Journal of Universal Computer Science, vol. 26, no. 2 (2020), 244-267
submitted: 5/6/19, accepted: 10/2/20, appeared: 28/2/20

Convolutional Neural Networks and Transfer Learning Based Classification of Natural Landscape Images

Damir Krstinić
(Universty of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Ruđera Boškovića 32, 21000 Split, Croatia
damir.krstinic@fesb.hr)

Maja Braović
(Universty of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Ruđera Boškovića 32, 21000 Split, Croatia
maja.braovic@fesb.hr)

Dunja Božić-Štulić
(Universty of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Ruđera Boškovića 32, 21000 Split, Croatia
dunja.gotovac@fesb.hr)

Abstract: Natural landscape image classification is a difficult problem in computer vision. Many classes that can be found in such images are often ambiguous and can easily be confused with each other (e.g. smoke and fog), and not just by a computer algorithm, but by a human as well. Since natural landscape video surveillance became relatively pervasive in recent years, in this paper we focus on the classification of natural landscape images taken mostly from forest fire monitoring towers. Since these images usually suffer from the lack of the usual low and middle level features (e.g. sharp edges and corners), and since their quality is degraded by atmospheric conditions, this makes the already difficult problem of natural landscape classification even more challenging. In this paper we tackle the problem of automatic natural landscape classification by proposing and evaluating a classifier based on a pretrained deep convolutional neural network and transfer learning.

Key Words: deep learning, transfer learning, convolutional neural networks, image classification, natural landscape images, wildfire smoke
Category: I.2.1, I.2.6, I.2.10, I.4.0, I.4.6, I.4.8, I.4.9, I.5.3

1 Introduction

Automatic image classification has been one of the ultimate goals of artificial intelligence since its inception. In recent years this area has shown a significant progress that can mainly be attributed to the increased usege of deep learning. Deep learning encompasses deep convolutional neural networks (DNN) with multiple hidden layers, and even though it is a novel label in computer science, it is not a novel concept. Despite this fact, however, deep learning techniques started to gain widespread popularity only in the last few years. One of the

causes of this renewed focus of attention can be found in the availability of large datasets needed for the development of deep learning based models, and in the advances in hardware development that are able to support the requirements of deep learning models.

In this paper we explore the possibility of using a pretrained deep convolutional neural network and transfer learning approach for a very specific problem of automatic classification of natural landscape images. Natural landscape images are very different from the ones usually encountered in general classification problems as they do not contain a large number of artificial structures and are generally poorer in low and middle level features like corners and edges, or typical and recognizable shapes. Furthermore, most of the images used in this paper are collected from forest fire surveillance cameras mounted on wildland locations that operate for 24 hours a day through the whole year and are prone to image quality degradation due to atmospheric and weather conditions. An example of a natural landscape image obtained from a wildfire surveillance camera is given in Figure 1, and the degradation in the image quality is clearly observable.

Figure 1: An example of a natural landscape image taken from a wildfire surveillance camera

Natural landscape image classification is commonly used in the video surveillance of the wide wildland areas. It is in the focus of this research because the

authors have a strong background in image processing based wildfire monitoring primarily oriented on early wildfire detection [Krstinić et al. 2009, Štula et al. 2012, Jakovčević et al. 2013, Bugarić et al. 2013], and are working on a wildfire monitoring system that is already deployed in a number of Croatian forests and National parks. The authors would like to improve the already high accuracy of the above mentioned wildfire monitoring system, so in this paper special attention is given to the evaluation of the proposed method on the images that contain wildfire smoke, which is the first visible sign of wildfires occurring in heavily wooded areas.

Primary goals of this research are the development of specific methods useful for natural landscape image classification, and the analysis of the impact and potential of deep learning techniques in the same context. This area of research differs significantly from the problem of classification of urban and close range scenes and is conceptually a very distinct problem that is further complicated by the non-existence of widely used natural landscape image datasets, the absence of strongly expressed image features (e.g. hard edges and typical shapes of various objects) and the vast similarity between different classes. Furthermore, some image regions can be semi-transparent (e.g. smoke or fog) and allow a different region to be seen through them, or they can contain reflections of other regions (e.g. sea surface). Therefore, it is not uncommon that even human experts responsible for creating ground truth segmentations of these images cannot draw a hard boundary between various regions or classify them into one of the predefined classes. These experts usually have no choice but to leave some parts of the image unclassified because they are often ambiguous and can be seen as belonging to a number of classes instead of belonging to just one class.

Even though the proposed method is geared towards natural landscape image classification, it can still be used in many other areas of research with little adjustment. Examples of potential areas of research include image processing based detection of climate change, forest biology and early detection of pathological changes in vegetation, tracking of flooded areas, surveillance of agricultural areas, automatic connection and calibration with geographical information systems, and the applications of augmented reality.

This paper is structured as follows. In Section 2 related work is discussed, with special attention being given to deep learning based approaches for image classification, segmentation and object detection. Proposed approach to classification of natural landscape images and research methodology is elaborated in section 3. In Section 4 the evaluation of the proposed method and its comparison to the existing state-of-the-art approaches for image analysis is presented. Finally, in Section 5 a conclusion is given and future work is discussed.

2 Related work

Deep learning based approaches to digital image analysis have become increasingly common and have been used in areas such as medicine [Sadanandan et al. 2017, Wang et al. 2018], biomedicine [Ronneberger et al. 2015], remote sensing [Liu et al. 2018, Hu et al. 2018, Zou et al. 2015, Gao et al. 2018, Hu et al. 2015, Chen et al. 2019] and path finding for visually impaired or blind people [Malūkas et al. 2018]. These approaches are quickly overshadowing the approaches based solely on traditional digital image analysis methods (such as template matching or histogram comparison) because they can usually offer more accurate results. One of the potentially negative aspects of deep learning based approaches to digital image analysis is its need for a large image dataset, much larger than the ones usually needed for digital image analysis that do not encompass deep learning techniques. This can present a problem in areas dealing with digital image analysis of uncommon occurrences, but one of the ways that this problem is being dealt with is image augmentation, i.e. obtaining a higher number of images from the images that are already available.

Rapid development in this field and high motivation to develop a standard evaluation methodology resulted in a standardized ImageNet [Deng et al. 2009] dataset for general image classification problem that contains more than 1.2 million hand labeled images. Each image is labeled with one of the 1000 predefined classes such as lemon, espresso, trombone, polaroid camera, poncho, microphone, castle, catamaran, golden retriever, etc. This image collection, divided in training, validation and testing dataset, is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al. 2015-1].

Beside this general classification dataset, task oriented datasets exist for some problems of high scientific and practical interest, e.g. MNIST dataset [LeCun and Cortes 2010] for the recognition of hand written characters. However, for most specific classification problems high costs of hand labeling of images by an expert in the field imposes a different approach based on transfer learning [Torrey and Shavlik 2009]. In this scenario, relatively small dataset of hand labeled images for a specific task is created and used to retrain model which was previously trained with a bigger dataset.

Traditional methods for digital image analysis include techniques such as contour detection, feature detection, template matching, histogram comparison, texture analysis, various classifiers (e.g. SVM or k-NN), etc., and until a recent increase in the popularity of deep learning techniques, traditional techniques were the go-to methods for digital image analysis. Examples of traditional methods for digital image analysis can be found in [Braović et al. 2017], [Vogel and Schiele 2004], [Fei-Fei and Perona 2005], etc. Even though various traditional methods for digital image analysis have existed for quite some time, there still does not seems to exist a commonly used measure for the evaluation of those

approaches, and the researchers are left to choose between a rather large ensemble of different metrics. This, of course, makes it difficult to compare the results of various methods and is a problem that should potentially be addressed in the near future.

In recent years deep learning methods have been commonly used for digital image analysis, and it is demonstrated in [Buscombe and Ritchie 2018] that deep convolutional neural networks can successfully be used in the classification of landforms and land cover in medium-range imagery acquired from UAS, aerial, and ground-based platforms. Examples of deep learning based image analysis methods can be found in [Liu et al. 2018], [Socher et al. 2011], [Lam et al. 2017] and [Krizhevsky et al. 2012].

Liu et al. [Liu et al. 2018] proposed a method for classification of high-resolution remote sensing images that is based on deep random-scale stretched convolutional neural network. Multiple views of one image were used so the image was classified multiple times and the final image label was obtained by a voting procedure.

Socher et al. [Socher et al. 2011] proposed a deep learning based method for segmentation and annotation of complex scenes. The input image was over-segmented into superpixels and the features (e.g. color, shape, texture, etc.) for these superpixels were extracted. A simple neural network layer was used to map the extracted features into the semantic n-dimensional space. Recursive neural network was later used for tree structure prediction.

Lam et al. [Lam et al. 2017] proposed a method for deep learning based fine-grained object recognition. Their method was based on the search for informative image parts, i.e. the ones that could make it easier to differentiate between similar object classes. This search was accomplished by searching the deep feature map produced by a convolutional neural network (CNN). The evaluation metric that was used in their paper was top-1 accuracy, where the predicted (automatic) classification is considered accurate if the ground truth label is present in the top 1 most confident predictions.

Krizhevsky et al. [Krizhevsky et al. 2012] proposed a method for deep learning based classification of images from ImageNet LSVRC-2010 dataset into 1000 classes. The CNN architecture that they used consisted of 8 layers: 5 convolutional and 3 fully connected. The last fully connected layer output was used as an input to a 1000-way softmax, which gave the distribution over the 1000 class labels.

Table 1 presents an overview of recent deep learning based approaches for digital image analysis. Results given in these tables have been obtained from the corresponding papers.

As can be seen from Table 1, results obtained by deep learning based digital image analysis are generally good. This makes it clear that deep learning has

Table 1: An overview of deep learning based methods for image classification, segmentation, annotation or object detection

Classes	**Image dataset**	**Method**	**Results**
Aeroplane, bike, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, plant, sheep, sofa, train, TV monitor, background (additional)	PASCAL VOC 2010 [Everingham et al. 2010]	Zheng et al. [Zheng et al. 2015]	Average intersection over union (IU): 75.7%
	PASCAL VOC 2011 [Everingham et al. 2011]	Zheng et al. [Zheng et al. 2015]	Average intersection over union: 75.0%
	PASCAL VOC 2012 [Everingham et al. 2012]	Zheng et al. [Zheng et al. 2015]	Average intersection over union: 74.7%
		Roy and Todorovic [Roy and Todorovic 2017]	Average mean intersection over union: 53.7%
		Hong et al. [Hong et al. 2015]	Mean intersection over union: 66.6%
		Chen et al. [Chen et al. 2018]	Mean intersection over union: 79.7%
		Islam et al. [Islam et al. 2016]	Mean intersection over union span: 62.1%-64.5%
		Yu and Koltun [Yu and Koltun 2016]	Mean intersection over union: 67.6%
Road, building, sky, tree, side walk, car, column pole, fence, pedestrian, bicycle, sign	CamVid [Brostow et al. 2009]	Kendall et al. [Kendall et al. 2015]	Average accuracy: 76.3%
		Mahasseni et al. [Mahasseni et al. 2017]	Accuracy span: 53.5%-73.3%
		Ardiyanto and Adji [Ardiyanto and Adji 2017]	Average class accuracy span: 69.8%-71.5%

Table 1: An overview of deep learning based methods for image classification, segmentation, annotation or object detection (cont.)

Classes	**Image dataset**	**Method**	**Results**
Agriculture, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, tennis court	UC Merced Land Use Dataset [Newsam, Yang and Newsam 2010]	Liu et al. [Liu et al. 2018]	Average accuracy: 95.57%
Meadow, pond, harbor, industrial, park, river, residential, overpass, agriculture, commercial, water, idle land	Google dataset of SIRI-WHU [Zhao et al. 2016, Zhong]	Liu et al. [Liu et al. 2018]	Average accuracy: 94.76% and 93.44% (when trained on 80% and 50% of the samples from the dataset, respectively)
Dense residential, idle, industrial, medium residential, parking lot, commercial, vegetation, water	The Wuhan IKONOS Dataset	Liu et al. [Liu et al. 2018]	Average accuracy: 85.00%
Wall, floor, cabinet, bed, chair, sofa, table, door, window, bookshelf, picture, counter, blinds, desk, shelves, curtain, dresser, pillow, mirror, floor mat, clothes, ceiling, books, fridge, TV, paper, towel, shower curtain, box, whiteboard, person, night stand, toilet, sink, lamp, bathtub, bag	SUN RGB-D benchmark dataset [Song et al. 2015]	Kendall et al. [Kendall et al. 2015]	Average accuracy: 45.92%
Types of clouds: cumulus, cirrus, altocumulus, clear sky, stratocumulus, stratus, cumulonimbus	Multimodal ground-based cloud (MGC) dataset	Liu and Li [Liu and Li 2018]	Average accuracy: 86.30%
Building, tree, sign, road, fence, pole, sidewalk	KITTI [Geiger et al. 2012]	Mahasseni et al. [Mahasseni et al. 2017]	Accuracy span: 78.2%-92.7%
Plane, bird, boat, car, cat, cow, dog, horse, motorbike, train	YouTube-Objects dataset v2.0 [Prest et al.]	Tripathi et al. [Tripathi et al. 2016]	Mean average precision: 37.413%
67 indoor scenes categories (classes)	MIT Indoor scene dataset [Quattoni and Torralba 2009]	Li et al. [Li et al. 2017]	Average accuracy span: 74.86%-87.97%

Table 1: An overview of deep learning based methods for image classification, segmentation, annotation or object detection (cont.)

Classes	Image dataset	Method	Results
397 scene categories (classes), including: coast, field wild, forest fire, forest broadleaf, forest needleleaf, hill, house, lake natural, mountain, ocean, pond, river, tree farm, village and woodland	Scene UNderstanding (SUN) database [Xiao et al. 2010, Xiao et al.]	Li et al. [Li et al. 2017]	Average accuracy span: 57.15%-72.01%
3 scene types are used in [Socher et al. 2011]: city, countryside and sea-side	Stanford background dataset [Gould et al. 2009, Stanford Background Dataset]	Socher et al. [Socher et al. 2011]	Pixel-level accuracy: 78.1%
200 species of birds	CUB-2011 [Wah et al. 2011]	Lam et al. [Lam et al. 2017]	Top-1 accuracy: 87.5%
196 types of cars	Cars-196 [Krause et al. 2013]	Lam et al. [Lam et al. 2017]	Top-1 accuracy: 93.9%
1000 object categories (classes), including: seashore, lakeside, coral, box turtle, tarantula, bee orchid and orange	ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC2010) [Russakovsky et al. 2015-2, Russakovsky et al. 2010]	Krizhevsky et al. [Krizhevsky et al. 2012]	Top-1 and top-5 test set error rates, respectively: 37.5% and 17.0%

great potential in image processing applications, and possibly represents a step towards achieving strong artificial intelligence.

3 Transfer learning and the Classification of Natural Landscape Images

The proposed method is based on a somewhat classical approach of transfer learning where a pretrained DNN is retrained with newly encountered images specific to a target task (e.g. [Cengil and Çinar 2019]). When dealing with deep convolutional neural networks, convolutional layers are commonly preserved from the pretrained model because it is assumed that they have acquired the ability to extract low and middle level features present in many various images from the large dataset that was used in the training process, while the last few layers are

retrained with newly encountered images from a task-specific dataset. Examples of algorithms for image analysis that are based on transfer learning can be found in [Cengil and Çinar 2019], [Chang et al. 2017], [Alsabahi et al. 2018], etc.

3.1 Dataset

In the development and validation of the proposed method a Mediterranean Landscape Image Dataset (FESB MLID) [Braović et al. 2017] was used. FESB MLID dataset contains 400 images of wide wildland open space areas. Each image is processed by a human observer and divided into regions marked as one of 11 predefined classes. These segmented images are called *ground-truth* (GT) images. The classes that were used in the FESB MLID dataset are *smoke, clouds and fog, sun and light effects, sky, water surfaces, distant landscape, rocks, distant vegetation, close vegetation, low vegetation and agricultural areas* and *buildings and artificial objects*. Additional class named *unknown* was used for the parts of the image that cannot be classified to one of the predefined categories by a human because of their ambiguity. Most of the images in the FESB MLID dataset were collected from the real operational forest fire surveillance cameras mounted on monitoring locations on the Croatian coastline and islands. These cameras are covering wide open-space areas that mainly consist of a diverse and heterogeneous Mediterranean landscape.

One of the implications of the fact that the images from the FESB MLID dataset mostly came from real high-ground mounted cameras surveilling wide open space areas is that it is very possible that most of the images may contain many of the predefined classes of natural landscape. It is even possible and not completely uncommon that all of the classes are present in a single image. On the other hand, some of the classes representing landscape or phenomena which is by nature less common in the overall data are present on a smaller number of images in the dataset and often represented by a much smaller surface area compared to the other classes. The best example of such a class is *smoke*. Favourably, smoke should be recognized in the early stages of its development, thus examples of this class in the training data are usually represented by small smoke plumes covering a very small area of the image. On the other hand, classes such as *sky* or *distant vegetation* are present in almost all of the images and are usually covering a dominant part of the image. Trivial solution to which training process could converge is to label all of the images with labels representing common classes, and declaring less common classes not present in any of the images, as this would result in low overall error rate and high accuracy.

To avoid this scenario, a dataset for this research is created from the original FESB MLID dataset by dividing each image from the original dataset into a set of smaller images. Each of the resulting images represents a small part of the landscape and contains a lower number of predefined classes than the image they

were extracted from. Furthermore, each of the classes in a smaller image occupies a sufficiently large portion of the image area. Additional advantage is that the resulting dataset is much larger in the number of images than the original FESB MLID dataset. The final dataset is created by following the steps outlined below:

1. FESB MLID dataset is divided in two parts, one containing 300 images used for training, and the other with 100 images used for testing.
2. The number of images in the training data set is increased by data augmentation: (a) each image is reflected horizontally, and (b) each image (original and reflected) is rotated by a random angle. To avoid unnatural landscape scenes, rotational angle is limited to a 30°.
3. Each image generated in step 2 (original image, reflected image and rotated images) is divided into a set of smaller sub images.
4. For each image in the FESB MLID dataset, appropriate GT segmentation created by a human expert is used, alongside with the parameters from the data augmentation step (rotation angles of the original and reflected images), to extract a list of classes of natural landscape present in each of the smaller sub images.

Using the technique outlined above, three data sets are created with image sizes of 32×32, 64×64 and 128×128 pixels. Data augmentation is used only in the generation of the train dataset. For test dataset step two is skipped and only the original images are divided into smaller samples. The first data set with the image size of 32×32 pixels contains 1688563 train images and 152889 test images. Dataset of size 64×64 pixels contains 383688 train images and 34934 test images. Dataset with the sample size of 128×128 pixels contains 77418 train and 7178 test images.

3.2 Classifier

The classifier is based on the Inception V3 convolutional neural network [Szegedy et al. 2015, Szegedy et al. 2016] trained on the ImageNet dataset [Deng et al. 2009]. Originally, Inception is trained for single-label image classification. To adapt Inception to a multi-label image classification where each image is labeled with more than one label, i.e. where each image contains more then one of the non-exclusive classes, we followed the approach proposed in [Bartyzal]. The main difference to the original Inception architecture is a modified fully-connected layer which is retrained with new images, and new evaluation method used to evaluate model prediction and compare its output with multi-class GT vectors that describe each image in the dataset. A diagram of the proposed classifier is shown in Figure 2.

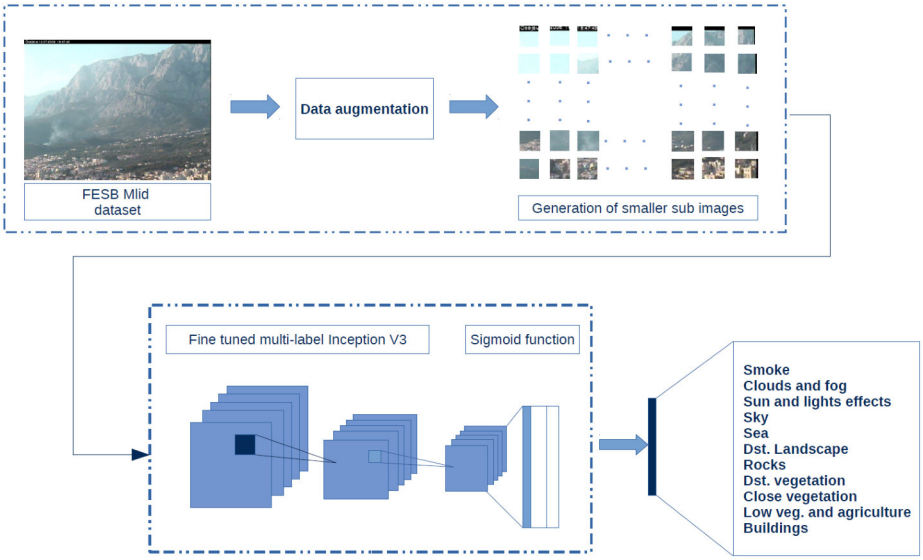
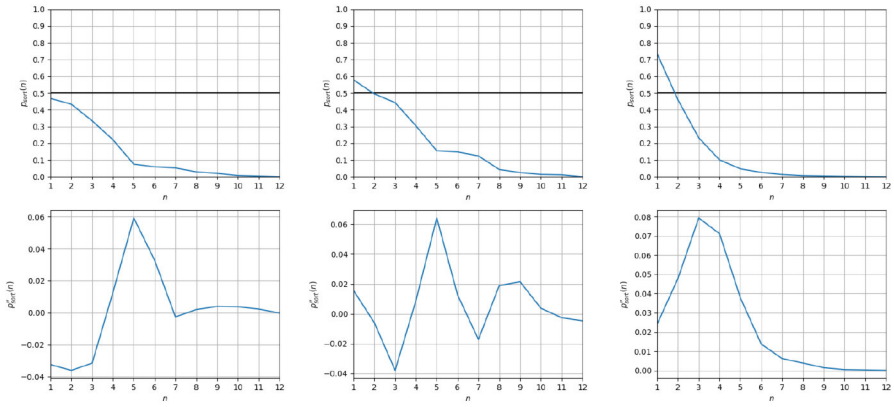


Figure 2: A diagram of the proposed classifier

The output of the model for one image is a set of probabilities for each of the 12 classes, including the class *unknown*. By introducing this class it is ensured that at least one class is present in each image. In extreme cases it is possible that all 12 classes are present in a single image. Therefore, in each image one or more classes are present and probabilities do not sum up to one. The final stage of the algorithm is to select which classes will be accepted as present in the image based on the output of the model. Straightforward solution is to accept all of the classes with probability above 0.5. However, by examining statistical distribution of the top probability on the train data set we found that the top probability is sometimes below 0.5 which is in contrast with the model premise that at least one class is present in each image. More detailed examination of the output of the model shows that any value of fixed threshold is too lenient for some images, i.e. it accepts classes that do not exist in GT segmentation, while too rigorous for others which would result in an output where no class is present in the image.

To determine a threshold for the acceptance or the rejection of the classes present in the image an adaptive criterion has to be established. We set this threshold to the first sharp fall in the sorted probabilities.

In Figure 3 examples of sorted probabilities obtained from the DNN model and their second derivatives for three different images are shown. In example (a) the first minimum of the second derivative of sorted probabilities $p''_{sort}(n)$ marks steep fall of the sorted probabilities after second highest probability, indicating



(a) classes present in image: 2, (b) classes present in image: 3, (c) classes present in image: 1

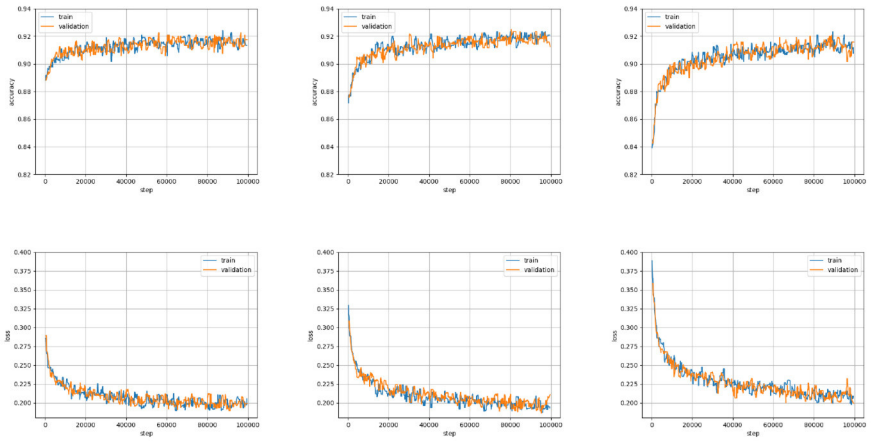
Figure 3: Probabilities for all 12 classes (including *unknown*) for three different images. First row: probabilities computed by the retrained neural network. Probabilities are sorted from the highest to the lowest. Second row: second derivative of the sorted probabilities. Last class to be accepted as present in the image corresponds with the first minimum of the second derivative of the sorted probabilities.

the presence of two classes in the image. This is a point where sorted probability function $p_{sort}(n)$ has a concave form, before point of inflection. After this, $p_{sort}(n)$ exhibits convex behavior and difference between subsequent probabilities becomes smaller and smaller. For image in example (b) first minimum of the $p''_{sort}(n)$ corresponds to a third highest probability, implicating that there are three classes present in the image. In the example (c) sorted probabilities exhibit convex shape from the first element, with strongest fall in subsequent probabilities between the first two classes. This suggests that only the class with the highest probability is present in the observed image.

Having in mind the above observations, adaptive criterion for accepting classes is created based on two simple rules:

(a) all classes with the probability above 0.5 are accepted,

(b) first k classes, sorted by their probabilities from highest to lowest, are accepted, where k is the location of the first minimum of the second derivative of sorted probabilities $p''_{sort}(n)$.



(d) image size 32×32 (e) image size 64×64 (f) image size 128×128

Figure 4: Training history - accuracy and cross entropy loss

4 Results and discussion

The proposed method of classification based on transfer learning and adaptive class acceptance is evaluated on three different datasets generated from the same FESB MLID set of images by using different sample image size, as depicted in subsection 3.1. For each of the three datasets with different image size, a classifier is trained using the train dataset. After this process is complete, trained classifier is evaluated on the dataset generated from 100 images from the FESB MLID dataset which were not used in the training stage. Except for the different input sample image size used for training and testing of the classifier, there is no difference between the three evaluated classifiers. All three classifiers are trained for equal number of iterations, with training logs given in Figure 4.

Several standard measures are calculated, namely *Accuracy, Balanced accuracy, Precision, Recall* and F_1-score. These measures are given in equations (1), (2), (3), (4) and (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Balanced\ accuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{FP+TN}}{2} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5}$$

The parameters of the above equations are defined as follows:

- TP (*True Positive*) - the number of times that the class is detected as present and exists in the GT segmentation,
- TN (*True Negative*) - the number of times that the class is detected as not present and does not exist in the GT segmentation,
- FP (*False Positive*) - the number of times that the class is detected as present, but does not exist in the GT segmentation,
- FN (*False Negative*) - the number of times that the class is detected as not present, but exists in GT segmentation.

Results achieved for each of the three sample image resolutions are given in Tables 2, 3 and 4. In all of these tables the results are shown for each of the predefined classes separately, thus depicting the ability of the trained classifier to accurately detect a specific class. Class *unknown*, used for areas of the image that cannot be clearly classified to any of the predefined categories by a human expert, was not taken into consideration in the evaluation of the classifier results.

As can be seen from tables 2, 3 and 4, *Accuracy* is relatively high for all of the classes. However, this measure does not give clear insight into the classifier performance. *Accuracy* does not give us information about FP to FN ratio. If a particular class is present only in a small number of images, the classifier could have high *False negative* rate and still have high *Accuracy* for the observed class. By using smaller sub images from the original dataset where only a small part of the surveilled area is contained in an image, as depicted in subsection 3.1, this assumption is correct for most, if not all classes. In fact, whenever a single class is present in only a limited number of samples, the classifier will achieve good accuracy as long as it keeps TN rate high. Even if *Precision* is also low, i.e. if the significant ratio of samples labeled with a particular label does not really contain that class (if $TP \simeq FP$), the accuracy will be high as long as $TP + FP \ll TN + FN$.

A clearer picture of the classifier performance on unbalanced data set for a different classes can be obtained from *Precision*, *Recall* and *Balanced accuracy*.

Precision measures the ratio of images from the test dataset that are labeled with a particular label by the classifier and actually contain this type of landscape (in human assigned GT segmentations). *Recall* shows the share of images that contain a particular type of landscape for which that class is accurately recognized by the classifier. *Balanced or normalized accuracy* [Brodersen et al. 2010] is the arithmetic mean of true positive rate (*Recall* or *Sensitivity*) and true negative rate (*Specificity*). F_1 *-score* is a balanced function of the *Precision* and *Recall*, and can be considered as an objective measure of *per class* classifier performance.

In Table 2 results are shown for the classifier that was trained and evaluated on image size of 32×32 pixels. It can be seen that there is a large discrepancy in the classifier performance for different classes. Best results are achieved on classes *Sun and light effects* and *Sky*. We assume that for these classes the classifier managed to learn a distinct typical class color which is very different from other colors usually present in natural landscape images. Other classes with relatively high results include classes with distinct details that the classifier can relay on, e.g. *Rocks*. The worst results are achieved for the class *Smoke*, where *Recall* is only 0.383, meaning that only 38.2% of the images containing smoke (in human assigned GT segmentation) are accurately recognized. *Precision* for class *Smoke* is 0.320, suggesting that from all images labeled with this label by the classifier, only 32% actually contain that class.

Table 2: Accuracy, balanced accuracy, precision, recall and F_1-score per classes, resolution 32×32 pixels

Class	**Accuracy**	**Bal.Acc.**	**Precision**	**Recall**	**F_1-score**
Smoke	0.961	0.679	0.295	0.383	0.333
Clouds and fog	0.914	0.792	0.591	0.636	0.612
Sun and light effects	0.952	0.857	0.841	0.732	0.782
Sky	0.963	0.854	0.736	0.727	0.731
Sea	0.975	0.709	0.631	0.425	0.508
Dst. landscape	0.928	0.744	0.625	0.520	0.568
Rocks	0.945	0.767	0.685	0.556	0.614
Dst. vegetation	0.827	0.662	0.511	0.408	0.454
Close vegetation	0.865	0.775	0.550	0.646	0.594
Low veg. and agriculture	0.919	0.695	0.546	0.424	0.477
Buildings	0.944	0.671	0.736	0.352	0.477
Average	**0.927**	**0.746**	**0.613**	**0.528**	**0.559**

More stable results are achieved when the classifier is trained on image size

64×64 pixels, as shown in Table 3. *Precision* is significantly better for all classes, and the improvement in *Recall* is visible for all classes for which the classifier achieved relatively weak results in 32×32 sample image resolution. *Recall* remains constant for classes with good results for image size 32×32 pixels and even falls for the best performing category in lower resolution.

Table 3: Accuracy, ballanced accuracy, precision, recall and F_1-score per classes, resolution 64×64 pixels

Class	**Accuracy**	**Bal.Acc.**	**Precision**	**Recall**	F_1**-score**
Smoke	0.961	0.754	0.476	0.530	0.502
Clouds and fog	0.916	0.797	0.673	0.639	0.655
Sun and light effects	0.943	0.828	0.872	0.671	0.758
Sky	0.964	0.853	0.809	0.721	0.763
Sea	0.977	0.808	0.728	0.624	0.672
Dst. landscape	0.919	0.785	0.710	0.605	0.653
Rocks	0.940	0.775	0.790	0.567	0.660
Dst. vegetation	0.820	0.722	0.621	0.541	0.579
Close vegetation	0.871	0.776	0.651	0.627	0.639
Low veg. and agriculture	0.912	0.731	0.613	0.500	0.551
Buildings	0.930	0.740	0.775	0.499	0.607
Average	**0.923**	**0.779**	**0.702**	**0.593**	**0.640**

The results for the classifier trained with the sample image size of 128×128 pixels are shown in Table 4. The classifier improves the performance for classes for which its performance was weaker on lower sample image size. The classifier performance remains relatively constant or even slightly decreases for classes for which it achieved good performance on smaller image size. The overall results for different classes are very stable. The ratio of the worst to best *per class* performance in F_1 score is 0.85, compared to the results in sample image resolution 32×32 where the classifier performance for class *Smoke* is only 0.44 of the performance for class *Sun and light effects* in F_1-score.

Even though the motivation for the work proposed in this paper was rooted in our intention of improving the accuracy of our image processing based wildfire smoke recognition system, we can see from Tables 2, 3 and 4 that the overall F_1-scores for class *Smoke* are lower than the F_1-scores for other classes. Even though these results seem counterintuitive for the task at hand, we can still use the proposed method to classify classes other than *Smoke*, and then use additional image processing techniques on the remaining regions to detect *Smoke*. In other words, the proposed method can be used as a preprocessing step for *Smoke*

Table 4: Accuracy, balanced accuracy, precision, recall and F_1-score per classes, resolution 128×128 pixels

Class	**Accuracy**	**Bal.Acc.**	**Precision**	**Recall**	F_1**-score**
Smoke	0.954	0.792	0.705	0.603	0.650
Clouds and fog	0.891	0.782	0.651	0.625	0.638
Sun and light effects	0.919	0.778	0.867	0.572	0.689
Sky	0.941	0.786	0.773	0.592	0.670
Sea	0.974	0.828	0.820	0.664	0.734
Dst. landscape	0.881	0.800	0.723	0.665	0.693
Rocks	0.926	0.782	0.868	0.579	0.695
Dst. vegetation	0.801	0.767	0.699	0.672	0.685
Close vegetation	0.853	0.793	0.697	0.679	0.688
Low veg. and agriculture	0.897	0.772	0.684	0.594	0.636
Buildings	0.905	0.795	0.810	0.622	0.704
Average	**0.904**	**0.789**	**0.754**	**0.624**	**0.680**

detection, i.e. it can be used as a method that can *clean* the input image by detecting the majority of the non-*Smoke* pixels.

Table 5 shows the comparison of results achieved in this work to other deep learning based image classification method discussed in Section 2. It should be noted that, although these problem share common characteristics, each of them has its own specificities. Thus the results given in this table are not directly comparable but are only illustrative to evaluate the method efficiency.

Another disadvantage of this comparison is that most of the authors give only accuracy as the measure of algorithm performance. Even though the average accuracy is often used as an evaluation measure of image classification algorithms, high accuracy does not necessarily mean high correctness of the algorithm predictions, as discussed previously in this paper. Other statistical measures should also be considered in the evaluation process in order to give a complete picture of the algorithm performance. An illustrative example of this are the results of the proposed classifier for different image resolutions. Highest accuracy is achieved on resolution 32×32. However, precision, recall and consequently F_1-score are much better for higher image resolutions.

A clearer picture of the effectiveness of the proposed method can be obtained by comparing it with the results achieved by Cogent confabulation based expert system for segmentation and classification of natural landscape images [Braović et al. 2017]. This expert system was developed and evaluated using the same FESB MLID dataset, and the evaluational results are given in F_1-score. The system achieved 0.441 in F_1-score, while the proposed method achieves average F_1-score from 0.575 for image size 32×32 pixels to 0.677 for image resolution of

Table 5: A comparison of the proposed work with various deep learning based image classification methods

Method	**Accuracy**	**Precision**	**Recall**	F_1**-score**
Kendall et al. [Kendall et al. 2015] (on CamVid dataset)	0.763	/	/	/
Kendall et al. [Kendall et al. 2015] (on SUN RGB-D dataset)	0.459	/	/	/
Liu et al. [Liu et al. 2018] (on UCM dataset)	0.956	/	/	/
Liu et al. [Liu et al. 2018] (on Google SIRI-WHU dataset when trained on 80% of the samples)	0.948	/	/	/
Liu et al. [Liu et al. 2018] (on Google SIRI-WHU dataset when trained on 50% of the samples)	0.934	/	/	/
Liu et al. [Liu et al. 2018] (on Wuhan IKONS dataset)	0.850	/	/	/
Liu and Li [Liu and Li 2018]	0.863	/	/	/
Braović et al. [Braović et al. 2017] (our previous work that does not use deep learning)	/	/	/	0.441
Proposed method (32 x 32)	0.929	0.619	0.548	0.575
Proposed method (64 x 64)	0.925	0.702	0.596	0.642
Proposed method (128 x 128)	0.902	0.750	0.623	0.677

128×128 pixels.

5 Conclusions and Future Work

In this paper we presented our research on the exploration of the possibility of application of convolutional neural networks to the analysis of natural landscape images. Automatic classification of this type of images comes with its own set of challenges and differs from the classification of urban, indoor or close range scenes. Specific challenges related to the natural landscape image classification make if difficult to develop an unambiguous evaluation methodology or to compare evaluation results to other classification methods. Many published methods

on natural landscape image classification offer only the average accuracy as a statistical measure of their performance. We found that this measure is not informative enough even if it is computed on a *per class* basis (i.e. if the accuracy is computed separately for each class). Furthermore, relaying solely on the measure of accuracy could even mislead the evaluation and suggest that the classifier performs acceptably well even though it could have high rates of both false positive and false negative errors for particular classes. Because of these shortcomings attached to the accuracy measure, in this paper we rely on *Precision*, *Recall* and F_1-score measures. These measures give us a more clear insight into the classifier performance for different classes and different input image sizes. Furthermore, we compared the results obtained by the proposed method to the results obtained by similar methods.

Comparative analysis of the classifier performance for different input image sizes highlights the fact that the best performance for different natural landscape image classes is achieved at different scales. These scales do not necessarily coincide with the sizes of landscape types or phenomena. For example, class *Smoke* is usually very small in size but is best recognized on larger input image sizes, while other classes that normally occupy larger areas on input images (e.g. *Sun and light effects* or *Sky*) are best recognized on smaller input image sizes. The reasons behind this can be found in the internal ability of the classifier to reveal contextual features that are not visible on smaller image samples. It is also possible that the phenomena that is small in size and covers only a small area of a high resolution image and coupled with a small sample image size results in a low number of training data samples containing this category. This consequently results in a low classifier performance for that class. For larger image sizes phenomena is contained in more training samples and this results in a better performance.

Achieved results strongly support further research in this direction as they are much higher than the ones obtained on the same dataset by a Cogent Confabulation based expert system. Based on the research presented in this paper, we have set the guidelines for further work. First of all, to continue research in this direction it is necessary to extend the dataset of natural landscape images. This is not only important for training neural network models, but also to establish a standardized testbed and methodology to be able to compare different methods. This methodology should comprise and reveal fine details of the *per class* classifier performance and give clear and comparable results for different classifiers.

Result achieved by a DNN model trained on general image dataset and transfer learning approach supports further research on the topic discussed in this work. However, further efforts will be steered towards the development of a novel neural network model specifically designed for natural landscape image

classification which will incorporate specificities of the problem. Regardless of the architecture, designed deep learning model should work on multi-scale resolution, detecting features and assigning class labels on different resolutions.

Acknowledgment

This work was partly supported by the Ministry of Science, Education and Sport of the Republic of Croatia under grant "ViO - Vision Based Intelligent Observers" (in Croatian: "ViO - Vidom temeljeni inteligentni opserveri"). This work is part of activities of ACROSS - Centre of Research Excellence for Advanced Cooperative Systems (http://across.fer.hr).

References

[Alsabahi et al. 2018] Alsabahi, Y. A. L, Fan, L., Feng, X.; "Image Classification Method in DR Image Based on Transfer Learning"; 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA), Xi'an (2018) pp. 1-4.

[Ardiyanto and Adji 2017] Ardiyanto, I., Adji, T. B.: "Deep residual coalesced convolutional network for efficient semantic road segmentation"; 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), (May 2017) 378-381.

[Bartyzal] Bartyzal, R.: "Multi-label image classification with Inception net"; `https://towardsdatascience.com/multi-label-image-classification-with-inception-net-cbb2ee538e30`.

[Braović et al. 2017] Braović, M., Stipaničev, D., Krstinić, D.: "Cogent Confabulation based Expert System for Segmentation and Classification of Natural Landscape Images"; Advances in Electrical and Computer Engineering, 17 (2017), 85-94.

[Brodersen et al. 2010] Brodersen, K.H., Ong, C.S., Stephan, K., Buhmann J. M. (2010) "The Balanced Accuracy and Its Posterior Distribution"; In Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR '10), 3121-3124; 23.-26. Aug. 2010.; Istanbul, Turkey

[Brostow et al. 2009] Brostow, G. J., Fauqueur, J., Cipolla, R.: "Semantic Object Classes in Video: A High-definition Ground Truth Database"; Pattern Recognition Letters, 30, 2 (Jan 2009) 88-97.

[Bugarić et al. 2013] Bugarić, M., Braović, M., Stipaničev, D.: "Augmented reality based segmentation of outdoor landscape images"; 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA) (Sep) 43-48.

[Buscombe and Ritchie 2018] Buscombe, D., Ritchie, A.C.: "Landscape Classification with Deep Neural Networks"; Geosciences, 8, 7 (2018), 244.

[Chang et al. 2017] Chang, J., Yu, J., Han, T., Chang, H., Park, E.; "A Method for Classifying Medical Images using Transfer Learning: A Pilot Study on Histopathology of Breast Cancer"; 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), Dalian (2017) pp. 1-4.

[Cengil and Çinar 2019] Cengil, E., Çinar, A.; "Multiple Classification of Flower Images Using Transfer Learning"; 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey (2019) pp. 1-6.

[Chen et al. 2018] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L.: "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs"; IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, 4 (2018) 834-848.

[Chen et al. 2019] Chen, G., Li, C., Wei, W., Jing, W., Woźniak, M., Blažauskas, T., Damaševičius, R.; "Fully Convolutional Neural Network with Augmented Atrous Spatial Pyramid Pool and Fully Connected Fusion Path for High Resolution Remote Sensing Image Segmentation"; Applied Sciences, 9 (2019), 1816.

[Deng et al. 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. : "ImageNet: A Large-Scale Hierarchical Image Database"; CVPR09 (2009).

[Everingham et al. 2007] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A.: "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results"; (2007) http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[Everingham at al. 2009] Everingham, M. and Van Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A.: "The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results"; (200) http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[Everingham et al. 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A.: "The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results"; (2010) http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[Everingham et al. 2011] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A.: "The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results"; (2011) http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[Everingham et al. 2012] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A.: "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results"; (2012) http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[Fei-Fei and Perona 2005] Fei-Fei, L., Perona, P.: "A Bayesian hierarchical model for learning natural scene categories"; 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2 (2005), 524-531.

[Gao et al. 2018] Gao, Q., Lim, S., Jia, X.: "Hyperspectral Image Classification Using Convolutional Neural Networks and Multiple Feature Learning"; Remote Sensing, 10, 2 (2018).

[Geiger et al. 2012] Geiger, A., Lenz, P., Urtasun, R.: "Are we ready for autonomous driving? The KITTI vision benchmark suite"; 2012 IEEE Conference on Computer Vision and Pattern Recognition, (Jun 2012) 3354-3361.

[Gökalp and Aksoy 2007] Gökalp, D., Aksoy, S.: "Scene Classification Using Bag-of-Regions Representations"; 2007 IEEE Conference on Computer Vision and Pattern Recognition (2007), 1-8.

[Google-SIRI-WHO] Google image dataset of SIRI-WHO, http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/Num/Google.html.

[Gould et al. 2009] Gould, S., Fulton, R., Koller, D.: "Decomposing a scene into geometric and semantically consistent regions"; 2009 IEEE 12th International Conference on Computer Vision (Sept 2009) 1-8.

[Hong et al. 2015] Hong, S., Noh, H., Han, B.: "Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation"; Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15), (2015) 1495-1503.

[Hu et al. 2015] Hu, F., Xia, G.-S., Hu, J., Zhang, L.: "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery"; Remote Sensing, 7, 11 (2015), 14680–14707.

[Hu et al. 2018] Hu, G., Yang, Z., Han, J., Huang, L., Gong, J., Xiong, N.: "Aircraft detection in remote sensing images based on saliency and convolution neural network"; EURASIP Journal on Wireless Communications and Networking, 1 (Feb 2018).

[Islam et al. 2016] Islam, Md. A., Bruce, N., Wang, Y.: "Dense Image Labeling Using Deep Convolutional Neural Networks"; 2016 13th Conference on Computer and

Robot Vision (CRV), (Jun 2016) 16-23.

[Jakovčević et al. 2013] Jakovčević, T., Stipaničev, D., Krstinić, D.: "Visual spatial-context based wildfire smoke sensor"; Machine Vision and Applications, 24, 4 (May 2013) 707-719.

[Kendall et al. 2015] Kendall, A., Badrinarayanan, V., Cipolla, R.: "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding"; (2015) `https://arxiv.org/abs/1511.02680`.

[Krause et al. 2013] Krause, J., Stark, M., Deng, J., Fei-Fei, L.: "3D object representations for fine-grained categorization"; Proceedings - 2013 IEEE International Conference on Computer Vision Workshops, ICCVW 2013 (2013) 554-561.

[Krizhevsky et al. 2012] Krizhevsky, A., Sutskever, I., Hinton, G. E.: "ImageNet Classification with Deep Convolutional Neural Networks"; Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12) (2012) 1097-1105.

[Krstinić et al. 2009] Krstinić, D., Stipaničev, D., Jakovčević, T: "Histogram-based smoke segmentation in forest fire detection system"; Information Technology and Control, 38, 2 (2009) 237-244.

[Lam et al. 2017] Lam, M., Mahasseni, B., Todorovic, S.: "Fine-Grained Recognition as HSnet Search for Informative Image Parts"; 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jul 2017) 6497-6506.

[LeCun and Cortes 2010] LeCun, Y., Cortes, C.: "MNIST handwritten digit database"; (2010) `http://yann.lecun.com/exdb/mnist/`.

[Li et al. 2017] Li, Y., Dixit, M., Vasconcelos, N.: "Deep Scene Image Classification with the MFAFVNet"; 2017 IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 5757-5765.

[Liu and Li 2018] Liu, S., Li, M. : "Deep multimodal fusion for ground-based cloud classification in weather station networks"; EURASIP Journal on Wireless Communications and Networking, 2018, 1 (Feb 2018).

[Liu et al. 2018] Liu, Y., Zhong, Y., Fei, F., Zhu, Q., Qin, Q.: "Scene Classification Based on a Deep Random-Scale Stretched Convolutional Neural Network"; Remote Sensing, 10, 444, 2018.

[Mahasseni et al. 2017] Mahasseni, B., Todorovic, S., Fern, A.: "Budget-Aware Deep Semantic Video Segmentation"; 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017) 2077-2086.

[Malūkas et al. 2018] Malūkas, U., Maskeliūnas, R., Damaševičius, R., Woźniak, M.; "Real Time Path Finding for Assisted Living Using Deep Learning"; Journal of Universal Computer Science, 24, 4 (2018).

[Newsam] Newsam, S. D.: "UC Merced Land Use Dataset"; `http://weegee.vision.ucmerced.edu/datasets/landuse.html`.

[Papadopoulos et al. 2007] Papadopoulos, G. Th., Mezaris, V., Kompatsiaris, I., Strintzis, M. G.: "Combining Global and Local Information for Knowledge-Assisted Image Analysis and Classification"; EURASIP Journal on Advances in Signal Processing, 2007, 1 (Jul 2007).

[Payet and Todorovic 2013] Payet, N., Todorovic, S.: "Hough Forest Random Field for Object Recognition and Segmentation"; IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, 5 (2013) 1066-1079.

[Prest et al.] Prest, A., Kalogeiton, V., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: "YouTube-Objects dataset v2.3"; `http://calvin.inf.ed.ac.uk/datasets/youtube-objects-dataset/`.

[Quattoni and Torralba 2009] Quattoni, A., Torralba, A.: "Recognizing indoor scenes"; 2009 IEEE Conference on Computer Vision and Pattern Recognition (Jun 2009) 413-420.

[Rončević et al. 2017] Rončević, T., Braović, M., Stipaničev, D.: "Non-Parametric Context-based object classification in images"; Journal of Information Technology and Control, 46, 1 (2017) 86-99.

[Ronneberger et al. 2015] Ronneberger, O., Fischer, P., Brox, T.: "U-Net: Convolutional Networks for Biomedical Image Segmentation"; Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015) (2015), 234-241.

[Roy and Todorovic 2014] Roy, A., Todorovic, S.: "Scene Labeling Using Beam Search under Mutex Constraints"; 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014) 1178-1185.

[Roy and Todorovic 2017] Roy, A., Todorovic, S.: "Combining Bottom-Up, Top-Down, and Smoothness Cues for Weakly Supervised Image Segmentation"; 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017) 7282-7291.

[Russakovsky et al. 2010] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L.: "ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC2010)"; http://image-net.org/challenges/LSVRC/2010/index,http://image-net.org/challenges/LSVRC/2010/browse-synsets (2010).

[Russakovsky et al. 2012] Russakovsky, O., Lin, Y., Yu, K., Fei-Fei, L.: "Object-Centric Spatial Pooling for Image Classification"; Computer Vision – ECCV 2012 (2012) 1-15.

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

[Russakovsky et al. 2015-2] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L.: "ImageNet Large Scale Visual Recognition Challenge"; International Journal of Computer Vision (IJCV), 115, 3 (2015) 211-252.

[Russell et al.] Russell, B., Torralba, A., Yuen, J., Way, D., Almazan, D. B., Torralba, A., Fernandez, X. P., Davies, S., Cooper, E., Mejia, J., Ayuso, A. : "LabelMe"; http://labelme2.csail.mit.edu/Release3.0/index.php.

[Sadanandan et al. 2017] Sadanandan, S. K., Ranefall, P., Le Guyader, S., Wählby, C.: "Automated Training of Deep Convolutional Neural Networks for Cell Segmentation"; Scientific Reports, 7 (Aug 2017).

[Shotton et al. 2008] Shotton, J., Johnson, M., Cipolla, R.: "Semantic texton forests for image categorization and segmentation"; 2008 IEEE Conference on Computer Vision and Pattern Recognition (Jun 2008) 1-8.

[Shotton et al. 2009] Shotton, J., Winn, J., Rother, C., Criminisi, A.: "TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context"; International Journal of Computer Vision, 81, 1 (Jan 2009) 2-23.

[Socher et al. 2011] Socher, R., Lin, C. C.-Y., Ng, A. Y., Manning, C. D.: "Parsing Natural Scenes and Natural Language with Recursive Neural Networks"; Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11) (2011) 129-136.

[Song et al. 2015] Song, S., Lichtenberg, S. P., Xiao, J.: "SUN RGB-D: A RGB-D scene understanding benchmark suite"; 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (Jun 2015) 567-576.

[Stanford Background Dataset] : "Stanford Background Dataset"; http://dags.stanford.edu/projects/scenedataset.html.

[Štula et al. 2012] Štula, M., Krstinić, D., Šerić, Lj.: "Intelligent forest fire monitoring system"; Information Systems Frontiers, 14, 3 (Jul 2012) 725-739.

[Szegedy et al. 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: "Going deeper with convolutions"; 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 1-9.

[Szegedy et al. 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: "Rethinking the Inception Architecture for Computer Vision"; 2016 IEEE Conference

on Computer Vision and Pattern Recognition (CVPR) (2016), 2818-2826.

[Torrey and Shavlik 2009] Torrey, L., Shavlik, J.: "Transfer Learning"; In Handbook of Research on Machine Learning Applications (2009).

[Tripathi et al. 2016] Tripathi, S., Belongie, S., Hwang, Y., Nguyen, T.: "Detecting temporally consistent objects in videos through object class label propagation"; 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), (Mar 2016) 1-9.

[Vogel and Schiele 2004] Vogel, J., Schiele, B.: "A Semantic Typicality Measure for Natural Scene Categorization"; Pattern Recognition (2004), 195-203.

[Wah et al. 2011] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: "The Caltech-UCSD Birds-200-2011 Dataset"; California Institute of Technology, CNS-TR-2011-001 (2011).

[Wang et al. 2018] Wang, G., Wenqi, L., Zuluaga, M. A., Pratt, R., Patel, P. A., Aertsen, M., Doel, T., David, A. L., Deprest, J., Ourselin, S., Vercauteren, T.: "Interactive Medical Image Segmentation using Deep Learning with Image-specific Fine-tuning"; IEEE Transactions on Medical Imaging (2018).

[Xiao et al. 2010] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., Torralba, A.: "SUN database: Large-scale scene recognition from abbey to zoo"; 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Jun 2010) 3485-3492.

[Xiao et al.] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., Torralba, A.: "SUN database"; `https://groups.csail.mit.edu/vision/SUN/`.

[Yang and Newsam 2010] Yang, L., Newsam, S.: "Bag-of-visual-words and Spatial Extensions for Land-use Classification"; Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10 (2010) 270-279.

[Yu and Koltun 2016] Yu, F., Koltun, V.: "Multi-Scale Context Aggregation by Dilated Convolutions"; International Conference on Learning Representations, (May 2016).

[Zhao et al. 2016] Zhao, B., Zhong, Y., Xia, G.-S., Zhang, L.: "Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery"; IEEE Transactions on Geoscience and Remote Sensing, 54, 4 (Apr 2016) 2108-2123.

[Zheng et al. 2015] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P. H. S.: "Conditional Random Fields As Recurrent Neural Networks"; Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (2015), 1529-1537.

[Zhong] Zhong, Y.: "The Google image dataset of SIRI-WHU"; `http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html`.

[Zou et al. 2015] Zou, Q., Ni, L., Zhang, T., Wang, Q.: "Deep Learning Based Feature Selection for Remote Sensing Scene Classification"; IEEE Geoscience and Remote Sensing Letters, 12, 11 (2015), 2321-2325.