Iterative self-organizing SCene-LEvel sampling (ISOSCELES) for large-scale building extraction

Benjamin Swan, Melanie Laverdiere, Dr. H. Lexie Yang and Dr. Amy Rose
Oak Ridge National Laboratory, Oak Ridge, TN USA

ABSTRACT
Convolutional neural networks (CNN) provide state-of-the-art performance in many computer vision tasks, including those related to remote-sensing image analysis. Successfully training a CNN to generalize well to unseen data, however, requires training on samples that represent the full distribution of variation of both the target classes and their surrounding contexts. With remote sensing data, acquiring a sufficiently representative training set is a challenge due to both the inherent multi-modal variability of satellite or aerial imagery and the general high cost of labeling data. To address this challenge, we have developed ISOSCELES, an Iterative Self-Organizing SCene-LEvel Sampling method for hierarchical sampling of large image sets. Using affinity propagation, ISOSCELES automates the selection of highly representative training images. Compared to random sampling or using available reference data, the distribution of the training is principally data driven, reducing the chance of oversampling uninformative areas or undersampling informative ones. In comparison to manual sample selection by an analyst, ISOSCELES exploits descriptive features, spectral and/or textural, and eliminates human bias in sample selection. Using a hierarchical sampling approach, ISOSCELES can obtain a training set that reflects both between-scene variability, such as in viewing angle and time of day, and within-scene variability at the level of individual training samples. We verify the method by demonstrating its superiority to stratified random sampling in the challenging task of adapting a pre-trained model to a new image and spatial domain for country-scale building extraction. Using a pair of hand-labeled training sets comprising 1,987 sample image chips, a total of 496,000,000 individually labeled pixels, we show, across three distinct model architectures, an increase in accuracy, as measured by F1-score, of 2.2–4.2%.

1 Introduction
1.1 Problem statement
Since Convolutional Neural Networks achieved state-of-the-art performance in computer vision, as demonstrated by their dominance in ImageNet competitions since 2012, they have been at the forefront of advancements in many domains, including electro-optical remote sensing (RS). In the past few years, there has been a great focus on adapting for RS applications, where, coupled with the rapid rise in the availability of high-resolution, spectrally or spatially, imagery from commercial satellites, there is the potential for operating at a scale and resolution that was impossible a decade ago (Zhu et al. 2017; Zhang, Zhang, and Du 2016; Hong et al. 2021a; Rasti et al. 2020; Hong et al. 2020), including optical, SAR, and hyperspectral data. During this time, the scale of imagery inferencing with CNNs has become increasingly refined, proceeding from patch-level, to bounding box localization, to semantic segmentation with full pixel-wise labeling (Shelhamer, Long, and Darrell 2016). CNN-based semantic segmentation models have already been shown to have great utility in RS tasks.

Traditional methods based upon hand crafted feature descriptors, such as gray level co-occurrence matrix (Haralick 1979), histogram of oriented gradients (Dalal and Triggs 2005), and scale invariant feature transform (Lowe 2004) have achieved acceptable accuracy in object extraction tasks; however, classifiers trained on these low-level features have shown difficulty handling the large intra-class variability and between-class similarity that are exhibited across multiple scenes and over large geographic extents (Lunga et al. 2018; Zhang, Zhang, and Du 2016; Cheng, Han, and Lu 2017). In contrast, the ability of CNNs to learn data-driven features at both low and high levels of
abstraction make them capable of achieving high accuracy over large spatial and image domains. Thus, CNNs not only offer a new level of performance, but also the ability to operate at unprecedented scale.

Successfully training a model for state-of-the-art performance presents considerable challenges which are consequences of the nature of both CNNs and RS imagery. As a supervised and data-driven approach, deep learners rely upon very large training sets, on the order of tens to hundreds of thousands of samples (Goodfellow, Bengio, and Courville 2016), to learn their features and avoid overfitting; however, RS imagery is costly and time consuming to label, all the more so when pixel-wise labels are required, as in semantic segmentation. In turn, common methods of data sampling used in the machine learning and RS communities are poorly suited to adequately capturing the complex, multi-modal variability inherent to large RS image domains, where such things as sensor characteristics, viewing geometries, and semantic content may all vary considerably (Zhang, Zhang, and Du 2016; Tuia, Persello, and Bruzzone 2016; G’omez-chova et al. 2015). Thus, a diverse sampling methodology is needed to reduce the cost of label production by focusing on the most salient examples while also maximizing out-of-sample accuracy by representing the full variability in the target domain (Hong et al. 2021b).

Standard sampling methods in RS include random sampling, sampling by expert selection, and sampling based upon the availability of reference data. Random sampling is clearly the most straightforward and can be adequate when working over small areas and with limited variability; however, it cannot be feasibly scaled to large areas. More commonly, samples are either chosen based upon expert judgment or reference data. Manual selection typically involves such things as a qualitative evaluation of target images, examination of scene and class level spectral histograms, and choosing samples based on uniformity and class separability of the resulting set (Jensen 2005; Campbell and Wynne 2011). On the other hand, sampling can be based upon where existing reference data are available, which can greatly speed-up the process of sample labeling. For example, commonly used reference data for building and road detection with remote sensing data include volunteered geographic information from OpenStreetMaps (Huang et al. 2016; Maggiori et al. 2017a; Kaiser et al. 2017), cadastral survey data (Florczyk et al. 2016; Bittner, Adam, Cui, K’orner, & Reinartz, 2018), and LiDAR-derived footprints (H. L. Yang et al. 2018; Yuan and Cheriyadat 2014).

Both of these sampling methods have considerable drawbacks when facing the aforementioned key challenges for training semantic segmentation CNNs. First, they simply do not scale well for use on large target domains. Manually selecting samples from hundreds, or even dozens, of images is impractical, leaving aside the potential for biases and omission creeping in as image analysts deal with increasingly unwieldy amounts of data. Reference data sets can have limited coverage or have temporal and/or spatial discrepancies relative to available imagery, making it more likely the resulting training set will not fully represent the target area and classes. Second, established methods were developed for per-pixel, not image object, sampling and can have difficulty both with selecting whole image chips for labeling and incorporating the spatial and semantic contexts of the target classes. These contexts introduce many implicit variables with complex distributions for manual sampling to simultaneously consider. Unless dense and comprehensive, reference data may have important gaps in representing such things as varying patterns of terrain, land cover, and anthropogenic features.

1.2 Previous work

To the best of our knowledge, we are the first work addressing sampling methods for establishing generalizable CNNs for RS applications at scale with a complete collection of real-world imagery. Existing work on improving sampling methods has largely focused upon drawing sets from data with preexisting labels; we note only two studies that give alternative methods for selecting samples for labeling from the whole target image domain. Working with the Indian Pine benchmark dataset, Rajadell et al. (2014) trained a kNN classifier using samples selected by employing kNN mode seeking unsupervised clustering with spatial coordinates and spectral values as descriptive features. They showed an increase in overall accuracy compared to random selection of 0.3 for small training sets decreasing to 0.15 for larger sets. With the goal of minimizing the amount of field survey effort needed for classification of tree species from
hyperspectral imagery, Dalponte et al. (2014) compared different approaches to sampling through clustering to both random sampling and the standard practices in RS for forestry management. Using green band reflectance as the sole descriptive feature, they tested sample sets based on either a single level of clustering done at the plot or individual tree level or a multilevel approach of first selecting optimal plots and then optimal trees within those plots. They found that the multilevel method achieved the greatest reduction in required training data with their model equaling the accuracy of a baseline model trained on all samples with only 1/5 the data.

Other work has used clustering to partition training and testing data after labeling on the premise that training and testing on different clusters will yield a more realistic assessment of model accuracy. Since randomly sampled pixels could be located near each other, Hänsch, Ley, and Hellwich (2017) contended that choosing the samples to withhold for testing from the same distribution as the training necessarily violates the requirement of sample independence and thus underestimates generalization error. They propose a method for splitting training and testing samples based upon clustering labeled samples on their spatial coordinates and selecting test and train from separate clusters. Applied to three, single-image benchmark datasets, their spatial sampling method generally yielded worse results than random sampling; however, not all results were reported. Extending the premise of clustering classes based on coordinates, Lange et al. (2018) examined the effects of adding either total contiguous area or spectral variance as features for clustering. Using a CNN and the Indian Pines dataset, they demonstrate a large decrease in calculated accuracy for both types of features even when training on 90% of available samples. In this way, these studies sought to provide a better idea of how well a model might generalize to new spatial domains, not to generate a better model.

2 Iterative self-organizing SCene-LEvel sampling (ISOSCELES)

We propose a method, dubbed ISOSCELES (Iterative Self-Organizing SCene LEvel Sampling), which uses a hierarchical, unsupervised data clustering approach to automatically select a sample set that represents the different modes and variations in spectral and spatial-semantic contexts needed for efficient CNN training to achieve better generalization. This is motivated by the fact that a CNN trained using supervised methods can only perform well when it is given a set of quality training data which capture intrinsic features inherent not only in training data but also unseen data. To this end, we need to develop a data-driven sampling approach requiring minimal prior assumptions. ISOSCELES is fully driven by target image data, making it independent of reference or ancillary data and usable anywhere in the world with imagery of any vintage, sensor type, resolution, or viewing angle. This fully data-driven approach minimizes the need for such things as time-consuming parameter tuning, a priori knowledge of the target area on the part of the user, or assumptions about variable distributions. It’s coarse-to-fine, hierarchical sampling, first selecting the most representative scenes and then the most representative subset chips from those scenes, allows for a training set that represents both within scene and between scene variability. The scales of the hierarchical sampling are illustrated in Figure 1 and an overview of the whole process is shown in Figure 2. Finally, it is computationally scalable to run on workstation-level hardware, which we defined as being able to process a dataset of 2,000 large (typically \(\sim 35,000 \times 38,000\) pixels), high-resolution images (e.g. a typical WorldView or QuickBird pan-sharpened scene) and produce a sample set ready for labeling within a standard work week.

2.1 Affinity propagation

ISOSCELES starts with an unsupervised data clustering, motivated by the need to represent image variability with a minimum number of samples. We use affinity propagation (AP), an exemplar-based clustering algorithm that eschews initialization of clusters in favor of simultaneously considering all samples as potential cluster exemplars. AP was first proposed by Frey and Dueck (2007), who showed that it could achieve superior results in less time than standard k-means based approaches. By avoiding the sensitivity to the initial choice of cluster centers that is inherent to k-means or similar methods, AP can arrive at a more accurate solution without the inefficiency of relying on large numbers of random initializations. The AP algorithm works by passing messages between each pair of samples about the suitability
of one being the exemplar of the other. At each iteration, more evidence is accumulated until the algorithm converges on a stable set of exemplars and corresponding clusters. Users can indirectly influence the number of clusters by choice of similarity preferences, however it is the underlying variability in the data that ultimately drives the number of clusters.

Since its introduction, AP has been shown to perform well in RS applications (C. Yang et al. 2013; Jia et al. 2012; Chehdi, Soltani, and Cariou 2014; Wen, Chen, and Guo 2008; Xia, Chen, and Guo 2009). Beyond general efficiency and accuracy in RS tasks, AP has several useful traits as a sampling method. First, it can easily handle large numbers of clusters, allowing it to capture local variability in large satellite images that can cover hundreds of square kilometers. Second, since its clusters can be of any size, small, yet important, portions of images can still be sampled. This is particularly useful when dealing with minority classes, such as buildings or stressed vegetation, which may occupy only a small fraction of an image. Instead of being lumped in with an inappropriate cluster, these minority classes can be preserved as distinct clusters, even if their cluster size is far smaller than the average in a given clustering. Finally, since the number of clusters is primarily data-driven, it can be run over large image sets without the user having to either choose to adjust parameters for each image or adopt a one-size-fits-all approach.

2.2 Descriptive features

We then define a set of descriptive features to be used in AP. The choice of descriptive features is crucial for successfully finding relevant patterns in image data. Commonly, image band statistics are used to guide sampling (Lillesand, Kiefer, and Chipman 2015; Jensen 2005; Campbell and Wynne 2011); however, this spectral information alone cannot adequately represent spatial and contextual variation. Spectral data are principally related to the material composition of the land cover, yet, to train a CNN for good performance, it must be exposed to the full variance in appearance and surrounding context of the target class. Thus, in addition to using spectral data, we employ image metadata and textural features.

Figure 1. Illustration of ISOSCELES hierarchical sampling. (A) 145 exemplar scenes selected from 2,593 WorldView-2 satellite image scenes. Size varies by collection parameters, but is generally 17 × 19 km, 35,000 × 38,000 pixels. (B) Exemplar image scene with 46 exemplar sample image chips. (C) 500 × 500 pixel exemplar sample chip. © 2010 Digital NextView License.
Since ISOSCELES uses a hierarchical approach to capture both between-scene and within-scene variability, we chose descriptive sets of descriptive features that are informative at these different scales. For both scene and chip-level clustering, we used band means and standard deviations as our spectral features. At the scene level of clustering, we add the off-nadir angle to account for different viewing perspectives and the sun elevation for differences in shadows and illumination. Both of these features can contribute to between-image variation in the appearance of classes, yet cannot be readily computed from pixel values alone. For the image chip clustering, we include Gabor filters as texture descriptors. Gabor filters have long been used in image processing for such things as feature extraction and texture analysis. Gabor filters are sensitive to orientation and spatial frequency in a manner similar to low-level mammalian vision (Bianconi and Fernández 2007; Petrou and García Sevilla 2006; McGlone and Lee 2013). For sampling RS images, Gabor filters can help increase the representation of classes that are similar spectrally but differ semantically. Examples include differentiating a tree plantation from a forest or a grassy levee from a natural hill. In the case of training a CNN, this has the advantage of increasing the number of semantic contexts to which the model is exposed.

### 2.3 Detailed description

Our proposed workflow begins with selecting exemplar image scenes for sampling. Starting with the full image dataset $I^T = \{I_1, I_2, \ldots, I_t\}$ where $t$ scenes are included, we calculate spectral band features $f_S$, including the band means and standard deviations, and metadata features $f_M$, including the satellite view angle and sun elevation at the time of acquisition, as the features for AP clustering. Working with coarse level, ISOSCELES is then run on each scene to create a potential training set at scene level. For each scene, we generate a seamless, non-overlapping set of $d \times d$ pixel chips, where $d \times d$ are the input dimensions for the model being trained. For a typical WorldView scene of $\sim 35,000 \times 38,000$ pixels, this results in around $5,300 \times 500 \times 500$ pixel image chips. We then zoom in to calculate image band statistics and texture features on the level of individual chips. For image texture, we...
use a bank of 16 Gabor filters, representing four orientations with two band widths at two frequency scales. The mean and variance of each of these filters are combined with the band means and variances to create a 40-dimensional feature vector for each chip. Each category of feature is normalized separately (e.g. mean band values) and a negative Euclidean distance similarity matrix is calculated, as recommended in (Frey and Dueck 2007). After performing AP clustering using this similarity matrix, this typically yields 30–300 exemplar chips per scene or 18–177 cluster members per cluster, depending on within-scene heterogeneity and choice of similarity preference. An example of ISOSCELES output from a satellite image of Taizz, Yemen, is shown in Figure 3 with a closer look at individual samples from the image in Figure 4. The 29 color-coded clusters shown in Figure 3 do not represent target classes but rather examples of the clusters identified by fundamental image variability in a single image scene based on ISOSCELES clustering. In this case and as shown in Figure 4, the clusters are meaningful to a human image analyst and include LU/LC contexts, such as dense modern settlement and large commercial structures, that make up only a small fraction of the whole image; however, a single target class of interest (building) can be represented by multiple clustering classes because there can be incredible variability within target classes. The goal of ISOSCELES is to represent between-class and within-class variability to ensure training data is as representative of the full target domain as possible so that the CNN is exposed to spectral and textural variability that a human analyst or random sampling may miss.

3 Testing on real world application

To show the efficacy of our proposed method on a real-world scenario, we chose the task of adapting pre-trained building-extraction models to a high-resolution satellite image set covering Yemen. Machine learning models usually generalize poorly when used on image and spatial domains beyond their training data (Wang et al. 2017; Magiori et al. 2017a); therefore re-training with samples drawn from the new target domain is necessary to maintain good performance. We compare models re-trained using a sample set selected using our ISOSCELES-based method to those re-trained on a stratified random sample of known
built-up areas based on reference data. To validate that our methods deliver uniform benefits, we tested the proposed method in three CNN architectures: SegNet (Badrinarayanan, Kendall, and Cipolla 2017), U-Net (Ronneberger, Fischer, and Brox 2015), and DeepLab (Chen et al. 2018). These three architectures have shown superior performance on semantic segmentation tasks.

### 3.1 Data and experimental settings

Our target image dataset consisted of 2,593 WorldView-2 satellite image scenes (DigitalGlobe 2017) orthorectified and pan-sharpened to a resolution of 45–70 cm. Once selected, all samples were manually digitized with all visible parts of buildings labeled, including any complete or partial building facades. Representative training samples are shown in Figure 6. While our source imagery is four
band and all four bands were used in ISOSCELES sampling, the CNN architectures we used in this study have a three channel input. As such, instead of using all bands or true color, we used a green (515 – 586 nm), red (632 – 692 nm), and near-infrared (NIR) (772 – 890 nm) band combination as the NIR band is generally more informative and less noisy than the blue band.

ISOSCELES-based sampling was carried out as described in Section 2 starting with those 2,593 image scenes. This yields 145 exemplar scenes (with Algorithm 1 scene selection) and then 989 image chips selected (with Algorithm 2 image chip selection).

For the baseline comparison model, we used a stratified random sampling method based on MGCP (J’ozsef and Ol’ivia 2009) building point data. In this dataset, each building is represented as a single vector point feature. Although there are open source options such as OpenStreetMap (OpenStreetMap contributors 2017) building footprints (polygons) that can be directly utilized for building extraction, unfortunately it exhibited large gaps in building data completeness across Yemen. Using MGCP, we first classified potential samples as belonging to one of five building density categories: 1–13; 14–30; 31–50; 51–76; and 77–176 buildings in a 500 × 500 pixel block. We then randomly chose 1% of samples from each of the density categories, resulting in 989 image chips in the final selection. Note that the MGCP point data was not used for model training. Instead, these points were only used as an indicator for the potential location for buildings to reduce the image area considered for random sampling.

We then manually labeled buildings as polygons associated with both the ISOSCELES and randomly selected chips. The combined 1,978 sample chips identified between the ISOSCELES and randomly selected image chips correspond to 58,700 individual buildings. Labeling these datasets took a team of experienced imagery analysts approximately 1,200 work hours, with an additional 200 hours required for quality control reviews and revisions. An overview of the experimental workflow can be found in Figure 5. Figure 7 illustrates the spatial distribution of scenes sampled using each method. The scenes that were sampled by both methods are shown in orange in Figure 7. We can see that the set of scenes selected through AP clustering is more evenly distributed and only 15 scenes are in common.
Our base models were created using a transfer learning approach to adapt models ImageNet pre-trained to aerial imagery, as described in (H. L. Yang et al. 2018). It has been found that models trained on very large image sets, such as ImageNet (a database of more than 14,000,000 hand-annotated images) learn basic image features that can be transferred to more specialized tasks by fine-tuning using a set of domain-specific samples (Penatti, Nogueira, and Dos Santos 2015; Tajbaksh et al. 2016; Bittner, Cui, and Reinartz 2017). Transfer learning has been successfully used in multiple applications and been shown to yield better results with less training time than training from scratch. The sample set used to adapt our base models to remote sensing imagery had 4,000 500 × 500 pixel image chips. The chips were drawn from 78 image scenes of the 2013–2014 vintages of the United States Department of Agriculture’s National Aerial Imagery Program (NAIP), which consists of 1-meter resolution 4-band (red, green, blue, near infrared) imagery typically taken in leaf-on conditions. Labeling was accomplished by applying an auto-shifting algorithm to align preexisting, LiDAR-derived building footprints with the NAIP imagery.

The three CNNs were initialized by the NAIP pre-trained models, respectively, and then were fine-tuned with selected samples. We trained each model for 100 epochs with batch size = 4 and learning rate = 10⁻⁵ and selected the best model for each combination of training sample set and architecture based on the highest validation accuracy. Because these models were pre-trained models for building mapping, even using relatively few samples can still achieve training convergence and reach satisfactory results.

3.2 Results and discussions

All models were validated against 300 samples that were withheld from the training data, 150 from each sample set. This allowed for both direct comparison of the performance of models trained with different data and ensured that each model would be tested against samples chosen using a different method from the one by which its own training set was chosen. For accuracy metrics, we created confusion matrices and calculated four common metrics: (1) precision, a measure of the reliability of predictions, (2) recall, a measure of the completeness of predictions, (3) F1-score, a harmonic mean of precision and recall, and (4) Intersection over Union (IoU), a measure of how well predicted objects line up with reference data. These are calculated using the equations:

\[
\text{Precision} = \frac{tp}{tp + fp} \tag{1}
\]

\[
\text{Recall} = \frac{tp}{tp + fn} \tag{2}
\]
\[ F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]  

\[ \text{IoU} = \frac{\text{tp}}{\text{tp} + \text{fp} + \text{fn}} \]  

where \( \text{tp} \) are true positives (correctly classified building pixels), \( \text{fp} \) are false positives (non-building classes – sized as building), and \( \text{fn} \) are false negatives (building pixels classified as non-building).

### 3.3 Quantitative evaluation

In Table 1, we present the evaluation scores for the model with the highest F1-score in each condition. Across all three architectures, the model trained using ISOSCELES sampling achieved a higher F1-score than its random sampling counterpart with an increase of between 2.2 and 4.2 percentage points. Similarly, the buildings predicted by ISOSCELES models were consistently more similar to the reference buildings than their random model counterparts, with IoU increases of 2.5–5.3 points. Examining all metrics, the ISOSCELES model outperformed the random model in all but precision in the U-Net model, which had a 0.6 difference, dwarfed by the 7.4 point gain in recall. Taken together, these results confirm that, with no additional labeling effort, our proposed stratified sampling method yields significant performance gains.

ISOSCELES sampling models showed not only increased performance, but a considerable difference in the spread between precision and recall scores. By taking the ratio of precision score to recall score, we can get an approximation of how well the model has learned to separate the building and non-building classes. An ideal model should have both a high accuracy and a ratio close to 1, indicating that performance is not due to biased predictions. Taking the median precision to recall ratio over the course of training, shown in Table 2, all models had scores above one, indicating a bias toward underpredicting building areas. Random sampling models showed an increase in precision-recall ratio over ISOSCELES models of 3, 17, and 21% for SegNet, U-Net, and DeepLab, respectively. Looking across all epochs, we can see that the ISOSCELES sampling models have a precision-to-recall gap that narrows over the training period, while the gap for the random sampling models generally maintains its width. This shows that the samples provided by ISOSCELES can indeed efficiently boost the training performance in both time and convergence.

### 3.4 Qualitative analysis

For a qualitative evaluation, Figures 8 and 9 compare the outputs from the SegNet and U-Net models which showed the smallest and largest differences in accuracy between ISOSCELES models and random sampling models. In both figures, we can see similar patterns. Row (A) shows the much more prominent pure commission errors of the random model in comparison to the ISOSCELES model. In (B), the random model missed many of the smaller buildings and had issues with mis-classifying ground adjacent to buildings. Finally, (C) shows how the ISOSCELES sampling yields better feature extractors to identify the high-density areas of multi-story (Figure 8) or single-story (Figure 9) structures, resulting in a markedly lower omission error relative to the random sample model.

### 3.5 Scalability

The ability to scale for use over large datasets without needing either large amounts of computing time or resources is an important requirement for automated methods for large-scale remote sensing imagery analysis. Given the sample size \( n \), the computational

| Model  | Sampling | Precision | Recall | F1   | IoU  |
|--------|----------|-----------|--------|------|------|
| SegNet | Random   | 0.719     | 0.655  | 0.685| 0.521|
| SegNet | ISOSCELES| 0.741     | 0.675  | 0.707| 0.546|
| U-Net  | Random   | 0.809     | 0.639  | 0.714| 0.555|
| U-Net  | ISOSCELES| 0.803     | 0.713  | 0.756| 0.608|
| DeepLab| Random   | 0.707     | 0.693  | 0.700| 0.538|
| DeepLab| ISOSCELES| 0.758     | 0.704  | 0.730| 0.575|

#### Table 2. Median precision-recall ratio of models trained using ISOSCELES sampling method and stratified random sampling.

| Model  | Sampling | P-R Ratio |
|--------|----------|-----------|
| SegNet | Random   | 1.207     |
| SegNet | ISOSCELES| 1.165     |
| U-Net  | Random   | 1.450     |
| U-Net  | ISOSCELES| 1.240     |
| DeepLab| Random   | 1.523     |
| DeepLab| ISOSCELES| 1.254     |
complexity of the affinity propagation is $O(n^2)$ (Furtlehner, Sebag, and Zhang 2010) where $n$ is all possible 500 pixels by 500 pixels image chips in a country size image collection.

As mentioned in Section 2, we set a goal that ISOSCELES should be able to sample a dataset of at least 2,000, large, high-resolution images within a week while running on workstation-level hardware.

Figure 8. Qualitative comparison of SegNet model outputs for ISOSCELES sampling (right) and random sampling methods (left) for varying settlement patterns. (A) Mixed density modern single-story, (B) medium density informal, single-story, (C) high density traditional, multi-story. © 2010 Digital NextView License.
ISOSCELES achieves scalability through two means. First, coarse-level scene selection minimizes the amount of data that needs to be processed at pixel levels for generating descriptive features. Second, calculation of descriptive features is performed in parallel, typically with 12–14 worker processes achieving maximum efficiency and Input/Output balancing. With this hierarchical design of ISOSCELES, we can avoid the potential computational issue in clustering whole image collection. With the Yemen

**Figure 9.** Qualitative comparison for U-Net model outputs with ISOSCELES sampling (right) and random sampling (left). (A) Rural medium density informal, multi-story, (B) low density informal, single-story, (C) high density single-story. © 2010 Digital NextView License.
dataset, we first handled \( m = 2 \), 593 scenes in Algorithm 1 and approximately \( n = 3 \), 100 chips in Algorithm 2 rather than \( n = 8, 000, 000 \) chips on the entire imagery collection covering Yemen with affinity propagation.

ISOSELES sampling for this study was carried out on a Dell Precision 5810 workstation equipped with an Intel Xeon E5-2630 v4 supporting up to 20 processing threads at 2.20 GHz, 32 GB of RAM, and solid-state hard drive. For the total 2,593 scenes in the Yemen dataset, coarse sampling took less than one minute with ISOSELES sampling to select the 145 exemplar scenes. The average size for a scene was 36,000 × 37,300 pixels and 10.22 GB. The 145 exemplar scenes were processed in 44.4 hours. The vast majority of processing time was taken up by reading/writing data (5.75 minutes/scene on average) and calculating chip-level features (7.46 minutes/scene on average).

4 Conclusions

CNNs are now a principle driver in the advancement of RS image analysis, capable of achieving favorable performance over domains of thousands of images covering hundreds of thousands of square kilometers. Effectively training a model for state-of-the-art performance, however, presents the serious challenge of obtaining a training data set that adequately represents the target image domain while dealing with the practical constraint of the time and expense required for sample labeling. Conventional methods of randomly selecting training samples based on the availability of data are poorly suited for this task as they do not consider the multi-modal variability of large image sets and do not account for spatial-textural features. Thus, there is a need to develop methods to efficiently identify highly informative samples from large target domains.

We proposed and demonstrated ISOSELES, a novel, hierarchical use of unsupervised data clustering with image spectral and textural features to address the challenge of obtaining a sample set that represents complex, multi-modal variation in remote sensing imagery. Without strategic sampling, the training samples can be both unrepresentative and redundant, which leads to poor generalization to unseen data and longer training times. Our proposed sampling strategy uses unsupervised clustering hierarchically, first to identify representative images within a target domain by their spectral and image acquisition characteristics, then to identify individual samples that can represent the spectral and textural variability within each image selected in the first stage. Thus, ISOSELES can efficiently obtain a set of samples for labeling that represents both between-image and within-image variability. Compared to random sampling, using ISOSELES we can ensure that during supervised training the CNN is exposed to the variability of target imagery while optimizing feature extractors to achieve better performance in the target domain with an equal number of samples.

With a large-scale, real-world dataset, we have shown that, with the same amount of labeling effort, our proposed sampling method can yield significantly better results than the sampling based on reference data. Our test case was the adaptation of a pre-trained semantic segmentation CNN model to a new spatial and sensor domain: from aerial imagery of the United States to high-resolution satellite imagery of Yemen. Compared to the models trained with random sampling of areas with available building point data, we show significant improvements using our proposed method that were consistent across three various model architectures. These results demonstrate the potential for bottom-up, data-driven methods such as ISOSELES to facilitate the successful application of deep neural networks to challenging large-scale remote sensing imagery analysis.

Acknowledgements

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Disclosure statement

No potential conflict of interest was reported by the authors.
ORCID

Benjamin Swan
http://orcid.org/0000-0002-2787-407X

Data and Code Availability

The imagery data used in this paper cannot be publicly shared due to licensing restrictions. Building labels used in this paper are available as polygon shapefiles at 10.6084/m9.figshare.16663744.v1. Pseudo-code and diagrams are provided in paper to allow for coding this sampling method in the lan- guage of the user’s choosing. Example python implementations with full documentation are available at https://github.com/btswan87/isosceles under an MIT Open Source Software license.

References

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. “Segnet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation.” IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (December): 2481–2495. doi:10.1109/TPAMI.2016.2644615.

Bianconi, F., and A. Fern’andez. 2007. “Evaluation of the Effects of Gabor Filter Parameters on Texture Classification.” Pattern Recognition 40 (12) (December): 3325–3335. doi:10.1016/j.patcog.2007.04.023.

Bittner, K., F. Adam, S. Cui, M. K’orné, and P. Reinartz. 2018. “Building Footprint Extraction from VHR Remote Sensing Images Combined with Normalized DSMs Using Fused Fully Convolutional Networks.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (8) (August): 2615–2629. doi:10.1109/JSTARS.2018.2849363.

Bittner, K., S. Cui, and P. Reinartz. 2017. “Building Extraction from Remote Sensing Data Using Fully Convolutional Networks.” International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 42.

Campbell, J. B., and R. H. Wynne. 2011. Introduction to Remote Sensing. 5th ed. ed. New York:Gulfiff Press. (OCLC: ocn666239609).

Chehdi, K., M. Soltani, and C. Cariou. 2014. “Pixel Classification of Large-size Hyperspectral Images by Affinity Propagation.” Journal of Applied Remote Sensing 8 (1) (August): 083567. doi:10.1117/1.JRS.8.083567.

Chen, L., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.” IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (4) (April): 834–848. doi:10.1109/TPAMI.2017.2699184.

Cheng, G., J. Han, and X. Lu. 2017. “Remote Sensing Image Scene Classification: Benchmark and State of the Art.” Proceedings of the IEEE 105 (10) (October): 1865–1883. doi:10.1109/JPROC.2017.2675998.

Dalal, N., and B. Triggs. 2005, June. “Histograms of Oriented Gradients for Human Detection.” C. Schmid, S. Soatto, and C. Tomasi, edited by. International Conference on Computer Vision & Pattern Recognition (CVPR ’05) Vol. 1. 886–893. San Diego, United States:IEEE Computer Society. doi: 10.1109/CVPR.2005.177.

Dalponte, M., L. T. Ene, H. O. Ørka, T. Gobakken, and E. Næsset. 2014. "Unsupervised Segmentation of Training Samples for Tree Species Classification Using Hyperspectral Data." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (8) (August): 3560–3569. doi:10.1109/JSTARS.2014.2315664.

DigitalGlobe. (2017). "Digitalglobe Nextview License Agreement." Retrieved 12 July 2017 from http://www.digitalglobe.com/legal/information

Florczyk, A. J., S. Ferri, V. Syriris, T. Kemper, M. Halkia, P. Soille, and M. Pesaresi. 2016. “A New European Settlement Map From Optical Remotely Sensed Data.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (5) (May): 1978–1992. doi:10.1109/JSTARS.2015.2485662.

Frey, B. J., and D. Dueck. 2007, February. “Clustering by Passing Messages between Data Points.” Science 315 (5814):972–976. DOI: 10.1126/science.1136800.

Furrlehner, C., M. Sebag, and X. Zhang. 2010, June. “Scaling Analysis of Affinity Propagation.” Phys. Rev. E 81:066102. DOI: 10.1103/PhysRevE.81.066102.

Gómez-chova, L., D. Tuia, G. Moser, and G. Camps-Valls. 2015. “Multimodal Classification of Remote Sensing Images: A Review and Future Directions.” Proceedings of the IEEE 103 (9) (September): 1560–1584. doi:10.1109/JPROC.2015.2449668.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. Deep Learning. (OCLC: 955778308), Cambridge, Massachusetts London, England: MIT Press.

Hansch, R., A. Ley, and O. Hellwich (2017, July). “Correct and Still Wrong: The Relationship between Sampling Strategies and the Estimation of the Generalization Error.” In 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 3672–3675). doi:10.1109/IGARSS.2017.8127795.

Haralick, R. M. 1979. “Statistical and Structural Approaches to Texture.” Proceedings of the IEEE 67 (5) (May): 786–804. doi:10.1109/PROC.1979.11328.

Hong, D., L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot. 2021a. "Graph Convolutional Networks for Hyperspectral Image Classification." IEEE Transactions on Geoscience and Remote Sensing 59 (7): 5966–5978. doi:10.1109/TGRS.2020.3015157.

Hong, D., L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang. 2021b. “More Diverse Means Better: Multimodal Deep Learning Meets Remote-sensing Imagery Classification.” IEEE Transactions on Geoscience and Remote Sensing 59 (5): 4340–4354. doi:10.1109/TGRS.2020.3016820.

Hong, D., N. Yokoya, G. Xia, J. Chanussot, and X. Zhu. 2020. “X-modalnet: A Semi-supervised Deep Cross-modal Network for Classification of Remote Sensing Data.” Isprs Journal of Photogrammetry and Remote Sensing 167: 12–23.
Huang, Z., G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan (2016, July). “Building Extraction from Multi-source Remote Sensing Images via Deep Deconvolution Neural Networks.” In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1835–1838). Beijing, China: IEEE. Retrieved 11 June 2019, from doi: 10.1109/IGARSS.2016.7729471.

József, C., and M. Ol’ivia. 2009. “Multinational Geospatial Co-production Program (Mgcp): Geodezia Es Kartografa 61 (January): 15–18.

Jensen, J. R. 2005. Introductory Digital Image Processing: A Remote Sensing Perspective. 3 ed. (OCLC: 604318046). Upper Saddle River, NJ: Pearson/Prentice Hall.

Jia, S., Z. Ji, Y. Qian, and L. Shen. 2012. “Unsupervised Band Selection for Hyperspectral Imagery Classification without Manual Band Removal.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5 (2) (April): 531–543. doi: 10.1109/JSTARS.2012.2218743.

Kaiser, P., J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler. 2017. “Learning Aerial Image Segmentation From Online Maps.” IEEE Transactions on Geoscience and Remote Sensing 55 (11) (November): 6054–6068. doi: 10.1109/TGRS.2017.2719738.

Lange, J., G. Cavallaro, M. G’ozt, E. Erlingsson, and M. Riedel (2018, July). “The Influence of Sampling Methods on Pixel-Wise Hyperspectral Image Classification with 3d Convolutional Neural Networks.” In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 2087–2090). doi: 10.1109/IGARSS.2018.8518671.

Lillesand, T. M., R. W. Kiefer, and J. W. Chipman. 2015. Remote Sensing and Image Interpretation. Seventh ed. Hoboken, N.J: John Wiley & Sons,

Low, D. G. 2004. “Distinctive Image Features from Scale-invariant Keypoints.” International Journal of Computer Vision 60 (2): 91–110.

Lunga, D., H. L. Yang, A. Reith, J. Weaver, J. Yuan, and B. Bhaduri. 2018. “Domain-Adapted Convolutional Networks for Satellite Image Classification: A Large-Scale Interactive Learning Workflow.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (3) (March): 962–977. doi: 10.1109/JSTARS.2018.2795753.

Maggiori, E., Y. Tarabalka, G. Charpillet, and P. Alliez (2017a, July). “Can Semantic Labeling Methods Generalize to Any City? the Inria Aerial Image Labeling Benchmark.” In 2017 ieee international geoscience and remote sensing symposium (igars) (p. 3226–3229). doi: 10.1109/IGARSS.2017.8127684.

Maggiori, E., Y. Tarabalka, G. Charpillet, and P. Alliez. 2017b. “Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification.” IEEE Transactions on Geoscience and Remote Sensing 55 (2) (February): 645–657. doi: 10.1109/TGRS.2016.2612821.

McGlone, J. C., and G. Y. G. Lee, Eds. 2013. Manual of Photogrammetry. 6 ed. Bethesda, Md:ASPRS, American Soc. for Photogrammetry and Remote Sensing. (OCLC: 931241194).

OpenStreetMap contributors. (2017). Planet Dump. retrieved from https://www.openstreetmap.org

Penatti, O. A. B., K. Nogueira, and J. A. Dos Santos (2015, June). “Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?” In 2015 ieee conference on computer vision and pattern recognition workshops (cvprw) (p. 44–51). doi: 10.1109/CVPRW.2015.7301382.

Petrov, M., and P. García Sevilla. 2006. Image Processing: Dealing with Texture. (OCLC: 255680294), Chichester: Wiley.

Rajadell, O., P. García-Sevilla, V. C. Dinh, and R. P. W. Duin. 2014. “Improving Hyperspectral Pixel Classification With Unsupervised Training Data Selection.” IEEE Geoscience and Remote Sensing Letters 11 (3) (March): 656–660. doi: 10.1109/LGRS.2013.2273983.

Rasti, B., D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson. 2020. “Feature Extraction for Hyperspectral Imagery: The Evolution from Shallow to Deep: Overview and Toolbox.” IEEE Geoscience and Remote Sensing Magazine 8 (4): 60–88. doi: 10.1109/MGRS.2020.2979764.

Ronneberger, O., P. Fischer, and T. Brox (2015). “U-net: Convolutional Networks for Biomedical Image Segmentation.” In International conference on medical image computing and computer-assisted intervention (pp. 234–241).

Shellhammer, E., J. Long, and T. Darrell (2016, May). “Fully Convolutional Networks for Semantic Segmentation.” arXiv:1605.06211 [cs]. (arXiv: 1605.06211).

Tajbakhsh, N., J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. 2016. “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?” IEEE Transactions on Medical Imaging 35 (5) (May): 1299–1312. doi: 10.1109/TMI.2016.2535302.

Tuia, D., C. Persello, and L. Bruzzone. 2016. “Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances.” IEEE Geoscience and Remote Sensing Magazine 4 (2) (June): 41–57. doi: 10.1109/MGRS.2016.2548504.

Wang, R., J. Camilo, L. M. Collins, K. Bradbury, and J. M. Malof (2017, October). “The Poor Generalization of Deep Convolutional Networks to Aerial Imagery from New Geographic Locations: An Empirical Study with Solar Array Detection.” In 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1–8). doi: 10.1109/AIPR.2017.8457965.

Wen, L., X. Chen, and P. Guo. 2008. “A Comparative Study on Clustering Algorithms for Multispectral Remote Sensing Image Recognition.” In Advances in Neural Networks - ISNN 2008, edited by F. Sun, J. Zhang, Y. Tan, J. Cao, and W. Yu, 610–617. Berlin, Heidelberg: Springer Berlin Heidelberg.

Xia, H., X. Chen, and P. Guo (2009, October). “A Shadow Detection Method for Remote Sensing Images Using Affinity Propagation Algorithm.” In 2009 IEEE International Conference on Systems, Man and Cybernetics (pp. 3116–3121). doi: 10.1109/ICSMC.2009.5346147.
Yang, C., S. Liu, L. Bruzzone, R. Guan, and P. Du. 2013. “A Feature-Metric-Based Affinity Propagation Technique for Feature Selection in Hyperspectral Image Classification.” IEEE Geoscience and Remote Sensing Letters 10 (5) (September): 1152–1156. doi:10.1109/LGRS.2012.2233711.

Yang, H. L., J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri. 2018. “Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (8) (August): 2600–2614. doi:10.1109/JSTARS.2018.2835377.

Yuan, J., and A. M. Cheriyadat (2014). “Learning to Count Buildings in Diverse Aerial Scenes.” In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 271–280). New York, NY, USA: ACM. Retrieved 30 April 2018, from doi: 10.1145/2666310.2666389.

Zhang, L., L. Zhang, and B. Du. 2016. “Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art.” IEEE Geoscience and Remote Sensing Magazine 4 (2) (June): 22–40. doi:10.1109/MGRS.2016.2540798.

Zhu, X. X., D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer. 2017. “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources.” IEEE Geoscience and Remote Sensing Magazine 5 (4) (December): 8–36. doi:10.1109/MGRS.2017.2762307.