PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Diversity-based active learning: creating a representative object detection dataset in 3D point clouds

Aimee Moses, Christopher Bogart, Kelsey O'Haire, Lidia Solorzano, Elinor Yeo

Aimee Moses, Christopher W. Bogart, Kelsey O'Haire, Lidia Solorzano, Elinor Yeo, "Diversity-based active learning: creating a representative object detection dataset in 3D point clouds," Proc. SPIE 12525, Geospatial Informatics XIII
, 125250F (15 June 2023); doi: 10.1117/12.2663179

SPIE.

Event: SPIE Defense + Commercial Sensing, 2023, Orlando, Florida, United States

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Diversity-Based Active Learning: Creating a Representative Object Detection Dataset in 3D Point Clouds

Aimee Moses[a], Christopher W. Bogart[a], Kelsey O'Haire[a], Lidia Solorzano[a], and Elinor Yeo[a,b]

[a]Expedition Technology, Inc, 13865 Sunrise Valley Drive, Suite 350, Herndon, VA USA, 20171
[b]UCLA Department of Psychology, Los Angeles, CA

ABSTRACT

Object detection in 3D point clouds is essential in fields such as geospatial intelligence and autonomous driving. The common machine learning problem of scarce labeled training data is even more acute with 3D point cloud data. Active learning provides a framework to prioritize the additional effort to manually annotate unlabeled training data. Most active learning methods for deep learning fall into one of two categories: uncertainty methods and diversity methods. Uncertainty methods select data by assessing model outputs for their confidence and consistency and are therefore dependent on the expected output of each deep learning task. These methods tend to select batches of informative yet highly similar samples to label. Diversity-based active learning aims to create a labeled dataset that is both varied and representative of the remaining unlabeled data. Diversity methods operate directly on the feature representations of the inputs and are thus more flexible with respect to the specifics of the deep learning task. Our current work explores applying diversity methods and uncertainty-diversity hybrid methods to 3D object detection. We evaluate various approaches to incorporate diversity, including K-Medoids Clustering, Core Set Selection, and Furthest Nearest Neighbors. We address the high dimensionality of the features extracted from a VoxelNet-based object detector by varying the distance metric used in the active learning algorithms. Furthermore, we compare our results to those obtained using only uncertainty methods. We assess the performance and efficiency of each active learning method in addition to the representativeness and diversity of the labeled datasets produced. We find that hybrid uncertainty-diversity methods outperform other methods in terms of object detection AP50 throughout active learning, annotation efficiency, and class balance.

Keywords: Active learning, deep learning, point clouds, object detection, feature representation, latent space, diversity methods, uncertainty estimation

1. INTRODUCTION

Deep learning has revolutionized the task of object detection and classification in computer vision. While deep learning produces state-of-the-art results,[1] it is resource intensive, requiring large amounts of data to achieve satisfactory performance. Thanks to readily available large-scale datasets, 2D object detection in images is generally considered a solved problem, with networks achieving error rates of less than 2.5% on the ImageNet dataset.[1–3] However, there is still a need for improvement in 3D object detection in point clouds, with the current leaders on one notable single-modal point cloud object detection benchmark performing at under 70% for their evaluation metric.[4] The largest obstacle in this domain is access to sufficient high-quality synthetic aperture radar (SAR) or light detection and ranging (LiDAR) 3D point cloud data. 3D point clouds also have unique characteristics that make object detection a challenge: point clouds are typically variable in point density, suffer from point occlusion, and present further challenges.[5] Even when point cloud data is available, labeling targets is often tedious, time consuming, and prone to human error.[5,6] Our work aims to ease the burden of manually labeling targets in 3D point clouds by leveraging active learning to select point clouds for training a high accuracy 3D object detector.

Active learning provides a framework for minimizing data required to achieve comparable performance results to a model trained on a full dataset.[6,7] Our work focuses on pool-based active learning,[8,9] identifying the most

Further author information: (Send correspondence to Aimee Moses)
Aimee Moses: E-mail: amoses@exptechinc.com

Geospatial Informatics XIII, edited by Kannappan Palaniappan, Gunasekaran Seetharaman, Joshua D. Harguess, Proc. of SPIE Vol. 12525, 125250F · © 2023 SPIE
0277-786X · doi: 10.1117/12.2663179

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

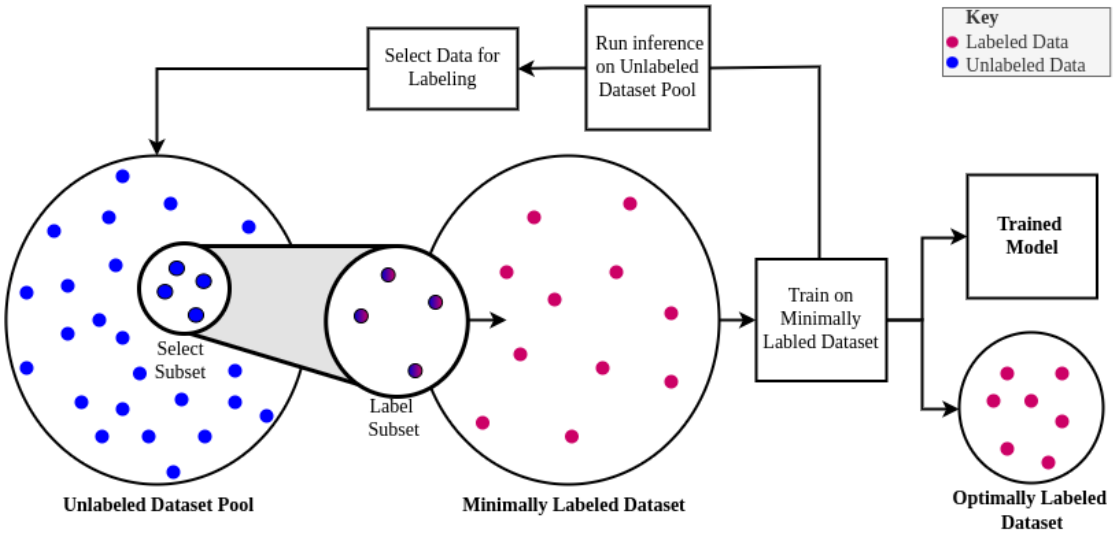

Figure 1. High level overview of active learning. Begin with an unlabeled pool of data and label a small subset. Train a model on this small labeled dataset in a supervised manner. Run inference with the trained model on the remaining unlabeled data. Based off the model outputs, active learning selects samples optimal for the next iteration of training. Those samples are manually labeled and added to the labeled dataset, the model is retrained, and the cycle continues until the chosen stopping criteria is met.

information-rich samples from a pool of unlabeled data to be labeled and used for training a deep learning model. An overview of the active learning pipeline is shown in Figure 1. The recent literature includes many variations on this framework across deep learning tasks, including self- and semi-supervised learning to utilize the unlabeled data during training,[10,11] modifications for domain adaptation[12–14] and open-set detection tasks,[15] and integrating data selection into the task model training.[16,17] These methods are not yet sufficiently mature for 3D point cloud data so were not included in this analysis. Prior work shows that the active learning pipeline shown in Figure 1 can select a subset as little as 33%[18] to 50%[19] of a dataset, depending on the complexity of the dataset, and train a model with the same performance as one trained on all available data. Our work is one of the first approaches implementing active learning for 3D point cloud data[6,7,20] and the first to our knowledge to use diversity methods for a dataset comprised only of 3D point clouds.

Training a deep learning model properly requires a robust dataset comprised of both abundant and representative data points. To achieve these requirements, there are three common active learning paradigms for data selection: diversity, uncertainty, and a hybrid approach.[9,21] Uncertainty methods define acquisition functions to select data samples that the model is least confident detecting. While this is an effective method for clarifying decision boundaries, selected samples tend to be highly similar to one another.[9] Diversity methods are used to alleviate this drawback. Diversity methods select scenes that improve the heterogeneity of the batch. Most diversity methods utilize learned feature representations to evaluate similarity among samples.[6,19,21–23] Based on a review of the current literature, only one prior work has explored diversity methods for object detection in 3D point clouds, but this study used a multimodal dataset with point clouds as only one of a variety of data inputs.[7] Hybrid methods leverage aspects of both diversity and uncertainty, often simply using a combination of the two methods.

In this work, we explore multiple diversity methods of active learning for 3D point cloud data. Specifically, we employ K-Medoids, Core Set Selection, and Furthest Nearest Neighbors for data selection in the feature latent space. We evaluate these methods on their own as well as when utilized in uncertainty-diversity hybrid methods. We compare these methods to our past work using uncertainty methods.[6] We use a modified version of VoxelNet as our deep learning object detection model. Details of our model can be found in Section 3.1.

We begin Section 2 by exploring prior advances made in both active learning and object detection tasks. In

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Section 3, we discuss our approach to object detection in 3D point clouds and implementation of diversity-based active learning algorithms. Sections 4 and 5 discuss the experiments and results of our trials. Our closing points and plans for future work can be found in Section 6.

2. BACKGROUND

2.1 Uncertainty-Based Active Learning

Most active learning methods operate by applying one or more forward passes of the deep learning model on the unlabeled data and selecting unlabeled samples that are expected to be the most favorable for labeling. As the name implies, the acquisition function for uncertainty-based sample selection computes a metric quantifying how certain the model is about its predictions and selects those samples where the model is least certain as the most advantageous prospective samples to be labeled.

The predicted probability of class membership output by a final softmax layer of the trained model is commonly used as input for the acquisition function. The simplest approach is taking the score for the most probable class for each of the unlabeled samples, treating this value as the confidence of the prediction, and labeling the samples whose detections have the lowest confidence scores among the unlabeled samples.[24–26] To take advantage of all available information, methods can also consider predicted probabilities for other classes, not just the best scoring class. One such approach is to treat the softmax values for all classes as random variables drawn from a probability distribution and compute the Shannon Entropy of that distribution.[23,25,27] If all classes are equally distributed, the Entropy will be maximized; if one class is highly probable, the Entropy will be small. Then, high Entropy unlabeled samples, about which the model is least decisive, can be selected for labeling. Entropy of the softmax vector is defined in Equation (1) for an input x, train data D_{train}, and output y with possible classes $c \in C$.

$$\mathcal{H}(y|x, D_{train}) = -\sum_{c} \mathrm{P}(y = c|x, D_{train}) \log \mathrm{P}(y = c|x, D_{train}) \tag{1}$$

The methods mentioned above rely on the assumption that predicted probability of class membership is a valid proxy for the confidence of the prediction. However, in practice, deep learning methods oriented toward classification tend to overestimate the confidence of class membership.[28] As a result, many active learning methods avoid taking this "confidence" at face value. One approach is introducing stochasticity to the same model at inference-time using techniques such as Monte Carlo dropout[29] or inference-time data augmentation.[26,30] These approaches will cause inference to produce non-deterministic outputs. Another more computationally intensive technique for obtaining variable outputs is to train an ensemble of models with randomized initializations on the same data. Averaging the outputs of stochastic inference runs can provide a more robust estimation of prediction confidence.[23,31,32] For T forward passes, where M_t is the t^{th} variation of the model, we can define the average softmax score as:

$$\mathrm{P}(\mathbf{y}|x, D_{train}) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{P}(y|x, M_t). \tag{2}$$

Other approaches compare the predictions of the collection of varying outputs for each unlabeled data point. Samples that show large variation between these stochastic outputs may lay near class decision boundaries and will thus be more valuable to label. These methods include query-by-committee methods which make use of disagreement between predictions by giving each a "vote" as to whether to label or not label a particular sample.[33–35] Another approach is to calculate the Mutual Information across the runs. An advantage of using Mutual Information is that it can make use of the Entropy of the distributions of softmax outputs across the collection of predictions.[23,27,31,32,36] We can write it as the Entropy of the average softmax vector minus the average of the Entropy of the individual softmax vectors:

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

$$\begin{aligned}\mathcal{MI}(\mathbf{y}, \mathbf{M}|x, D_{train}) &= -\sum_c \mathrm{P}(\mathbf{y}=c|x, D_{train}) \log \mathrm{P}(\mathbf{y}=c|x, D_{train}) \\ &\quad - \frac{1}{T}\sum_{t=1}^{T}\sum_c -\mathrm{P}(y=c|x, M_t) \log \mathrm{P}(y=c|x, M_t) \\ &= \mathcal{H}(\mathbf{y}|x, D_{train}) - \mathbb{E}_{\mathrm{p}(M|D_{train})}[\mathcal{H}(\mathbf{y}|x, M)].\end{aligned} \tag{3}$$

For object detection tasks in images or point clouds, object classification probability is not the only consideration in selecting samples to label in uncertainty-based active learning. The number of objects detected as well as the positions of the objects vary in each image. In this scenario, acquisition functions should incorporate information about the uncertainty in spatial position and orientation of the detected objects in addition to their classifications and class probabilities. The simplest approach is to ignore the spatial variation, instead using the classification metrics and aggregating across the detections found in each sample.[37–39] Another approach is to consider classification uncertainty by pixel, using a mix of local and global aggregation to attempt to incorporate spatial uncertainty.[40,41] New metrics have also been developed to address the uncertainty in object position and orientation, with mixed success.[6,20,26,39,40,42–46] Our previous work on LOCalization-Based Active Learning (LOCAL) draws on the most successful of these approaches.[6] The main innovations of LOCAL over prior approaches are the method of matching detections across stochastic inference outputs for uncertainty calculations, and a new metric for evaluating uncertainty that takes into account the spatial uncertainty of the predicted bounding boxes as well as the softmax values of matched detections.

2.2 Diversity-Based Active Learning

Although uncertainty-based active learning with deep learning networks has been successfully demonstrated on both 2D and 3D image and point cloud data, there are drawbacks that limit its performance and applicability. Uncertainty methods tend to select samples that are close to the decision boundary between classes, exploiting known unknowns to train a model in order to fine-tune this boundary.[47] While these samples can be strong candidates for labeling, they can be quite similar and, thus, redundant for training a model.[22,47–49] Consequently, samples that represent the full diversity of the latent space but are further from the decision boundary may be missed. This limits the amount of variation in the dataset, reducing the generalization of the model. Since most uncertainty methods operate on the model predictions for the unlabeled data, they are also dependent on the format and characteristics of the expected outputs for the specific task.

Diversity methods provide an alternative approach to sample selection. Diversity methods explore the sample space by selecting data points that are far apart from one another on a hyperplane. The most common approaches to diversity-based data selection occur in a latent domain. Data is projected onto a latent space and samples are selected by either clustering or traversing data across the hyperplane. Feature extraction followed by sample selection is the method used in this work; it is task agnostic and fully independent of the model outputs. We define a data sample as a single point cloud tile that has been projected into the latent feature space of our 3D object detector.

There are several approaches to selecting samples with a diversity focus, many of which are derived from classical clustering algorithms. We focus on three common diversity selection methods: K-Medoids, Core Set, and Furthest Nearest Neighbor. Xu et al.[48] were the first to introduce clustering as a means of selecting samples for active learning. Using a Support Vector Machine with logistic regression as their task model, they employ K-Medoids clustering to select data for labeling. K-Medoids selects samples as centers for clusters, and then iteratively selects new samples as centers in order minimize the distance of all points in a cluster to their center.[50] When using K-Medoids for active learning, the final cluster centers are considered exemplars in the data and are selected for labeling.

Sener and Savarese[22] define an active learning method, Core Set, as a mathematical cover set selection task. When choosing samples to select, the Euclidean distance is taken between activations of the final fully connected layer of their network.[22] Core Set chooses a set of data samples that minimizes the distance between each element of the unlabeled set and its nearest selected sample. Their results show that Core Set is an extremely effective

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

algorithm for data selection when training a fully supervised model, achieving up to 5% accuracy improvement compared to other state-of-the-art algorithms across the CIFAR-10,[51] CIFAR-100,[51] and SVHN[52] datasets.

Furthest Nearest Neighbors (FNN) is a similar algorithm to Core Set in that it is a distance-based approach. In contrast to Core Set, FNN considers the currently labeled data when selecting unlabeled samples to label next. Informally, samples selected are a maximum distance from the labeled set. Geiffman and El-Yaniv[47] were the first to implement a Furthest Nearest Neighbors-style algorithm in the context of active learning. By labeling only 50% of the CIFAR-10 dataset, they only lost 1.5% accuracy compared to a model trained on the 100% of the dataset.

Other works have introduced novel approaches to diversity-based data selection. Rather than strictly focusing on the feature space to select samples, Liang et al.[7] also incorporate temporal and spatial diversity when selecting candidate points from their multimodal dataset. Gissin and Shalev-Shwartz[53] introduce their methodology, Discriminative Active Learning (DAL), which relies on a binary classifier to determine if a sample comes from the labeled or unlabeled distribution. The goal of their work is to select samples such that the labeled and unlabeled distributions look indistinguishable. If the features are divergent from the labeled subset distribution, they are predicted to be coming from the unlabeled dataset, thus, should be labeled. Additional works show that it is possible to select methods using a gradient-based approach to diversity-based data selection.[9,49,54] Guo and Schuurmans[54] focus on a binary logistic regression problem. They optimize the log-likelihood of labeled examples and pseudo-labels of a selected subset to minimize the Entropy of the unlabeled samples without the subset. Adversarial approaches to selecting data samples have also been explored.[17,55,56] Shui et al.[56] propose a method to data selection where a variational autoencoder is trained to fool an adversarial network into predicting that all data is from a labeled dataset. As a result, the adversarial network learns the dissimilarities between the labeled and unlabeled sets in the latent space. This adversarial approach produces a labeled dataset with a slight improvement from classical approaches. While these are strong approaches to diversity-based active learning, significant changes are needed to the task model to implement these methods. Therefore, they are not evaluated in this work but may be the subject of future research.

2.3 Uncertainty-Diversity Hybrid Active Learning Methods

The purpose of a hybrid active learning approach is to leverage the strengths of both diversity and uncertainty methods. An effective dataset for training an object detector must include samples that help to fine-tune an improved decision boundary with respect to encoded features, which is the strength of uncertainty methods. In turn, the strength of diversity methods is selecting samples that reflect the full range of variation in the dataset. There are a wide variety of hybrid approaches proposed in the literature. Unfortunately, due to the uncertainty component these methods are not agnostic to the task or model outputs.

A naïve approach to uncertainty-diversity active learning is to use a weighted combination of uncertainty scores and diversity metrics to select samples.[57–59] In their work, Rodriguez et al.[59] propose a novel acquisition function for batches of data that contain a sum of both uncertainty and diversity weights. Kirsch, Van Amersfoort, and Gal[60] instead use a modified version of uncertainty scoring, taking the Mutual Information of the whole batch in order to maximize diversity of samples selected together. Their work shows improved diversity and performance of the model compared to ordinary Mutual Information. Zhdanov[61] takes a more integrated approach, comparing three different hybrid active learning methods that combine K-Means and uncertainty scoring. In the first proposed approach, the cost function of K-Means is altered to include both distance and uncertainty scores, where the centroids are selected as points to be labeled. In the second, latent features are filtered by uncertainty scores and then clustered, selecting points near each cluster center. The final approach uses normal K-Means to cluster and then picks the most informative example from each cluster rather than the centroid. Results from these papers show that across the proposed approaches, hybrid methods performed better than the uncertainty baselines.

Other hybrid methods consist of successively applying uncertainty and diversity methods across the feature space. One of the earliest of these approaches was introduced by Smailagic et al.[62] with their framework, MeDAL. They introduce a novel approach to data selection in a deep learning framework, where they filter data using Entropy of the predictions. Then, they use a trained object detector to project data into the feature space and select samples with the highest average distance to the other samples for labeling. More recently, Wu et al.[63]

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

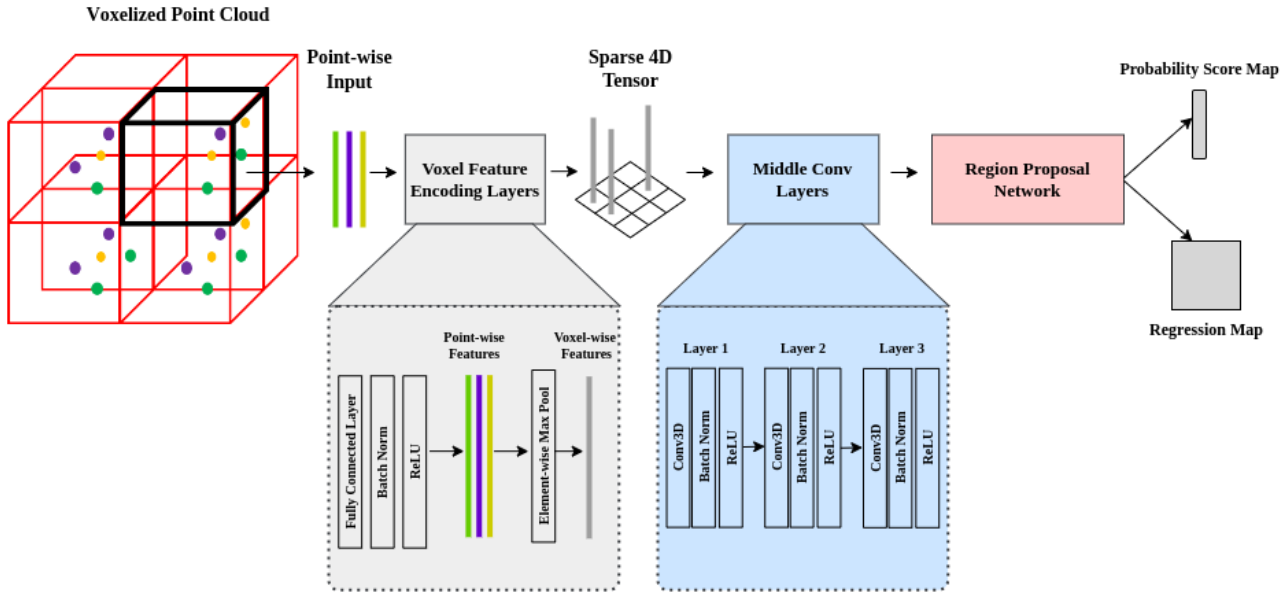

Figure 2. High level overview of our modified VoxelNet object detector. Initially, we voxelize our point clouds into evenly sized 3D voxels. Point-wise Inputs (raw data) are embedded using Voxel Feature Encoding layers into Voxel-wise Features. The Voxel-wise embeddings populate a sparse 4D tensor representative of the initial point cloud. This tensor is passed into the Middle Convolutional Layers and then the Region Proposal Network stages of the architecture. The Region Proposal Network outputs a probability map and regression map, creating the object predictions.

employ a similar approach in their active learning framework, ReDAL. In contrast to MeDAL, ReDAL begins by using a diversity method to explore the latent codes using K-Means, followed by ranking samples by uncertainty and weighting those scores by a cluster weight. In our work, we employ a methodology similar to MeDAL, where we first filter for high uncertainty and then select samples via diversity methods. For 3D point cloud data, we believe that filtering by uncertainty and then evaluating diversity will improve the decision boundary while selecting representative samples. Additionally, this methodology naturally expands on the successes from our earlier research.[6]

3. TECHNICAL APPROACH

3.1 Deep Learning Architecture

We use a modified version of VoxelNet[5] as our object detector. VoxelNet was chosen because of its demonstrated performance on 3D point clouds. Prior to training, we augment and voxelize our point cloud data. Doing these steps offline decreases training time without a major impact on network performance. Voxelization is the process of dividing a point cloud into evenly sized 3D voxels with variable amounts of points in each.

Similar to VoxelNet, our network contains three main sections: a Feature Learning Network, Convolutional Middle Layers, and a final Region Proposal Network.[5,6] The Feature Learning section of our network embeds voxels containing raw points into voxel-wise features using modified voxel feature encoding (VFE) layers originally proposed by Zhou and Tuzel.[5] Our VFE layers consist of Batch Norm, ReLU, and max-pool to create voxel-wise features. VoxelNet proposes to concatenate voxel-wise features with point-wise features after max-pool. However, to decrease the number of required computations and decrease training time, this concatenation step was dropped from our VFE architecture with no impact on model performance. The voxel-wise feature vectors are spaced into a 4D sparse tensor, where the convolutional middle layers convolve the tensors down to a bottleneck. A Region Proposal Network deconvolves a combination of the middle layer features and the bottleneck to generate a probability score map and a regression map.[5]

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Contrary to VoxelNet, we utilize Focal Loss[64] and weighted sampling to improve performance on our class-imbalanced dataset. Additionally, we employ L_2 weight regularization and stochastic weight averaging[65] to prevent overfitting. We utilize Adam[66] as our optimizer because of its fast convergence time and improved performance on our dataset compared to standard stochastic gradient decent. We evaluate our object detector by its mean average precision per class at an Intersection Over Union (IoU) threshold of 0.5 (AP50). A high-level overview of our object detector can be seen in Figure 2. The output of the model is a list of predicted detections defined by a 7-value representation of the bounding box for targets. Each detection also has a vector of classification scores that sum to 1 over the target classes plus a background class.

3.2 Active Learning Methods

We focused on approaches that find a representative selection of samples in the feature space of our 3D object detector. The methods we explored are K-Medoids clustering, Core Set, and Furthest Nearest Neighbors, as well as a hybrid uncertainty-diversity approach in which the diversity methods evaluate only the most uncertain point cloud tiles as determined by the Entropy of their detections.

3.2.1 Feature Space

All the diversity methods we investigated operate on a feature representation of the point cloud tiles. We extract the features from the latent space of our neural network after the Region Proposal Network, but prior to our softmax. At this point in the network, we have 1152 features for each of the 400 voxels per tile. We perform max pooling across features on a per-voxel basis. We then use the resulting features to determine similarity and distance between each tile and to select a diverse subset of tiles to label. This space is still high-dimensional and, thus, Euclidean Distance may not be the best indicator of qualitative similarity in the space.[67] In addition to Euclidean Distance, we evaluate the performance of the diversity methods with Cosine Distance and Correlation Distance, defined as follows. Given two feature vectors, $\mathbf{a}$ and $\mathbf{b}$, and their mean values, μ_a and μ_b,

$$\text{Cosine Distance } = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{4}$$

$$\text{Correlation Distance } = 1 - \frac{(\mathbf{a} - \mu_a) \cdot (\mathbf{b} - \mu_b)}{\| (\mathbf{a} - \mu_a) \| \| (\mathbf{b} - \mu_b) \|} \tag{5}$$

3.2.2 Diversity Approaches

In our first diversity approach, we use K-Medoids to cluster the unlabeled data in the feature space with the various distance metrics. We initialize the medoids with K-Medoids++ and then use the alternating algorithm to optimize: while the cost is decreasing, assign samples to clusters based on the closest medoid, find the medoids of the resulting clusters, and repeat.[68,69] Xu et al.,[48] who introduced K-Medoids as a method of selecting representative samples to label, set the number of clusters to be the number of samples desired for labeling and selected the medoids to label. As we are selecting a large batch of samples to label per active learning loop and we are clustering in a high-dimensional space, we take an alternative approach. We set the number of clusters to be the number of classes in our dataset and select the medoids as well as the points closest to the medoids to label. We evenly distribute the selections among the clusters, redistributing among the more populated clusters when others are not sufficiently populated.

We use the same algorithms for Core Set and FNN as outlined in Sener and Savarese[22] and Geifman and El-Yaniv,[47] respectively. The implementation of Core Set can be reduced to an optimization of the K-Center problem.[70] We start by selecting the tile closest to the center of the unlabeled data in the feature space. We then compute the distance from that "center" to the rest of the unlabeled tiles and select the furthest tile in the feature space as the next "center" to select for labeling. We iterate, selecting each successive "center" to be the sample furthest from its nearest neighbor amongst the already selected "centers". As the language suggests, the implementation of FNN is quite similar. The primary difference is that instead of selecting the first tile as the center of the unlabeled data, we initialize our "centers" as feature representations of the labeled data. Doing so reduces redundancy between the unlabeled tiles we select and the existing labeled data, so we expect better performance. We use all three of these algorithms with each of our distance metrics.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

3.2.3 Hybrid Approaches

Prior knowledge of our dataset indicates that most point cloud tiles contain only background points, with no objects. We consider these tiles to be negative examples and populated tiles to be positive examples. A truly representative subset of this dataset will similarly contain many tiles without objects and a limited number of tiles with labeled objects. Since a representative dataset is the objective of diversity methods, more loops of these methods will be necessary to create a dataset with enough positive examples to train an object detector with sufficient performance. On the other hand, uncertainty methods aim to identify data that the network is unsure about, generally by examining the predictions for each sample. In order to select data that is both representative and informative to the network, we explored hybrid approaches that select a diverse subset of data from among the samples that the network is most uncertain about. In particular, we filter by uncertainty scores based on an objective criteria before applying the diversity methods outlined in Section 3.2.2.

In our prior work,[6] we determined that summing the Entropy of the softmax outputs of the predictions in each tile produced a labeled dataset that contained many objects, resulting in the greatest performance gains for our object detector. Thus, we chose Entropy as the uncertainty method to use in our hybrid active learning approach. A hyperparameter search indicated that to attain an optimal amount of filtering for our dataset, the Entropy filtering should reduce the number of samples provided to the diversity method to twice the number of samples intended to be selected for labeling. The complete method is formalized in Algorithm 1.

Algorithm 1: Uncertainty-Diversity Hybrid Active Learning Approach

stopping criteria $\leftarrow$ performance threshold, number of loops, annotation cost, etc.
$\mathcal{H} \leftarrow$ Entropy of softmax vector, as defined in Equation 1
$D_d \leftarrow$ diversity method with distance metric d
$n \leftarrow$ selection batch size
$f \leftarrow$ filtering scale parameter
$L \leftarrow \{l_0, l_1, ..., l_n\}$; `// randomly selected initial labeled data`
$U \leftarrow \{u_0, u_1, ...\}$; `// remaining unlabeled data`
while *not stopping criteria* **do**
  $Q_L \leftarrow$ network trained on L ;
  $S_0 \leftarrow \{\}$;
  while $|S_0| < f * n$ **do**
    $S_0 \leftarrow S_0 \cup \{\underset{u_i \in U \setminus S_0}{\text{argmax}}\ \mathcal{H}(Q_L(u_i)|u_i, L)\}$;
  end
  $S \leftarrow D_d(S_0)$; `// selected subset of size` n `from` S_0
  human oracle labels S ;
  $L \leftarrow L \cup S$;
  $U \leftarrow U \setminus S$;
end
return *Labeled dataset, L*

4. EXPERIMENTS

4.1 Dataset

We use the same dataset to evaluate our diversity active learning methods as we used in our prior work exploring uncertainty methods.[6] The dataset comprises 41 LiDAR point clouds with approximately 7900 objects each belonging to one of seven different classes. To appropriately assess model performance, the dataset is split into train, validation, and test sets containing 25, 9, and 7 point clouds, respectively. We preprocess the data by dividing the point clouds into overlapping tiles as input for the object detector. Class distributions within the dataset are shown in Figure 3 and Table 1.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

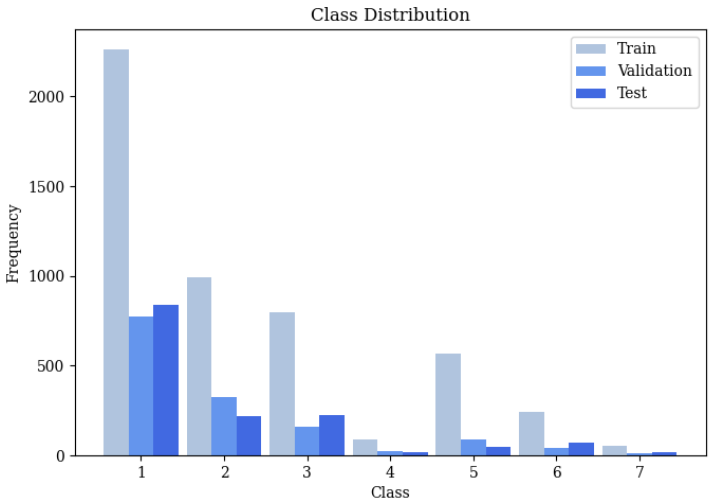

Figure 3. Histogram of object classes

Table 1. Dataset class counts

Class	Counts
Class 1	3874
Class 2	1534
Class 3	1181
Class 4	125
Class 5	706
Class 6	359
Class 7	84

4.2 Experimental Setup

We randomly select 550 point cloud tiles from the data in the train split (7.2% of the training data) to label initially. Through experimentation, we found this was the minimum amount of training data to learn an initial object detector that is sufficiently performant to produce nontrivial insight during sample selection. We use the same initial labeled tiles as our starting train data for every trial of active learning. In each trial, we randomly disregard 90% of the background-only point cloud tiles to discourage the object detector from learning false negatives. We complete six loops of active learning for each trial. In each active learning loop, we select an additional 550 point cloud tiles from the remaining unlabeled train data to annotate. After this data is annotated, the object detector is fully retrained on the new and improved train set.

We ran this experiment for each of the following diversity-based methods of identifying data to label: K-Medoids, Core Set, and Furthest Nearest Neighbors (FNN). For each method, we ran five trials with each of the following definitions of distance between point cloud tile features: Euclidean, Cosine, and Correlation. We then ran a series of experiments using hybrid uncertainty-diversity methods using the most effective distance metric across all three diversity methods. For these experiments, we follow the selection process outlined in Algorithm 1, first selecting the 1100 most uncertain unlabeled point cloud tiles by their Entropy scores, then selecting the final 550 tiles to label using a diversity method. We refer to these hybrid methods as Entropy/K-Medoids, Entropy/Core Set, and Entropy/FNN. For comparison, we ran active learning with random tile selection as well as uncertainty-based trials considering only the Entropy of the point cloud tiles. Additionally, we determined a baseline for the object detector by training the model for five trials on 100% of the train dataset.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

4.3 Results

We evaluate the data selection methods in terms of model performance, annotation cost and efficiency, and diversity and representativeness of the datasets. These criteria correspond to the desired outcomes of active learning, regardless of the model task. After acquiring an unlabeled dataset, we want to spend as little time and resources as possible annotating the data. However, we desire a labeled dataset that contains enough examples of positive and negative samples to train a model that performs well on unseen data, e.g., the validation set. Additionally, we want to avoid leaving valuable information unutilized in the remaining unlabeled data.

4.3.1 Object Detector Performance

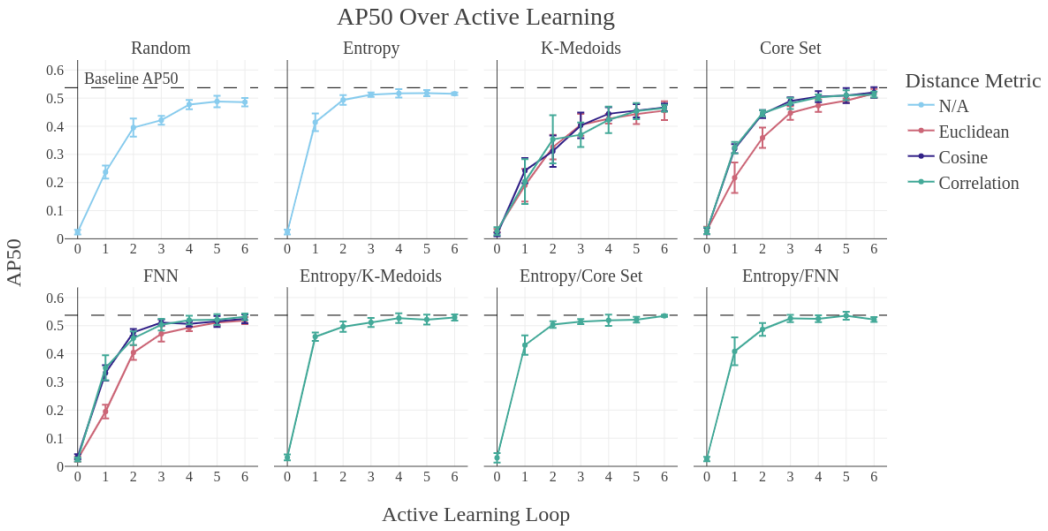

Figure 4. Average object detector performance during active learning with standard deviation error bars for each data selection approach. Final models were trained on 3850 tiles (approximately 50.4% of the total train set). The distance measure used in the feature space is indicated by line color, and the AP50 of the model baseline, 0.537, is denoted by a dotted black line.

We measure the object detector performance by the mean average precision at an IoU threshold of 0.50 (AP50) of the object detectors trained on the labeled datasets at each loop of active learning. AP50 is shown over the course of active learning for each selection method in Figure 4. We also compare the conditions on the final object detectors' improvement from the initial object detector, or ΔAP50. The ΔAP50 accounts for variability in the models trained on the consistent initial train dataset. This variability is independent of the active learning data selection method. In addition to the performance of the terminal object detector, we can further assess the active learning methods' efficacy at selecting a dataset that trains well-performing models by observing how quickly the performance improves over active learning. Less data annotation is necessary when fewer loops of active learning are required to produce a sufficient object detector. To quantify this characteristic of the active learning methods, we calculate the area under the curve (AUC) of the AP50 over active learning graphs. This value as well as the AP50 and ΔAP50 of the final object detector for each method are shown in Table 2.

The baseline object detector trained on the complete train dataset reached a final AP50 of 0.537, which we consider an approximate upper bound for the active learning trials. All active learning methods outperformed the random control by the final loop except for K-Medoids. With their most successful distance metrics, both other diversity methods, Core Set and FNN, selected datasets resulting in better performing final object detectors than Entropy. As expected, FNN outperformed the other diversity methods. Euclidean distance was the least effective distance measure for all three diversity methods. Correlation distance and Cosine distance were not significantly different in terms of object detection performance. For the hybrid methods, we used Correlation distance in the feature space for the diversity stage.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 2. Object detection performance for each active learning method.

Method	**Distance Metric**	**Final AP50**	**Final ΔAP50**	**AUC AP50**
Random	N/A	0.486	0.462	2.249
Entropy	N/A	0.516	0.491	2.712
K-Medoids	Euclidean	0.455	0.426	2.013
	Cosine	0.467	0.452	2.080
	Correlation	0.464	0.437	2.012
Core Set	Euclidean	0.516	0.487	2.258
	Cosine	0.520	0.493	2.532
	Correlation	0.512	0.484	2.519
FNN	Euclidean	0.518	0.493	2.339
	Cosine	0.524	0.490	2.596
	Correlation	0.531	**0.507**	2.610
Entropy/K-Medoids	Correlation	0.529	0.498	**2.777**
Entropy/Core Set	Correlation	**0.535**	0.505	2.755
Entropy/FNN	Correlation	0.522	0.496	2.732

Table 3. The amount of annotation required and its expected cost as defined in Equation 6 with each active learning method. All diversity methods are shown using Correlation distance.

Method	**# Populated Tiles Selected**	**# Objects in Selected Tiles**	**Annotation Cost**
Random	807.6	2250.0	552.5
Entropy	1518.6	4434.6	639.9
K-Medoids	736.6	1675.8	**529.6**
Core Set	1264.2	3869.8	617.3
FNN	1341.4	3969.0	621.3
Entropy/K-Medoids	1469.2	4157.0	628.8
Entropy/Core Set	**1557.2**	**4466.2**	641.2
Entropy/FNN	1528.2	4382.2	637.8

K-Medoids and Core Set produced notably better performing final object detectors when employed in our hybrid methods, selecting from only the subset of unlabeled tiles with the highest Entropy, than when used in diversity-only approaches, with increases in AP50 of 0.065 and 0.023, respectively. Ultimately, the best object detector trained on the final dataset came from Entropy/Core Set with a final AP50 of 0.535, within one standard deviation of the baseline object detector performance. FNN alone produced the next best object detector and demonstrated the most improvement in object detection performance, as measured by ΔAP50.

Though the models trained on the final datasets selected by Core Set with Cosine distance and FNN with all distance metrics perform better than that of Entropy, Entropy produces more successful object detectors in the first few loops than the diversity methods do at that stage. The hybrid methods display a behavior similar to Entropy in this regard, with the early additions to the datasets providing substantial performance gains for the object detectors, as shown in Figure 4 and indicated by the AUC AP50 in Table 2. For the subsequent analyses, we restricted the diversity methods to only using Correlation distance.

4.3.2 Annotation Cost and Efficiency

To calculate cost of annotation, we adopt the formal definition given by Liang et al.[7] They define the annotation cost of an active learning method as

$$\text{Cost} = c_t * n_t + c_b * n_b \tag{6}$$

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

where c_t and c_b are the costs per tile selected and per bounding box labeled in the selected tiles, respectively, and n_t and n_b are the numbers of tiles and boxes. Based on subject matter expert input, they set the cost coefficients to be $c_t = 0.12$ and $c_b = 0.04$. We calculate the cost of labeling the data selected by each method in Table 3.

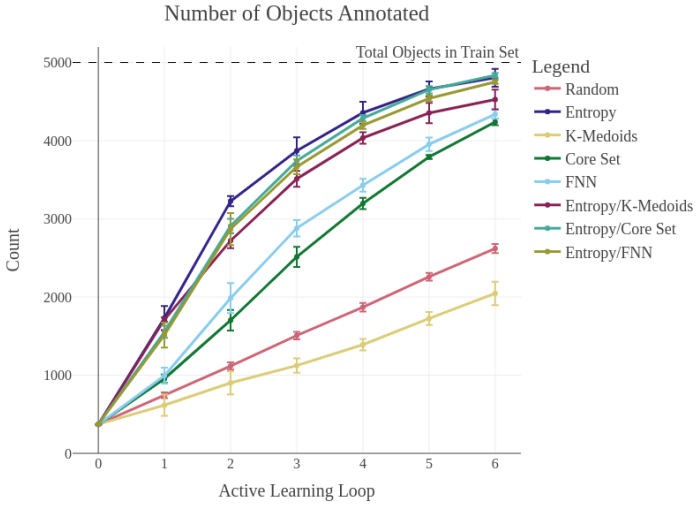

Figure 5. Average number of objects in the labeled train set throughout active learning with standard deviation error bars for each data selection method. All diversity methods are shown using Correlation distance.

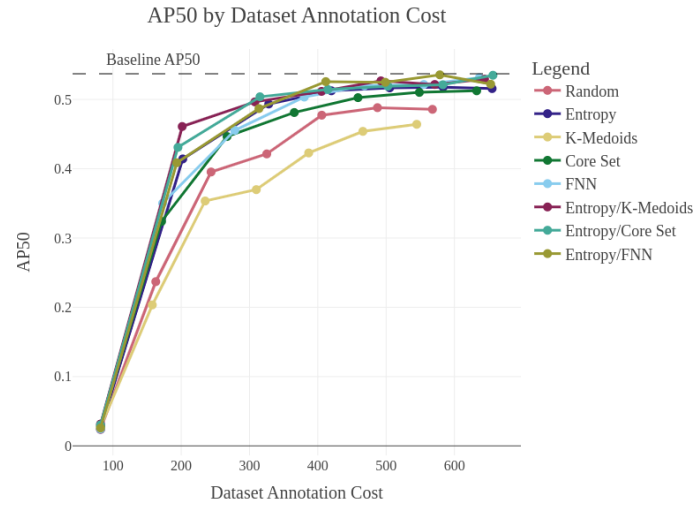

Figure 6. Average performance (AP50) of the object detectors as a function of the average annotation costs for the datasets they were trained on for each data selection method. All diversity methods are shown using Correlation distance.

As we observed in our previous work,[6] Entropy selects tiles with a high density of objects. As such, it has a high total cost of annotation for the same or worse final object detection performance as Core Set and FNN. Random selection and K-Medoids, on the other hand, select subsets of tiles that mirror the large proportion of background-only tiles in the unlabeled data. They select the least number of populated tiles and the least number of total objects in those tiles. The result is lower overall annotation costs, but poor performance. In our hybrid methods, the datasets produced contain more populated tiles and more labeled objects than the respective diversity-only methods. In particular, the Entropy/Core Set hybrid method identified the most tiles containing

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

objects and the final dataset contained the most labeled objects of the methods. Furthermore, despite only annotating 50.4% of the train data, Entropy alone, Entropy/Core Set, and Entropy/FNN all produced subsets containing over 95% of the objects in the train set. The number of objects in the datasets produced by each method over the course of active learning is shown in Figure 5.

Figure 6 shows the performance of the object detectors as a function of the annotation costs for the datasets they were trained on. This plot allows us to consider the annotation cost associated with each active learning method that is required to achieve the same object detection performance and, conversely, the performance one can expect from each method given a fixed annotation budget. We see that the higher cost associated with selecting object-dense tiles is minor relative to the overall costs. Further, all active learning methods other than K-Medoids require less than half of the annotation cost as random selection to achieve the maximum AP50 reached by the random control. At any fixed cost, the object detectors trained on datasets produced by these methods outperform those trained on randomly selected data by as much as 0.15 AP50. FNN and the hybrid methods are the most efficient in this regard.

4.3.3 Class Distribution of Labeled Dataset

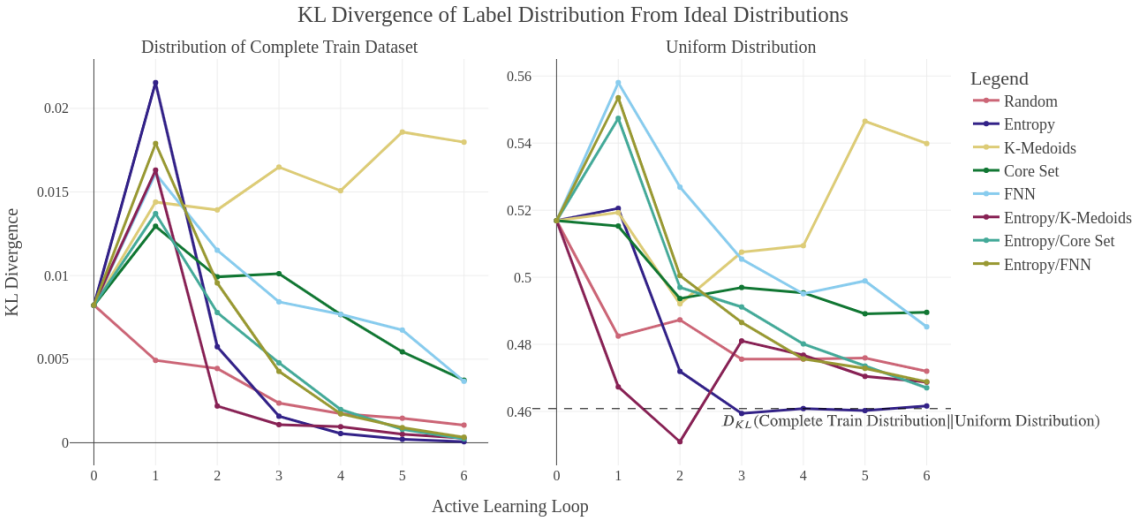

Figure 7. KL Divergence of the class distribution in the labeled dataset throughout active learning as compared to each of two "ideal" distributions: the class distribution of the complete train set and a uniform distribution.

In using diversity or hybrid uncertainty-diversity methods as opposed to uncertainty methods, we aim to reduce redundancy in the selected data and identify a more representative subset of the train data. In other words, we want to increase both the diversity and the representativeness of the selected data. To evaluate how the various active learning methods perform with respect to these goals, we compare the class distribution in the labeled dataset at each loop of active learning against two "ideal" distributions: the empirical class distribution of the complete train data and a uniform distribution of class populations. We assess the label distributions via the Kullback-Leibler (KL) Divergence, which is defined in Equation 7 where P is a discrete probability distribution derived from a selected subset and Q is the discrete ideal distribution.

$$D_{KL}(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{7}$$

The KL Divergence from each of the ideal distributions is shown for the datasets selected with each active learning method in Figure 7 and the area under each curve is logged in Table 4. The respective magnitudes of the KL Divergence for the two ideal distributions indicate that the class distributions of the selected subsets are much more similar to the distribution of the complete train dataset than a uniform distribution. In both

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 4. Area under the curve (AUC) of the KL Divergence of the class distribution in the labeled datasets throughout active learning as compared to each of the class distribution of the complete train set and a uniform distribution.

Method	**AUC KL Divergence**	
	Train	**Uniform**
Random	0.011	2.822
Entropy	0.012	2.788
K-Medoids	0.083	3.019
Core Set	0.042	2.945
FNN	0.043	3.005
Entropy/K-Medoids	**0.009**	**2.749**
Entropy/Core Set	0.020	2.893
Entropy/FNN	0.021	2.880

cases, the final labeled datasets selected by diversity-only methods are furthest from the ideal distributions, followed by random selection. The methods involving Entropy produce final datasets that are closest to the ideal distributions. The intermediate datasets are more mixed. The distances from the distributions of datasets produced to the distribution of the whole train dataset are largely monotonically decreasing after the first loop, with the exception of K-Medoids whose dataset further diverges from the class distribution of the complete train set.

We expect the datasets to approach the distribution of the whole train dataset, and thus the KL Divergence to approach 0, as they add populated tiles throughout active learning. On the other hand, the datasets' similarity to the uniform distribution is limited by the class imbalance in the dataset. As more objects are labeled, the KL Divergence of the labeled dataset from the uniform distribution will converge to the KL Divergence of the complete train dataset with the uniform distribution, 0.461. Since the methods utilizing Entropy label more objects than diversity-only methods and random selection, we see that their final datasets are closer to this value. By the third loop, the Entropy-only method has stabilized around this value. The hybrid Entropy/K-Medoids method is the only other method whose selected datasets are ever closer to the uniform distribution than the whole train distribution. This is reflected in the AUC of the KL Divergence over active learning in Table 4. Based on these values, Entropy/K-Medoids, Entropy alone, and random selection remained the closest to the ideal distributions over the course of active learning.

5. DISCUSSION

We have shown the potential of leveraging active learning for 3D point cloud annotation. To the authors' knowledge, our work is the first to deploy diversity-based active learning methods on a dataset comprised solely of 3D point clouds. With 3D data comes domain-specific obstacles, leading to a challenging dataset for training an object detector. Further, our dataset is especially challenging due to the high number of scenes without target objects and heavy class imbalance. Regardless, we show definitive results that diversity-based active learning methods are an effective method for data selection, especially when used in an uncertainty-diversity hybrid approach.

In the object detection task that is the focus of our work, the number of objects in each tile may be zero, one, or many. Uncertainty methods operate on the detections output by the model and are, therefore, dependent on the number of objects detected in each tile. Therefore, a decision is required for how to score samples without any predictions. As in our previous work,[6] we score these samples as least uncertain, creating a bias towards selecting positive examples. By summing the object-wise scores into a per-sample score, we further encourage selection of high occupancy tiles. The results shown in Section 4.3.2 demonstrate the bias in uncertainty methods, with Entropy and the hybrid methods that filter with Entropy selecting the most densely populated tiles throughout the duration of active learning. In fact, by the sixth loop, these methods selected tiles containing over 90% of the objects in the train dataset, despite the datasets only containing 50.4% of the tiles in the train dataset. This demonstrates the efficacy of model predictions as a proxy for finding samples containing objects, even when the detections do not meet the stricter criteria used for calculating performance metrics. These methods

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

are successful in not leaving valuable information unlabeled and unused. However, the uncertainty and hybrid methods' dependence on the contents of each tile adds a layer of complexity, as the user must decide how to aggregate uncertainty across multiple objects in a tile and how to treat unpopulated tiles. It poses a further logistical challenge for comparing predictions across forward passes in stochastic uncertainty methods, as we explored in our previous work.[6]

Object detector performance is a key indicator of active learning's success. Entropy outperforms the diversity methods when considering AUC AP50, indicating higher object detector performance in the earlier iterations of active learning as we see in Figure 4. Thus, it can be concluded that the increased number of objects in the train data selected by Entropy has a large positive impact on performance early on in training and saturates over time. The diversity methods do not achieve the number of objects selected as early on as Entropy. However, the performance of FNN's final object detector is superior to Entropy and that of Core Set is approximately equivalent to Entropy as measured by AP50 and ΔAP50, showing that they are selecting more information-rich tiles that have a greater impact on training. Figure 6 indicates that FNN selects particularly valuable, low-cost tiles, achieving better performance than Entropy for annotation costs over 400. FNN is a task agnostic active learning algorithm superior to Entropy in performance and efficiency.

Although FNN alone had the highest overall ΔAP50, the hybrid methods consistently outperform FNN in terms of AUC AP50. Combining the object selection tendencies of Entropy with diversity-seeking techniques in our hybrid methods resulted in better object detection performance than Entropy alone across all measures of performance in Table 2. The hybrid methods also had lower costs of annotation for the same object detection performance than Entropy, as shown in Figure 6. By using Entropy and then a diversity method, the hybrid active learning methods focus on exploring the uncertain scenes, leading to robustness and consistent annotation efficiency that the other methods fail to offer. In particular, the Entropy/Core Set hybrid method achieved the highest final AP50 across all methods: 0.535. The AP50 of our baseline object detector trained on the full train dataset peaked at 0.537. Therefore, the model trained on the 50.4% of the train dataset that the Entropy/Core Set hybrid active learning method selected came within 1% of the performance of the object detector trained on the full dataset.

When the data is projected into the feature space using our VoxelNet model, our diversity and hybrid approaches successfully traverse the hyperplane and select scenes that are representative of the remaining data and are informative to our object detector. For our trials, the Correlation and Cosine distance measures significantly improve performance of our diversity methods, as seen in Table 2. In a high-dimensional space, the Euclidean distance between features becomes inconsequential due to the sparsity of the hyperplane.[67] Clearly, diversity method performance is directly impacted by the topology of the latent space and the appropriateness of the distance metric utilized within it.

Of the diversity methods, K-Medoids selected the fewest populated tiles and objects—even fewer than random selection. One likely explanation for this behavior is that due to the high number of unpopulated tiles in the train dataset, at least one of the seven clusters identified by K-Medoids is consistently composed primarily of unpopulated tiles. K-Medoids selects the seven cluster medoids and the tiles nearest to them, in contrast to Core Set and FNN whose selected points are each representative of distinct areas in the feature space. When a high-density region of background-only tiles exists, Core Set and FNN will select one representative, while K-Medoids will choose one seventh of the total selected tiles to be from that region. Thus, the dataset selected by K-Medoids has fewer labeled objects and less diversity among both positive and negative examples. Moreover, since we only train our object detector on a random 10% of the unpopulated tiles in the train dataset, the models trained on the dataset produced by K-Medoids are not only training on fewer positive examples of the object classes, but also fewer tiles overall. While this results in poorer object detector performance, it also lowers annotation costs and requires less computational resources to train the model. Depending on the trade-offs important to the use case, K-Medoids may thus still be a valuable option when selecting an active learning approach.

Entropy/K-Medoids selects the least populated tiles and the fewest total number of objects of the hybrid methods, resulting in the lowest total annotation cost. Nonetheless, Entropy/K-Medoids outperforms the other hybrid methods in terms of AUC of the AP50 curve. As shown in Figure 5, Entropy/K-Medoids selects just as many objects as the other Entropy-based methods in the first loop and is within one standard deviation from selecting as many in the second loop. In these earlier loops, the first stage of the hybrid method—filtering

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

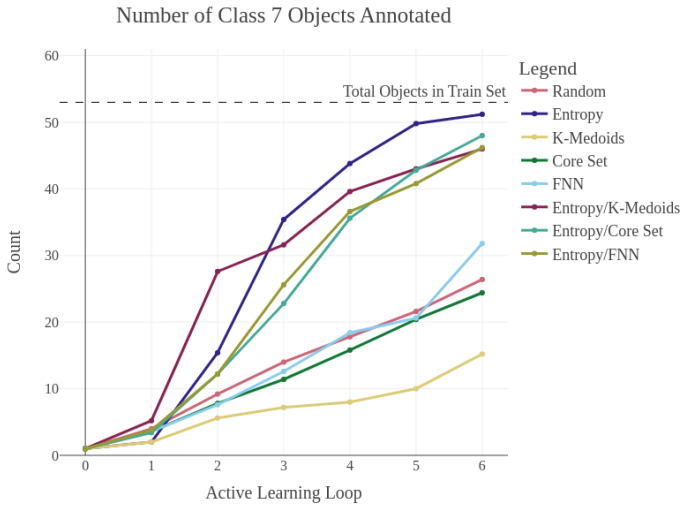

Figure 8. Average number of objects from Class 7 in the labeled train set throughout active learning for each data selection method. All diversity methods are shown using Correlation distance.

with Entropy—reduces the search space for the diversity methods to the subset of the unlabeled train data that contains the most high-uncertainty object detections. This eliminates the problem that K-Medoids faced when utilized on its own: picking one or more clusters containing primarily unpopulated tiles. Instead, as a stage of the hybrid method, K-Medoids behaves as we intended it to when we set the number of clusters to be the number of classes in the dataset; it selects as even a distribution as possible of objects from the seven classes. This is evidenced by Figure 8 which shows the number of objects from Class 7, our least represented class in the total dataset, over active learning. Figure 6 indicates that the diversity in these early datasets produced by Entropy/K-Medoids leads to better performance at low annotation costs. By the later loops, all Entropy-related methods have already selected most of the objects in the train data, so the remaining tiles that score high in Entropy are a mix of empty and populated tiles. Entropy/K-Medoids likely has some clusters of each, and its superior ability to distinguish tiles containing underrepresented object classes from a group of populated tiles provides a lesser advantage.

We see further evidence of even class sampling by Entropy/K-Medoids in the early loops through the low KL Divergence between the early Entropy/K-Medoids datasets and both ideal distributions. As mentioned in Section 4.3.3, the KL Divergence from the uniform distribution to the class distribution of any subset of the train dataset will converge to the KL Divergence from the uniform distribution to the class distribution of the full train dataset. However, in the intermediate datasets before all the objects are labeled, it is possible for a method to create a dataset that is more uniform than the complete train set by selecting a greater percentage of the objects in underrepresented classes and a lower percentage of objects in overrepresented classes. This is precisely what the hybrid Entropy/K-Medoids method does to achieve its low KL Divergence with the uniform distribution. Labeling objects from underrepresented classes also reduces the KL Divergence from the complete train distribution. As such, this measure is an evaluation of representativeness that puts value on oversampling the smaller classes. By selecting more objects overall, Entropy quickly creates a diverse and representative dataset. The diversity-only methods are less successful in this regard due to the imbalances in our dataset. However, by limiting the search space to high density tiles, the hybrid methods, particularly Entropy/K-Medoids, select more diverse and representative datasets than random selection.

Across the board, the hybrid methods proved to be the most successful active learning methods for object detection on our 3D point cloud dataset. These methods offer strong object selection at low cost as well as sufficient class distribution for a well-balanced 3D point cloud dataset consisting of both populated and empty point cloud scenes. Networks trained on the final and intermediate datasets produced by hybrid methods generally outperform those of the random control, uncertainty-only methods, and diversity-only methods.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

6. CONCLUSIONS

We show empirically that our diversity-based active learning framework selects an informative dataset that can be used to train an exceptionally performing object detector. Further, we demonstrate that our methods select data that is representative of the training dataset and distributed among object classes, both key indicators of an informative subset of data. We are able to train a high-achieving VoxelNet model with only 50.4% of our available dataset by selecting only the most information-rich data samples. We maintain quality results while reducing the quantity of labeled data required. By narrowing down an unlabeled dataset to a focused subset, the effort and time required by data analysts to label data is cut in half. This is especially relevant for 3D point cloud data, where annotating objects is particularly tedious and time consuming.

Diversity-based active learning methods are an effective means to select the minimum amount of data needed to train an accurate object detector. For our dataset, the top diversity-only method in terms of both performance and cost efficiency is FNN. By considering AUC AP50 and performance at fixed costs, we found significant evidence that uncertainty-diversity hybrid approaches produce the most informative datasets throughout the entirety of active learning. The Entropy/Core Set hybrid method resulted in object detection performance comparable to our baselines trained on the complete train set. Additionally, the hybrid methods show strong object selection capabilities, selecting nearly all of the objects in the train dataset by the sixth iteration of active learning. For 3D point cloud data, particularly with imbalanced classes, it is evident that FNN and uncertainty-then-diversity successive methods are more effective than strictly uncertainty methods.

Our work shows the advantages of leveraging active learning techniques to annotate the minimum number of data samples required to train a robust 3D object detector. Uncertainty methods provide a method of data selection focusing around the decision boundary. While this methodology is beneficial in theory, implementations must be adapted to each specific task model and samples selected tend to be highly similar. Our results show that Entropy on 3D point cloud data suffers from this problem. Entropy selects the greatest number of objects during training, but lags in AP50, implying redundancy in the selected data. In this work, we explored diversity methods to combat this inherent drawback. Diversity methods show promise and are task agnostic, however they are limited by the vast number of empty scenes in our dataset. The successive hybrid approaches are the optimal methods for our dataset, achieving impressive final AP50 and AUC AP50 values with low annotation costs by selecting samples which are both representative of uncertain scenes and inter-class diversity.

In future work, further analysis of the feature space would be beneficial. This may include alternative feature extraction architectures or dimensionality reduction of the embedded feature space using methods such as PCA, tSNE, or UMAP to reduce redundancy and noise in the data. Our results indicate that assessing unlabeled data in the feature space provides valuable information to the active learning selection methods, and we expect that a lower dimensional feature space will be more conducive to comparison between samples. With the success of our current hybrid approaches, a broader focus should also be made to explore more complex methods of hybrid selection, such as MeDAL and ReDAL. In addition, methods of selecting the appropriate active learning algorithm for each unique task and dataset should be explored.

ACKNOWLEDGMENTS

This work was supported by the US Air Force Research Laboratory, and the National Geospatial-Intelligence Agency.

REFERENCES

[1] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)* **115**(3), 211–252 (2015).

[2] Kumar, N., Kaur, N., and Gupta, D., "Major convolutional neural networks in image classification: a survey," in [*Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India*], 243–258, Springer (2020).

[3] Hu, J., Shen, L., and Sun, G., "Squeeze-and-excitation networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 7132–7141 (2018).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

[4] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O., "nuscenes: A multimodal dataset for autonomous driving," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 11621–11631 (2020).

[5] Zhou, Y. and Tuzel, O., "Voxelnet: End-to-end learning for point cloud based 3d object detection," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 4490–4499 (2018).

[6] Moses, A., Jakkampudi, S., Danner, C., and Biega, D., "Localization-based active learning (LOCAL) for object detection in 3D point clouds," in [*Geospatial Informatics XII*], Palaniappan, K., Seetharaman, G., and Harguess, J. D., eds., **12099**, 1209907, International Society for Optics and Photonics, SPIE (2022).

[7] Liang, Z., Xu, X., Deng, S., Cai, L., Jiang, T., and Jia, K., "Exploring diversity-based active learning for 3d object detection in autonomous driving," *arXiv preprint arXiv:2205.07708* (2022).

[8] Settles, B., "Active learning literature survey," (2009).

[9] Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A., "Deep batch active learning by diverse, uncertain gradient lower bounds," *arXiv preprint arXiv:1906.03671* (2019).

[10] Gao, F., Yue, Z., Wang, J., Sun, J., Yang, E., and Zhou, H., "A Novel Active Semisupervised Convolutional Neural Network Algorithm for SAR Image Recognition," *Computational Intelligence and Neuroscience* **2017**, e3105053 (Oct. 2017).

[11] Rottmann, M., Kahl, K., and Gottschalk, H., "Deep Bayesian Active Semi-Supervised Learning," (Mar. 2018). arXiv:1803.01216 [cs, stat].

[12] Rai, P., Saha, A., Daumé III, H., and Venkatasubramanian, S., "Domain adaptation meets active learning," in [*Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*], 27–32 (2010).

[13] Prabhu, V., Chandrasekaran, A., Saenko, K., and Hoffman, J., "Active domain adaptation via clustering uncertainty-weighted embeddings," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 8505–8514 (2021).

[14] Su, J.-C., Tsai, Y.-H., Sohn, K., Liu, B., Maji, S., and Chandraker, M., "Active adversarial domain adaptation," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 739–748 (2020).

[15] Meyer, B. J. and Drummond, T., "The importance of metric learning for robotic vision: Open set recognition and active learning," in [*2019 International Conference on Robotics and Automation (ICRA)*], 2924–2931, IEEE (2019).

[16] Yoo, D. and Kweon, I. S., "Learning loss for active learning," (2019).

[17] Sinha, S., Ebrahimi, S., and Darrell, T., "Variational adversarial active learning," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 5972–5981 (2019).

[18] Tang, M., Luo, X., and Roukos, S., "Active learning for statistical natural language parsing," in [*Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*], 120–127 (2002).

[19] Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D. Z., "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in [*International conference on medical image computing and computer-assisted intervention*], 399–407, Springer (2017).

[20] Schmidt, S., Rao, Q., Tatsch, J., and Knoll, A., "Advanced active learning strategies for object detection," in [*2020 IEEE Intelligent Vehicles Symposium (IV)*], 871–876, IEEE (2020).

[21] Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X., "A survey of deep active learning," *ACM computing surveys (CSUR)* **54**(9), 1–40 (2021).

[22] Sener, O. and Savarese, S., "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489* (2017).

[23] Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M., "The power of ensembles for active learning in image classification," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 9368–9377 (2018).

[24] Lewis, D. D. and Gale, W. A., "A sequential algorithm for training text classifiers," in [*SIGIR'94*], 3–12, Springer (1994).

[25] Wang, K., Zhang, D., Li, Y., Zhang, R., and Lin, L., "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology* **27**(12), 2591–2600 (2016).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

[26] Kao, C.-C., Lee, T.-Y., Sen, P., and Liu, M.-Y., "Localization-aware active learning for object detection," in [*Asian Conference on Computer Vision*], 506–522, Springer (2018).
[27] Gal, Y., Islam, R., and Ghahramani, Z., "Deep bayesian active learning with image data," in [*International Conference on Machine Learning*], 1183–1192, PMLR (2017).
[28] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., "On calibration of modern neural networks," in [*International Conference on Machine Learning*], 1321–1330, PMLR (2017).
[29] Gal, Y. and Ghahramani, Z., "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in [*international conference on machine learning*], 1050–1059, PMLR (2016).
[30] Ayhan, M. S. and Berens, P., "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," (2018).
[31] Pop, R. and Fulop, P., "Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles," *arXiv preprint arXiv:1811.03897* (2018).
[32] Atighehchian, P., Branchaud-Charron, F., and Lacoste, A., "Bayesian active learning for production, a systematic study and a reusable library," *arXiv preprint arXiv:2006.09916* (2020).
[33] Seung, H. S., Opper, M., and Sompolinsky, H., "Query by committee," in [*Proceedings of the fifth annual workshop on Computational learning theory*], 287–294 (1992).
[34] Melville, P. and Mooney, R. J., "Diverse ensembles for active learning," in [*Proceedings of the twenty-first international conference on Machine learning*], 74 (2004).
[35] Ducoffe, M. and Precioso, F., "Active learning strategy for cnn combining batchwise dropout and query-by-committee.," in [*ESANN*], (2017).
[36] Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M., "Bayesian active learning for classification and preference learning," *arXiv preprint arXiv:1112.5745* (2011).
[37] Brust, C.-A., Käding, C., and Denzler, J., "Active learning for deep object detection," *arXiv preprint arXiv:1809.09875* (2018).
[38] Desai, S. V., Chandra, A. L., Guo, W., Ninomiya, S., and Balasubramanian, V. N., "An adaptive supervision framework for active learning in object detection," *arXiv preprint arXiv:1908.02454* (2019).
[39] Roy, S., Unmesh, A., and Namboodiri, V. P., "Deep active learning for object detection.," in [*BMVC*], 91 (2018).
[40] Haussmann, E., Fenzi, M., Chitta, K., Ivanecky, J., Xu, H., Roy, D., Mittel, A., Koumchatzky, N., Farabet, C., and Alvarez, J. M., "Scalable active learning for object detection," in [*2020 IEEE intelligent vehicles symposium (iv)*], 1430–1435, IEEE (2020).
[41] Aghdam, H. H., Gonzalez-Garcia, A., Weijer, J. v. d., and López, A. M., "Active learning for deep detection neural networks," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 3672–3680 (2019).
[42] Feng, D., Rosenbaum, L., and Dietmayer, K., "Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection," in [*2018 21st International Conference on Intelligent Transportation Systems (ITSC)*], 3266–3273, IEEE (2018).
[43] Miller, D., Nicholson, L., Dayoub, F., and Sünderhauf, N., "Dropout sampling for robust object detection in open-set conditions," in [*2018 IEEE International Conference on Robotics and Automation (ICRA)*], 3243–3249, IEEE (2018).
[44] Miller, D., Dayoub, F., Milford, M., and Sünderhauf, N., "Evaluating merging strategies for sampling-based uncertainty techniques in object detection," in [*2019 International Conference on Robotics and Automation (ICRA)*], 2348–2354, IEEE (2019).
[45] Morrison, D., Milan, A., and Antonakos, E., "Uncertainty-aware instance segmentation using dropout sampling," in [*Proceedings of the Robotic Vision Probabilistic Object Detection Challenge (CVPR 2019 Workshop), Long Beach, CA, USA*], 16–20 (2019).
[46] Blok, P. M., Kootstra, G., Elghor, H. E., Diallo, B., van Evert, F. K., and van Henten, E. J., "Active learning with maskal reduces annotation effort for training mask r-cnn," *arXiv preprint arXiv:2112.06586* (2021).
[47] Geifman, Y. and El-Yaniv, R., "Deep active learning over the long tail," *arXiv preprint arXiv:1711.00941* (2017).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

[48] Xu, Z., Yu, K., Tresp, V., Xu, X., and Wang, J., "Representative sampling for text classification using support vector machines," *Advances in Information Retrieval: 25th European Conf on IR Research ECIR 2003: 2003; Italy* **2633**, 11–11 (04 2003).

[49] Wang, T., Li, X., Yang, P., Hu, G., Zeng, X., Huang, S., Xu, C.-Z., and Xu, M., "Boosting active learning via improving test performance," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **36**(8), 8566–8574 (2022).

[50] Bishop, C. M. and Nasrabadi, N. M., [*Pattern recognition and machine learning*], vol. 4, Springer (2006).

[51] Krizhevsky, A., Hinton, G., et al., "Learning multiple layers of features from tiny images," (2009).

[52] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y., "Reading digits in natural images with unsupervised feature learning," (2011).

[53] Gissin, D. and Shalev-Shwartz, S., "Discriminative active learning," *arXiv preprint arXiv:1907.06347* (2019).

[54] Guo, Y. and Schuurmans, D., "Discriminative batch mode active learning," *Advances in neural information processing systems* **20** (2007).

[55] Sun, C., Sun, H., and Liu, X., "Robust adversarial active learning with a novel diversity constraint," in [*2020 IEEE International Conference on Big Data (Big Data)*], 226–231, IEEE (2020).

[56] Shui, C., Zhou, F., Gagné, C., and Wang, B., "Deep active learning: Unified and principled method for query and training," in [*International Conference on Artificial Intelligence and Statistics*], 1308–1318, PMLR (2020).

[57] Li, X. and Guo, Y., "Active learning with multi-label svm classification.," in [*IjCAI*], **13**, 1479–1485, Citeseer (2013).

[58] Hekimoglu, A., Schmidt, M., Marcos-Ramiro, A., and Rigoll, G., "Efficient active learning strategies for monocular 3d object detection," in [*2022 IEEE Intelligent Vehicles Symposium (IV)*], 295–302, IEEE (2022).

[59] Rodríguez, A. C., D'Aronco, S., Schindler, K., and Wegner, J. D., "Mapping oil palm density at country scale: An active learning approach," *Remote Sensing of Environment* **261**, 112479 (2021).

[60] Kirsch, A., Van Amersfoort, J., and Gal, Y., "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," *Advances in neural information processing systems* **32** (2019).

[61] Zhdanov, F., "Diverse mini-batch active learning," *arXiv preprint arXiv:1901.05954* (2019).

[62] Smailagic, A., Costa, P., Noh, H. Y., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P., et al., "Medal: Accurate and robust deep active learning for medical image analysis," in [*2018 17th IEEE international conference on machine learning and applications (ICMLA)*], 481–488, IEEE (2018).

[63] Wu, T.-H., Liu, Y.-C., Huang, Y.-K., Lee, H.-Y., Su, H.-T., Huang, P.-C., and Hsu, W. H., "Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 15510–15519 (2021).

[64] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*Proceedings of the IEEE international conference on computer vision*], 2980–2988 (2017).

[65] Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G., "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407* (2018).

[66] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[67] Aggarwal, C. C., Hinneburg, A., and Keim, D. A., "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in [*Database Theory — ICDT 2001*], Van den Bussche, J. and Vianu, V., eds., *Lecture Notes in Computer Science*, 420–434, Springer, Berlin, Heidelberg (2001).

[68] Maranzana, F. E., "On the location of supply points to minimize transportation costs," *IBM Systems Journal* **2**(2), 129–135 (1963).

[69] Park, H.-S. and Jun, C.-H., "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications* **36**, 3336–3341 (Mar. 2009).

[70] Biazaran, M. and SeyediNezhad, B., "Center Problem," in [*Facility Location: Concepts, Models, Algorithms and Case Studies*], Zanjirani Farahani, R. and Hekmatfar, M., eds., *Contributions to Management Science*, 193–217, Physica-Verlag HD, Heidelberg (2009).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 20 Jun 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use