Automated segmentation and classification of land use from overhead imagery

Richard A, Benbihi A, Pradalier C, Perez V, Durand P, Van Couwenberghe R

Georgia Tech Lorraine, UMI 2958 GT-CNRS, Metz, France
Université de Lorraine, AgroParisTech, INRA, Silva, Nancy, France

A paper from the Proceedings of the
14th International Conference on Precision Agriculture
June 24 – June 27, 2018
Montreal, Quebec, Canada

Abstract.

Reliable land cover or habitat maps are an important component of any long-term landscape planning initiatives relying on current and past land use. Particularly in regions where sustainable management of natural resources is a goal, high spatial resolution habitat maps over large areas will give guidance in land-use management. We propose a computational approach to identify habitats based on the automated analysis of overhead imagery. Ultimately, this approach could be used to assist experts, policy and decision makers who promote sustainable agroecology by evaluating habitat services and prioritizing land uses.

The overall objective of our project is to classify the evolution of land usage since the advent of aerial imagery. In this paper, our goal is to bring automatic habitat classification to the level achieved by a human expert performing a high spatial resolution classification. This classification consists in identifying habitats such as hedges, lakes, fields, pastures or forests. Therefore, we train a machine vision algorithm to segment an overhead imagery into a dozen of expert-specified land use classes. Relying on the recent developments in machine learning, and in particular deep learning, the best machine vision model appears to be convolutional neural networks (e.g. SegNet, DeepLab).

The training was performed using data from a hand-labelled high-resolution (0.5m/pixel) database around the Orne River (Moselle, France – 2000km²). Aerial orthophoto are available for two time periods: 2015 and 1955. In addition, we also generated artificial 1955 data from 2015 imagery and used them as learning base for the 1955 imagery as the data available in 2015 provides more quantity and more diversity.

The paper highlights the performances of these state-of- the-art machine learning algorithms for land use recognition and segmentation. It shows their potential in the context of studies in environment sciences and environmental decisions. The automatic approach presents an alternative for detailed and accurate land cover maps acquired manually, which are labor intensive and time consuming.

The paper also illustrates the potential benefits of generating artificial imagery to pre-train the machine vision model and requires less annotated data. This approach may prove useful for time periods where there is few labeled data.

Keywords. *pixel-wise classification, deep learning, environment monitoring, conservation ecology*

Introduction

Land classification is used in many applications that need to monitor the land state and its changes such as environment monitoring, agriculture and town planning. For instance, urban master plans rely on the evolution of such classification maps to assess the ecological impact and the global carbon budget of new town projects. These maps are usually built from aerial or satellite images which a human expert annotates with the land category such as meadows, fields, hedges. This task is tedious and prone to ambiguities even for geomatics specialists given the image high resolution of up to 10cm/pixel and the expert hand and eye precision. Furthermore, it is very time consuming: our human expert takes about 48 hours to annotate an aerial image of 4 square kilometers at 50 centimeters per pixel resolution.

An alternative is to use machine vision to automate this task. There is wide literature on the automatization of land map classification reviewed by Ma, L. et al (2017). Our method differs in that it can process images at the highest current resolution of 50 cm per pixel (Madden, M. 2009), classify at least 15 land categories against up to 11 categories for most of the state-of-art (Ma, L. et al 2017). Also, it uses the most recent machine vision techniques involving Deep Convolutional Neural Network (DCNN) which only needs RGB images contrary to state-of-the-art methods that rely on multi-spectral images. Recent works (Sherrah, J. (2016)., Scott, G. J. et al. (2017), Liu, T., and Abd-Elrahman, A. (2018)) have proven the performance of DCNN for land classification. We extend it to cover more land categories with category definitions that not only depend on their visual appearance but also on the town plan definitions.

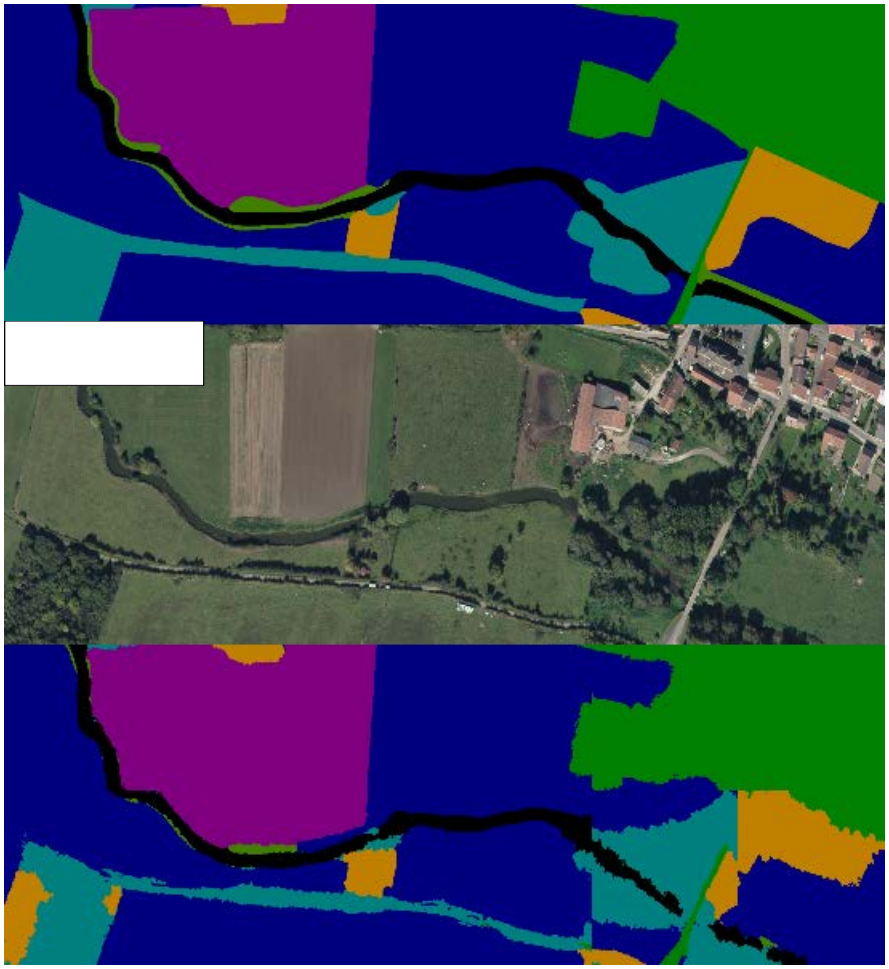

Fig 1. An example of classification map autonomously generated by a deep convolutional neural network. Top-down: photointerpreter work or expert classification, input image, DCNN output Map legend: (constructed areas: green), (surface waters: black), (Heatlands and scrub: orange), (grassland: blue), (arable land: purple), (broadleaved trees: cyan)

Most of the state-of-the-art for land classification uses multi-spectral images to manually compute

discriminative features such as NDVI between land categories (Bhandari, A. K. et al. 2012). Such features capture chemical and biological properties which are then used to discriminate between different vegetation species (Yu, Q. et al. 2006). However, these features can also introduce variability when they should not. The image surveys of a given area with different climatic conditions will have different feature maps. For example, the near infrared feature map of the same area before and after rain changes a lot. The use of DCNN solves several issues regarding multi-spectral images. First, the DCNN autonomously learns the relevant features to discriminate between given classes so there is no need to manually compute features from spectral images. Also, our experiments show that RGB channel images are enough for the network to discriminate between classes, so data acquisition can be done with a common vision camera. This also implies that DCNN classification can be used on images acquired before the launch of monitoring satellites. This make DCNN agnostic to the image acquisition method as most modern images have at least RGB channels. We show that the performance of the DCNN does not depend on the data acquisition method by running experiment on data acquired in 2015 and data acquired in 1955. Another advantage of RGB images is that they cost less to acquire than multi-spectral images, which solves one the challenge for airborne land monitoring.

The type of sensor data required for land classification is another challenge faced by airborne surveying. Recent works (Bryson, M. et al. 2010, Hung, C. et al. 2014) aims at replacing the data usually collected with satellite imagery with data collected with UAVs. However, multi-spectral sensors are typically used in satellite and the associate cost is high compared to common vision camera. Since our method only requires RGB images, the satellite image acquisition can be replaced with UAV surveying. This reduces the data acquisition cost and simplifies the collection of high resolution images. However, for this paper, we follow the guidelines of Franklin, S. E and Wulder, M. A. (2002) and rely on a regional geomatics partner (IGN 2016) who provides rectified and aligned images through time. The images are recorded with a plane equipped with a camera which technology depends on the years. The image acquisition and processing method is further described in the next section.

Previous work on autonomous land classification relied on classic machine learning methods such as random forests (Rodriguez-Galiano et al. 2007) and support vector machines (Huang, C. et al. 2002). Even though these methods can classify many categories with good accuracy, they rely on hand-crafted features computed from the multi-spectral images. DCNN bypasses this step by autonomously learning features only from RGB images. Also, at the time of this writing, state-of-the-art performance for most machine vision problems are achieved by DCNN. Scott et al. 2017 uses DCNN to classify the aerial image of the UCM dataset (Yang, Y. and Newsam, S (2010)) which are made of some land categories but also of object classes such as airplanes, buildings or parking lots. These categories differ from ours as described in the next section. One challenge of our dataset is that some categories are more difficult to discriminate than others: for example, it is easier to distinguish coniferous trees from constructed sites than from broadleaved trees. Another challenge is that some categories have a very sparse pixel distribution such as riparian groves. Also, the category definition not only relies on its visual appearance but also on definition set by town-planners. For example, a group of trees may be classified either as a chopping area, a broadleaved tree forest or a coniferous tree forest. To the author's knowledge, this is the first work that uses state-of-the-art DCNN for agricultural land classification with such refined category definition.

Methods and material

Data collection

The images are collected above the Orne watershed in the Grand Est region which covers $1200\ km^2$ in the north east of France. It is made of two main type of lands: a clay basin of the Woëvre with mainly cultures and forests, and a lime plateau of the Pays-Haut region with many cities and extraction sites. (French national Institute of Geographical and forestry Information (IGN)) surveys the region regularly and provided us with the 1955 and 2015 datasets. They are

made of 10.000x10.000 pixels ortho-images: this means that the terrain elevation is artificially removed. 2015 images were acquired with a plane equipped with a multi-spectral camera. For each image, we are provided with red, green, blue and near-infrared channels even though DCNN can classify land maps only with RGB channels. The pictures from 1955 are analog black and white photography that have been digitized and corrected. This yields a similar resolution as the 2015 images but the images are blurry, and a typical analog photography grain can be seen. Also, the contrast and luminosity of the images are poor, and they contain large artifacts.

“Hand-made” mapping

Using geographic information system (GIS) software, we produced a land-use map based on the interpretation of ortho-images. Areal units were delineated, with a minimum mapping unit of $2500\ m^2$. Linear elements such as roads, rivers and vegetation features were delineated when their width was greater than 4 m. Linear elements of woody nature, such as hedgerows, were also mapped when they were wider than 10 m.

The classification methodology followed a hierarchical interpretation key with 6 main classes: built areas, agricultural lands, moor and bushes, forest areas, and hydrographical areas. All the classes were declined in subclasses with up to 4 hierarchical levels. Each class was described by supporting text and images. The key was built to study the impact of land use on water quality and therefore, it was focused on classes of wetland. For example, it included classes such as hems and riparian groves. A total of 25 classes were retained. In addition to the images, we also used water network maps (Sandre 2018), forest maps (IGN forest 2018), culture maps (IGN RPG 2018) and Google Street View to better classify the units.

The nomenclature of the units was inspired by the EUNIS habitat classification (LOUVEL, Justine and GAUDILLAT, Vincent 2013), which is a pan-European system describing habitats across Europe. This system complies with the European Environment Agency guidelines.

In average, it took 8 hours to annotate a 10 000x10 000 images with a resolution of 50 cm per pixel.

Data processing

Our objective is for the DCNN to classify even the smallest scale land elements, so it must be able to detect small details such as hedges, transportation infrastructures and rivers. These details can be extremely narrow with a width of less than 2 meters. The DCNN must also find much larger structures such as quarries, or commercial centers, which can easily exceed hundreds of meters in size. So we chose to keep the original resolution of the provided images of 50 centimeters per pixel so as to allow the DCNN to detect small enough objects in the image. The size of the images fed to the DCNN is set to 300x300 pixels as it seems to be the best tradeoff between images small enough to fit on the GPU RAM and large enough to have a sufficient field of view.

Some of these 25 originals classes show visual similarity and the only thing that discriminates them is their direct neighboring environment. Therefore, we aggregate some of the classes. Some classes are a combination of two classes: mixed trees are a mix of coniferous and deciduous trees. Such classes are removed from the dataset and their occurrences are blanked so that they do not confuse the networks as they are a combination of multiple classes. Other classes, such as gardens and sport infrastructures are removed and blanked as they also confuse the DCNN due to their poor representation and similarity with other classes such as meadows. In the end the datasets used to train the DCNN has a total of 14 classes (Table 1).

Table 1: Class categories

0	Littoral zone of inland surface waterbodies	7	Coniferous woodland
1	Surface standing waters	8	Tree farms
2	Constructed areas	9	Fruit orchards
3	Extractive industrial sites	10	Riparian vegetation
4	Grasslands	11	Heathlands, scrub and tundra
5	Arable lands	12	Chopping areas
6	Broadleaved woodland	13	Vineyards

The datasets are not originally intended to be used as a learning base. As such, the distribution of pixels per classes contained in the whole datasets is extremely heterogeneous. We overcome this issue with data augmentation. In our case, the scale of the images never changes, the quality of the images is always the same and the parameters used to take the pictures are invariant. This means that augmentation techniques that take part of the images, such as crops, are not relevant. Techniques that relies on saturation changes, image and color distortion cannot be applied. This leaves us (Wang, J., & Perez, L.) with isometric transformations such as rotations, translations or image flips. However, translations are not used as they would require masking. The remaining transformations keep the image visually unaltered while making them different for the network as they cannot be made using combinations of linear operations. We use an online optimization process to select which image to augment. It minimizes a cost function based on the variance in pixel per class distribution in the whole dataset. If an image is suitable for augmentation, the process will augment it till it is no longer suitable, or if it runs out of transformations to apply. The results can be found in Table 2, as can be observed, the overall distribution is more homogeneous. This result is achieved with a maximum of seven transformations applicable by the algorithm. The more operations available, the better the result.

***T*able 2: Class distribution in the 1955 dataset before and after augmentation.**

	Class distribution											
Before augmentation (%)	1.2	1.1	1.7	0.3	41.7	42.3	4.1	1.3	0	3.3	0.3	2.1
After augmentation (%)	4.0	4.1	8.2	1.1	27.5	26.9	10	4.8	0	7.0	0.5	5.4

The augmentation process is applied on both the 2015 dataset and the 1955 dataset. As mentioned earlier, the pictures from 1955 are analog black and white photography that have been digitized and corrected. They yields a similar resolution to the one of 2015. However, the details are not as sharp: they are slightly blurred, and a typical analog photography grain can be seen. Also, the contrast and luminosity of the image is poor. Figure 2 illustrates those issues: the edges of the houses are hardly distinguishable from roads because of the poor contrast, the textures in the trees are completely blurred, and the grain can be seen in the meadows and fields. In some cases, rare artifacts can be found. These images are taken in grey scale and the digitized version covers 255 levels of grey.

Fig 2. Difference between 1955 images and 2015.

The generation of the datasets are made by manually selecting some large labeled images (10000x10000pixels) for the validation set and leaving the rest to the training set. Once those images were selected, an algorithm extracts learning data by taking patches of 300x300 pixels images in the large images of 10000x10000pixels striding from left to right, top to bottom by 150 pixels.

Classification

State of the art performance in dense classification is achieved with Deep Convolutional Neural Networks. DeepLab (Chen, L. C. et al. 2018) and SegNet (Badrinarayanan, V. et al. 2017) are two specific network architectures that show the best performances on indoor and outdoor scene dense classification. This section describes the specifics of each architecture and how to generalize them for ortho-image dense classification. The training method is invariant to the choice of the network architecture.

A DCNN is a stack of convolutional filters that can be seen as visual filters which each reacts to some specific feature of the images. Rather than hand-crafting the relevant image features for the classification and the corresponding filter, the DCNN autonomously learns the relevant filters to compute the features that best discriminates the land categories. DeepLab and SegNet are originally trained on two bench datasets of indoor and outdoor scenes PASCAL VOC12 (Everingham, M. Et al. 2010) and CityScapes (Cordts, M. et al. 2016). They prove to generalize well to ortho-image classification. DeepLab and SegNet outstanding performances rely on specific arrangement of convolutional filters which are described next.

Network Architectures

DeepLab is originally made of three parallel Resnet (He, K. et al. 2016) which each process a scaled version of the input image. The scaled outputs are then fused to produce the final result: this allows the network to decide how much attention to pay to features at different positions and scales (Chen, L. C et al. 2016). Resnet uses skip connections between feature maps: this has the advantage to better propagate the information along the network even when there is a high number of filters: the deeper the network, the more complex filters it can build. Given the complexity of the category definition, we chose a Resnet made of 101 convolutional filters. Another specificity of DeepLab is the use of atrous convolution and spatial pyramid pooling that allow the filters to have larger field of view. This is especially relevant to the land classification

dataset as the definition of a category may depend on the neighboring category: for example, riparian groves are always along water. The network also uses Conditional Random Fields (Krähenbühl, P. and Koltun, V. (2011)) to smoothen the classification around category boundaries.
All DeepLab specifics are used except the multi-scale processing. In the specific case of land classification of aerial images, it is better to keep the image elements at their original dimension as it may induce confusion for the network. For example, a zoomed-in bush may be misclassified as a broad leave tree. Also, this brings a threefold reduction of the size of the network: the network needs only 3GB of RAM instead of 8GB.

The second DCNN we are considering – SegNet – is made of less layers but has an encoder-decoder architecture: the encoder is the part of the network that builds a semantic representation of the image in a low spatial dimensional space and the decoder is the part that projects this representation back into the image space. The encoder is made of stack of convolutional filters interleaved with max-pooling layers that can be seen as a high-pass filter: the feature maps go through that layer and only the features with the highest values are kept whereas the one with lower values are discarded. This allows building a feature representation of the image in a lower dimensional space than the image. It also increases the field of view of the features maps: one pixel of a feature map now holds information about its location and the location of its neighbor. This can be seen as the counterpart of the atrous convolution and spatial pyramidal pooling of DeepLab. To build the classification map, the decoder projects this semantic representation back into the image space. The dimension is augmented with unpooling layers, which are the inverse of the encoder pooling layer, and the projection is learned with convolutional layers.

Loss optimization

Instead of optimizing a classic cross-entropy loss as DeepLab does, SegNet minimizes a sum of weighted cross-entropy loss per class to account for the unbalanced distribution of pixel in one image. For example, let X_1 and X_2 be two images of 100 pixels with the following category distribution (Table 3). Let us say that the network has poor performance at classifying vineyards and misclassify all the pixels in that category with a penalty of 1 per misclassified pixel. With the classic loss, the network is optimized differently depending on the image distribution it is given: given the first image, the loss is only 10 so the network may think it is performing well. Given the second image, the loss becomes 90 so it knows that it must improve to better classify vineyards. The weighted loss is the same for both images which makes the optimization agnostic to the image distribution and only to the network performance. Whichever the image, the network knows that it is misclassifying the vineyard pixels.

Table 3. Example of weighted loss

Category	Vineyard pixels	Chopping areas pixel	Loss	Weights	Weighted loss
X_1	10	90	10	(5.0, 0.56)	50 = 5.0 * 10
X_2	90	10	90	(0.56,5.0)	50.4 = 90 * 0.56

The class weights are computed as follow. Let pix_c the number of pixels for the class c over the full dataset and im_c be the number of images of height *h* and width *w*, with pixels of class c. For each class c, we compute the average pixel density per image d_c and store into an array D. Then we compute the weight:

$$c = \frac{pix_c}{im_c * h * w} \quad (1)$$

$$D \triangleq [d_c]_c \quad (2)$$

$$w_c = \frac{median(D)}{d_c} \quad (3)$$

And for one pixel, the weighted loss is

$$L(x) = -\sum_{c} w_c * log\ P(c|x) \quad (4)$$

Validation method

Given the ground truth classification map and the network output, we compute two metrics: the accuracy and the Jaccard Index also known as mean Intersection Over Union (mIOU).

For a given class C, let TP be the number of true positive, FP the number of false positive and P the total number of pixel of C. The accuracy of class C is defined as

$$acc = \frac{TP}{TP+FP} (5)$$

The global accuracy is defined as the average of class accuracy. This value indicates how correct the model classification the class C (when TP increases) and whether the model confuses other classes with C (when FP increases).

For a given image and a given class C, let B_{gt} and B be the ground truth class boundary and the predicted one. The *mIOU* is defined as

$$mIOU = \frac{(B_{gt} \cap B)}{(B_{gt} \cup B)} (6)$$

This value indicates whether the model locates C correctly ($B_{gt} \cap B$) and whether the location is accurate ($B_{gt} \cup B$). The global mIOU is the average of class mIOU.

Results and discussions

Experiments

The networks are implemented using the Caffe library (Jia, Y. et al. 2014) and the open source implementations from the DeepLab and SegNet projects. The evaluation metrics are computed using the MATLAB code of DeepLab.

Each network is trained following the guidelines of its respective paper. Both networks reach convergence after 30 000 steps. DeepLab is trained using stochastic gradient descent (SGD) with a "poly" learning rate policy which means that the learning rate decreases following the rule:

$$\alpha \leftarrow \alpha * \left(1 - \frac{iter}{max_{iter}}\right)^{power} \quad (7)$$

The initial learning rate $is\ 2.5\ 10^{-4}$ and$power = 0.9$. The gradient is weighted with a weight decay of 0.0005 and the SGD momentum is set to 0.9. One gradient step is computed over a batch size of 8. The weights are initialized with the weights of the multi-scale Resnet of Chen, L. C. et al. 2018 trained on PASCAL VOC12.

SegNet is also trained using SGD but a constant learning rate of 1.10^{-3} and a batch size of 12. The weights are initialized with the weights of the VGG network (Simonyan, K. and Zisserman, A. (2014)) trained on the classification of the ImageNet dataset (Deng, J. et al. 2009).

2015 Results

Both models have the same performance on the 2015 dataset with a 73% global accuracy. DeepLab proves to better locate the class distribution with a 75% mIOU which is 10% higher than SegNet. Figure 3 show the confusion matrix M. This should be read as follows: M(i,j) is the proportion of pixels of class *i* classified as class *j*.

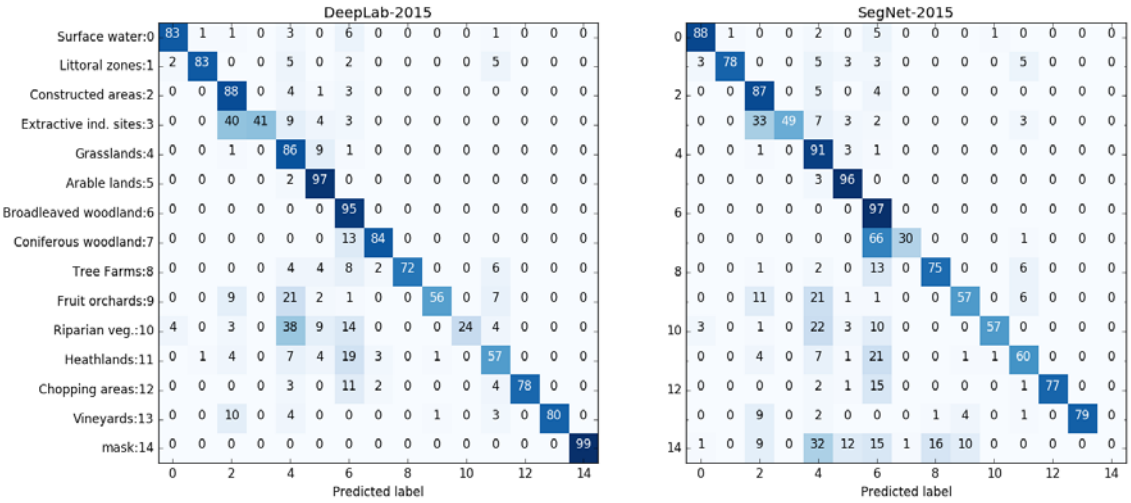

Fig 3. Per class accuracy.
Global accuracy DeepLab: 73% - Global accuracy SegNet: 73%

Table 4. Per class mIOU on the 2015 dataset.
Global mIOU DeepLab: 65% - Global mIOU SegNet: 57%

Model \ Class	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
DeepLab	79	75	77	39	77	95	92	74	56	46	19	44	73	63	98
SegNet	78	69	72	35	74	95	89	30	14	30	21	48	69	72	NA

The individual class accuracies are almost equal for both models except for the coniferous trees and the riparian groves. Both network reach accuracies near or above 80% for categories with many examples such as the as the constructed habitats (buildings of towns and villages) and water category (surface running or standing waters) or with very discriminative features such as vineyards. The networks exhibits different classification accuracy for coniferous trees and riparian vegetation. DeepLab's accuracy is 50% higher on coniferous woodlands and SegNet is 30% higher on riparian vegetation.

DeepLab confuses riparian vegetation (Riverine and fen scrubs) with grasslands or some cultivated agricultural habitats. This can be explained by the fact these classes are visually similar but riparian groves are spatially narrow (Figure 4) and the network did not get enough examples of this kind of habitats to learn how to differentiate them. Riparian vegetation is one of class with the lowest density per image. So, we assume that the cost sensitive loss of SegNet compensates the lack of representation per imageby assigning the second highest weight to the riparian vegetation class: the network is highly penalized when it misclassifies the riparian vegetation even though it sees these examples less than the grassland examples.

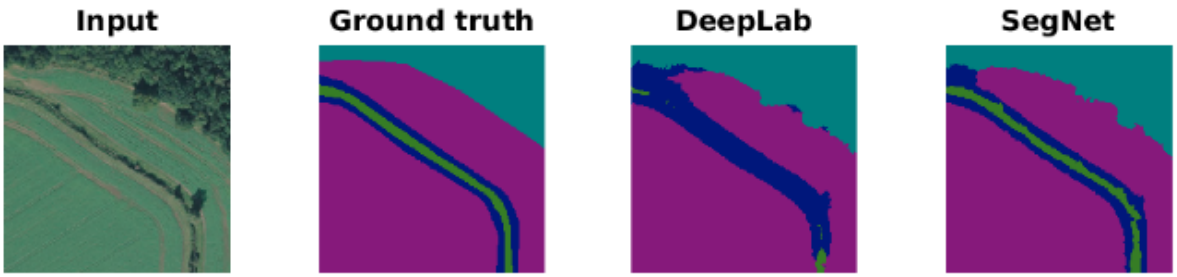

Fig 4. DeepLab does not learn to recognize riparian vegetation (green) because the pixel distribution is sparse and it

doesn't see enough examples. The cost-sensitive loss of SegNet solves this issue. Legend: (riparian vegetation:green), (grasslands: blue), (cultivated agricultural habitats: purple), (broadleaved woodland:cyan)

SegNet's cost-sensitive classification can compensate for unbalanced class pixel distribution but shows its limits when the variance of the pixel distributions is too high i.e. when the order of magnitude between over-represented and under-represented class is too high. For example, coniferous and broadleaved woodlands have similar appearance but there are 21 times more images with broadleaved trees than with coniferous ones. One would expect the weight for misclassifying coniferous woodlands to be higher than for broadleaved woodlands. However, the weight formula of SegNet gives equivalent weights to both classes of 0.58 and 0.63 because they have the same average density per image: the trees occupy the same space in an image no matter their specie. In this case, the loss penalizes misclassifying each class the same way. Since there are more examples of broadleaved trees, the network ends up classifying all trees with a similar visual aspect as broadleaved woodland. We assume that DeepLab can better differentiate these classes thanks to its higher representative space: a higher number of layers allows to learn more discriminative features than SegNet (Figure 5). This way it learns the features that discriminate coniferous trees from broadleaved trees. However, it needs a certain amount of example to learn these additional features.

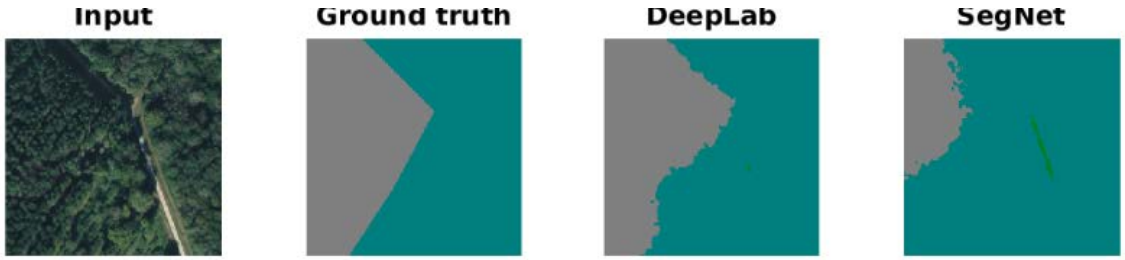

Fig 5. Confusion between broad leaves trees (cyan) and coniferous trees (grey). DeepLab better discriminates these two classes than SegNet.

Identifying land uses is a challenging task for machines, much more than identifying buildings or road networks, because it involves understanding context and environments. If we have a deeper look into which classes are hard for both architectures to recognize, three classes stands out: the heatlands, the waste deposits and extractive sites, and the riparian vegetation. For the heatlands it is because they are visually hard to differentiate even for humans. Indeed, GIS experts uses much more data than the simple orthophotos, they can rely on the land register or on Google Street View to have higher resolution images. This helps them classify the heatlands properly and not mistake them with woodlands like the networks do. As for the riparian vegetation, it is most probably due to the fact the network has not properly learned the concept which is subtle: riparian vegetation can be described as Riversides, lakesides, fens and marshy floodplains dominated by woody vegetation less than 5 m high. The waste deposits and extractive sites are problematic for the machine for different reasons. First the size of the features: often they are much larger than the images fed to the network a quarry can be kilometers large when the network is fed images with a scale of the 150 meters. Then there is the extreme diversity of those sites, indeed inside a quarry one can find grass or trees which confuses the network. This shows that the networks have difficulties extracting context when fed with small images. This phenomenon is illustrated in Figure 6. Finally, it should be noted that lots of the classification errors happen because both network did not manage to learn the minimal collection area: the minimal collection is the smallest surface that a land feature must measure to be classified by the expert.

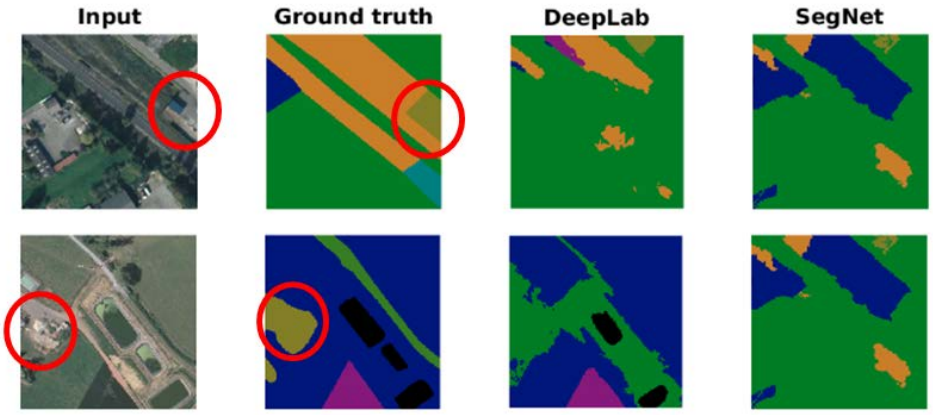

Fig 6. The network only sees a small part of the dump site which makes it hard to classify it correctly. (constructed areas: green), (surface waters: black), (Heatlands and scrub: orange), (grassland: blue), (arable land: purple)

1955 results

On the smaller 1955 dataset, DeepLab global accuracy reaches 65% of accuracy and 55% of mIOU. The network performance decrease of 10% which we consider still satisfying given the fact that the 1955 dataset is only half the size of the 2015 dataset. DeepLab training converged with the same setting as for the 2015 dataset but SegNet required pre-training to converge. Once again, the performances of both networks are equivalent but this time, the training methods differs.

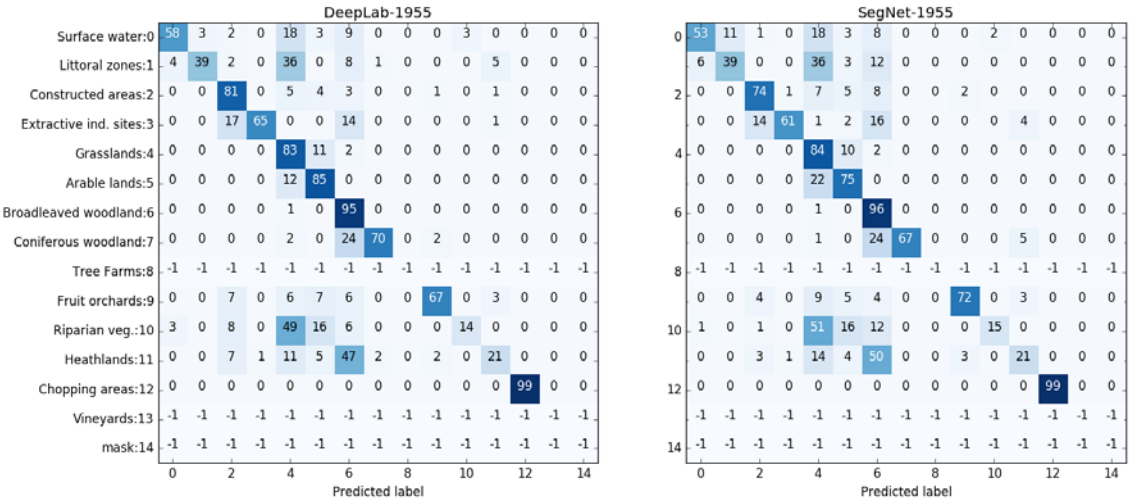

Fig 7. Per class accuracy.
Global accuracy DeepLab: 65% - Global accuracy SegNet: 63%

Table 5. Per class mIOU on the 1955 dataset.
Global mIOU DeepLab: 55%
Global mIOU SegNet: 53%

Model\Class	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
DeepLab	51	34	68	49	73	74	90	51	NA	51	11	14	96	NA	NA
SegNet	47	27	67	44	68	66	90	58	NA	54	13	14	98	NA	NA

DeepLab is initialized with weights provided by the authors of (Chen, L. C. et al. 2018), as described in section **Experiments**. It is then trained on the 1955 dataset for 30 000 steps. This procedure does not work for SegNet because of the weight distribution for the 1955 dataset. A

solution is to pre-train SegNet on a black and white version of 2015 with the weights of the 2015 dataset, then finetune on the 1955 dataset while keeping the 2015 weights.

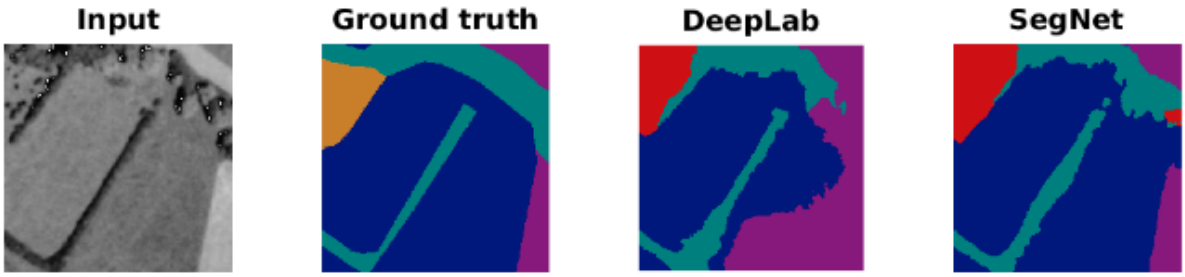

Fig 8. The networks reach high performances even when the image is blurry (broad leaves trees: cyan), (tree farms: red), (bushes: orange), (meadow: blue), (cultures: purple)

The 1955 weights prevent the network from converging because their variance is too high: classes with high pixel density per image have very low weights ($\sim 10^{-2}$) and the others have very high weights ranging from 1 to 275. This prevents the network from getting a loss consistent with its performance: when it misclassifies many pixels from the first category the loss stays low and when it misclassifies a few pixels from the second category the loss becomes high. This weight distribution can be explained by the fact that the 1955 dataset is very unbalanced: for example, there are only 35 images with tree farms. Similarly, to the 2015 dataset, the weight formula of the SegNet network cannot handle very unbalanced pixel distributions. We assume that the 2015 weights allow the network to converge because they are better balanced. Even though the 2015 and 1955 pixel distributions are not equal, the classes with a high pixel density are the same so the 2015 weights still are meaningful for the 1955 dataset.

This training method for SegNet shows the benefits of using synthetic data to compensate for the lack of annotated data. For SegNet, synthetic 1955 data are generated by converting the 2015 dataset into black and white image. This dataset has a lot more instances than the 1955 one so it is easier to train on it. However, such a network cannot be used on the 1955 directly as it needs to be fitted to the real 1955 dataset. Table 5 show the performance of SegNet trained only on the black and white 2015 dataset and after finetuning on the real 1955 data. The metrics are computed on the real 1955 data.

For both models, a network trained on fake 1955 performs poorly on the real 1955 dataset with only 17.12% of accuracy at best with SegNet (Table 6). One of the explication is that the synthetic data does not approximate well the 1955 data. Converting the 2015 data into black and white images is not enough as there are also variations of saturation and resolution between the two datasets. The performance of the models without finetuning can be improved by putting more effort in the generation of synthetic data. The transformation between the 2015 dataset and the 1955 one is mainly color intensity, resolution and saturation changes. This can be qualified as the style of the image. Even though it is complex for a human to explicit these transformations and their range, a DCNN can learn to project the style of the 1955 images onto the 2015 dataset following the work of Johnson, J et al. (2016) (Figure 9). Without finetuning on the real 1955 dataset, a network trained on the second synthetic dataset reaches up to 23.92% of global accuracy against 9.78% with only black and white conversion. After finetuning, the performances are higher than when the networks are only trained on 1955 data (Table 6).

Fig 9. Comparison of synthetic data. (left to right)
Top: 2015 data and the conversion to BW.
Bottom: matching 1955 data and the stylized 2015 data.

Table 6. Global accuracy.
Left columns: Network trained of on fake 1955 data before finetuning on real 1955 data
Right columns: Same network after finetuning on real 1955 data

	Black and white				Style			
	Before finetuning (BF)		After finetuning (AF)		BF		AF	
	acc	mIOU	Acc	mIOU	acc	mIOU	acc	mIOU
Segnet	17.12	8.91	63.60	53.90	14.46	8.26	**64.28**	**55.67**
DeepLab	9.78	4.77	67.28	57.33	23.92	12.40	**70.37**	**58.58**

In 1955, the algorithms can no longer rely on color, this makes their job much harder especially for the classes with smooth textures such as water and water shore classes. The two classes in addition to bushes and riparian groves are the hardest to classify for the network. This can be explained by the fact that the network only has a very small field of view and this effect is crystalized by the loss of the colors. Here again it is our belief that the algorithms are penalized by the dimensions of the input images. Larger input or different input should be investigated.

Conclusions

All in all, machine learning has delivered impressive results on our land-use classification task, even though the dataset that was used to train them was not optimized for machine learning. This shows that DCNN can learn from complex data as long as there are discriminative features between the land categories. On the 1955 dataset, even if the performances are not as good as

in 2015, the results are still satisfying enough showing that even with less information (1 channel instead of 3) the networks are able to recognize most classes easily.

As was demonstrated in the result section, off-the-shelves auto-encoders can reach high accuracy on large datasets without modifications. They are not only accurate but also dramatically quick. Once trained, our model can label 15 land categories within 25 square kilometers with a 50cm per pixels resolution in 15 minutes. In comparison, it took our GIS expert 3 working days to label 25 the same area. Even though, our model classifies fewer categories, it speeds up the classification process. Also, it only needs to be trained once and can be used on any new ortho-imagery sampled with the same camera. Our method is much more cost-efficient since the GPU we use is worth about 600 euros and cloud solutions such as AWS EC2 Elastic GPUs cost about 0.40 euros per hour. This is less than the cost of a GIS expert paid at least 17 euros per hours in France. Using cloud solutions also allows parallelizing the computation which decreases the classification time. Furthermore, machines are consistent in the way they label the data.

However, machine-based labelling lacks the high-level context and is not as accurate as an expert, in particular for very rare class occurrences. In its current state, our model can be used for computer-aided labelling for GIS expert, making their work more comfortable while helping to complete the same task faster. In the future, it would most definitely be interesting to investigate ways to provide the network with both high-level context and low-level information so that it can perceives the maps like the experts do. Moreover, when creating the dataset, we realized that choosing the right period of the year to realize the surveys could increase the overall accuracy of the machines.

Acknowledgement

We thank Gabriel Hurtado (master student) for the generation of fake 1955 data using neural style transfer methods.

References

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

Bhandari, A. K., Kumar, A., & Singh, G. K. (2012). Feature extraction using Normalized Difference Vegetation Index (NDVI): a case study of Jabalpur city. *Procedia Technology*, *6*, 612-621.

Bryson, M., Reid, A., Ramos, F., & Sukkarieh, S. (2010). Airborne vision-based mapping and classification of large farmland environments. *Journal of Field Robotics*, *27*(5), 632-655.

Chen, L. C., Yang, Y., Wang, J., Xu, W., & 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).

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & 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), 834-848.

Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & 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.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

Fisher, J. R., Acosta, E. A., Dennedy-Frank, P. J., Kroeger, T., & Boucher, T. M. (2017). Impact of satellite imagery spatial resolution on land use classification accuracy and modeled water quality. Remote Sensing in Ecology and Conservation.

Franklin, S. E., & 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), 173-205.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of remote sensing*, *23*(4), 725-749.

Hung, C., Xu, Z., & Sukkarieh, S. (2014). Feature learning based approach for weed classification using high resolution aerial images from a digital camera mounted on a UAV. *Remote Sensing*, *6*(12), 12037-12054.

IGN (2016) BD ORTHO® Version 2.0 et ORTHO HR® Version 1.0 – Descriptif de contenu – Février 2016, http://professionnels.ign.fr/doc/DC_BDORTHO_2-0_ORTHOHR_1-0.pdf

IGN forests http://professionnels.ign.fr/bdforet

IGN RPG http://professionnels.ign.fr/rpg

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678). ACM.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016, October). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision* (pp. 694-711). Springer, Cham.

Krähenbühl, P., & Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In Advances in neural information processing systems (pp. 109-117).

Liu, T., & Abd-Elrahman, A. (2018). Deep convolutional neural network training enrichment using multi-view object-based analysis of Unmanned Aerial systems imagery for wetlands classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *139*, 154-170.

LOUVEL, Justine et GAUDILLAT, Vincent. EUNIS, European Nature Information System, Système d'information européen sur la nature: classification des habitats: traduction française: habitats terrestres et d'eau douce. MNHN, 2013.

Lu, D., Batistella, M., Li, G., Moran, E., Hetrick, S., Freitas, C. da C., … Sant'Anna, S. J. S. (2012). Land use/cover classification in the Brazilian Amazon using satellite images. Pesquisa Agropecuaria Brasileira, 47(9), 10.1590/S0100–204X2012000900004. http://doi.org/10.1590/S0100-204X2012000900004

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., & Liu, Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 277-293.

Madden, M. (2009, May). GeoEye-1, the world's highest resolution commercial satellite. In *Conference on Lasers and Electro-Optics* (p. PWB4). Optical Society of America.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93-104.

Sandre http://www.sandre.eaufrance

Scott, G. J., England, M.R., Starms, W.A., Marcum, R.A., Davis, C.H. (2017). Training deep convolutional neural networks for land-cover classification of high-resolution imagery. IEEE Geoscience and Remote Sensing Letters, 14(4), 549-553.

Sherrah, J. (2016). Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*.

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003, August). Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR* (Vol. 3, pp. 958-962).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Qian, Y., Zhou, W., Yan, J., Li, W., & Han, L. (2014). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. Remote Sensing, 7(1), 153-168.

Wang, J., & Perez, L. (2017). *The effectiveness of data augmentation in image classification using deep learning* (No. 300). Technical report.

Yang, Y., & Newsam, S. (2010, November). Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 270-279). ACM.

Yu, Q., Gong, P., Clinton, N., Biging, G., Kelly, M., & Schirokauer, D. (2006). Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogrammetric Engineering & Remote Sensing*, *72*(7), 799-811.