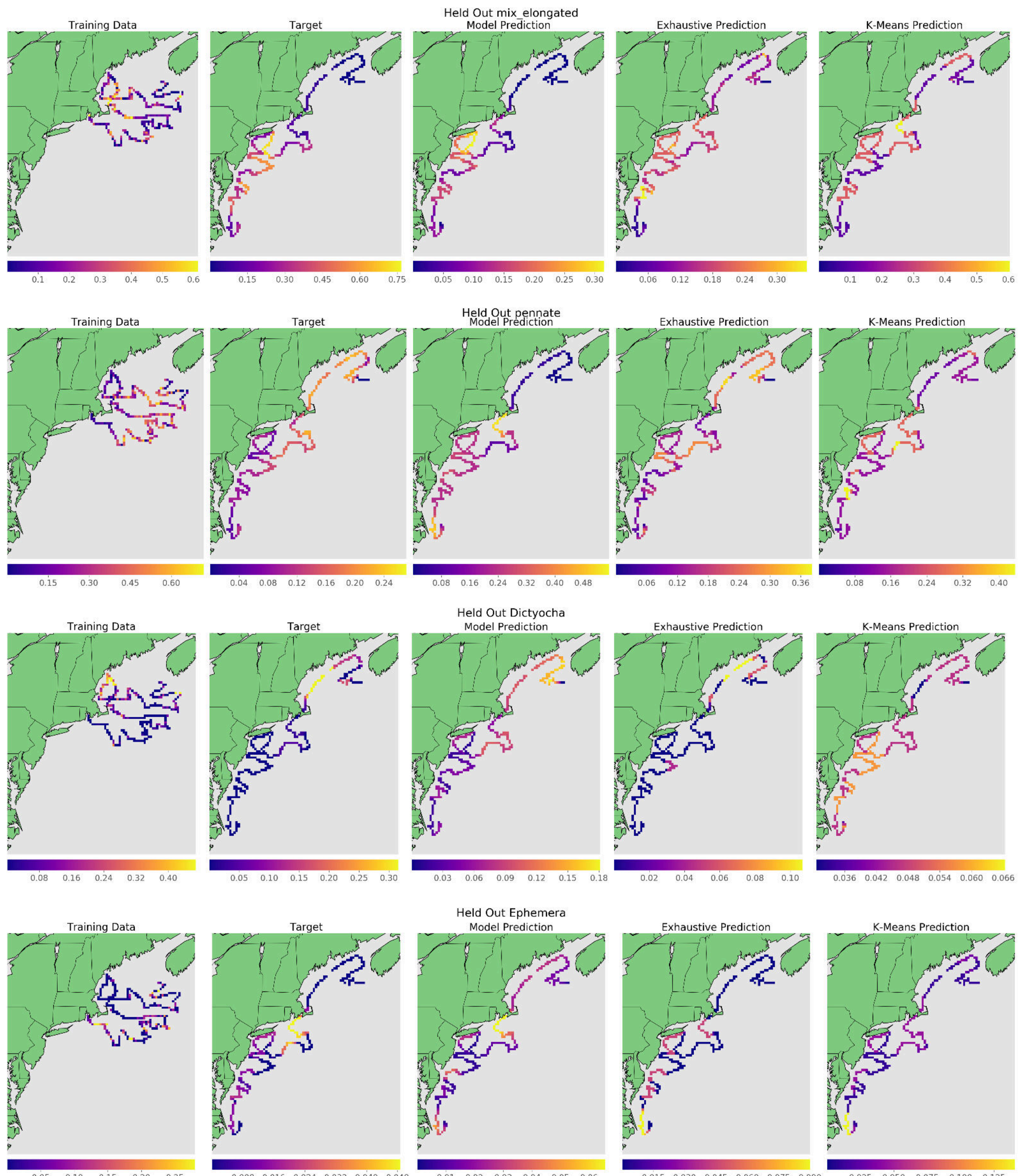


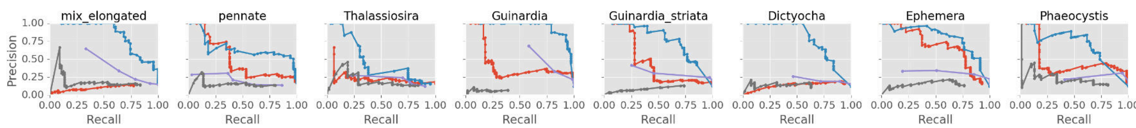
Figure 4–11: Spatial distribution for four target classes (rows) in split training/testing samples. The columns correspond to training data (col. 1), held-out target locations (col. 2) , and the three models under evaluation (col. 3-5). The proposed plankton topic model provides predictions that agree better with the held-out observations than do the simpler k-means based plankton community model or the exhaustive nearest neighbor search.

Our findings show that the prediction problem is relatively straightforward for the interleaved experiment (Fig. 4–10). In contrast, the problem is much more difficult when training and testing locations are in different parts of the world. (Fig. 4–11). Despite this, for three (Fig. 4–10 and 4–11, rows 1, 2, 4) of the four target classes shown here, the spatial location of maxima of our model's predictions are consistently near the maxima in the target distributions.

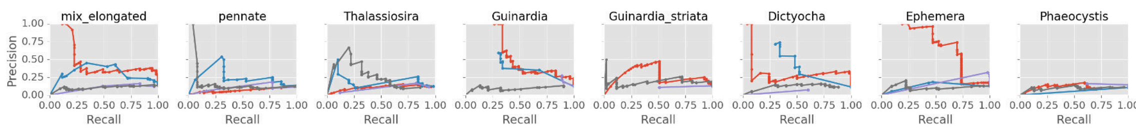
Varying τ for each strategy and parameter setting we can count the true positive, true negative, false positive, and false negative hotspot predictions compared to the top 50 examples in the held-out data. These counts give the precision and recall for each parameter choice, for each $v^{\star}$. We also accumulate these counts across all $v^{\star}$ to compute the overall precision and recall for each choice of parameters. We assign each set of parameters a score given by the area under its aggregated precision-recall curve and select the parameter set with the maximum score for further comparisons. For the interleaved experiment, best performance was achieved with $\alpha = 0.1, \beta = 0.1, \gamma = 10^{-5}, \sigma = 25km$ and for the split experiment, $\alpha = 0.1, \beta = 1.0, \gamma = 10^{-5}, \sigma = 35km$. We chose the number of centroids for the k-means strategy to be the same as the number of topics in the best performing topic model, $K = 9$ for the interleaved experiment, and $K = 6$ for the split experiment.

We compare the aggregated and individual class precision-recall curves for the best parameters for each strategy (Fig. 4–12). From the aggregated precision-recall curves, we find that our model significantly outperforms the exhaustive nearest-neighbor and the k-means strategies on the split-samples regime, especially when the required precision is high. This indicates that the top few predictions of our model were more likely to be true hotspots than those of the other strategies. The exhaustive nearest-neighbor strategy barely performs better than random guessing on the split regime, yet it performs extremely well on the interleaved regime. This result is expected as the exhaustive strategy does not reason at all about the underlying association between plankton types. Instead, it depends on having observed a training point whose distribution is similar to every test point. In contrast, our model performs nearly as well on the split regime as the interleaved regime.

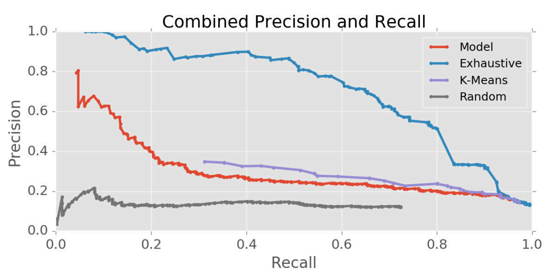
Our model also outperforms the k-means strategy on the split-samples regime. Note that the k-means strategy is exactly equivalent to the exhaustive search strategy in the limit where K is the number of training points. Both these strategies rely on a distance measure over the class distributions. The high dimensionality of the distributions acts to the detriment of the divergence measure. As the dimensionality of a space increases, the discriminating power of distance metrics within that space decreases. The amount of data needed to find meaningful clusters grows exponentially with the number of dimensions, a phenomenon sometimes called *the curse of dimensionality*. As a result, the two search-based strategies perform well when test points are very near to training points in taxon distribution space, but when test points are further away, a distance measure is less informative and performance is negatively impacted.



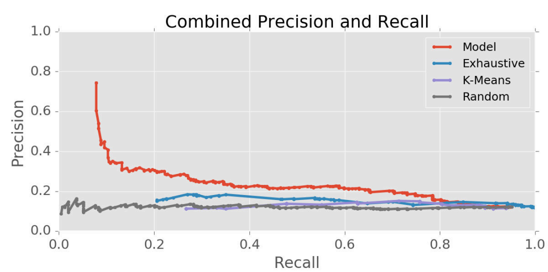
(a)



(b)



(c)



(d)

Figure 4–12: Precision vs recall curves for the community model on each of 8 held-out taxa. (a), (c) Interleaved train and test data. (b), (d) Split train and test data.

CHAPTER 5
Spatial Awareness for Robot Cameras: Predicting images from topic context

In this chapter we consider learning a word distribution function which is suitable for topic modeling as well as a function to map such a word distribution back to the original domain. Such a feature representation combined with a spatio-temporal topic model opens a broad range of applications where we would like to develop a high-level understanding of data in its original domain rather than in a word-like representation. More concretely, in this chapter we consider using this approach to model the views available to a robotic Pan, Tilt, Zoom (PTZ) camera.

Our word distribution function is a novel convolutional autoencoder (CAE), which takes an image as input, produces a distribution of words as its encoding, and then attempts to reconstruct the original image from this encoding. This novel autoencoder, which we call the Categorical Information Boost Autoencoder (CIB-AE), is designed to enhance the differences between the encoding distributions of different images, creating an encoding space well-suited to topic modeling. By fitting a topic model to word distributions that can be decoded as images, we enable a user of the PTZ camera to inspect the topic distributions Φ directly as images. These topic images provide an intuitive decomposition of the high-level parts of the scene as well as a novel interface to explore views by combing them in different proportions.

Further, we employ a similar approach to Ch. 4.2 to recover the word distribution for a neighborhood given its estimated topic prior. Together with the CIB-AE's word-distribution decoding function, this enables us to directly view the data captured by the topic model as images, rather than obscure feature distributions. Finally, we develop an energy-based topic-prior mapping model that exploits the spatial smoothness of the learned topic priors to make predictions for unobserved views. Combining all of these pieces together, we develop a system to predict the images the PTZ camera would encounter if it were to configure itself for a novel view. Fig. 5–1 shows a diagram of the components of this system.

5.1 Learning an invertible word distribution function

5.1.1 Background

Our work on learning an invertible feature function draws from the extensive literature on Deep Convolutional Autoencoders. The simplest version of an autoencoder is a parameterized model, trained to reproduce its inputs with minimal error. Most often, autoencoders adopt a 'bottleneck' intermediate representation with many fewer dimensions than the input space; this helps ensure that the model learns a useful representation of the data rather than a trivial one. The bottleneck layer divides an autoencoder into an encoding function e and a decoding function d, while the activations for a particular input at the bottleneck layer are its encoding. We will write I_x for the image at point x, $h_x = e(I_x)$ for its hidden encoding, and $\hat{I}_x = d(h_x)$ for the 'autoencoded' representation of the image through the network. The standard account of autoencoder training considers the random variables $\mathbf{I}$ and $\hat{\mathbf{I}}$ and seeks

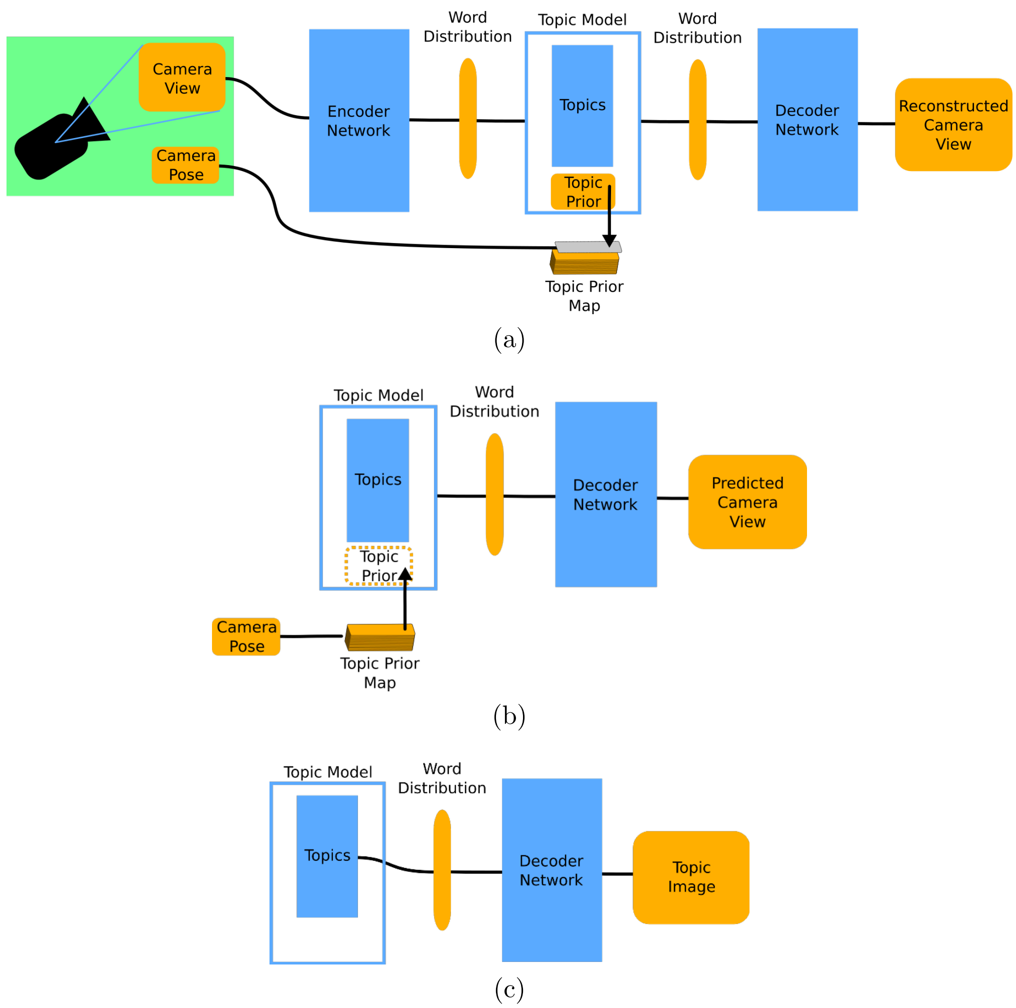


Figure 5–1: Illustration of our proposed spatial image prediction system. (a) We train an autoencoder to produce word distributions as its latent representation, which we model in turn with a spatio-temporal topic model trained for a specific scene. We also train a separate model to predict the topic prior based on the camera pose (its pan, tilt, zoom configuration) and previously observed topic assignments. (b) At test time, we predict a topic prior for a new camera pose. The topic model then allows us to interpret this as a word distribution, which we can decode as an image using our autoencoder. (c) The topics themselves are also word distributions. Viewing these directly by passing them through our decoder allows us to visualize the high-level components of a scene.

to minimize $L_{ae} = -\log P(\mathbf{I}|\hat{\mathbf{I}})$ given a modeling choice for $P(\mathbf{I}|\hat{\mathbf{I}})$. When I is considered an unconstrained real-valued tensor, it is common to assume $\mathbf{I}|\hat{\mathbf{I}} \sim \mathcal{N}(\hat{\mathbf{I}}, \Sigma)$, which leads to the standard MSE loss $\mathcal{L}_{ae} = \sum_x \|I_x - \hat{I}_x\|_2^2$, as an unbiased estimator for L_{ae} (Vincent et al., 2010).

A key strength of autoencoder models is that the only training data they require are the images themselves, and therefore large datasets are relatively easy to produce. Because of this, as larger, more elaborate, supervised CNN architectures have improved and the availability of high-performance GPUs has increased, it has become tempting to use similarly large networks to get more exact autoencoders without increasing bottleneck sizes. It is often the case, however, that a very powerful decoder will learn to largely ignore the inputs and their encodings and instead memorize a few indistinct outputs to use for every input, eventually learning identical encodings for all inputs (Vincent et al., 2010; Chen et al., 2016)[1] . In the applications discussed in this chapter, avoiding such collapsed encodings is crucial; for the topic model to learn the spatial arrangement of the scene, the learned feature representation should encode as much about the differences between images as possible rather than deferring to the decoder network. Many recent models partially solve this problem by using recurrent networks rather than single-shot encodings to progressively encode and refine their representations, for example (van den Oord et al., 2016; Chen et al.,

[1] This problem has been discussed most prominently under the name 'posterior collapse' within the Variational Autoencoder literature, referring to models that learn e as an approximate posterior $q(h|I) \approx p(h)$.

2016; Gregor et al., 2016). These approaches have achieved some of the most visually convincing results to date, especially in the domain of high-resolution image encodings. These methods are complimentary to ours, however we seek to demonstrate our approach with a more traditional convolutional autoencoder architecture to simplify the training process.

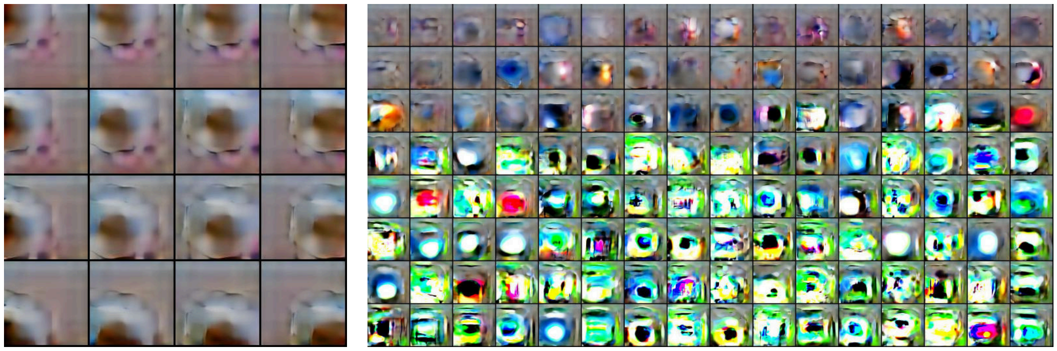
Beyond the common collapsed encoder issue, the most notable novel requirement for our feature encoding is that features must be discrete-valued, whereas most autoencoders produce real-vector valued encodings. Discrete-valued autoencoder representations have been previously considered in the context of image compression, however these approaches have relied on vector-quantization based on a fixed codebook to discretize the encoding space (Agustsson et al., 2017). The Vector Quantized Variational Autoencoder (VQ-VAE), is a notable recent improvement on this approach that simplifies implementing this style of compression by learning the codebook at the same time as fitting the parameters of the feature network (van den Oord et al., 2017). Its quantization still relies on the Euclidean distances between feature vectors however, despite the Euclidean distance being only marginally meaningful in very high-dimensional spaces (Aggarwal et al., 2001).

In contrast, we design an autoencoder model where the latent space is taken to be the parameter of a categorical or multinomial random variable. In other words, we use the dimensions of the encoding as discrete feature values directly, rather than trying to discretize a high-dimensional space with a codebook. This has the advantage of fitting directly with the histogram of words document model assumed by LDA, while simultaneously considerably simplifying the required modifications to standard autoencoder architectures to achieve discretization, as well as sidestepping the issue of choosing a distance function for quantization.

5.1.2 Categorical Information Boost Autoencoder – CIB-AE

In order to learn a representation compatible with topic modeling, we develop a modified CAE, the Categorical Information Boost Autoencoder (CIB-AE), where the encoding can be seen as the parameter of a V-dimensional multinomial distribution, providing a natural link to the topic modeling methodology we have discussed throughout this thesis. We consider encoding an image by the normalized histogram of N 'words' – the learned visual primitives which we can use to reconstruct an image. Similar to DNNs used for supervised classification tasks, we simply use a softmax layer to convert activations to a PMF parameterizing a multinomial distribution. In doing so, we discretize the feature space by considering each dimension of the representation as a discrete 'word', and preserve detail in our model by modeling the full word distribution for an image. Fig. 5–2 presents decodings of single words from our trained model.

Through its Dirichlet priors, our topic modeling approach makes the assumption that each topic-word distribution and neighborhood-topic distribution is sparse (recall Fig. 2–1). Consequently, the topic model's implicit assumption is that neighborhood-word distributions are somewhat sparse as well. We have found in practice that simply re-normalizing the features of a conventional autoencoder produces word-distributions that are too uniform to be modeled productively by LDA. CIB-AE employs an additional regularization term in its loss to ensure images are encoded



(a) (b)

Figure 5–2: CIB-AE decodings of single 'words', i.e. one-hot encoding vectors. An image is reconstructed as a normalized word-histogram. (a) Each word is duplicated at 4×4 spatial locations within a 128×128 input image. (b) The top 128 most common words from the training set, at their central position.

using a variety of sparse word distributions:

$$L_{CIB} = -\log P(\mathbf{I}|\hat{\mathbf{I}}) - \lambda MI\,(\mathbf{I}, W) \tag{5.1}$$

where $W \sim e(I)$ and MI stands for the mutual information, a measure of the degree of dependence between two random variables. Choosing a large value for λ will result in representations that favor high mutual information between images and encoding distributions, i.e. those where training images are as distinguishable as possible in encoding space. In other words this regularization term ensures that encoder collapse does not occur as it does with many other autoencoders [2] [3] .

To illustrate how our loss works, consider another way to rewrite the mutual information:

$$MI\,(\mathbf{I}, W) = H[W] - H[W|\mathbf{I}] \tag{5.2}$$

where $H[W|\mathbf{I}]$ signifies the conditional entropy. If we subtract the mutual information from the loss, the $H[W|\mathbf{I}]$ term ensures that individual encodings must have low entropy, and are therefore sparse, while the $-H[W]$ term ensures that the overall distribution of encodings is high entropy, and therefore as uniform as possible. The combination of these two terms means that the encodings must use different dimensions of the feature space from one another.

Reorganizing the mutual information in this also way reveals how to estimate it while performing stochastic gradient descent (SGD). In SGD, the gradient of the loss function over the whole dataset is estimated by taking the gradient of the loss over a portion of the dataset, called a minibatch. Given a minibatch of M training images chosen at random from the dataset, we use a simple Monte-Carlo estimate

[2] We note an interesting connection to the literature on VAEs, which use an approximate prior $q(W|\mathbf{I}) \approx p(W)$. Rewriting $MI\,(\mathbf{I}, W)$ as $D_{KL}[(I, W)\|IW]$, our objective and the VAE objective take extremely similar forms: In particular, $L_{CIB} = -\log p(\mathbf{I}|\hat{\mathbf{I}}) - \lambda D_{KL}[(\mathbf{I}, W)\|IW]$ while $L_{VAE} = -\log p(\mathbf{I}|\hat{\mathbf{I}}) + \lambda D_{KL}[q(W|\mathbf{I})\|p(W)]$.

[3] Note that in the context of a Gaussian or Bernoulli model for $P(I|\hat{I})$, denoising autoencoders (Vincent et al., 2010), i.e. autoencoders with dropout, can also be considered to enhance the MI between images and encoding.

to approximate the mutual information by rewriting it in terms of expectations:

$$\begin{aligned} -MI\left(\mathbf{I}, W\right) = H[W|\mathbf{I}] - H[W] = \\ \mathbb{E}_{w\sim e(I)}\left[-\log P(w)\right] - \mathbb{E}_{\mathbf{I}}\left[\mathbb{E}_{w\sim e(I)}\left[-\log P(w)\right]\right] \approx \\ -\frac{1}{M}\sum_{m=0}^{M-1}\sum_{d=0}^{V} h_{m,d}(\log(h_{m,d})) + \sum_{d=0}^{V}\overline{h}_d\log(\overline{h}_d) \end{aligned} \tag{5.3}$$

Recall that we write $h_m = e(I_m)$, and here we use $\overline{h}$ to signify the average encoding for the M images in the minibatch. Clearly, estimating an expectation over the space of all possible training images with a minibatch of few dozen or at most a few hundred examples is crude. More precisely, this estimator is likely to have extremely high variance, and thus lead to slow convergence while optimizing the loss function. Nevertheless, the terms of this approximation naturally fit our goal of producing sparse, distinct encodings for as much of the training data as possible.

To summarize our approximation of the CIB loss

$$\begin{aligned} &L_{CIB} \approx \mathcal{L}_{CIB} = \\ &\frac{1}{M|I|}\left(\sum_{m=0}^{M-1} I_m - \hat{I}_m\right)^2 - \frac{\lambda}{M}\left(\sum_{m=0}^{M-1}\sum_{d=0}^{V-1} h_{m,d}\log(h_{m,d})\right) + \lambda\sum_{d=0}^{V-1}\bar{h}_d\log(\bar{h}_d) \end{aligned} \tag{5.4}$$

where $|I|$ denotes the (constant) number of pixels in an input image. In other words our loss is the MSE plus the average entropy of all encodings, minus the entropy of the average encoding. We find in practice that when the image space is normalized to the range $[-a, a]$, $\lambda = \frac{(2a)^2}{\log V}$ is a good choice, where V is the dimension of the encodings [4] .

Our specific choice of architecture is inspired by the feature layers of VGG (Simonyan and Zisserman, 2014). In particular, following the approach of Segnet (Badrinarayanan et al., 2015) we choose a VGG-like encoder, with a symmetric decoder of transposed convolutional layers. This style of encoder uses a sequence of convolution 'blocks', which each ultimately either downsample (Max Pool) or upsample (Nearest Neighbor) the activations by a factor of two (See Fig. A–1). We add an additional dropout and an additional softmax layer to the encodings to constrain the real-valued activations of a Segnet style encoding to a simplex. Our feature representation downsamples the image spatially by a factor of 32, while producing a 512 channel feature activation for each spatial region. All experiments are conducted with $128 \times 128 \times 3$ input images, resulting in $4 \times 4 \times 512 = 8192$ dimensional encodings. The details of our architecture are found in App. A, however we emphasize that by adding these two extra layers and the additional loss term, our approach can potentially be transferred to many existing autoencoder architectures and other domains.

---

[4] Note $\log V$ is the maximum possible entropy for a categorical distribution of dimension V, so this choice represents letting the MSE and MI terms have 'equal' weight when both are in their worst cases

5.2 Predicting topic priors from camera poses

In this work, our goal is not only to encode and decode observed images, but also to use the encodings of observed images to predict the encodings of images at new locations, and decode these to produce a prediction of what the camera might see if it went to that location. We have seen in Chs. 3 and 4 how a spatio-temporal topic model may be used to learn spatially coherent topic-prior maps which encode the high-level categories of a dataset, as well as how to interpret the maximum likelihood word distribution associated with a particular topic prior. In this section, we focus on how to predict topic priors for a new location given those produced by our topic model.

In this work we consider the location associated with an observation to be the Pan, Tilt, and Zoom coordinates of a robotic camera (i.e. a PTZ camera). Rather than fully exploiting the overlap between views, we seek a more general method to predict topic priors, and therefore only make limited assumptions beyond those that the spatio-temporal topic model makes itself. Recall that our topic model learns topic priors that are spatially smooth. This means that without observing the image associated with a location, a reasonable guess for its topic prior is that it will be similar to the topic priors for nearby locations.

The simplest reasonable prediction for a new location would be to predict the topic prior for the nearest point which has been observed by the topic model. If our system were allowed to observe images for many locations, this strategy could be very accurate, however reconfiguring the camera and taking photos many times is time consuming, and we envision training our model with only a few dozen locations.

To better accommodate test locations that may be far from training locations, we adopt an energy-based approach to topic prior prediction. We predict the topic prior for a new location using a weighted average over the topic priors we have observed so far, where far observations have low weight and near observations have high weight. Rather than performing such a prediction with a direct mixture of observed topic-priors and re-normalizing, we choose to implement the weighted sum in log-probability space, and then return the result to the natural probability space with a softmax. We find that in practice this approach helps to maintain sparsity of predicted topic priors, a property assumed by the topic model. Note that if the mixture of log-probabilities belongs to a valid probability distribution (*eg.* all the weight is placed on a single training location), the softmax function returns exactly that distribution.

We define the weights in terms of a Boltzmann distribution (i.e. the softmax of negative energy), using the distance to each observed topic prior as the energy. Let $D_X(x)$ denote the distance vector from point x to each point in the training set $x_i \in X$. This choice of weight function has the appealing property that if any entry of $D_X(x)$ is much smaller than the rest, all of the weight will be put on that training point. Further, with this weight function, by controlling the scale of the distances we can control the sparsity of the weight vector; when the average distance is low the weights will be fairly uniform, while when the average distance is high, the weights will be sparser. We choose to use the L_2 distance between pan-tilt-zoom points as our distance function. In order to appropriately scale the distances and balance dimensions which may not truly incur equal loss of relevance for higher distances, we first scale the input points by the element-wise product with the vector w [5]. To summarize our prediction model, given N training points X each paired with an observed topic prior in Θ_X:

$$
\begin{aligned}
D_X(x) &= \Big[\|w \odot (x - x_1)\|_2, \quad \|w \odot (x - x_2)\|_2, \ldots, \|w \odot (x - x_N)\|_2\Big]^T \\
W(x) &= \sigma\left(-D_X(x)\right) \\
\theta(x) &= \sigma\left(\log\left(\Theta_X\right) W(x)\right)
\end{aligned}
\tag{5.5}
$$

Finally, we note that setting w by hand is difficult as the best choice may vary from environment to environment, and is unnecessary, as both fitting and prediction are extremely efficient with this model. Instead, given a validation set $X^\star, \Theta^\star$ held out of training, we consider the information lost by using our model, measured by the mean Jensen-Shannon Divergence on the validation set (Lin, 1991):

$$
L_w = \frac{1}{|X^\star|} \sum_{x \in X} D_{JS}\left[\theta(x) \| \Theta^\star_x\right] = \frac{1}{|X^\star|} \sum_{x \in X} \frac{1}{2}\left(D_{KL}\left[\theta(x) \| m(x)\right] + D_{KL}\left[\Theta^\star_x \| m(x)\right]\right)
\tag{5.6}
$$

where $m(x) = \frac{\theta(x) + \Theta^\star_x}{2}$. While $D_{KL}\left[\theta(x) \| \Theta^\star_x\right]$ would most directly measure the information lost by approximating the true topic distribution with the prediction, D_{JS} is additionally symmetric, positive, and bounded above, making it a more convenient loss function. This loss and our model are differentiable, so we are able to fit w using gradient descent and a 2-fold cross-validation procedure. As a final note, although a

---

[5] Our distance could also be called a Mahalanobis distance (normally formulated in terms of the covariance matrix $\Sigma = ww^T$)

more complex distance model may better match how topic-priors are spatially distributed, because we aim to train this model with only a few dozen examples we prefer a simpler model that is less prone to overfitting.

5.3 Experiments

Training our view prediction system requires several steps. First, we train the CIB-AE on a large dataset of views extracted from panorama images. Then, we train our topic model and spatial topic prior prediction model for a specific scene. This process involves alternating between observing new images and refining the topic model for a fixed period of one second. At the end of each refinement phase the topic prior map is refit based on the counts of topic assignments for each location that has been observed so far. Then, the locations for the next training images are chosen uniformly at random (we discuss our efforts to use the partially trained model to select subsequent training points in Sec 5.4.1). Finally, after a fixed number of iterations the models are frozen, and query locations are passed through the topic prior map and the topic model to produce word distributions, which are then decoded as images by the CIB-AE.

5.3.1 Simulated Pan-Tilt-Zoom Camera Dataset

We demonstrate our approach with a simulated outdoor pan-tilt-zoom camera, using data from the 'street' category of the SUN360 dataset (Xiao et al., 2012). Specifically SUN360/street comprises a dataset of 161 360 degree panoramas projected into rectangular 9104×4552 pixel images (Fig. 5–3). These panoramas are taken in various cities and towns across the world, from street corners and from car-mounted cameras. The data are preprocessed so that the horizon is horizontal and

Figure 5–3: Example 360 degree panoramas from the SUN360/street dataset.

vertically centered in all panoramas. We parameterize views in terms of tilt angle T, pan angle P, zoom factor Z. Given a base field-of-view (fov) A_0 and a target camera resolution, we compute the projection to the desired perspective $R = R_T R_P$ where

$$R_T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(T) & -\sin(T) \\ 0 & \sin(T) & \cos(T) \end{bmatrix}$$
$$R_P = \begin{bmatrix} \cos(P) & \sin(P) - \sin(T) & \sin(P)\cos(T) \\ \sin(P)\sin(T) & \cos(P)\cos(T)^2(1-\cos(P)) & \cos(T)\sin(T)(1-\cos(P)) \\ \sin(P)\cos(T) & \sin(T)\cos(T)(1-\cos(P)) & \cos(P)+\sin(T)^2(1-\cos(P)) \end{bmatrix} \tag{5.7}$$

Using the fov $A = A_0/Z$ and fixed output image resolution, we represent each output pixel location as a pan-tilt rotation from the center P, T. Then, projection proceeds as rotation by R and a lookup in the full panorama image for each pixel in the

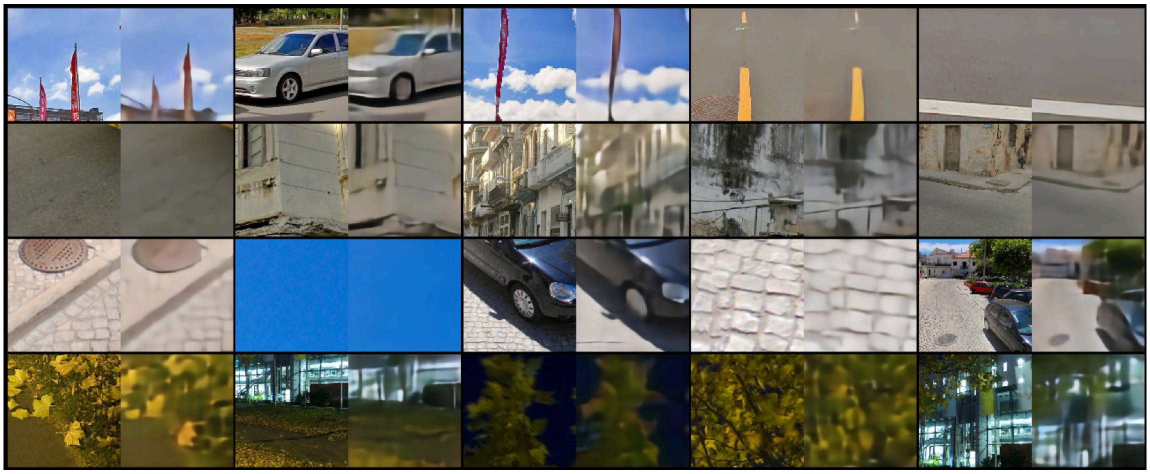


Figure 5–4: Random example input images (left) and CIB-AE reconstructions (right) for 4 worlds (rows) from the SUN360/street validation set.

output image. In all our experiments we consider a base fov of 76 deg and zoom factors $Z \in [1, 12]$, similar to a standard commodity camera. Further, to avoid wrapping angles, we consider only $T \in [-60, 60]$ and $P \in [-120, 120]$ degrees.

5.3.2 CIB-AE Training

We trained our feature model using epochs of 5000 random views from amongst 80 SUN360/street panoramas. We found that choosing a new set of random training views for each epoch was crucial, however we kept the validation set fixed throughout training. The projection described in the previous section is time consuming for high-resolution images, especially for those that do not fit in GPU memory. Therefore we trained for 175 epochs on random square crops from the warped panorama images, and then 25 more epochs on projections to fine-tune the results. We achieved our results using the adaptive learning rate optimizer Adam (Kingma and Ba, 2014), however found that a learning rate decay by a factor of 0.9 every 10 epochs was

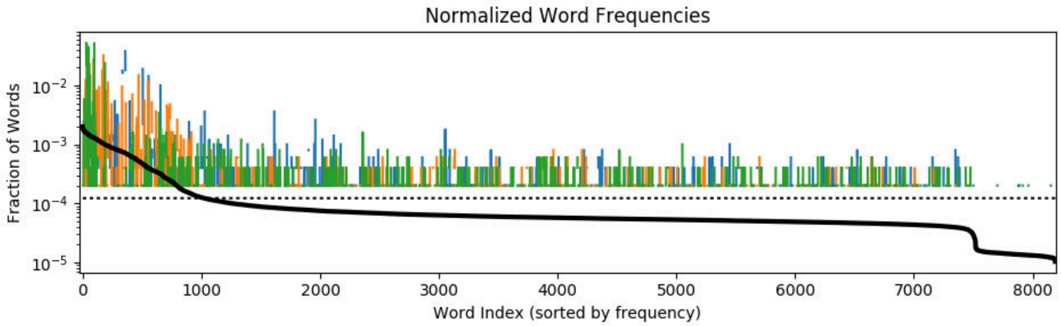


Figure 5–5: Word-distribution, sorted by frequency, of the average encoding over 1000 test images (heavy black line) and normalized word-histograms for 3 random images (colored lines). A uniform histogram over the 8192 dimensions (dashed black line) represents the ideal average encoding and the worst-case individual encoding.

also necessary. Example input images and their autoencoded counterparts can be seen in Fig. 5–4. We found the mean entropy of image encodings from 1000 random views from the test set to be 6.170 nats [6] , or approximately equivalent to a uniform PMF over just 470 dimensions. On the other hand, the entropy of the mean encoding distribution from these 1000 examples was 8.192 nats, whereas the maximum possible entropy where all feature dimensions were used uniformly would have had an entropy of 9.011 nats. Typical word distributions and the average word distribution are illustrated in Fig. 5–5.

5.3.3 View Prediction Training

After training the word distribution function, we next train a spatio-temporal topic model and topic prior map specific to a single scene. A key advantage of our

[6] While bits is the unit for information measured by entropy using log_2, nats is the unit when the entropy is calculated using ln

approach is that it can be trained online, whereas this is not always practical for large deep-networks. This means that while the CIB-AE must be trained to encode images from the entire dataset, our most compressed representation of the scene, the topic priors, only has to capture images from a single scene, a considerable simplification.

When the image corresponding to a new view is extracted, its word distribution is computed based on the trained CIB-AE. Because our Gibbs sampling inference method samples from the posterior for the topic assignment of a single word given all the others, it is necessary to discretize the word distribution by sampling from it rather than work with it directly. Note that despite an encoding dimension of 8192, it is possible to use a relatively small number of discrete observations without losing detail because individual images have low entropy encodings. In our experiments we define a constant encoding resolution of 3000 words per image; initial experimentation showed some degradation of performance below this level and little improvement above, however clearly this 'document size' parameter defines a trade-off between the required number of sampling iterations and the amount of detail modeled.

We choose a spatio-temporal topic model with $K = 25$ topics and neighborhoods of size 10 degrees $\times$ 20 degrees $\times$ 6 (tilt $\times$ pan $\times$ zoom). This neighborhood size choice corresponds to the intuition that environments are semantically 'smoother' in pan than tilt, and much smoother while zooming than moving the camera. We chose concentration parameter values $\alpha = 0.2$, $\beta = 0.2$. We performed extensive cross validation to select these hyperparameter values and certain other choices were found to perform much better for particular environments, however we did not find a combination of hyperparameter values that was definitively much more performant than the others on average.

5.3.4 Encoding and Prediction Results

Since we aim to demonstrate a real-time system that can learn about its surroundings relatively quickly, we stop learning after 50 observations. As a first inquiry into what the model has learned, we directly decode the point estimates of the topic distributions $\hat{\Phi}$ by passing them through the CIB-AE decoder (See Figs. 5–7a, 5–8a, 5–9a). Most other applications of visual topic models employ non-invertible feature functions and choose to inspect topics by looking at the most representative training images. To our knowledge, this is the first application of a visual topic model where the topics can be fully decoded and inspected directly as images. Interestingly, we find that for our autoencoder network, sparser encodings produce clearer images, while more uniform encodings produce muddled, mostly gray, blurry images (See Fig. 5–6). This motivates us to use a *maximum a posteriori* (MAP) point-estimate for $\hat{\Phi}$ (as well as for $\hat{\Theta}$) as opposed to the MLE discussed in Eqn. 4.1

$$\begin{aligned}
\hat{\Theta}_{g(x),k} &= \begin{cases} \frac{N^k_{g(x)}+\alpha-1}{\sum_{j=1}^{K} N^j_{g(x)}+\alpha-1} & if\ N^k_{g(x)} > 0 \\ 0 & otherwise \end{cases} \\
\hat{\Phi}_{k,w_i} &= \begin{cases} \frac{N^{w_i}_k+\beta-1}{\sum_{v=1}^{V} N^v_k+\beta-1} & if\ N^k_{g(x)} > 0 \\ 0 & otherwise \end{cases}
\end{aligned} \tag{5.8}$$

Prediction with our spatial topic prior model is extremely efficient, so we predict the topic priors for 10,000 random P, T, Z views. In Figs. 5–7b, 5–8b, and 5–9b we present the resulting topic prior maps (one scatter plot per topic, in the

(a)

(b)

Figure 5–6: MAP (a) vs MLE (b) point estimates for Φ trained with identical datasets, directly decoded as images. Note that the MAP estimate features more distinctive topic images.

same arrangement as the topic images), with the point x_i omitted for topic k if $\theta(x_i)_k < 0.01$, blue indicating $\theta(x)_k = 0.01$ and red $\theta(x)_k = 1$. Together, the decoded topics and view-topic maps provide an evocative graphical representation of how low-level textures combine into a higher level scene structure. After training, we collected 50 more views along with their inferred MAP topic priors (without performing any more refinement of Φ). Among these observed views for each of the 35 test panoramas, we found $D_{JS}[\theta(x), \hat{\theta}]$ to be on average 0.1953 nats, while $D_{JS}[\text{Unif}(K), \hat{\theta}]$ was on average 0.5122 nats[7] . This indicates that the topic mapping procedure is imperfect as is expected because of its weak geometric assumptions,

[7] Recall $0 \leq D_{JS} \leq \ln(2) \approx 0.6931$

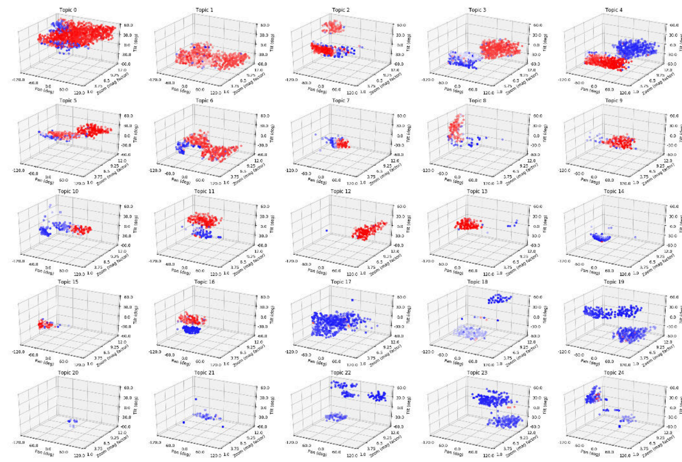
however it does capture a significant portion of the information about the spatial distribution of topics.

In addition to decoding the topic distributions themselves, we can predict the topic prior for a view and the corresponding encoding distribution given the topic distributions, and decode these predicted encodings. We evaluate our system based on the root mean squared error (RMSE) pixel value between the predicted images and the true image at that view. We call views where we have run topic inference and topic map training but not topic refinement 'observed'. At observed views our model predicts exactly the topic prior observed in that location, and so is able to represent the information encoded in the topic model without external loss.

As performance baselines, we trained two additional autoencoders with similarly constrained latent spaces on the full SUN360/street training set used to train the CIB-AE. First, we trained a plain CAE, without a categorical distribution latent space or additional loss terms beyond the MSE. We copy the architecture of the CIB-AE up to the softmax layer, and then add 3 additional fully connected layers of sizes 4096, 2048, and 1024, with ReLU and BatchNorm nonlinearities and with a final latent representation of 25 dimensions to complete the encoder. As with CIB-AE, the decoder is symmetric with the encoder. We initialized the weights of the encoder by copying them from the pretrained CIB-AE where possible, and trained the CAE using the same procedure as the CIB-AE.

In addition, we trained the previously mentioned vector-quantized variational autoencoder (VQ-VAE) of (van den Oord et al., 2017) with a codebook of size 25×8192, identical to the size of the topic model's Φ parameter. The VQ-VAE is a state of

(a)

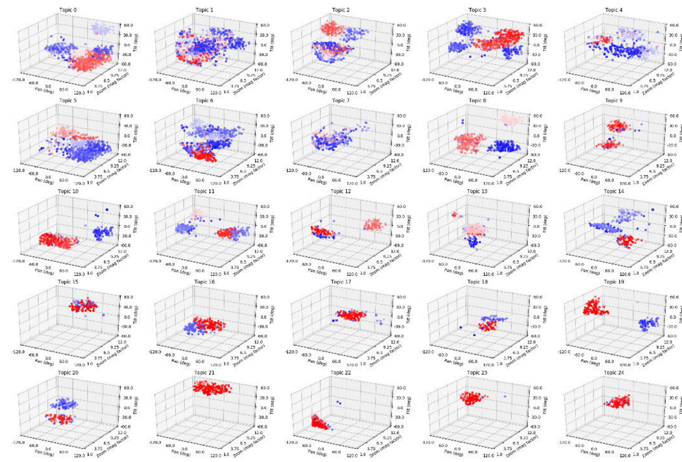


(b)

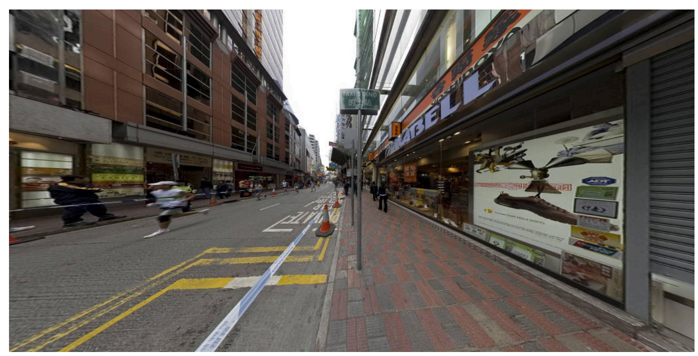
(c)

Figure 5–7: Example spatial prediction model after 50 training observations. (c) Fish-eye view of the entire panorama (a), Direct decoding of each topic-distribution (Φ_k), And (b) predicted mixture of each topic.

(a)

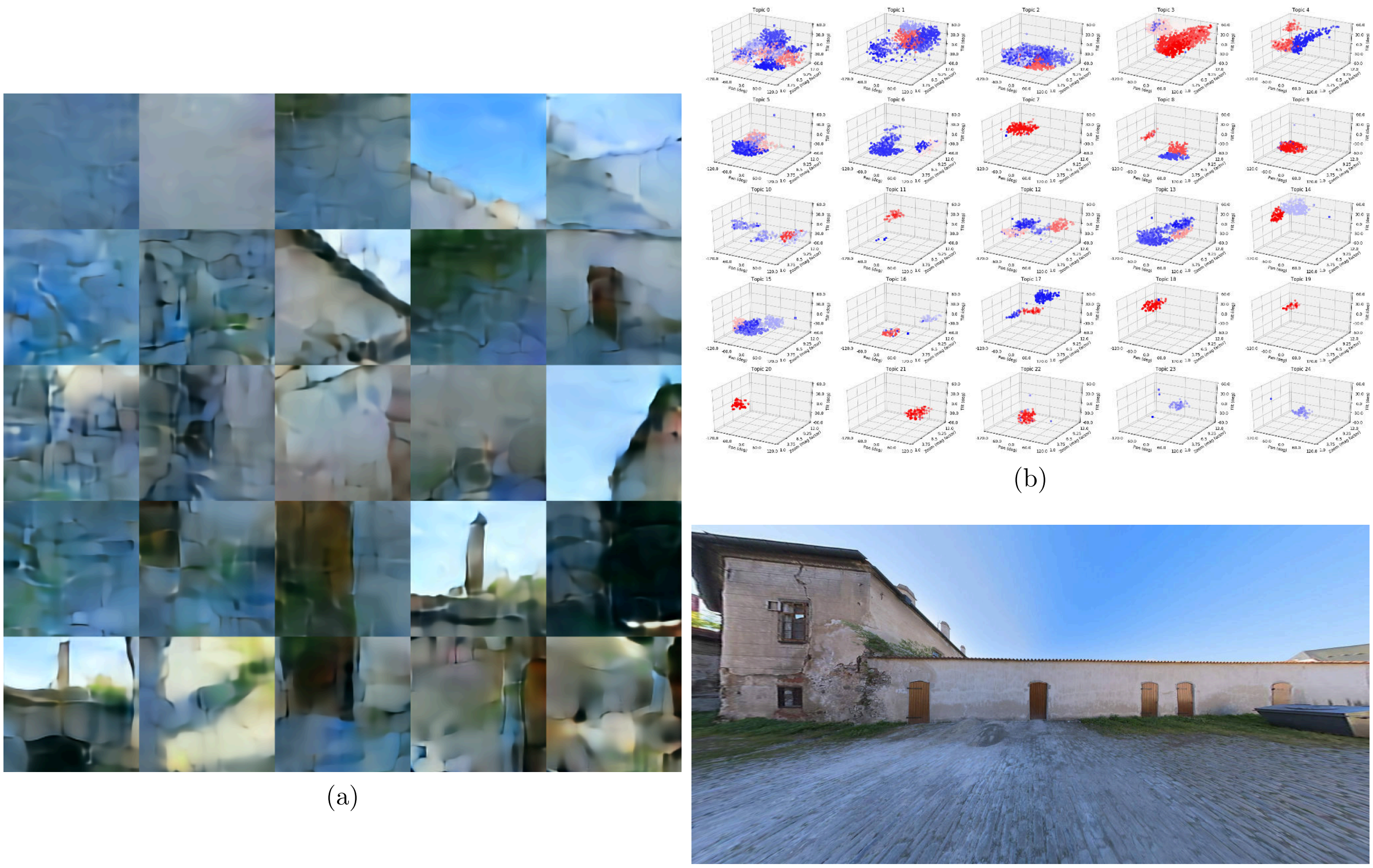


(b)



(c)

Figure 5–8: Example spatial prediction model after 50 training observations. (c) Fish-eye view of the entire panorama (a), Direct decoding of each topic-distribution (Φ_k), And (b) predicted mixture of each topic.



(a) (b) (c)

Figure 5–9: Example spatial prediction model after 50 training observations. (c) Fish-eye view of the entire panorama (a), Direct decoding of each topic-distribution (Φ_k), And (b) predicted mixture of each topic.

the art discrete-latent autoencoder model designed to address many of the same issues as the topic model, however it cannot be trained online like the topic model, and so it must fit a globally salient representation into the fully compressed 25-dimensional encoding space. We found that VQ-VAEs are very sensitive to architecture choices. For a fair comparison, it would be preferable to use a VQ-VAE with spatial resolution 4×4 and feature size 512 for each spatial block, however we could not escape encoder collapse using this representation. In fact, the only spatial resolution we found that attained reasonable results was the one the original authors demonstrated, therefore our comparisons are against a VQ-VAE with spatial resolution 32×32 and a feature size of just 8 for each spatial block.

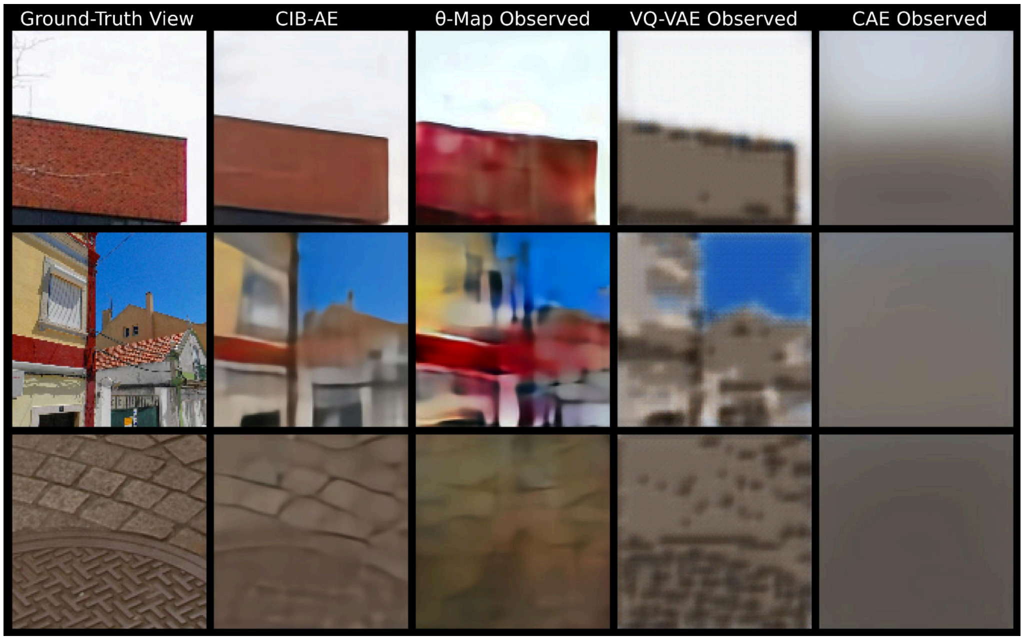
For the baseline approaches, we apply a similar energy-based prediction model as for the topic-priors (Eqn. 5.5), however as the encodings are not probability distributions, we perform the weighted sum of encodings directly rather than in log-probability space. In addition, the cross-validation metric for w is the MSE of the predicted encodings rather than D_{JS}. Note that the baselines are not trained with spatial information, so we would not expect this method to achieve the same performance as with the spatially smooth topic model.

Fig. 5–10a shows examples of the autoencoding performance of the various systems through their 25 dimensional representations. The leftmost column is the observed image and the second column is the CIB-AE autoencoding of that image. The subsequent columns are the decoded neighborhood-word distribution, the VQ-VAE autoencoding, and the CAE autoencoding. The CAE is not able to capture any detail in this regime, the gray images seen are characteristic of the encoder collapse problem mentioned previously. The VQ-VAE, unsurprisingly because of its high-spatial resolution and low feature dimensionality, captures the spatial structure of the image relatively well, but loses much of the color and distorts the texture. Finally, our approach is competitive with the VQ-VAE, featuring images that are more spatially distorted, but interestingly, that enhance the color contrast of its input images. Fig. 5–10b is similar, however it shows unobserved views, predicted through the topic prediction model. For many more examples, see App. B.

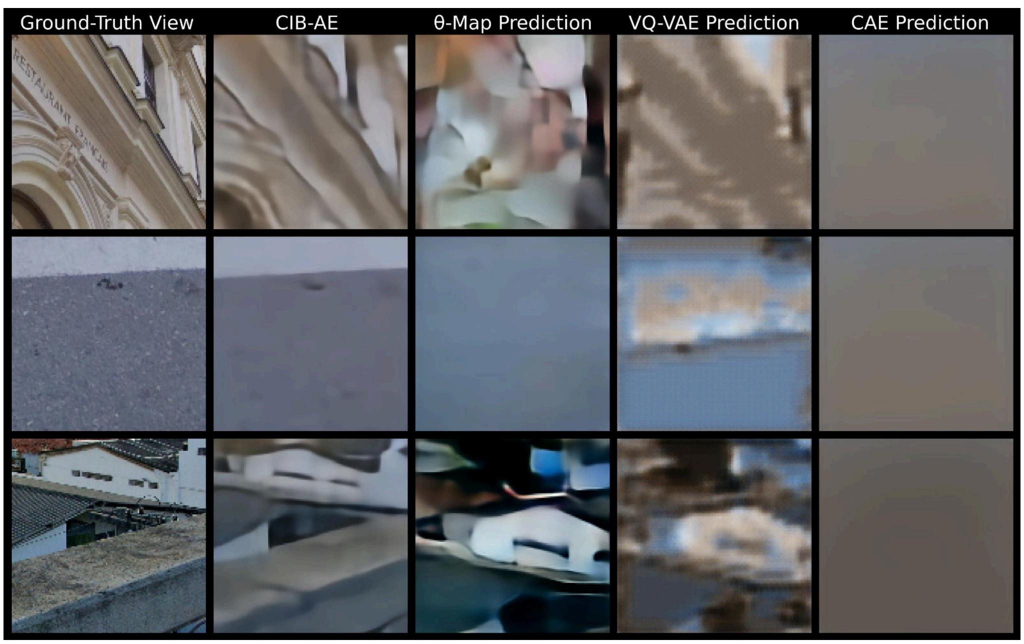
Fig. 5–11 shows the RMSE for observed and predicted topic priors decoded as images compared to the real images at the corresponding views. The low extent of the CAE whisker reveals that the conservative collapsed strategy of consistently predicting uniform grayish images is not necessarily bad in terms of pure pixel error, drawing out an important flaw of the pixel error as a performance measure at this level of compression. Nevertheless, to the extent that the RMSE distributions measure performance in our view prediction task, the data confirm our qualitative assessment that our method performs approximately as well as the VQ-VAE on average, and the extent of its typical errors is lower. Further, the more colorful, higher contrast images produced by our method are reflected by its higher variance than the more conservative VQ-VAE approach which tends to encode roughly the right shape, but obscure much of the color and texture.

5.3.5 Topic Search Interface

A second outcome of our work is a tool for exploring views in an environment through mixtures of topics. By directly decoding the topics, we are able to visually present a summary of the scene. Then, by providing an interface to specify a mixture



(a)



(b)

Figure 5–10: Observed (a) and predicted encodings (b) for random views from SUN360/street. From left to right the columns represent true image, CIB-AE autoencoding, topic prediction autoencoding, VQ-VAE latent prediction autoencoding, and plain CAE autoencoding. Many more examples are found in App. B).

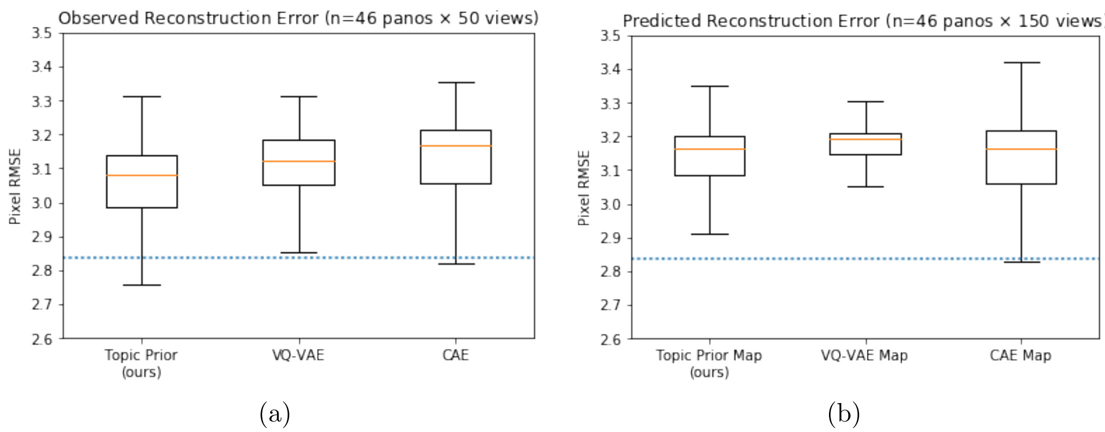


Figure 5–11: Performance comparison of our topic prior map with a VQ-VAE and a standard CAE. (a) Reconstruction performance for observed images (held out of training), (b) View prediction performance after training. The dashed line shows the median reconstruction error of the CIB-AE on this dataset. The box in this plot represents the interquartile range, while the whiskers represent the range of typical data (1.5 times IQR). RMSEs are computed for each image with respect to pixel values in the range [0,255].

Figure 5–12: Screenshot of visual topic-mixture search interaction: Counter-clockwise from Top-left: Direct decodings of 25 topics; numbers in each decoded topic indicate the proportion of the target mixture given to that topic. Image of the target-mixture. Sample from the view with the closest topic distribution given by our spatial prediction model: topic div gives the divergence between the predicted encoding and the target encoding. Spatial topic map. Full extent of the possible views. True image at the closest topic distribution. Demo video at `http://cim.mcgill.ca/~akalmbach/thesis/demos.html`

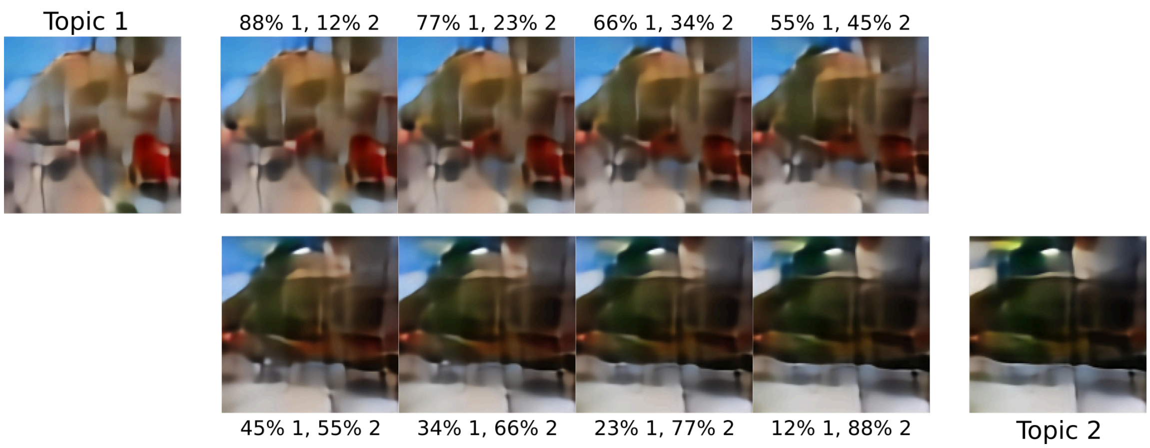
of these topics, and performing a search within the topic-prior map, we enable a unique interaction that enables a user to visually explore an environment, as well as the information encoded by the topic model and the topic-prior map (See Fig. 5–12).

This interface leads us to a few interesting qualitative assessments of the topic model. For instance, in Fig. 5–13a we visualize interpolating topic distributions connecting two topics. This experiment confirms that the powerful decoder network is doing more than linearly mixing the topic images as we transition between two topics. Instead, it reveals colored segments of the images gradually changing in shape and intensity. Secondly, in Fig. 5–13b, we visualize a predicted word distribution, as well as samples from the multinomial defined by that distribution. Recall that to train our model we resorted to discretizing the feature distribution in the same way, using a fixed number of samples from the word distribution. These samples are thus more akin to the training data our topic model sees. Further, they illustrate the true diversity of view predictions implied by our model. Each individual sample features slightly different shapes compared to each other and compared to the prediction based on the mean word distribution. Further, the samples often contain more intense colors while being less smooth than the mean.

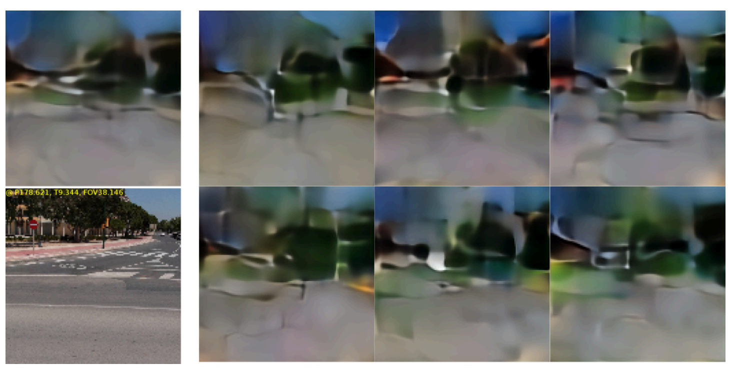
5.4 Discussion

5.4.1 Active learning

As a complimentary task to learning the view prediction model, we also spent significant effort on the problem of *efficiently* learning the view prediction model, that is, trying to accurately predict views in novel locations with as few training observations as possible. Although our specific view prediction model is novel, there



(a)



(b)

Figure 5–13: (a) Setting all other topic probabilities to 0, we can visualize the transition between two topics. (b) Top left: A target image specified as a mixture of topics. Bottom left: The real image with the closest predicted topic distribution to the target. Right: Multinomial ($N = 3000$ words) samples from the predicted word distributions.

is a rich literature on active learning (well summarized by (Settles, 2012)), including many applications of efficient spatial sampling strategies drawing from the Gaussian Process (GP) models of (Seo et al., 2000; Guestrin et al., 2005). Recent work has touched on applications very similar to ours, such as determining where to take images to build a mosaic using a GP-based approach (Manjanna et al., 2016), and which cells in a 'view-grid' to observe to be able to predict the rest using a deep reinforcement learning based approach (Jayaraman and Grauman, 2018).

We experimented with using GP-based upper confidence bounds on both the cross-entropy $H\left[e(I_x)\|\theta_x\Phi\right]$ and encoding divergence $D_{KL}\left[e(I_x)\|\theta_x\Phi\right]$ as well as using the predicted topic-prior entropy $H\left[\theta(x)\right]$ from our model as acquisition functions determining the most problematics regions in the training data, and therefore the best candidate locations for further training. We applied these acquisition functions to both the view-prediction problem and training a semantic segmentation model, however found that neither outperformed a uniform sampling strategy. We did observe that these policies resulted in more observations of complex images as opposed to simple ones (such as all blue sky or gray asphalt), however spending more time learning from complex examples did not improve performance as measured by predicted pixel MSE. It is our hypothesis that this was caused by a combination of biasing the topic model to use all its topics to describe a few small parts of the image (*eg.* the cross-entropy method was very attracted to leaves if they were in the scene), and the limitations of MSE as a performance measurement for images with complex textures. b

5.4.2 Getting more from geometry

Of note is the level of performance our model is able to achieve with a very simplistic geometric model. We use a Euclidean distance to place a prior on the similarity between views in both the topic model and the spatial prediction model. Clearly, however, there is more information to be gained from geometry in this problem. For instance, our feature does contain spatial information, so an edge in one location and the same edge translated or rotated to another location do not necessarily have similar word distributions. This means that the natural categories in a panorama may need to be described by multiple very location specific topics with our model. Further, the spatial prediction model could benefit from more sophisticated geometry. A similar operation to our projection of rectangular panorama images to rectified views could be used to ensure that training images are used to fill in information for all overlapping views. Even the simple addition of upsampling the center of a wide fov training image to get information about a more zoomed in test view would significantly improve our predictions.

5.4.3 End-to-End Variational LDA Autoencoder

While the CIB-AE provides slight modifications that can be made to many standard networks to produce encodings roughly compatible with the Dirichlet prior of LDA, it is natural to ask whether the full model can be learned end-to-end, and fine-tuned for particular environments instead of training a separate model. In fact, (Srivastava and Sutton, 2017) explores implementing the LDA prior for the latent space of a Variational autoencoder, and although the Dirichlet distribution resists exact re-parameterization with a simple distribution as is used in other Variational autoencoders, they use a Laplace approximation to achieve comparable results to the standard mean-field variational approximation with much more flexibility. Although the extra dependencies in the spatio-temporal topic model are not compatible with the re-parameterization used, we envision pre-training an LDA inference network similar to Srivastava and Sutton (2017). Nevertheless, our world-specific, online implementation has the advantage of only needing to be locally salient. Further, it is unclear precisely how loose or tight a variational approximation to the dependencies between topic priors between nearby views would be, while the Gibbs sampling implementation is known to converge within a reasonable number of samples. Therefore we emphasize that the end-to-end approach is best suited to a pre-training phase to determine a word-distribution function and initialize Φ, while our approach is more well-suited to view prediction from a reasonable number of observations.

CHAPTER 6
Conclusion

In this thesis we have demonstrated the versatility of the spatio-temporal topic model as a tool for understanding complex data in terms of coherent spatio-temporal semantic maps. We have shown their utility for many application domains beyond the textual ones normally associated with the topic modeling literature. More specifically, we have found that a prior that combines sparseness and spatial smoothness leads to a model that can learn categories closely related to the natural human-generated ones with extremely limited supervision.

In the process, we explored how the choice of feature function impacts the resulting topic prior maps, and how to use this to our advantage to get topics that match human-centric labels (Ch. 3). Further, we investigated projecting topic priors back into the space of features (Ch. 4.1). Because our topic priors are low-dimensional and spatio-temporally smooth, we found that feature distributions are easier to predict through predicting topic priors than they are directly. Combining these two concepts, we were able to predict feature distributions from very limited training data. Finally, we demonstrated an autoencoding method to learn a feature distribution function suitable for topic modeling (Ch. 5.1). This method allows inverting a feature distribution back into the raw data domain – images. Adding this last technique, we are able to directly inspect and evaluate the spatial predictions and topics themselves in a natural, human-interpretable domain.

In addition to these general contributions, we have specifically demonstrated the utility of the spatio-temporal topic model for a variety of applications First, we showed that with careful design of a feature function, a spatial topic model can be used to identify deep-sea substrate types from repurposed ROV videos (Ch. 3.1). Then, we showed that in addition, a temporal topic model can be used as an audio similarity metric for a challenging ambient sound dataset, using only off-the-shelf audio features (Ch. 3.2). Next, we developed new insight for phytoplankton ecologists approaching a dataset of over 3,000,000 phytoplankton classifications from 47 taxa. We used the spatial topic model to simplify understanding interactions between taxa and the environment, improving phytoplankton species distribution predictions from weather and oceanographic data (Ch. 4.2). In addition, we used a spatio-temporal topic model to understand the interactions between taxa, and predict the presence of unobserved plankton species given a set of observed ones (Ch. 4.3. Finally, we showed that a spatial topic model can be used to visually model a scene or environment, giving a robot camera the ability to predict what it will see if it looks in a given direction. In the process we explored how to make visual topic models more interpretable by developing an autoencoding feature function that can be learned, and a loss function that helps ensure that images are distinguishable in feature space. This allowed us to view topics, image predictions based on the topic model, and image predictions based on predicted topics directly as images instead of pure feature distributions (Ch. 5.3.4).

In all these applications, our model's hand-coded geometric knowledge is limited to an assumption that observations close to one another are likely to be semantically similar. Despite the fact that each application, most of all our final one, could be improved by more specific information about how observations in one spatio-temporal location affect what we expect to observe in others by limiting the prior information, we find a single model with minimal modification is able to achieve an impressive variety of tasks.

6.1 Future Work

In highlighting the versatility of topic models beyond the textual domain, and beyond simple unsupervised classification, we hope to enable and stimulate further work in areas that could benefit from sparse, smooth priors that we have not yet fully investigated. Firstly, our approach to learning features good for topic modeling is very general. In principle it should be simple to implement for any autoencoder architecture, in any domain where spatio-temporal data is available. Even within the domain we explored, image modeling, the space of possible autoencoder architectures is large, and our choices were based on simplicity rather than the most performant current models. Autoregressive convolutional autoencoders are an extremely good match for our method, naturally leading to multiple words with their sequential approach to image modeling, as well as being some of the top performing methods in terms of preserving detail for high-resolution images. Although our current autoencoding approach loses some detail when compressed through the topic representation, it seems likely that a more powerful feature model such as the autoregressive approach could improve this aspect. Further, beyond simple autoencoding, our extra loss term could be applied to feature layers of a supervised learning problem, and by doing so it may be possible to learn to exert a form of weak supervision on a later unsupervised topic modeling problem, akin to our hand-designed feature functions in Ch. 3.

Along similar lines, although we have focused on online, limited data regimes, sparse, spatially smooth representations are equally interesting for extremely large 'web-scale' problems where approximate inference through variational methods is preferable to MCMC. These problems are particularly interesting for end-to-end learning of the feature distribution function, the topic model, and the feature distribution decoding function. Nevertheless, to our knowledge, approximate inference for deep LDA-like models has only been attempted for text – arguably a domain that needs feature learning for LDA much less than others.

Finally, given recent notable successes in inverse reinforcement learning (IRL), we note that topic space is a potentially interesting feature space for such problems. In particular, many IRL approaches such as those descended from Maximum Entropy IRL (Ziebart et al., 2008) are considerably simplified when the reward is linear in some feature function, essentially learning the degree to which an agent 'likes' each feature dimension compared to the others. As a result, IRL applications often involve extensive supervised training of a pre-trained classifier or other feature function. As we have shown, with the right assumptions, topic-priors align well with intuitive semantic labels (that an 'agent' who we are trying to imitate may really care about) and are easy to model from limited samples in downstream learning tasks. Therefore, we propose that topics could be used in the place of such a model, alleviating the need for supervision in the pre-training steps.

6.2 Final Word

In this thesis we presented an approach to learn spatially coherent maps of complex data without any supervision. Our key insight is that interpretable maps feature a combination of semantic sparseness and spatio-temporal smoothness for a wide variety of domains. We implemented these assumptions in a Bayesian probabilistic framework, allowing our approach to be robust to different conditions associated with different problems by considering them as soft prior information rather than as hard constraints. We have shown that in comparison to prevailing unsupervised learning techniques, our approach produces representations that are both simpler to understand for humans and easier to use for later learning tasks in the common scenario where training data is limited. Our technique represents a unique tool to learn meaningful representations of spatio-temporal datasets, enabling intuitive understanding and informing further exploration for scientists, end-users, and robots alike.

APPENDIX A
Categorical Information Boost Autoencoder Architecture Details

Similar to Segnet (Badrinarayanan et al., 2015), our implementation uses a symmetric sequence of downsampling convolution modules (DownConvModule Fig. A–1a) and upsampling convolution modules (UpConvModule Fig. A–1b), where the spatial resolution changes by a factor of two at the end of each module. All convolutions are with kernel size 3, stride 1, and padding 1 so that the input and output spatial resolutions are identical.

Listing A.1: CIB-AE Architecture Detail

```
CIB_AE(
    Encoder(
        DownConvModule(3, 64, 2),
        DownConvModule(64, 128, 3),
        DownConvModule(128, 256, 3),
        DownConvModule(256, 512, 3),
        DownConvModule(512, 512, 3),
        Reshape((4, 4, 512) -> (8192)),
        Dropout(p=0.25),
        Softmax()
    ),
```

```
    Decoder(
        Reshape((8192) -> (4, 4, 512)),
        UpConvModule(512, 512, 3),
        UpConvModule(512, 256, 3)
        UpConvModule(256, 128, 3),
        UpConvModule(128, 64, 2),
        UpConvModule(64, 3, 2)
    )
)
```

In our experiments we found that the single Dropout layer before the Softmax in the encoder was crucial. Without dropout, the encodings become overly sensitive to differences in the input images, and approximate encodings through our topic model and topic mapping algorithms are too different from real encodings to produce reasonable predictions.

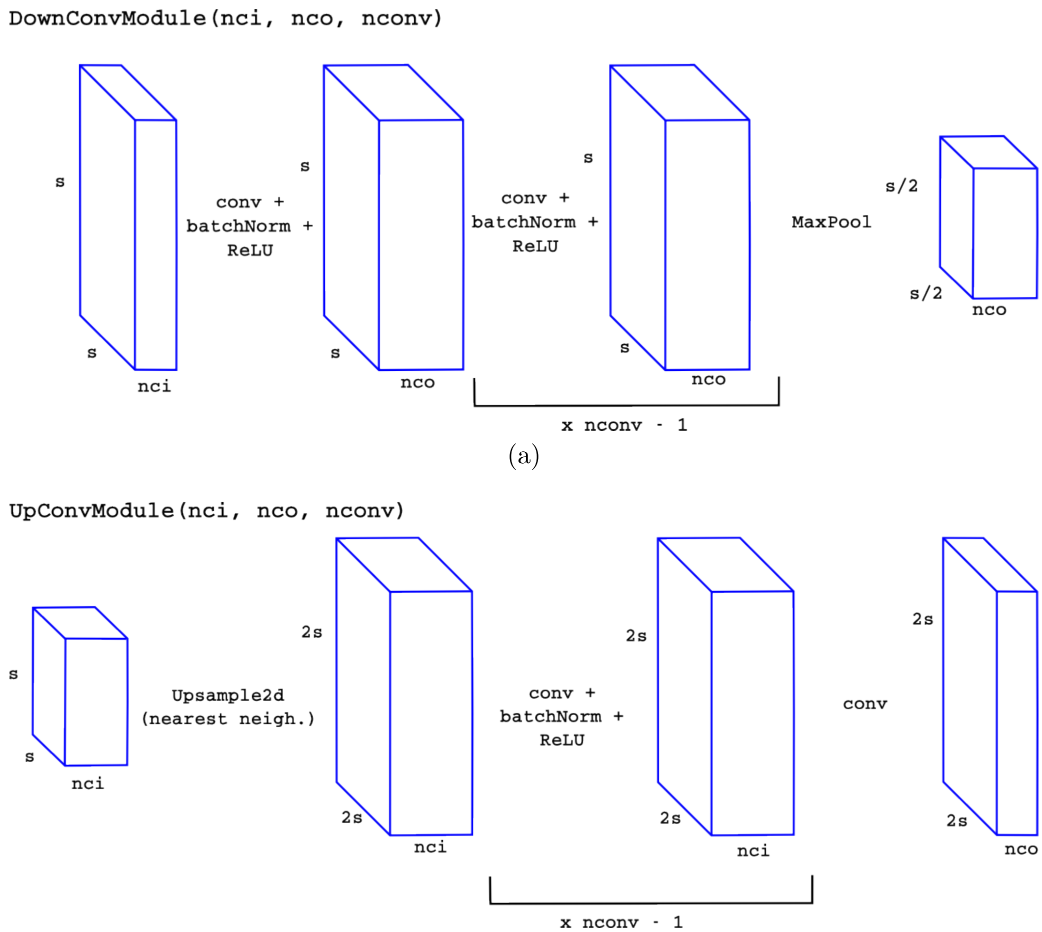


Figure A–1: (a) DownConv and (b) UpConv modules used in our architecture. nci stands for number of channels in, nco number of channels out, and s the spatial resolution of the input representation. In our model, $nco = 2nci$ for DownConvModules and $nco = nci/2$ for UpConvModules. Each block is made of a sequence of convolution, batch norm, and nonlinear activation blocks. The last convolution of the UpConvModule does not include the batch norm and nonlinearities because the training images are standardized (i.e. pixel values can be both negative and positive).

APPENDIX B
Supplemental View Predictions

B.1 Observed Views

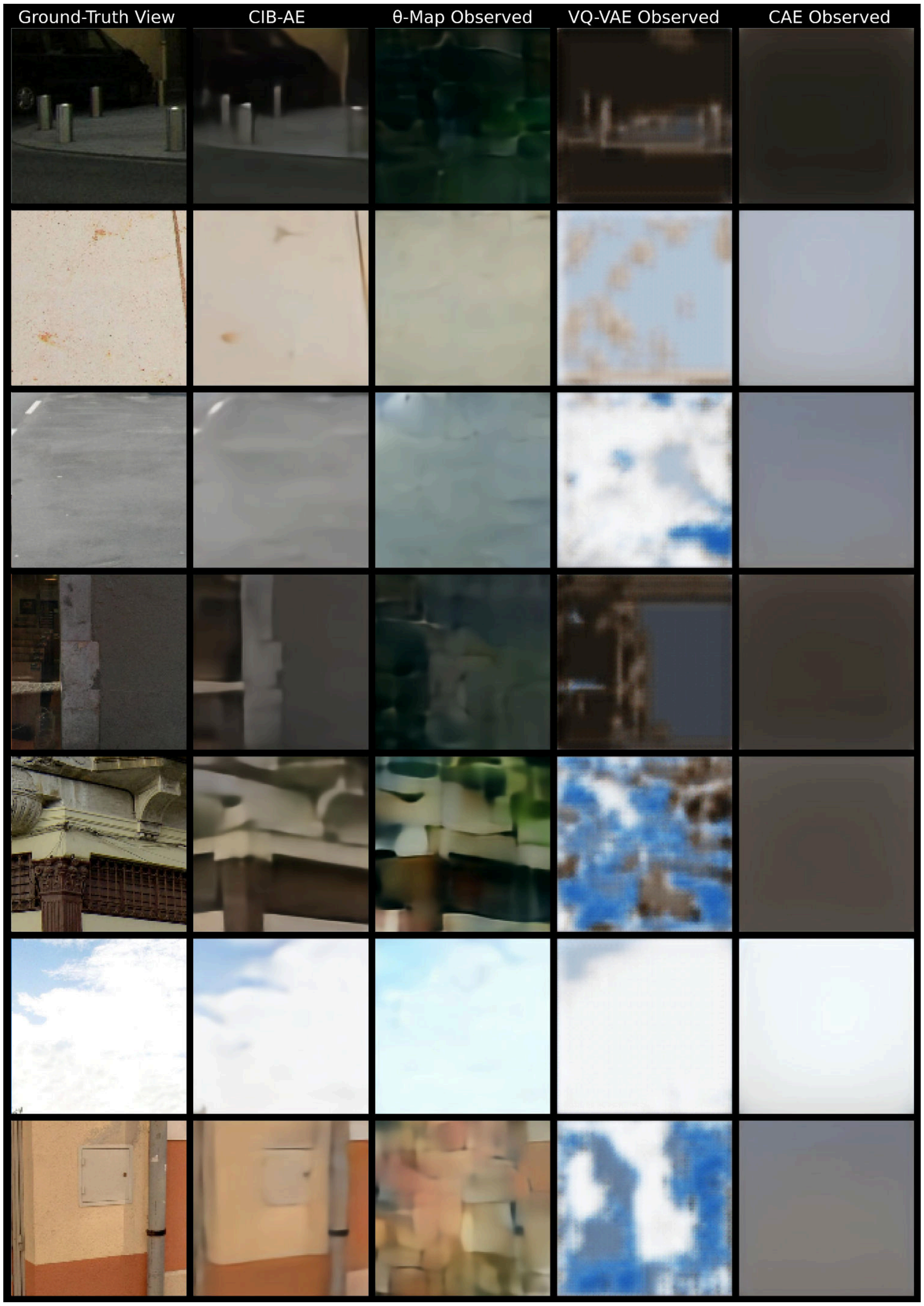


Figure B–1: Supplemental autoencoding examples

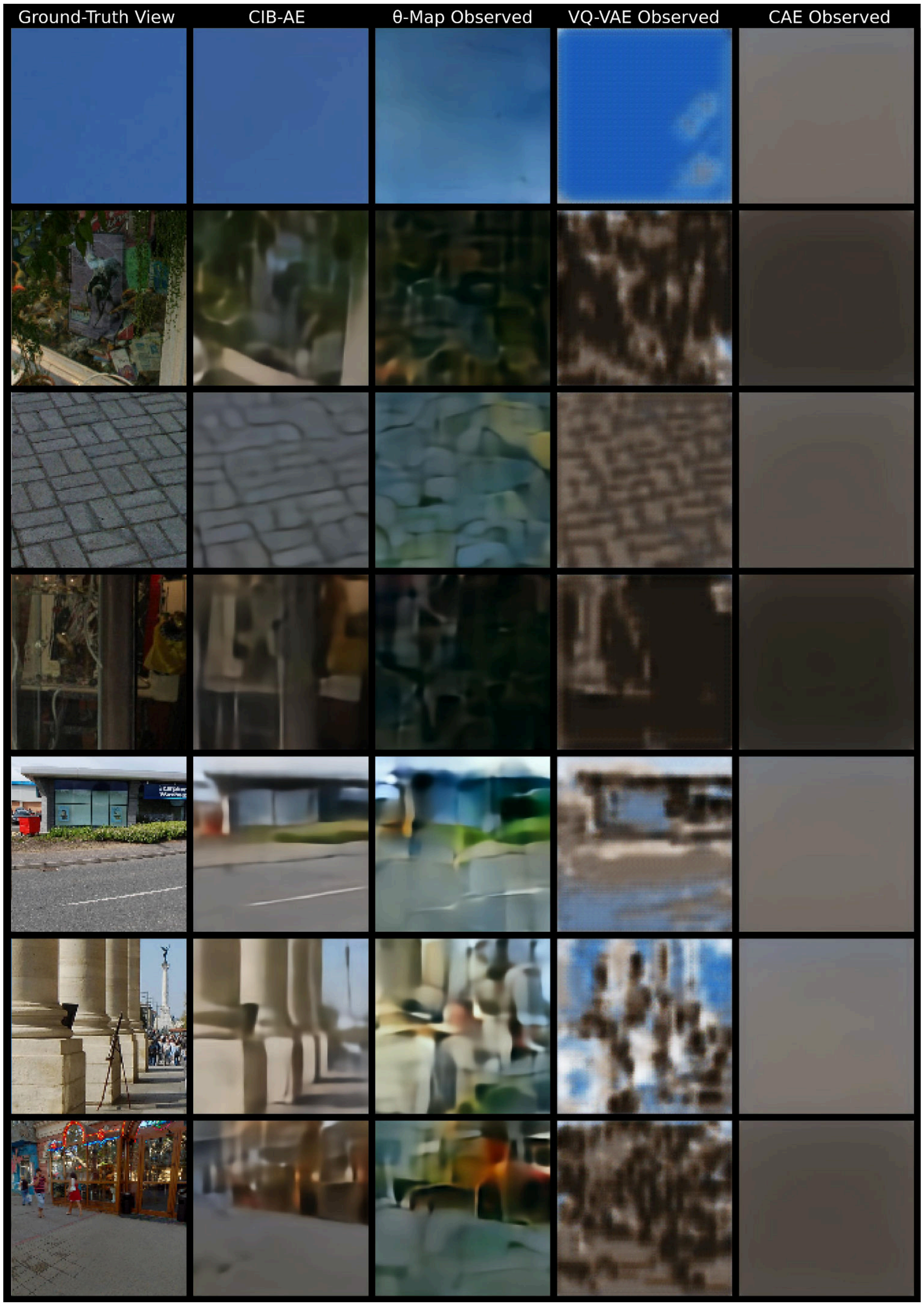


Figure B–2: Supplemental autoencoding examples (cont.)

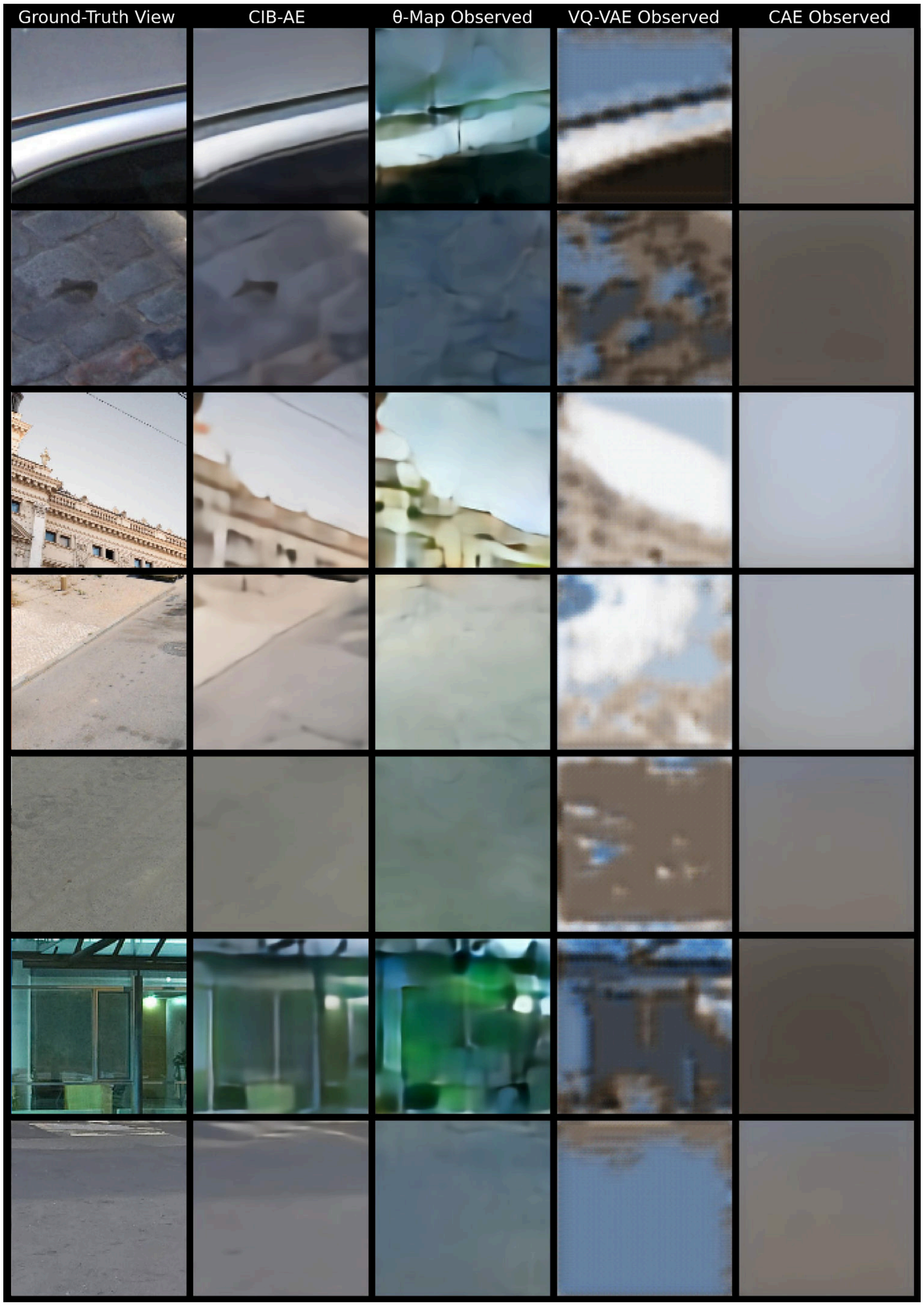


Figure B–3: Supplemental autoencoding examples (cont.)

Figure B–4: Supplemental autoencoding examples (cont.)

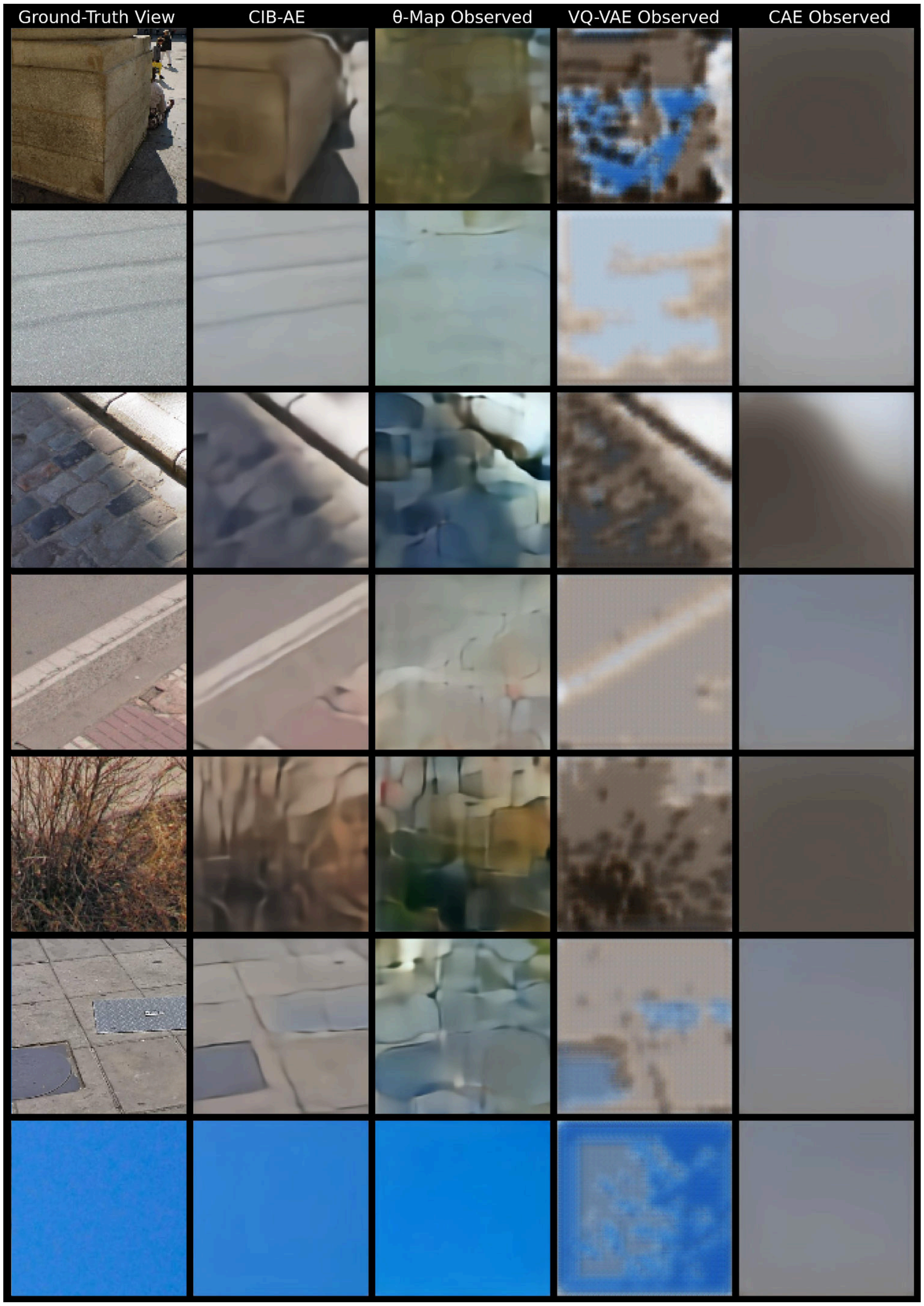


Figure B–5: Supplemental autoencoding examples (cont.)

B.2 Predicted Views

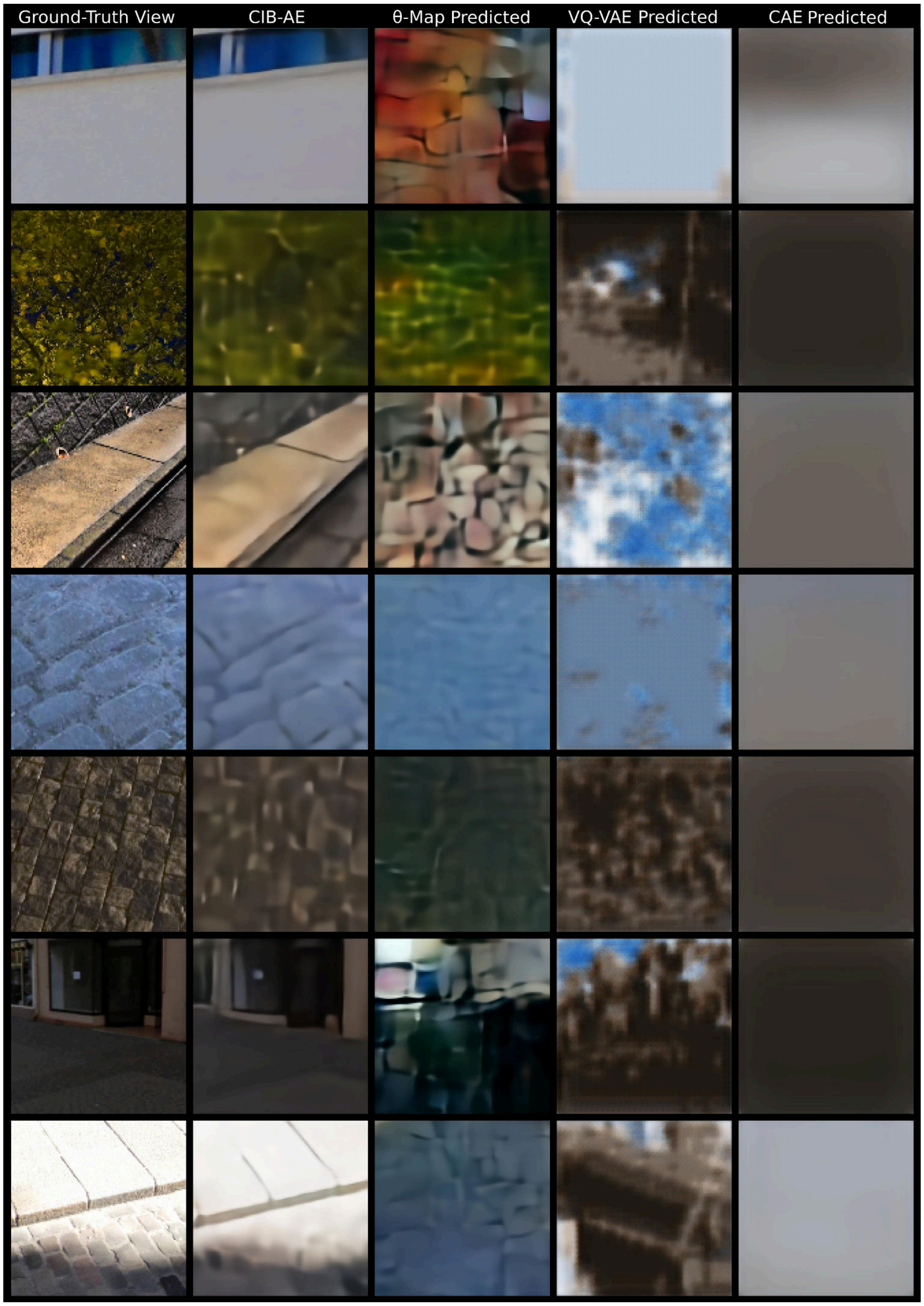


Figure B–6: Supplemental prediction examples

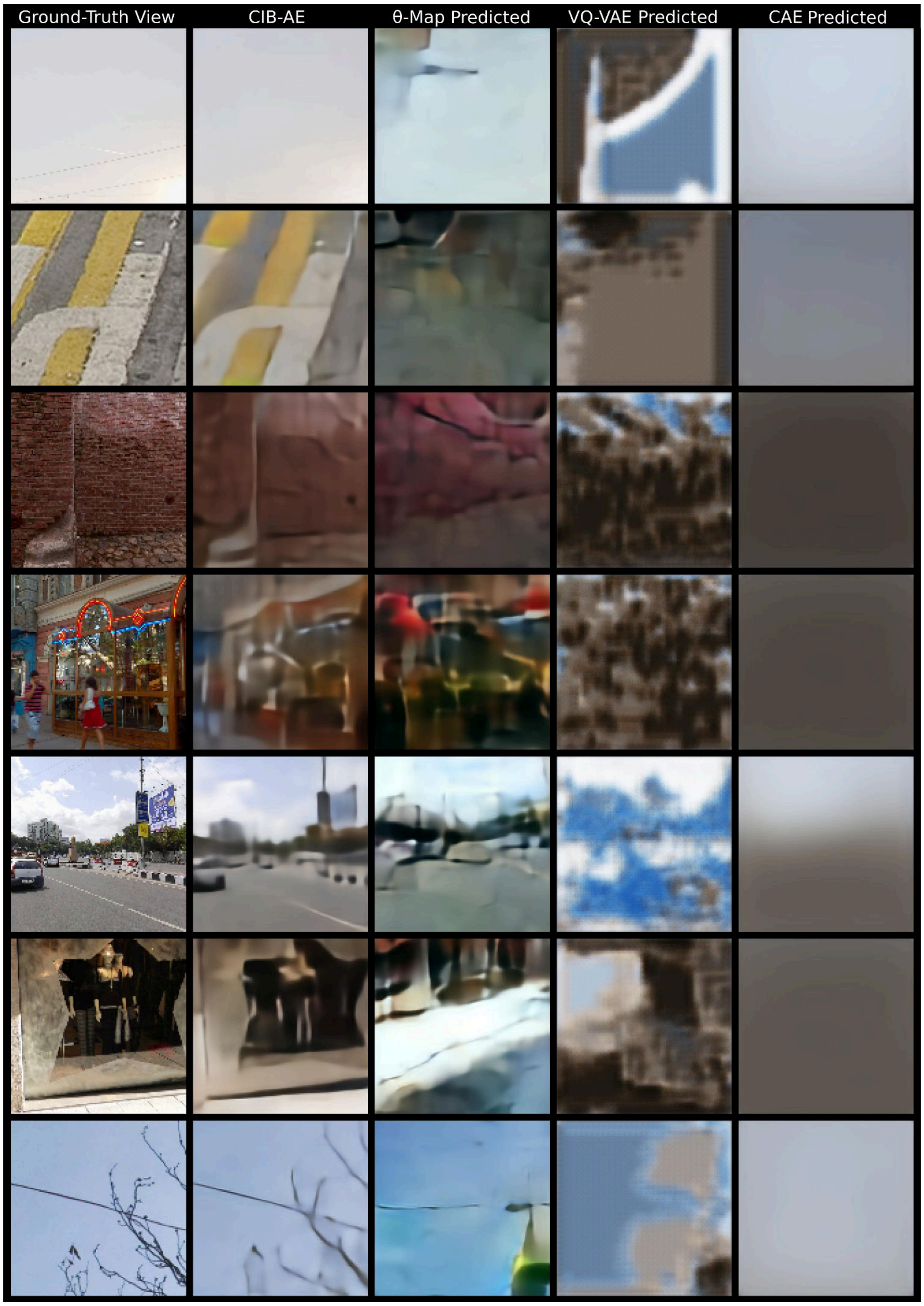


Figure B–7: Supplemental prediction examples (cont.)

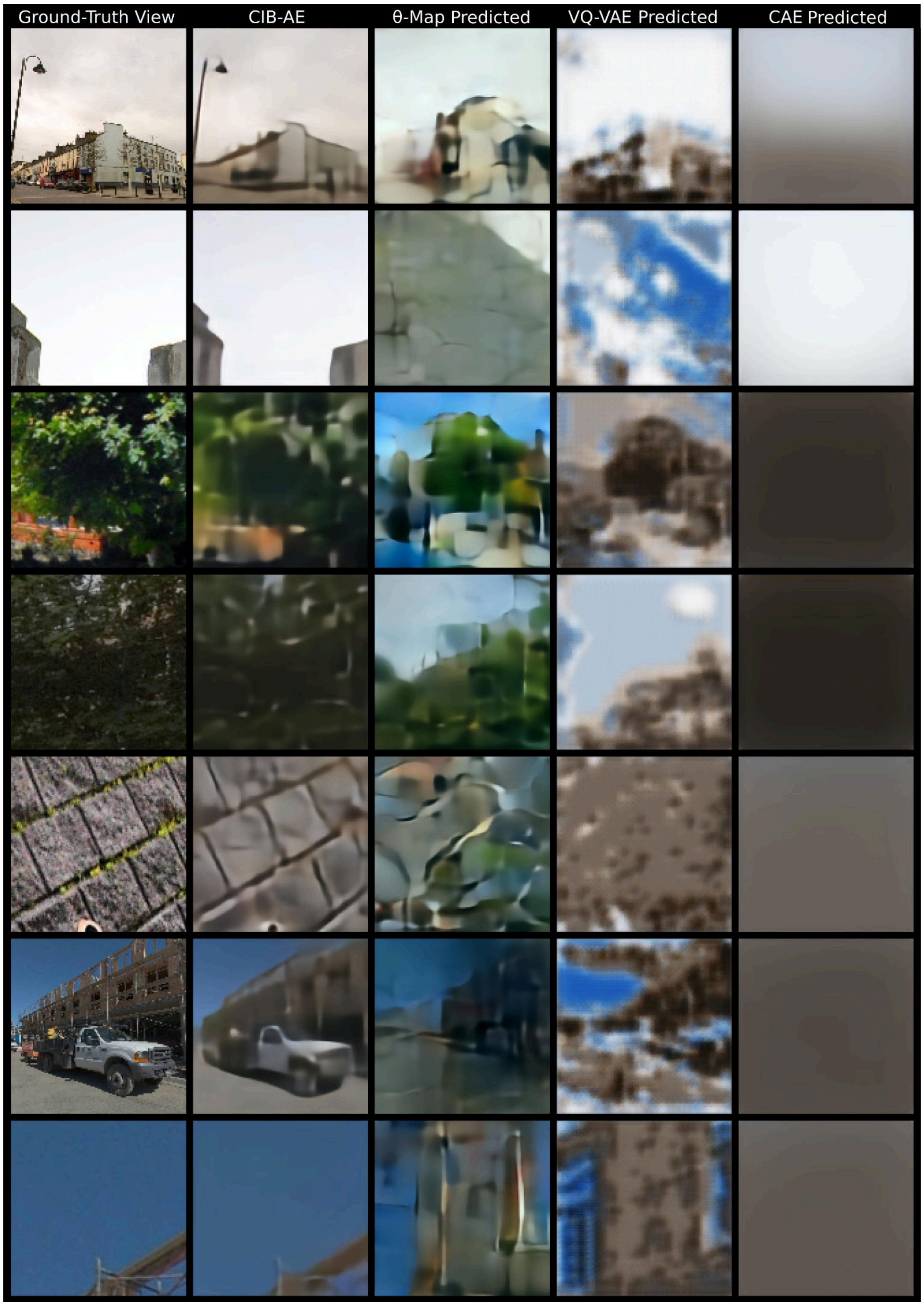


Figure B–8: Supplemental prediction examples (cont.)

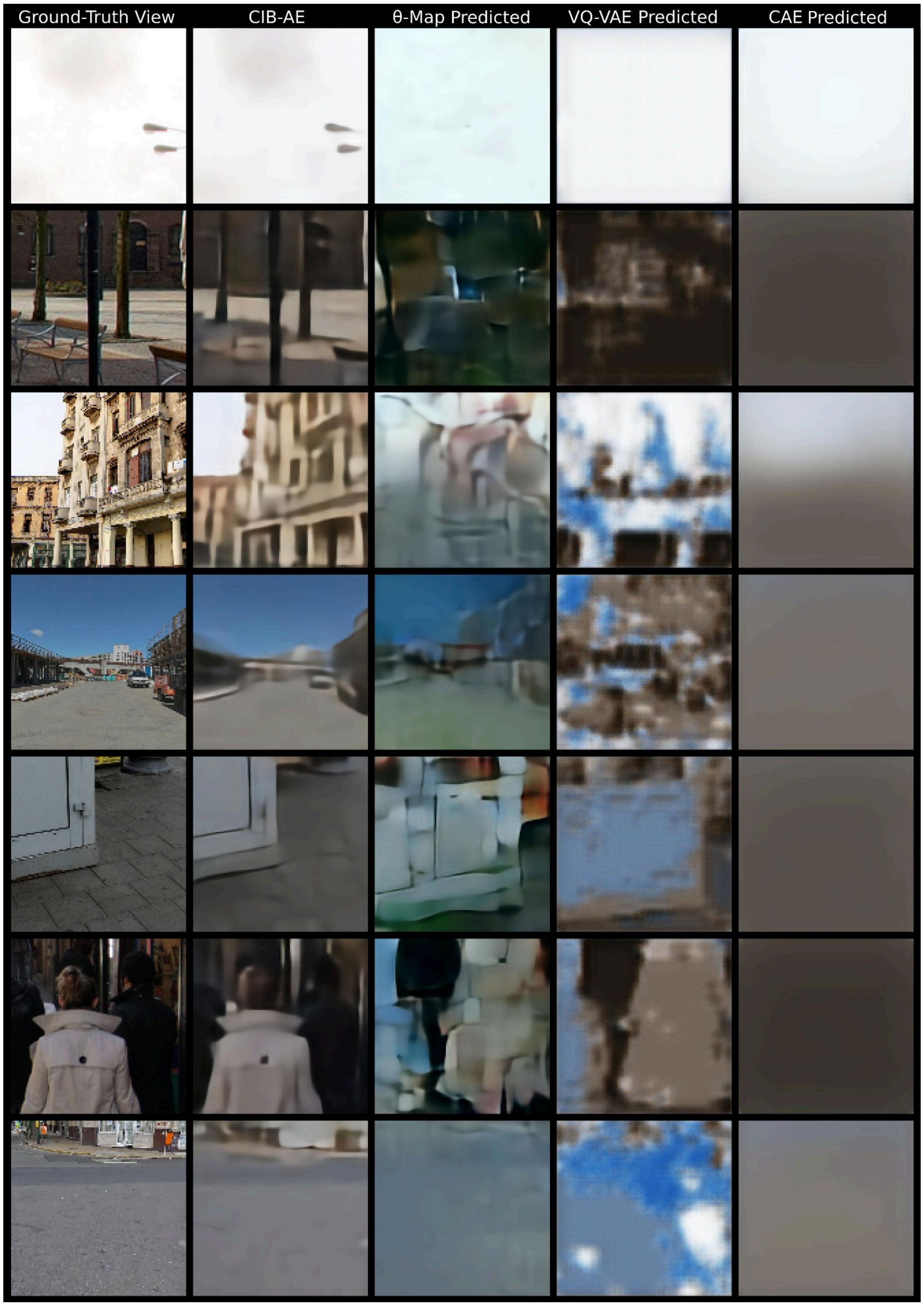


Figure B–9: Supplemental prediction examples (cont.)

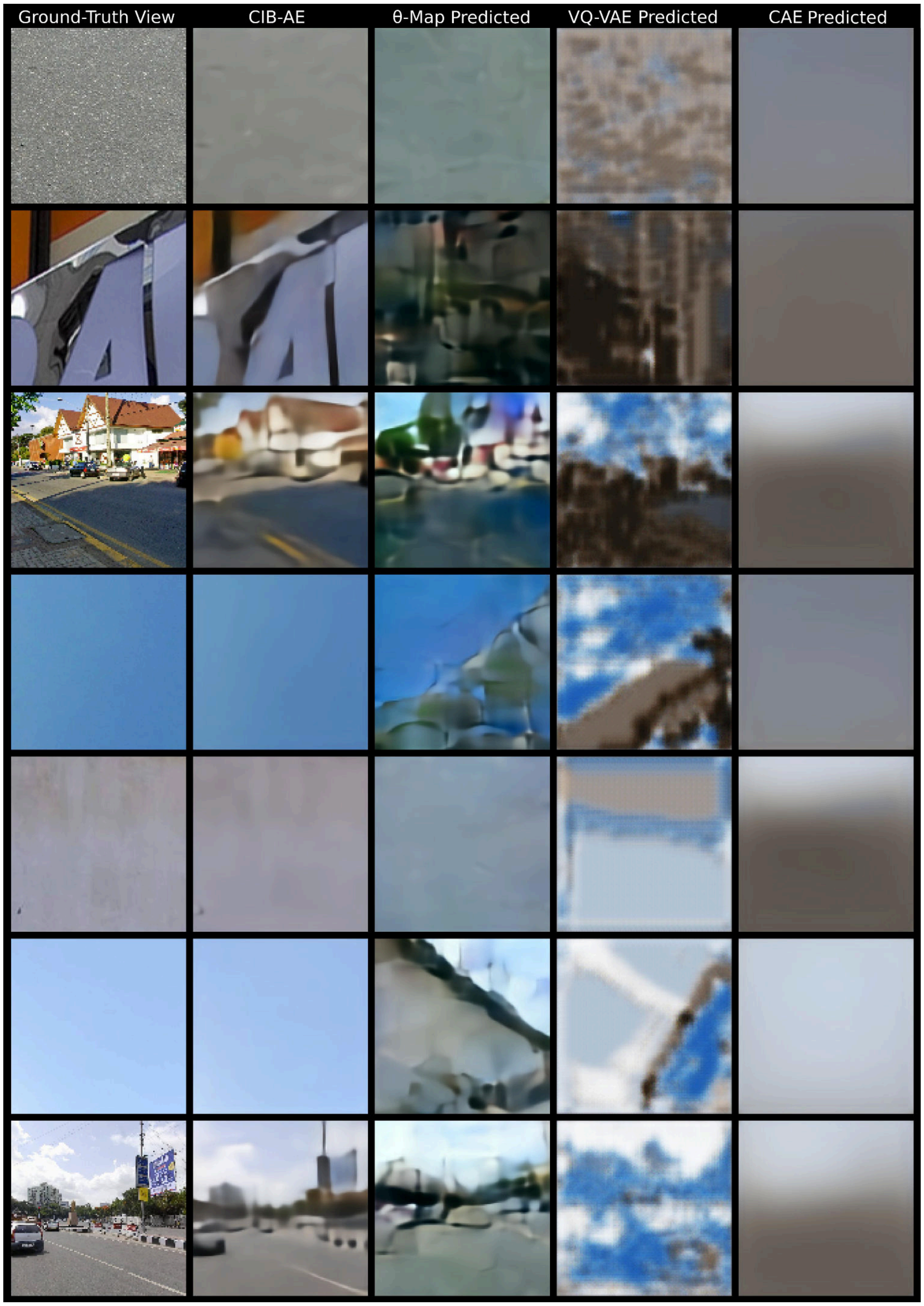


Figure B–10: Supplemental prediction examples (cont.)

References

Aggarwal, C. C., A. Hinneburg, and D. A. Keim (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pp. 420–434. Springer.

Agustsson, E., F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool (2017). Soft-to-hard vector quantization for end-to-end learned compression of images and neural networks. *CoRR abs/1704.00648*.

Badrinarayanan, V., A. Kendall, and R. Cipolla (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR abs/1511.00561*.

Bart, E., M. Welling, and P. Perona (2011, Nov). Unsupervised organization of image collections: Taxonomies and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*(11), 2302–2315.

Bengio, Y., A. Courville, and P. Vincent (2013, Aug). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(8), 1798–1828.

Bishop, C. (2006a). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Chapter 9, pp. 423–460. Berlin, Heidelberg: Springer-Verlag.

Bishop, C. (2006b). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Chapter 2.2, pp. 74–78. Berlin, Heidelberg: Springer-Verlag.

Bishop, C. M. (1999). Variational principal components. *9th International Conference on Artificial Neural Networks ICANN 99 1999*(470), 509–514.

Blei, D., L. Carin, and D. Dunson (2010, Nov). Probabilistic topic models. *IEEE Signal Processing Magazine 27*(6), 55–65.

Blei, D., A. Ng, and and M. Jordan (2003). Latent Dirichlet Allocation. *Jmlr 3*, 993–1022.

Blei, D. M. and M. I. Jordan (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, New York, NY, USA, pp. 127–134. ACM.

Chen, X., D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel (2016). Variational lossy autoencoder. *CoRR abs/1611.02731*.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 41*(6), 391–407.

Department of Fisheries and Oceans Canada (2009). Endeavour hydrothermal vents marine protected area management plan 2010-2015. Available online at `http://publications.gc.ca/pub?id=9.694447&sl=0`.

Douglas, K., S. K. Juniper, R. Jenkyns, M. Hoeberechts, P. Macoun, and J. Hillier (2017, Sept). Developing spatial management tools for offshore marine protected areas. In *OCEANS 2017 - Anchorage*, pp. 1–7.

Fei-Fei, L. and P. Perona (2005, June). A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition (CVPR'05), Volume 2, pp. 524–531 vol. 2.

Geigle, C. (2016, October). Inference methods for latent dirichlet allocation. Course notes (cs598cxz advanced topics in information retrieval), Department of Computer Science, University of Illinois at Urbana-Champaign.

Girdhar, Y. (2014). *Unsupervised Semantic Perception, Summarization, and Autonomous Exploration for Robots in Unstructured Environments*. Ph. D. thesis, McGill University.

Girdhar, Y. (2016). Modeling curiosity in a mobile robot for long-term autonomous exploration and monitoring. *Autonomous Robots 40*(7), 1267–1278.

Girdhar, Y. and G. Dudek (2012). Efficient on-line data summarization using extremum summaries. *Proceedings - IEEE International Conference on Robotics and Automation*, 3490–3496.

Girdhar, Y., D. Whitney, and G. Dudek (2014). Curiosity based exploration for learning terrain models. *2014 IEEE International Conference on Robotics and Automation*, 578–584.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc.

Gregor, K., F. Besse, D. Jimenez Rezende, I. Danihelka, and D. Wierstra (2016). Towards conceptual compression. In D. D. Lee, M. Sugiyama, U. V. Luxburg,

I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29*, pp. 3549–3557. Curran Associates, Inc.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National academy of Sciences 101*(suppl 1), 5228–5235.

Guestrin, C., A. Krause, and A. P. Singh (2005). Near-optimal sensor placements in gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pp. 265–272. ACM.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc.

Jayaraman, D. and K. Grauman (2018, June). Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Jordan, M. I. (2005). Dirichlet process, chinese restaurant process and all that. Tutorial presentation at NIPS Workshop on Nonparametric Bayesian Methods.

Kalmbach, A., Y. Girdhar, and G. Dudek (2013). Unsupervised environment recognition and modeling using sound sensing. *Proceedings - IEEE International Conference on Robotics and Automation*, 2699–2704.

Kalmbach, A., Y. Girdhar, H. M. Sosik, and G. Dudek (2017). Phytoplankton hotspot prediction with an unsupervised spatial community model. In *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 4906–4913.

Kalmbach, A., M. Hoeberechts, A. Albu, H. Glotin, S. Paris, and Y. Girdhar (2016). Learning deep-sea substrate types with visual topic models. In *2016 IEEE Winter*

Conference on Applications of Computer Vision, WACV 2016.

Kalmbach, A., H. M. Sosik, G. Dudek, and Y. Girdhar (2017). Learning Seasonal Phytoplankton Communities with Topic Models. In *Oceans Mts/Ieee*.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *CoRR abs/1412.6980*.

Kingma, D. P. and M. Welling (2014). Auto-Encoding Variational Bayes. In *ICLR*, Number Ml, pp. 1–14.

Krumbein, W. C. and E. J. Aberdeen (1937). The sediments of barataria bay [louisiana]. *Journal of Sedimentary Research 7*(1), 3.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2*(1-2), 83–97.

Liao, S., X. Zhu, Z. Lei, L. Zhang, and S. Li (2007). Learning Multi-scale Block Local Binary Patterns for Face Recognition. *Advances in Biometrics*, 828–837.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory 37*(1), 145–151.

Lorenzen, C. J. (1966). A method for the continuous measurement of in vivo chlorophyll concentration. *Deep-Sea Research 13*, 223–227.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 1–28.

Manjanna, S., N. Kakodkar, M. Meghjani, and G. Dudek (2016, June). Efficient terrain driven coral coverage using gaussian processes for mosaic synthesis. In *CRV '16: Proceedings of the 2016 International Conference on Computer and Robot Vision*. IEEE Computer Society.

Niebles, J. C., H. Wang, and L. Fei-Fei (2008, Sep). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision 79*(3), 299–318.

Oliva, A. and A. Torralba (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision 42*(3), 145–175.

Olson, R. J. and H. M. Sosik (2007, jun). A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods 5*(6), 195–203.

Paris, S., X. Halkias, and H. Glotin (2012). Efficient Bag of Scenes Analysis for Image Categorization. *International Conference on Pattern Recognition Applications and Methods*.

Peacock, E. E., R. J. Olson, and H. M. Sosik (2014). Parasitic infection of the diatom guinardia delicatula, a recurrent and ecologically important phenomenon on the new england shelf. *Marine Ecology Progress Series 503*, 1–10.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics 155*(2), 945–959.

Seo, S., M. Wallat, T. Graepel, and K. Obermayer (2000). Gaussian process regression: active data selection and test point rejection. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Volume 3, pp. 241–246 vol.3.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning 6*(1), 1–114.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556*.

Sosik, H. M., J. Futrelle, E. F. Brownlee, E. Peacock, T. Crockford, and R. J. Olson (2016, September). hsosik/ifcb-analysis: IFCB-Analysis software system, initial formal release at v2 feature stage.

Sosik, H. M. and R. J. Olson (2007, jun). Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods 5*(6), 204–216.

Sosik, H. M., E. E. Peacock, and E. F. Brownlee (2014). Whoi plankton dataset – annotated plankton images. `https://hdl.handle.net/10.1575/1912/7341`.

Srivastava, A. and C. Sutton (2017). Autoencoding variational inference for topic models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sturm, B., M. Morvidone, and L. Daudet (2010). Musical instrument identification using multiscale mel-frequency cepstral coefficients. In *18th European Signal Processing Conference (EUSIPCO-2010)*, Number 1, pp. 0–4.

Teh, Y. W. (2010). *Encyclopedia of Machine Learning: Dirichlet Process*, pp. 280–287. Boston, MA: Springer US.

van den Oord, A., N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu (2016). Conditional image generation with pixelcnn decoders. *CoRR abs/1606.05328*.

van den Oord, A., O. Vinyals, and K. Kavukcuoglu (2017). Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 6306–6315. Curran Associates, Inc.

Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research 11*(Dec), 3371–3408.

Wallach, H. M., I. Murray, R. Salakhutdinov, and D. Mimno (2009). Evaluation Methods for Topic Models. *International Conference on Machine Learning* (d), 1–8.

Xiao, J., K. A. Ehinger, A. Oliva, and A. Torralba (2012). Recognizing scene viewpoint using panoramic place representation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2695–2702. IEEE. See also supplemental material on the geometry `http://people.csail.mit.edu/jxiao/SUN360/supp_final/panorama.pdf`.

Ziebart, B. D., A. L. Maas, J. A. Bagnell, and A. K. Dey (2008). Maximum entropy inverse reinforcement learning. In *AAAI*, Volume 8, pp. 1433–1438. Chicago, IL, USA.

Zilberstein, S. (1996). Using anytime algorithms in intelligent systems. *AI magazine 17*(3), 73.