Working title: **How good are retrained deep neural networks at classifying images of landscapes?**

Authors: Dan, Jon, Andy, ?

(want open access and rapid publication ahead of workshops and use in upcoming papers): Geosciences?

**Abstract**

There is a growing need for fully automated pixel-scale classification of large datasets consisting of color digital photographic imagery, in the analysis and interpretation of natural landscapes and geomorphic processes. The application of deep learning, specifically deep convolutional neural networks (DCNNs), to the classification of remotely sensed imagery has the potential to both outperform and simplify such tasks, compared to traditional approaches. However, the general usefulness of deep learning applied to conventional photographic imagery at the landscape scales is, at yet, largely unproven. DCNNs are computationally intensive to train and deploy, very data hungry (often requiring millions of examples to train from scratch), and require expert knowledge to design and optimize. However, these issues are mostly overcome for general applications using ‘transfer learning’, in which existing DCNN architectures, built by specialists using massive image databases and designed to recognize features/objects in images in a generic sense, and with well-studied properties, are retrained to classify specific objects and features. If DCNN-based image classification is to gain wider application and acceptance with the geoscience community, demonstrable successes need to be coupled with accessible tools with which to transfer deep neural networks to landforms and land uses in images of landscapes. In this paper, we present an efficient approach to train/apply DCNNs with/on sets of photographic images, based on a powerful and efficient graphical method, called a conditional random field (CRF), to generate DCNN training and testing data using minimal manual supervision. We apply the method to several sets of images of natural landscapes, consisting of imagery from satellites, aircraft, unmanned aerial vehicles, and fixed camera installations, synthesizing the findings to examine the general effectiveness of transfer learning to landscape scale image classification. Finally, we show how DCNN predictions on small regions of images might be used in conjunction with a CRF for highly accurate pixel level classification of images.

**1. Introduction**

**1.1. The growing use of image classification in the geosciences**

The task of classifying natural objects and textures in images of landforms is increasingly widespread in a wide variety of geomorphological research (e.g. *Franklin and Wulder, 2002*; *Smith and Pain, 2009*; *Mulder et al., 2011*; *Sekovski et al., 2014*; *Ma et al., 2017*; *Cheng et al., 2017*; *O’Connor et al., 2017*), providing impetus for the development of completely automated methods to maximize speed and objectivity. The task of labeling image pixels into discrete classes is called object class segmentation or semantic segmentation, whereby an entire scene is parsed into object classes at a pixel level (e.g. *Long et al., 2015*; *Volpi and Tuia, 2017*).

There is a growing trend in studies of coastal and fluvial systems for using automated methods to extract information from time-series of imagery from ﬁxed camera installations (e.g. *Holman and Stanley, 2007; Bertoldi et al., 2010; Hoonhout et al., 2015; Bergsma et al., 2016; Almar et al., 2016; Benacchio et al., 2017; Grams et al., 2018*), UAVs (*Turner et al., 2016*; *Su and Gibeaut, 2017; Sturdivant et al., 2017*) and other aerial (e.g. *Warrick et al., 2016*) platforms. Fixed camera installations are designed for generating time-series of images for assessment of geomorphic change in dynamic environments. Many UAV and aerial imagery data sets are collected for building digital terrain models using Structure-from-Motion (SfM) photogrammetry (REFS). Numerous complementary or alternative uses of such imagery for the purposes of geomorphic research include facies description and grain size calculation, geomorphic and geologic mapping, vegetation structure description, and physical habitat quantification (REFS).

In this paper, we utilize and evaluate two emerging themes in computer vision research, namely deep learning and structured prediction, that, when combined, are shown to be extremely effective in application to pattern recognition, and semantic segmentation of highly structured, complex objects in images of natural scenes.

**1.2. Application of deep learning to landscape scale image classification**

Deep learning is the application of artificial neural networks with more than one hidden layer, to the task of learning and subsequently recognizing patterns in data (*LeCun et al., 2015; Goodfellow et al., 2016*). A class of deep learning algorithms called deep convolutional neural networks (DCNNs) are extremely powerful at image recognition, resulting in a massive proliferation of their use (*Szegedy et al., 2017*; *Chen et al., 2018*), across almost all scientific disciplines (e.g. *Litjens et al., 2017*; *Lui et al., 2017*; *Maggiori et al., 2017*). A major advantage to DCNNs over conventional machine learning approaches to image classification is that it does not require so-called ‘feature-engineering’ or ‘feature extraction’, which is the art of transforming the image data so that it is either more amenable to a speciﬁc machine-learning algorithm, or providing the algorithm more data by computing derivative products from the imagery, such as rasters of texture or alternative colorspaces (e.g. *Hoonhout et al., 2015; Belgiu and Drăgut, 2016; Cheng et al., 2017*). In deep learning, features are automatically learned from data using a general-purpose procedure. Another reputed advantage is that DCNN performance tends to keep improving with more and more data, whereas machine learning performance tends to plateau (*Dauphin et al., 2014*). For these reasons, DCNN techniques will find numerous applications where automated interpretation and quantification of natural landforms and textures are used to solve geomorphological problems.

However, many claims about the efficacy of DCNNs for image classification are largely based upon analyses of conventional photographic imagery of familiar, mostly anthropogenic objects (*Garcia-Garcia et al., 2017; Cheng et al., 2017*), and it remains to be seen that this is still always the case for classification of images of natural textures and objects. Aside from the relatively large scale, images of natural landscapes that are generally collected for geomorphological objectives tend to be taken from the air or at high vantage, with a nadir or obtuse perspective. In contrast, images that make up many libraries upon which DCNNs are trained and evaluated tend to be taken from the ground, with a straight or reflex perspective. In addition, variations in lighting and weather greatly affect distributions of color, contrast and brightness; certain land covers change appearance due to changing seasons (such as deciduous vegetation); and geomorphic processes alter the appearance of land covers and landforms causing large intra-class variation, for example, clear, turbid and aerated water. Finally, the distinction of certain objects and features may be difficult against similar backgrounds, for example groundcover between vegetation canopies.

The most popular DCNN architectures have been designed and trained on large generic image libraries such as Imagenet (*Deng et al., 2009*), mostly developed as a result of international computer vision competitions (*Russakovsky et al., 2015*) and primarily for application to close-range imagery with small spatial footprints (*Garcia-Garcia et al., 2017*), but more recently have been used for landform/landuse classification tasks in large spatial footprint imagery such as that used in satellite remote sensing (e.g. *Hu et al., 2015*; *Langkvist et al., 2016; Palafox et al., 2017*; *Lu et al., 2017; Marmanis et al., 2017*). These applications have involved design and implementation of new or modified DCNN architectures, or relatively large existing DCNN architectures. Also, these applications have largely been limited to satellite imagery only. In this contribution, one major objective is to examine the accuracy of DCNNs for oblique and nadir conventional medium-range imagery. Another objective is to evaluate the smallest, most lightweight existing DCNN models, retrained for specific land use/land cover purposes, with no retraining from scratch and no modification or fine-tuning to the data. We utilize a concept known as ‘transfer learning’, where a model trained on one task is re-purposed on a second related task (*Goodfellow et al., 2016*). Though powerful, DCNNs are also computationally intensive to train and deploy, very data hungry (often requiring millions of examples to train from scratch), and require expert knowledge to design and optimize. Collectively, this may impede widespread adoption of these methods within the geoscience community. Fortunately, several open-source DCNN architectures have been designed for general applicability to the task of recognizing objects and features in non-specific photographic imagery. Here, we use existing pre-trained DCNN models that are designed to be transferable for generic image recognition tasks, which facilitates rapid DCNN training when developing classifiers for specific image sets. Training is rapid because only the final layers in the DCNN need to be retrained to classify a specific set of objects.

**1.3. Pixel-scale image classification strategies**

Automated classification of pixels in digital photographic images involves predicting labels, ***y***, from observations of features, ***x***, which are derived from relative measures of color in red, green and blue spectral bands in imagery. In the geosciences, the labels of interest naturally depend on the application but may be almost any type of surface land cover (such as specific sediment, landforms, geological features, vegetation type and coverage, water bodies, etc) or description of land use (rangeland, cultivated land, urbanized land, etc). The relationships between ***x*** and ***y*** are complex and non-unique, because the labels we assign depend nonlinearly on observed features, as well as on each other. For example, neighboring regions in an image tend to have similar labels (i.e. they are spatially autocorrelated). Depending on the location and orientation of the camera relative to the scene, labels may be preferentially located. Some pairs of labels (e.g. ocean and beach sand) are more likely to be proximal than others (e.g. ocean and arable land).

A natural way to represent the manner in which labels depend on each other is provided by graphical models (e.g. *Sutton and McCallum, 2006*) where input variables (in the present case, image pixels and their associated labels) are mapped onto a graph consisting of nodes, and edges between the nodes describe the conditional dependence between the nodes. Whereas a discrete classifier can predict a label without considering neighboring pixels, graphical models can take this spatial context into account, which makes them very powerful for classifying data with large spatial structure, such as images. Much work in learning with graphical models has focused on generative models that explicitly attempt to model a joint probability distribution *P*(***x***,***y***) over inputs, ***x***, and outputs, ***y***. However, this approach has important limitations for image classification where, not only is the dimensionality of ***x*** potentially very large, but also the features may have complex dependencies, such as the dependencies or correlations between multiple metrics derived from images. Therefore, modeling the dependencies among ***x*** is difficult and leads to unmanageable models, but ignoring them can lead to poor classifications.

A solution to this problem is a discriminative approach, similar to that taken in classifiers such as logistic regression. The conditional distribution *P*(***y***|***x***) is modeled directly, which is all that is required for classification. Dependencies that involve only variables in ***x*** play no role in *P*(***y***|***x***), so an accurate conditional model can have much simpler structure than a joint model, *P*(***x***,***y***). The posterior probabilities of each label are modeled directly, so no attempt is made to capture the distributions over ***x***, and there is no need to model the correlations between them. Therefore, there is no need to specify an underlying prior statistical model, and the conditional independence assumption of a pixel value given a label, commonly used by generative models, can be relaxed.

This is the approach taken by conditional random fields (CRFs), which are a combination of classification and graphical modeling known as structured prediction (*Lafferty et al., 2001; Sutton and McCallum, 2006*). They combine the ability of graphical models to compactly model multivariate data (the continuum of landcover and landuse labels) with the ability of classification methods to leverage large sets of input features, derived from imagery, to perform prediction. In CRFs based on ‘local’ connectivity, nodes connect adjacent pixels in ***x*** (*Lafferty et al., 2001*; *Kumar and Herbert, 2006*), whereas in the fully connected definition, each node is linked to every other (*Tappen et al., 2007*; *Krahenbuhl and Koltun, 2011*). CRFs have recently been used extensively for task-specific predictions such as in photographic image segmentation (*Zhu et al., 2001*; *He et al., 2004*; *Chen et al., 2016*; *Garcia-Garcia et al., 2017*) where, typically, an algorithm estimates labels for sparse (i.e. non-contiguous) regions (i.e. supra-pixel) of the image. The CRF uses these labels in conjunction with the underlying features (derived from a photograph), to draw decision boundaries for each label, resulting in a highly accurate pixel-level label image (*Krahenbuhl and Koltun, 2011; Garcia-Garcia et al., 2017*).

**1.4. Paper purpose, scope, and outline**

In summary, the purpose of this paper is to evaluate the utility of DCNNs for both image recognition and semantic segmentation of images of natural landscapes. Whereas previous studies have demonstrated the effectiveness of DCNNs for classification of features in satellite imagery, we specifically use examples of high-vantage and nadir imagery that are commonly collected during geomorphic studies. In addition, whereas many previous studies have utilized DCNNs either specifically designed to recognize landforms, land cover or land use, or trained existing DCNN architectures from scratch using a specific dataset, the comparatively simple approach taken here is to repurpose an existing DCNN to a specific task. Previous studies have tended to use relatively large DCNN architectures, whereas here we use comparatively small, very fast Mobilenets v.2 framework. Further, we demonstrate how structured prediction using a fully connected CRF, can be used, in a semi-supervised manner, to efficiently generate ground truth label imagery and DCNN training libraries. Finally, we propose a hybrid method for accurate semantic segmentation based on combining 1) the recognition capacity of DCNNs to classify small regions in imagery, and 2) the ﬁne grained localization of fully connected CRFs for pixel level classification.

The rest of the paper is organized as follows. First, we outline the CRF method, and its use in the generation of ground truth label images and DCNN training libraries. Then we detail the transfer learning approach taken to DCNN model repurposing, and how DCNN model predictions on small regions of an image may be used in conjunction with a CRF for semantic classification. Four data sets for image classification are introduced. The first is a large satellite data set consisting of various natural land covers and landforms, and the final three are from high-vantage or aerial imagery. Those three are also used for semantic classification. In either case, some data is used for training the DCNN, and some for testing classification skill (out-of-calibration validation). For each of the datasets, we evaluate the ability of the DCNN to classify regions of images or whole images correctly. The skill of the semantic segmentation is assessed. Finally, we discuss the utility of these findings and broader application of these methods for geomorphic research, before conclusions are drawn.

**2. Methods**

**2.1. Fully connected Conditional Random Field**

We use the fully connected CRF approach detailed in *Krahenbuhl and Koltun (2011)*, which is summarized briefly below. The probability of a labeling ***y*** given the image-derived features, ***x***, is

where θ are a set of hyperparameters, *Z* is a normalization constant, and *E* is an energy function that is minimized, obtained by

where *i* and *j* are pixel locations in the horizontal and vertical dimensions. The vectors *fi* and *fj* are features created from ***x***, which are a function of both the relative position as well as intensities of the image pixels. The term indicate so-called ‘unary potentials’, which depend on the label describing the image associated with a label at a single pixel location, whereas ‘pairwise potentials’, , depend on the labels of the image at a pair of separated pixel locations. The unary potentials represent the cost of assigning label *yi* to grid node *i*. The pairwise potentials are the cost of simultaneously assigning label *yi* to grid node *i* and *yj* to grid node *j*, and are deﬁned:

where *l* = 1:*L* are the number of features derived from ***x***, and where the function Λ quantiﬁes label ‘compatibility’, by imposing a penalty for nearby similar grid nodes that are assigned different labels. Each is the sum of two Gaussian kernel functions that determines the similarity between connected grid nodes by means of a given feature

The first Gaussian quantifies the observation that nearby pixels, with a distance controlled by θα (standard deviation for the location component of the color-dependent term), with similar color, with similarity controlled by θβ (standard deviation for the color component of the color-dependent term), are likely to be in the same class. The second Gaussian is a ‘smoothness’ kernel that removes small isolated label regions, according to θγ, the standard deviation for the location component. This penalizes small pieces of segmentation that are spatially isolated, enforcing more spatially consistent classification.

**2.2. Generating DCNN training libraries**

We developed a user-interactive program that segments an image into smaller chunks. On each chunk, cycling through a pre-defined set of classes, the user is prompted to draw (using the cursor) example regions of the image that correspond to each label. Unary potentials are derived from these manual on-screen annotations on imagery. Using this information, the CRF algorithm estimates the class of each pixel in the image (Fig. 1). Finally, the image is divided up into tiles of a specified size, *T*. If the proportion of pixels within the tile is greater than a specified amount, *Pclass*, then the tile is written to a file in a folder denoting its class. We prepared a video of this process that is included as Supplemental data A. This simultaneously and efficiently generates both ground-truth label imagery (to evaluate classification performance) and sets of data suitable for training a DCNN. A single photograph typically takes 5-30 minutes to process in this way, depending on the complexity and size of the image, so all the data required to retrain a DCNN (see section below) may take only up to a few hours to generate.

**2.3. Retraining a deep neural network (transfer learning)**

Among many suitable popular and open-source frameworks for image classification using deep convolutional neural networks, we chose MobileNets v. 2.0 (*Sandler et al., 2018*) because it is relatively small and efficient (computationally faster to train and execute) compared to many of its competitor architectures designed to be transferable for generic image recognition tasks, such as Inception (*Szegedy et al., 2016*), Resnet (*He et al., 2016*), and NASnet (*Zoph et al., 2017*), and it is smaller and more accurate than Mobilenets v. 1.0 (*Howard et al., 2017*). It also is pretrained for various tile sizes (image windows with horizontal and vertical dimensions of 96, 128, 192, and 224 pixels) which allows us to evaluate the effect of that on classifications. However, all of the aforementioned models are implemented within Tensorflow-Hub (*Tensorflow-Hub, 2018*), which is a library specifically designed for reusing pre-trained TensorFlow (*Abadi et al., 2015*) models on new tasks. Therefore, the interested reader may modify our data and code (provided as Supplemental data X) and explore variation in classification accuracies among multiple DCNN architectures.

For all datasets, we only used tiles (in the training and evaluation) where 90% of the tile pixels were classified as a single class (that is, *Pclass* > 0.9). This avoided including tiles depicting mixed landcover/use classes. We chose tile sizes of *T* = 96x96 pixels and *T* = 224x224 pixels, which is the full range available for Mobilenets, in order to compare the effect of tile size. All model training was carried out in python using Tensorflow library version 1.7.0 and Tensorflow-hub version 0.1.0. For each dataset, model training parameters (1000 training steps, and a learning rate of 0.01) were kept constant, but not necessarily optimal. For most data sets, there are relatively small numbers of very general classes (water, vegetation, etc) which in some ways is a more difficult classification task than much more specific classes, owing to the greater within-class variability to be expected from having broadly defined categories.

**2.4. CRF-based semantic segmentation**

We developed a method that harnesses the classification power of the DCNN, with the discriminative capabilities of the CRF, for pixel-scale semantic segmentation of imagery. An input image is windowed into small regions of pixels, the size of which is dictated by the size of the tile used in the DCNN training (here, *T*=96x96 or *T*=224x224 pixels). Some windows, ideally with an even spatial distributed across the image, are classified with a trained DCNN. Collectively, these predictions serve as unary potentials (known labels) for a CRF to build a probabilistic model for pixelwise classification given the known labels and the underlying image (Fig. 2).

Adjustable parameters are 1) the proportion of the image to estimate unary potentials for (controlled by both *T* and the number/spacing of tiles), and 2) a threshold probability, *Pthres*, larger than which a DCNN classification was used in the CRF. Across each dataset, we found that using 50% of the image as unary potentials, and *Pthres* = 0.5, resulted in good performance. CRF hyperparameters were also held constant across all datasets. We found that good performance across all datasets was achieved using θα= 60, θβ = 5, and θγ = 60. Holding all of these parameters constant facilitates comparison of the general success of the proposed method, however, for individual datasets it should be noted that accuracy could be further improved by optimizing the parameters for that specific data. This could be achieved by minimizing the discrepancy between ground truth label images and model-generated estimates using a validation data set.

**2.5. Metrics to assess classification skill**

Standard metrics of precision, *P*, recall, *R*, accuracy, *A*, and F1 score, *F*, are used to assess classification of image regions and pixels. Where *TP*, *TN*, *FP*, and *FN* are, respectively, the frequencies of true positives, true negatives, false positives, and false negatives:

True positives are image regions/pixels correctly classified as a certain class by the model and false negatives are regions/pixels correctly classified. False positives are those regions/pixels correctly classified as not belonging to a certain class. Precision and recall are useful where the number of observations belonging to one class is significantly lower than those belonging to the other classes. These metrics are therefore used in evaluation of pixelwise segmentations, where the number of pixels corresponds to each class vary considerably. The F1 score is an equal weighting of the recall and precision and quantifies how well the model performs in general.

**3. Data**

The chosen data sets that encompass a variety of environments (coastal, fluvial and lacustrine) and collection platforms (oblique stationary cameras, oblique aircraft, nadir UAV, nadir satellite)

**3.1. NWPU**

To evaluate the Mobilenets v.2 DCNN with a conventional satellite-derived land use/land cover dataset, we chose the NWPU-RESISC45, which is a publicly available benchmark for REmote Sensing Image Scene Classification (RESISC), created by Northwestern Polytechnical University (NWPU). The entire dataset, described by *Cheng et al. (2017)*, contains 31,500 high-resolution images from Google Earth imagery, in 45 scene classes with 700 images in each class. The majority of those classes are urban/anthropogenic. We chose to use a subset of 11 classes corresponding to natural landforms and landcover (Fig. 3), namely: beach, chaparral, desert, forest, island, lake, meadow, mountain, river, sea ice, and wetland. All images are 256x256 pixels. We randomly chose 350 images from each class for DCNN training, and 350 for testing.

**3.2. Seabright beach, CA.**

The dataset consists of 13 images collected from a fixed-wing aircraft in February 2016, of which a random subset of seven were used for training, and six for testing. Training and testing tiles were generated for seven classes (Table S1 and Fig. 2, 3, 4).

**3.3. Lake Ontario, NY.**

The dataset consists of 48 images obtained from a camera mounted on a 3DR Solo quadcopter, a small unmanned aerial system (UAS) in July 2017, flying 80-100 meters above ground level, in the vicinity of Braddock Bay, New York, on the shores of southern Lake Ontario. A random subset of 24 were used for training, and 24 for testing. Training and testing tiles were generated for X classes (Table S2 and Fig. 5).

**3.4. Grand Canyon, AZ.**

Riparian environment

The dataset consists of 14 images collected from a stationary autonomous camera system, from 7 sites, of which 1 from each site were used for training, and 1 from each site for testing. Imagery came from various seasons. The camera system, sites and imagery is described in *Grams et al. (2018)*. Table S3 and Fig. 6

**4. Results**

**4.1. DCNN transfer learning**

Table 2

Figures S1, S2, S3, S4

**4.2. CRF-based semantic segmentation**

Table 3

Figures 7 and 8

Figures S5, S6, S7

**5. Discussion and Conclusions**

DCNN-CRF technique works best for spatially extensive classes

Average F1 score correlates fairly well with average spatial extent of label (Figure 8)

**Acknowledgments**

Thanks …

All code and data available at …

**References**

1. Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., and 35 others. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/about/bib Software available from <https://www.tensorflow.org>
2. Almar, R., Larnier, S., Castelle, B., Scott, T., 2016. On the use of the radon transform to estimate longshore currents from video imagery. Coastal Engineering 114, 301–308.
3. Belgiu, M. and Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, pp.24-31.
4. Benacchio, V., Piégay, H., Buffin-Bélanger, T. and Vaudor, L., 2017. A new methodology for monitoring wood fluxes in rivers using a ground camera: Potential and limits. Geomorphology, 279, pp.44-58.
5. Bergsma, E., Conley, D., Davidson, M., O’Hare, T., 2016. Video-based nearshore bathymetry estimation in macro-tidal environments. Marine Geology 374, 31–41.
6. Bertoldi, W., Zanoni, L. and Tubino, M., 2010. Assessment of morphological changes induced by flow and flood pulses in a gravel bed braided river: The Tagliamento River (Italy). Geomorphology, 114(3), pp.348-360.
7. Y. Boykov, and V. Kolmogorov. An experimental comparison of mincut/max-ﬂow algorithms for energy minimization in vision. IEEE T. Pattern Anal., vol. 26, no. 9, pp. 1124—1137, 2004.
8. Chen, L.C., Yang, Y., Wang, J., Xu, W. and Yuille, A.L., 2016. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3640-3649).
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L., 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), pp.834-848.
10. Cheng, G., Han, J. and Lu, X., 2017. Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 105(10), pp.1865-1883.
11. Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S. and Bengio, Y., 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Advances in neural information processing systems (pp. 2933-2941).
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.
13. Franklin, S.E. and Wulder, M.A., 2002. Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. Progress in Physical Geography, 26(2), pp.173-205.
14. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857, 2017.
15. Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., 2016. Deep learning (Vol. 1). Cambridge: MIT press.
16. Grams, P.E., Tusso, R.B. and Buscombe, D., 2018. Automated remote cameras for monitoring alluvial sandbars on the Colorado River in Grand Canyon, Arizona (No. 2018-1019). US Geological Survey Open File Report.
17. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
18. Holman, R.A., Stanley, J., 2007. The history and technical capabilities of Argus. Coastal engineering 54, 477–491.
19. Hoonhout, B., Radermacher, M., Baart, F., Van der Maaten, L., 2015. An automated method for semantic classiﬁcation of regions in coastal images. Coastal Engineering 105, 1–12.
20. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
21. Hu, F., Xia, G.S., Hu, J. and Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing, 7(11), pp.14680-14707. 8-56.
22. P. Krahenbuhl, and V. Koltun. Efﬁcient inference in fully connected CRFs with Gaussian edge potentials. In Adv. Neur. In. (NIPS), Granada, Spain, pp. 109—117, Dec. 2011.
23. S. Kumar, and M. Hebert. Discriminative random ﬁelds. Int. J. Comput. Vision, vol. 68, no. 2, pp. 179—201, Jun. 2006.
24. J. Lafferty, A. McCallum, and F. Pereira. Conditional random ﬁelds: Probabilistic models for segmenting and labeling sequence data. Int. Conf. Mach. Learn. (ICML), Williamstown, USA, pp. 282—289, July 2001.
25. M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, “Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks,” Remote Sensing, vol. 8, no. 4, pp. 329, 2016.
26. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature, 521(7553), p.436.
27. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B. and Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis, 42, pp.60-88.
28. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y. and Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing, 234, pp.11-26.
29. Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).
30. Lu, H., Fu, X., Liu, C., Li, L.G., He, Y.X. and Li, N.W., 2017. Cultivated land information extraction in UAV imagery based on deep convolutional neural network and transfer learning. Journal of Mountain Science, 14(4), pp.731-741.
31. Ma, L., Li, M., Ma, X., Cheng, L., Du, P. and Liu, Y., 2017. A review of supervised object-based land-cover image classification. ISPRS Journal of Photogrammetry and Remote Sensing, 130, pp.277-293.
32. Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 55(2), pp.645-657.
33. D. Marmanis, M. Datcu, T. Esch, and U. Stilla, “Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks,” IEEE Geosci. Remote Sens. Lett., vol. 13, no. 1, pp. 105-109, 2016.
34. Mulder, V.L., De Bruin, S., Schaepman, M.E. and Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping—A review. Geoderma, 162(1-2), pp.1-19.
35. O’Connor, J., Smith, M.J. and James, M.R., 2017. Cameras and settings for aerial surveys in the geosciences: Optimising image data. Progress in Physical Geography, 41(3), pp.325-344.
36. Palafox, L.F., Hamilton, C.W., Scheidt, S.P. and Alvarez, A.M., 2017. Automated detection of geological landforms on Mars using Convolutional Neural Networks. Computers & geosciences, 101, pp.4
37. Olga Russakovsky\*, Jia Deng\*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (\* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.
38. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C., 2018. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. arXiv preprint arXiv:1801.04381.
39. Sekovski, I., Stecchi, F., Mancini, F. and Del Rio, L., 2014. Image classification methods applied to shoreline extraction on very high-resolution multispectral imagery. International journal of remote sensing, 35(10), pp.3556-3578.
40. Smith, M.J. and Pain, C.F., 2009. Applications of remote sensing in geomorphology. Progress in Physical Geography, 33(4), pp.568-582.
41. Sturdivant, E.J., Lentz, E.E., Thieler, E.R., Farris, A.S., Weber, K.M., Remsen, D.P., Miner, S. and Henderson, R.E., 2017. UAS-SfM for Coastal Research: Geomorphic Feature Extraction and Land Cover Classification from High-Resolution Elevation and Optical Imagery. Remote Sensing, 9(10), p.1020.
42. Su, L., Gibeaut, J., 2017. Using UAS hyperspatial RGB imagery for identifying beach zones along the south texas coast. Remote Sensing 9, 159.
43. Sutton, C. and McCallum, A., 2006. An introduction to conditional random fields for relational learning (Vol. 2). Introduction to statistical relational learning. MIT Press.
44. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2818-2826).
45. Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A., 2017, February. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI (Vol. 4, p. 12).
46. M. Tappen, C. Liu, E. Adelson, and W. Freeman. Learning Gaussian conditional random ﬁelds for low-level vision. IEEE. Conf. Comp. Vision (CVPR), Minneapolis, USA, pp. 1—8, June 2007.
47. Tensorflow-Hub, 2018. <https://www.tensorflow.org/hub/modules/image>. Accessed: 5/23/2018
48. Turner, I.L., Harley, M.D., Drummond, C.D., 2016. UAVs for coastal surveying. Coastal Engineering 114, 19–24.
49. Volpi, M. and Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 55(2), pp.881-893.
50. Warrick, J.A., Ritchie, A.C., Adelman, G., Adelman, K., Limber, P.W., 2016. New techniques to measure cliﬀ change from historical oblique aerial photographs and structure-from-motion photogrammetry. Journal of Coastal Research 33, 39–55.
51. H. Zhu, F. Meng, J. Cai, and S. Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. J. Vis. Commun. Image R., vol. 34, pp .12—27, 2016.
52. Zoph, B., Vasudevan, V., Shlens, J. and Le, Q.V., 2017. Learning transferable architectures for scalable image recognition. arXiv preprint arXiv:1707.07012.

**Tables**

Table 1: Out-of-calibration whole tile classification accuracies and F1 scores for each data set and tile size

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***T* = 96** | | ***T* = 224** | |
| **Data set** | **Mean accuracy** | **Mean F1 score** | **Mean accuracy** | **Mean F1 score** |
| 1. NWPU | 87% | 93% | 89% | 94% |
| 2. Seabright | 94% | 97% | 96% | 97% |
| 3. Ontario | 83% | 91% | 96% | 98% |
| 4. Grand Canyon | 92% | 96% | 94% | 97% |

Table 2: Mean out-of-calibration whole tile classification accuracies (%), per class, for each of the non-satellite data sets (T=96 / T=224)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Seabright** | **Ontario** | **Grand Canyon** | **Mean** |
| **Sediment/sand** | 0.93 / 0.98 | 0.76 / 0.93 | 0.94 / 0.89 | 0.89 |
| **Shadow** |  |  |  |  |
| **Terrain/rock** | 0.91 / 0.91 | 0.78 / 0.91 | 0.89 / 0.95 | 0.84 |
| **Vegetation** | 0.89 / 0.95 | 0.96 / 0.98 | 0.94 / 0.90 | 0.94 |
| **Water** | 0.99 / 0.98 | 0.94 / 0.97 | 0.92 / 0.99 | 0.94 |
| **Wood** |  |  |  |  |
| **Anthropogenic** | 0.95 / 0.98 | 0.72 / 0.94 |  | 0.90 |
| **Foam** | 0.97 / 0.96 |  |  |  |
| **Road** | 0.96 / 0.98 |  |  |  |

Table 3: Mean out-of-calibration P/R/F/A (all %) per class for pixelwise classifications using each of the non-satellite data sets (T=96)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Seabright** | **Ontario** | **Grand Canyon** |
| **Sediment/sand** | 98/92/95/92 | 72/72/74/67 | 76/79/80/78 |
| **Shadow** |  |  |  |
| **Terrain/rock** | 44/51/46/50 | 32/32/30/41 | 80/97/87/96 |
| **Vegetation** | 63/41/48/42 | 90/93/89/91 | 92/31/46/43 |
| **Water** | 95/92/93/91 | 95/95/95/89 | 94/92/93/94 |
| **Wood** |  |  |  |
| **Anthropogenic** | 87/95/90/94 | 78/59/64/55 |  |
| **Foam** | 87/93/90/94 |  |  |
| **Road** | 86/81/83/79 |  |  |

**Supplemental Tables**

Table S1: Classes and number of tiles used for the Seabright data set

|  |  |  |
| --- | --- | --- |
| **Class** | **Number of training tiles (*T*=96/224)** | **Number of evaluation tiles (*T*=96/224)** |
| Anthropogenic | 23,566 / 4,548 | 15,575 / 3,031 |
| Road and pavement | 314 / 60 | 525 / 103 |
| Sand | 38,250 / 6,887 | 25,318 / 5,802 |
| Vegetation | 386 / 76 | 240 / 38 |
| Other terrain | 77 / 24 | 117 / 22 |
| Water | 11,394 / 1,723 | 14,360 / 2,251 |
| Foam | 5,076 / 735 | 5,139 / 843 |
| Total: | 76,063 / 14,053 | 61,274 / 12,090 |

Table S2: Classes and number of tiles used for the Ontario data set

|  |  |  |
| --- | --- | --- |
| **Class** | **Number of training tiles (*T*=96/224)** | **Number of evaluation tiles (*T*=96/224)** |
| Anthropogenic/buildings | 467 / 219 | 3,216 / 333 |
| Sediment | 2,856 / 289 | 3,758 / 407 |
| Vegetation | 33,871 / 5,139 | 33,421 / 5,001 |
| Other terrain | 1,596 / 157 | 1,094 / 92 |
| Water | 80,304 / 13,332 | 77,571 / 12,950 |
| Total: | 119,094 / 19,136 | 119,060 / 18,783 |

Table S3: Classes and number of tiles used for the Grand Canyon data set

|  |  |  |
| --- | --- | --- |
| **Class** | **Number of training tiles (*T*=96/224)** | **Number of evaluation tiles (*T*=96/224)** |
| Rock/scree/terrain | 15,059 / 2,405 | 12,151 / 1,999 |
| Sand | 751 / 39 | 1,069 / 91 |
| Riparian vegetation | 2,971 / 408 | 2,158 / 305 |
| Water | 8,568 / 1,462 | 5,277 / 1,130 |
| Total: | 27,349 / 4,314 | 20,655 / 3,525 |

Table S3: Classes and number of tiles used for the Grand Canyon data set

|  |  |  |
| --- | --- | --- |
| **Class** | **Number of training tiles (*T*=96/224)** | **Number of evaluation tiles (*T*=96/224)** |
| Beach | 48,281 / 7,451 | 53,172 / 8,250 |
| Anthropogenic/buildings | 46,155 / 7,390 | 52,194 / 8,462 |
| Cliff | 33,474 / 5,107 | 21,757 / 3,303 |
| Road | 7,152 / 936 | 4,050 / 477 |
| Sky | 42,370 / 6,683 | 26,240 / 4,267 |
| Surf/foam | 19,828 / 2,640 | 25,220 / 3,428 |
| Swash | 11,016 / 1,338 | 7,581 / 681 |
| Other terrain | 107,506 / 17,179 | 59,261 / 9,198 |
| Vegetation | 119,242 / 19,818 | 56,111 / 8,819 |
| Water | 145,383 / 21,610 | 78,779 / 11,142 |
| Total: | 580,407 / 90,152 | 384,365 / 58,027 |