Staircase Detection Using a Lightweight Look-Behind Fully Convolutional Neural Network

Dimitrios E. Diamantis(✉), Dimitra-Christina C. Koutsiou, and Dimitris K. Iakovidis

Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
{didiamantis,dkoutsiou,diakovidis}@uth.gr

Abstract. Staircase detection in natural images has several applications in the context of robotics and visually impaired navigation. Previous works are mainly based on handcrafted feature extraction and supervised learning using fully annotated images. In this work we address the problem of staircase detection in weakly labeled natural images, using a novel Fully Convolutional neural Network (FCN), named LB-FCN *light*. The proposed network is an enhanced version of our recent Look-Behind FCN (LB-FCN), suitable for deployment on mobile and embedded devices. Its architecture features multi-scale feature extraction, depthwise separable convolutions and residual learning. To evaluate its computational and classification performance, we have created a weakly-labeled benchmark dataset from publicly available images. The results from the experimental evaluation of LB-FCN *light* indicate its advantageous performance over the relevant state-of-the-art architectures.

1 Introduction

Staircases can be found almost everywhere in different colors, shapes and sizes in both indoor and outdoor environments. Staircases are useful in everyday life; however, they can be seen also as an obstacle for the navigation of humans with disabilities, as well as the navigation of artificial, robotic, agents. The detection of a staircase can be even more difficult in unknown environments, especially for the visually impaired, where there is no previous knowledge about the surroundings, and they can become hazardous. Therefore, staircase detection can be considered as an important component of any system aiming to provide navigational assistance in either indoor or outdoor environments. In controlled, indoor environments, markers, such as augmented reality markers can be used to provide high success rate of staircase detection [1]. The detection problem usually becomes much harder in outdoor, uncontrolled environments, where different types of staircases of various sizes can be found under various illumination conditions, and can be observed from different viewpoints.

In this paper we address image-based staircase detection as a pattern recognition problem in the context of embedded and mobile devices. The main challenge is to be able to provide sufficient detection accuracy by utilizing the limited computational

J. Macintyre et al. (Eds.): EANN 2019, CCIS 1000, pp. 522–532, 2019.
https://doi.org/10.1007/978-3-030-20257-6_45

resources of such devices, especially in outdoor environments with low latency and limited network accessibility. To address this challenge, we propose a novel lightweight Fully Convolutional neural Network (FCN) architecture as a modification of our recent Look-Behind FCN (LB-FCN) architecture [2]. This novel architecture, named LB-FCN *light*, has significantly fewer free parameters and requires fewer Floating Point Operations (FLOPs) compared to the previous LB-FCN and state-of-the-art architectures for mobile devices. This was achieved by implementing depthwise separable convolutions throughout the convolutional layers of the network. Also, it enables multi-scale feature extraction and residual learning, making it suitable for multi-scale staircase detection in both indoor and outdoor environments. To evaluate the performance of LB-FCN *light* we created a weakly labeled image dataset, with staircases found in natural images collected from publicly available datasets, i.e., a dataset with semantically labeled images as containing or not containing staircases.

The rest of the paper consists of four sections. In Sect. 2 the related work focusing on staircase detection is presented. In Sect. 3 we describe the proposed architecture and its advantages. In Sect. 4 we describe our weakly annotated staircase dataset, and the results of the experiments performed. The last section summarizes the conclusions that can be derived from this study along with our plans for future work.

2 Related Work

Staircase detection has been an active research topic in computer vision and robotics, with an increasing interest nowadays as we are going through the era of ubiquitous computing and pervasive intelligence. One of the first relevant works [3] was based on Gabor filters and concurrent line grouping for distant and close staircase detection respectively. In the context of autonomous vehicle navigation, an outdoor descending staircase detection algorithm was presented by [4], based on texture energy, optical flow, and scene geometry features. In the context of computer aided navigation of visually impaired in outdoor environments using a wearable stereo camera, [5] utilized Haar features and Adaboost learning providing real-time detection performance. A similar approach that utilizes Haar-like features and an improved staircase specific Viola-Jones detector was proposed in [6].

Frequency domain features obtained by ultrasonic sensors were investigated in [7], to detect and recognize floor and staircases in electronic white cane. A wearable RGB-D camera mounted on the chest of a visually impaired individual, was used in [8], where an indoor environment for staircase detection and modeling was proposed. Their approach is capable of providing information for the presence and location along with the number of steps of staircases. Recently an indoor staircase detection framework was proposed in [9], utilizing depth images, capable of running on mobile devices. The approach is based on the detection and clustering of image patches that have the surface vectors pointing to the top direction. In addition, information from the Inertial Measurement Unit (IMU) sensor of the device is used to calibrate the surface vectors with the camera orientation. Most of the current staircase detection approaches are supervised, requiring fully annotated training images from controlled environments, i.e., images indicating the location of the staircases within the images. Furthermore, to the

best of our knowledge the staircase detection has not been previously investigated to a sufficiently generic extent.

Although deep learning and more specifically Convolutional Neural Networks (CNNs) [10] have demonstrated impressive performance in computer vision applications, especially in natural image classification [11], staircase detection approaches have not been previously reported. While they are effective, conventional deep CNNs such as [12], suffer from high computational complexity mainly due to their large number of free parameters. As a result, high-end computational equipment such as Graphical Processing Units (GPUs) is needed for both training and testing time, limiting their use in indoor workstations. Recent studies such as [13–15] focus their interest in computational complexity reduction of CNN architectures, aiming to enable their usage in mobile and embedded devices. In this context, the tradeoff between computational efficiency and detection performance has been investigated, resulting in a state-of-the-art architecture called MobileNet-v2 [16], extending the original MobileNet-v1 proposed in [14]. More specifically this architecture keeps the basic principles of depthwise convolutions for the original design enhances it by adding linear bottleneck layers and shortcut connections between each bottleneck. Linear bottleneck layers were utilized as experimental evidence that the non-linear ones were damaging the extracted features between the bottlenecks. As a result of these changes the architecture contains 30% less parameters than MobileNet-v1 while providing a higher accuracy. Recently, we presented LB-FCN [2] architecture in the context of abnormality detection in medical images. The architecture featured multi-scale feature extraction modules composed of conventional convolutional layers, to better represent the different scales of abnormalities. In addition look-behind connections were used, which connect the input features to the output of each multi-scale feature extraction module. This was required, so that the high-level features will propagate throughout the network, allowing the network to converge faster and increasing the overall detection accuracy.

The core of LB-FCN *light* architecture is inspired by LB-FCN [2] and includes modification to enable efficient computations on mobile and embedded devices, while providing a sufficient staircase detection accuracy. More specifically, LB-FCN *light* extends the original LB-FCN design by replacing the multi-scale conventional convolutional layers with depthwise convolutional layers [17]. Key features of this architecture include the utilization of multi-scale depthwise separable convolution layers [17] and residual learning [18] connections which help to maintain relatively low number of free parameters, without sacrificing the detection accuracy.

3 Architecture

The design of the LB-FCN *light* architecture follows the FCN [19] network design, where only convolutional layers are utilized throughout the network. By replacing the fully connected layers, usually found in the classification layer of conventional CNN architectures such as [11, 12], a significant reduction of the number free parameters of the architecture can be achieved. Inspired by the MobileNet architecture, proposed in [14], depthwise separable convolutions [17] are implemented throughout the network

to further reduce the complexity of the overall architecture. While in conventional convolution the filters are connected on the entire depth of the input channels, in depthwise separable convolution the filter is applied separately on each channel. To connect the separate filters, the layers are followed by a 1×1 conventional convolution.

The main component of LB-FCN *light* is the Multi-Scale Depthwise Convolution module (Fig. 1) which follows the principles established in [2]. This module is capable of extracting features from parallel depthwise separable convolution layers, each one with a different filter size. More specifically the layers extract features at three different scales: 3×3, 5×5 and 7×7 respectively. The feature maps from each layer are then concatenated forming a multi-scale feature representation of the input which is then followed by 1×1 convolution layer. The architecture features residual connections, which connect the input volume of the multi-scale module using adding operator aggregation with the output of it. This is done in order to preserve the higher level features extracted from the previous multi-scale blocks throughout the network.

Following the FCN [19] approach which shows that conventional max pooling operation can be replaced with a convolutional based, we utilized convolutional pooling with filter size 3×3 and stride 2. This introduces another level of non-linearity to the network while keeping the overall architecture logically unified. After each pooling operation the number of extracted filters of each convolutional layer is doubled. In total four multi-scale depthwise convolution modules are utilized in the network with three residual connections as illustrated in Fig. 2. For the staircase detection, a softmax layer of two neurons is used as the output of the network.

Throughout the architecture all convolution layers use ReLU activations followed by output batch normalization. The normalization is used so that the output of the convolution layers are centered on zero mean with the unit standard deviation. It has been empirically confirmed that output normalization can contribute in a faster network converge while reducing overfitting phenomenon. As a result of the above no Dropout layer [20] was used.

While we maintained the multi-scale feature extraction characteristics established in the original LB-FCN [2] architecture, the change in original filter size selection block increased the overall accuracy of the network. Furthermore we utilized conventional ReLU activation functions throughout the network instead of Parametric ReLU that were used in original LB- FCN architecture, which resulted in lower computational complexity without any significant detection performance overhead. The overall improvements made in original LB-FCN architecture, resulted in a significant increase in computational efficiency. As a result, LB-FCN light architecture is capable to efficiently run on mobile and embedded devices.

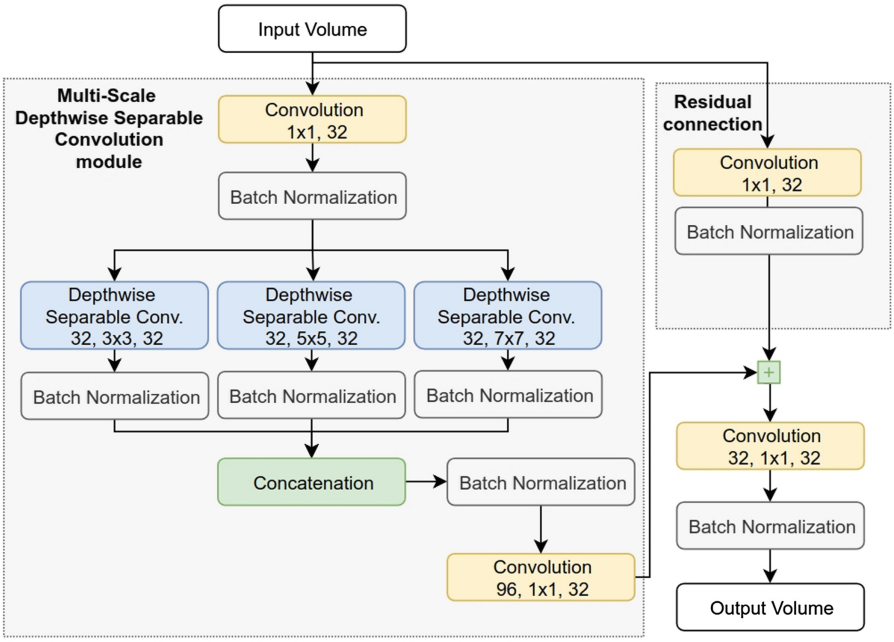

Fig. 1. The main building block of LB-FCN *light* architecture.

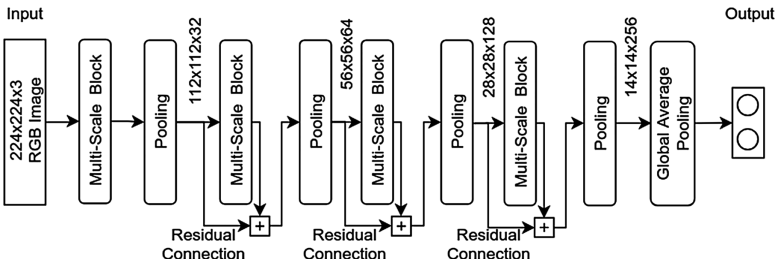

Fig. 2. The complete LB-FCN *light* architecture composed of four multi-scale blocks and three residual connections.

4 Experiments and Evaluation

4.1 Dataset

To evaluate the performance of the proposed architecture in the context of natural image staircase detection we have considered two publicly available datasets. The first dataset, named LM+Sun [21], is a fully annotated natural image dataset obtained from the combination of LabelMe Database [22] and SUN dataset [23]. The dataset consists of 45,676 images from 232 categories, found in indoor and outdoor environment under various conditions and sizes. For the purpose of our experiment we utilized a subset of

LM+Sun dataset which includes natural images found in urban and street areas. While the full LM+Sun dataset contains 314 staircase labeled images, most of them are found in indoor environments. Images containing staircases were also found in the urban and street subsets of this dataset, e.g., staircases of buildings that can be directly recognized by a human observer, considering: (a) staircases that have at least two steps, and (b) staircases covering >15% of the image (in staircases of smaller coverage the steps are not distinguishable; therefore, they cannot be perceived directly as such, without contextual information). To minimize the possibility of a human error in the annotation process, two reviewers separately reviewed and annotated the dataset, and found in total 245 images that include outdoor staircases. To further increase the number of outdoor staircase images, we have created a second dataset named "StairFlickr" which extends LM+Sun staircases with a total of 524 outdoor staircase images. StairFlickr dataset images were obtained from the popular photo management and sharing web application Flickr [24].

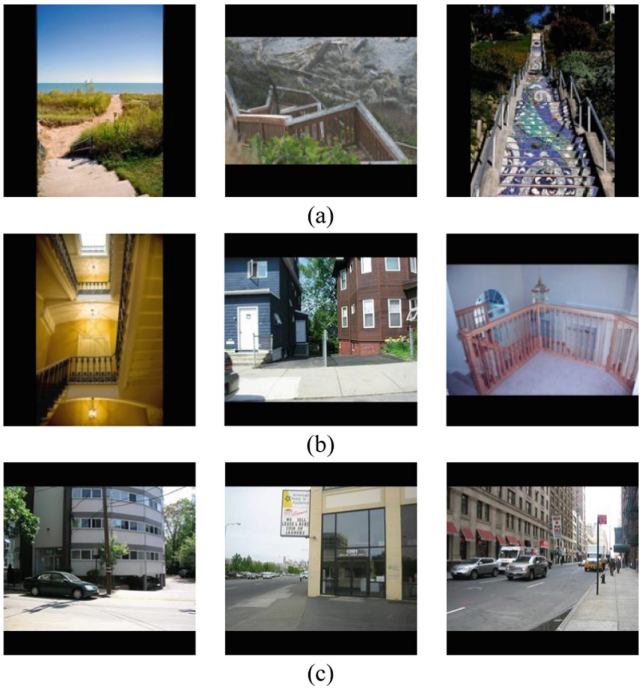

Fig. 3. Top: staircases found in StairFlickr dataset. Middle: staircases found in LM+Sun dataset. Bottom: non-staircases images from LM+Sun dataset.

For the purposes of our research, we omitted the fully annotated metadata provided about the staircases in the original LM+Sun dataset. This was performed as our architecture aims for staircase detection on solely weakly-labeled natural images. In total the described dataset includes 5,539 images from which 1,083 images contain

staircases[1]. Indicative images from this dataset are illustrated in Fig. 3. As it can be observed, the dataset includes various types of staircases found in various positions, sizes, capture from different viewpoints.

4.2 Evaluation Methodology

To evaluate the detection performance of the proposed architecture we followed the stratified 10-fold cross-validation (CV) procedure. The dataset was partitioned into 10 stratified subsets from which 9 were used for training and 1 for testing. This was repeated 10 times, each time selecting a different subset, until all folds have been tested. For each evaluation we calculated the accuracy (ACC), specificity (SPC), and sensitivity (TPR) of the trained model following the Eqs. (1–3), where true positives are denoted as TP, true negatives as TN, false positives as FP and false negatives as FN.

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

$$SPC = \frac{TN}{TP + FP} \tag{2}$$

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$FPR = 1 - SPC \tag{4}$$

To better evaluate the classification performance of the trained network, we utilized the Area Under ROC (AUC) measure. AUC measure is a reliable classification performance measure that is insensitive to imbalanced class distributions [25]. This was chosen as the total number of images containing staircases was significantly fewer than the rest of the rest natural images in the dataset.

4.3 Results

We trained the LB-FCN *light* architecture using the images from both Flickr and LM+Sun datasets. As the images differ from each other in both size and aspect ratio we rescaled the dataset to the standardized input size of the network which is 224 × 224 pixels. To maintain the original aspect ratio of the images, they were padded with zeros to match the network's input dimensions. It is worth mentioning that no further preprocessing step was applied to the images. As the proposed architecture focuses on weakly labeled images, the detailed annotations for the staircases provided by LM+Sun [21] dataset were ignored. We utilized only the semantic annotations of the images which indicate the presence or absence of staircases.

For the training of the network we utilized the Adam [26] optimizer with initial learning rate alpha = 0.001 and first and second moment estimates exponential decay

[1] A link to the dataset will be provided in the final manuscript.

rate beta1 = 0.9 and beta2 = 0.999 respectively. For the implementation of the architecture we utilized the Python Keras [27] library and the Tensorflow [28] tensor graph framework. The network was trained with mini-batch size of 32 samples on NVIDIA TITAN X GPU, equipped with 3584 CUDA [29] cores, 12 GB of RAM and base clock speed of 1417 MHz. On each fold we utilized the early-stopping technique where a small subset of the training fold was utilized as a validation dataset.

To evaluate the effectiveness in both detection accuracy and computational complexity reduction of LB-FCN *light* architecture we used the MobileNet-v2 [16] as a state-of-the-art architecture for comparison. The results obtained by the two architectures are illustrated in Table 1. The confusion matrix of LB-FCN *light* classification performance is illustrated in Table 3.

While the detection performance is slightly higher in case on LB-FCN *light*, the noticeable difference between the two architectures is the computational complexity requirements. Table 2 includes a comparison between the architectures in terms of both the number of trainable free parameters and the total number of required FLOPs. The improvements made on the original LB-FCN design, resulted in a significant reduction of the overall number of FLOPs, from 1.3×10^7 down to 0.6×10^6, and reduction of the free parameters of the network, from 8.2×10^6 down to 0.3×10^6 respectively.

Table 1. Detection performance comparison, using 10-fold cross-validation, between state-of-the-art MobileNet-v2 [16] and our LB-FCN light architecture

Architecture	AUC (%)	Accuracy (%)	Specificity (%)	Sensitivity (%)
LB-FCN *light*	**88.93 ± 1.86**	**91.89 ± 2.12**	**93.80 ± 2.61**	**84.05 ± 3.51**
MobileNet-v2 [16]	87.86 ± 2.11	89.99 ± 2.37	93.58 ± 2.45	83.78 ± 3.22

Table 2. Computation complexity comparison between state-of-the-art MobileNet-v2 [16] and our LB-FCN *light* architecture

Architecture	FLOPs ($\times 10^6$)	Trainable free parameters ($\times 10^6$)
LB-FCN *light*	**0.6**	**0.3**
MobileNet-v2 [16]	4.7	2.2

Table 3. Confusion matrix of LB-FCN *light* classification performance.

	Staircases *actual*	Non-Staircases *actual*
Staircases *predicted*	910	276
Non-Staircases *predicted*	173	4180

5 Conclusions

We proposed a novel lightweight multi-scale FCN architecture that copes with the problem of staircase detection in natural images. To evaluate the performance of the architecture we extended the LM+Sun [21] natural image dataset with staircase images

obtained from Flickr [24] social network. To the best of our knowledge there has been no existing work in this field that utilize solely weakly-labeled images to detect staircases in the natural images. The key features of the proposed LB-FCN *light* architecture can be summarized as follows:

- It has a relatively low number of free parameters requiring an also low number of FLOPs, which makes it suitable to be used on mobile and embedded devices;
- It features multi-scale feature extraction design allowing the architecture to detect staircases of various sizes and under difficult conditions, such as natural images;
- Following the FCN [12] architecture approach it offers a lightweight and logically unified design;
- Compared to MobileNet-v2 [16] network, the proposed architecture offers a relatively lower number of FLOPs and free parameters and a slightly higher detection performance. This makes it attractive for lower-end mobile and embedded devices.

In our future work we are planning to evaluate the performance of the proposed architecture in larger weakly-labeled staircase natural image datasets, to further explore the potential of the architecture. Furthermore we plan to extend the purpose of LB-FCN *light* architecture to include the localization of the staircases within the images, by following a weakly supervised approach.

Acknowledgments. This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code:T1EDK-02070). It was also supported by the Onassis Foundation - Scholarship ID: G ZO 004-1/2018-2019. The Titan X used for this research was donated by the NVIDIA Corporation.

References

1. Yu, X., Yang, G., Jones, S., Saniie, J.: AR marker aided obstacle localization system for assisting visually impairedn. In: IEEE International Conference on Electro/Information Technology (EIT), pp. 271–276 (2018)
2. Diamantis, D.E., Iakovidis, D.K., Koulaouzidis, A.: Look-behind fully convolutional neural network for computer-aided endoscopy. Biomed. Signal Process. Control **49**, 192–201 (2019)
3. Se, S., Brady, M.: Vision-based detection of staircases. In: Fourth Asian Conference on Computer Vision ACCV, vol. 1, pp. 535–540 (2000)
4. Hesch, J.A., Mariottini, G.L., Roumeliotis, S.I.: Descending-stair detection, approach, and traversal with an autonomous tracked vehicle. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5525–5531 (2010)
5. Lee, Y.H., Leung, T.-S., Medioni, G.: Real-time staircase detection from a wearable stereo system. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 3770–3773 (2012)
6. Maohai, L., Han, W., Lining, S., Zesu, C.: A robust vision-based method for staircase detection and localization. Cogn. Process. **15**(2), 173–194 (2014)

7. Bouhamed, S.A., Kallel, I.K., Masmoudi, D.S.: Stair case detection and recognition using ultrasonic signal. In: 2013 36th International Conference on Telecommunications and Signal Processing (TSP), pp. 672–676 (2013)
8. Pérez-Yus, A., López-Nicolás, G., Guerrero, J.J.: Detection and modelling of staircases using a wearable depth sensor. In: Agapito, L., Bronstein, Michael M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 449–463. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16199-0_32
9. Ciobanu, A., Morar, A., Moldoveanu, F., Petrescu, L., Ferche, O., Moldoveanu, A.: Real-time indoor staircase detection on mobile devices. In: 2017 21st International Conference on Control Systems and Computer Science (CSCS), pp. 287–293 (2017)
10. LeCun, Y., et al.: LeNet-5, convolutional neural networks, p. 20 (2015). http://yann.lecun.com/exdb/lenet
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014)
13. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size, arXiv preprint arXiv:1602.07360 (2016)
14. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017)
15. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. ArXiv e-prints, July 2017
16. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
17. Chollet, F.: Xception: Deep learning with depthwise separable convolutions, arXiv preprint, pp. 1610–2357 (2017)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
21. Tighe, J., Lazebnik, S.: SuperParsing: scalable nonparametric image parsing with superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_26
22. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. Int. J. Comput. Vision **77**(1–3), 157–173 (2008)
23. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: IEEE Conference on 2010 Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492 (2010)
24. Flickr Inc., "Find your inspiration. | Flickr." 2019
25. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
26. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

27. Chollet, F.: Keras. (2015)
28. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI, vol. 16, pp. 265–283 (2016)
29. Sanders, J., Kandrot, E.: CUDA by example: an introduction to general-purpose GPU programming. Addison-Wesley Professional, (2010)