Fully Convolutional Neural Networks for Polyp Segmentation in Colonoscopy

Patrick Brandao[1], Evangelos Mazomenos[1], Gastone Ciuti[2], Renato Caliò[2], Federico Bianchi[2], Arianna Menciassi[2], Paolo Dario[2], Anastasios Koulaouzidis[3], Alberto Arezzo[4], and Danail Stoyanov[1]

[1]Centre for Medical Image Computing, University College London, London, UK
[2]The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy
[3]Endoscopy Unit, The Royal Infirmary of Edinburgh, Edinburgh, UK
[4]Department of Surgical Sciences, University of Turin, Turin, Italy

ABSTRACT

Colorectal cancer (CRC) is one of the most common and deadliest forms of cancer, accounting for nearly 10% of all forms of cancer in the world. Even though colonoscopy is considered the most effective method for screening and diagnosis, the success of the procedure is highly dependent on the operator skills and level of hand-eye coordination. In this work, we propose to adapt fully convolution neural networks (FCN), to identify and segment polyps in colonoscopy images. We converted three established networks into a fully convolution architecture and fine-tuned their learned representations to the polyp segmentation task. We validate our framework on the 2015 MICCAI polyp detection challenge dataset, surpassing the state-of-the-art in automated polyp detection. Our method obtained high segmentation accuracy and a detection precision and recall of 73.61% and 86.31%, respectively.

1. INTRODUCTION

Colorectal cancer (CRC) is the most frequent pathology of the gastrointestinal tract. In 2012, the International Agency for Research on Cancer registered 1,360,000 new cases and 694,000 deaths worldwide. Furthermore, the 5-year survival rate of CRC patients is lower than 7% when the disease reaches an advanced stage, however, in cases of early diagnosis, it increases to more than 90% with successful treatment.[1]

Conventional colonoscopy is considered the reference standard for CRC screening and diagnosis. It provides direct visualization of the inner surface of the colon, acquire biopsies and perform therapeutic procedures on early stage neoplastic lesions. Despite the popularity of the method, the success of the exam highly depends on the operator skills, as a high level of hand-eye coordination is required to examine the majority of the colon wall.[1] Leufkens et al. reported that missed polyp detection rates could reach values as high as 25%.[2] One effective way to increase the detection rate in colonoscopy exams is the incorporation of computer-aided diagnostic systems.

Numerous methods for automatic polyp detection in colonoscopy have been proposed,[3,4] the majority of which can be roughly grouped in two categories: texture/colour based and shape based. Recently, deep learning approaches were successfully applied for polyp detection in colonoscopy videos.[5,6] However, the proposed methods were tested in different datasets, which limits the comparison between the reported results. To combat this issue, a validation framework was proposed in the 2015 MICCAI sub-challenge on automatic polyp detection.[7]

The problem of automatic polyp detection is quite challenging. As shown in Figure 1, polyps have a large variety of shapes, sizes, colours and textures, which in conjunction with specular reflections, blood vessels and endoluminal folds, impedes very high automatic detection accuracy in colonoscopy recordings. Despite significant progress in recent years, the best method in the 2015 MICCAI sub-challenge on automatic polyp detection, the

Further author information:
Patrick Brandao: patrick.brandao.15@ucl.ac.uk
Evangelos Mazomenos: e.mazomenos@ucl.ac.uk
Danail Stoyanov: danail.stoyanov@ucl.ac.uk

Medical Imaging 2017: Computer-Aided Diagnosis, edited by Samuel G. Armato III, Nicholas A. Petrick,
Proc. of SPIE Vol. 10134, 101340F · © 2017 SPIE · CCC code: 1605-7422/17/$18 · doi: 10.1117/12.2254361

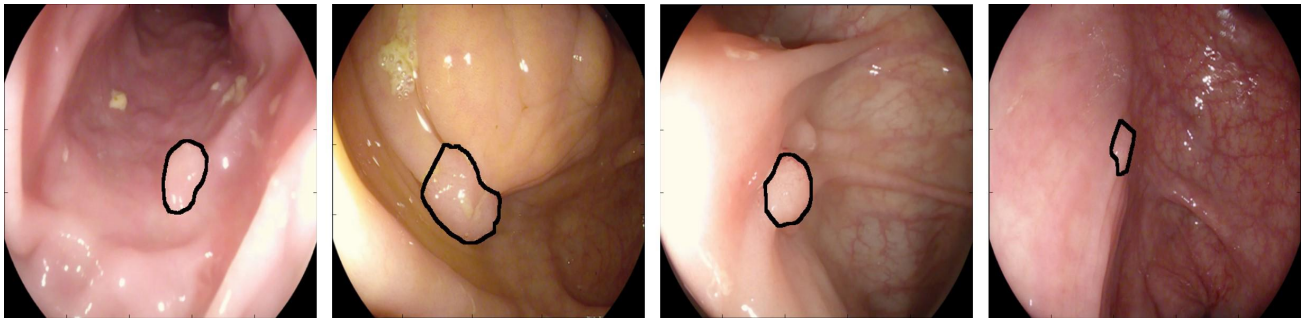

Figure 1: Examples of polyp images from the 2015 MICCAI sub-challenge on automatic polyp detection.[7] The contours of the manual polyp segmentations are overlaid in black lines.

CUMED submission, achieved only a 69.2% detection recall and 72.3% precision, which is still far from routine clinical use.[7]

In this study, we propose an automatic polyp detection system based on a convolutional neural network (CNN) framework, for assisting clinicians in accurately detecting polyps and high-risk regions in colorectal exams. We show that fully convolution network (FCN) architectures can be refined and adapted to recognize specific structures in colonoscopy images. We validate our approach with public datasets and significantly outperform current state-of-the-art methods.

2. MATERIALS AND METHODS

Propelled by large scale challenges, such as ImageNet,[8] deep learning revolutionized many fields in computer vision, surpassing traditional methods in classification, segmentation, detection and tracking problems. Neural networks are traditionally applied in image classification problems, however, some approaches were able to use CNNs for coarse inference by labelling each pixel with the class of its enclosing object. This can be achieved with post-processing by super-pixel projection, multi-scale approaches or patch-wise training. Alternatively, Long et al. proposed a fully convolution neural network (FCN) learned end-to-end, where dense prediction is obtained with in-network deconvolution layers.[9] We employ an adapted versions of these networks, configured specifically for polyp segmentation.

2.1 Dataset

In this work, we used the public datasets from the MICCAI 2015 polyp detection challenge for training and testing.[7] These are composed of three different databases obtained with different imaging systems and resolutions:

- CVC-CLINIC: 612 SD training frames with at least one polyp each;
- ETIS-Larib: 196 HD testing frames with at least one polyp each;
- ASU-Mayo: 36 small SD and HD videos sequences, containing training frames with and without polyps.

The method presented in our current work was developed after the 2015 polyp detection MICCAI challenge, so it could not be submitted as an entry. However, for comparison, our methodology follows the same data guidelines and restrictions.

2.2 Fully Convolution Networks for polyp detection

CNNs are bio-inspired models that follow the connectivity pattern of brain neurons. Each neuron in a artificial CNN consists of a pre-learned filter that is convoluted with the input. Depending of the strength of the filter response, an activation function decides if the neuron fires an output or not.

The neurons of a neural networks are grouped in layers, where the outputs of the neurons in the same layer are used as the input of the neurons of the next one. This creates a feed-forwarded interconnected network of convolutions that, due to the activation function, is able to create complex nonlinear predictive models. Beyond

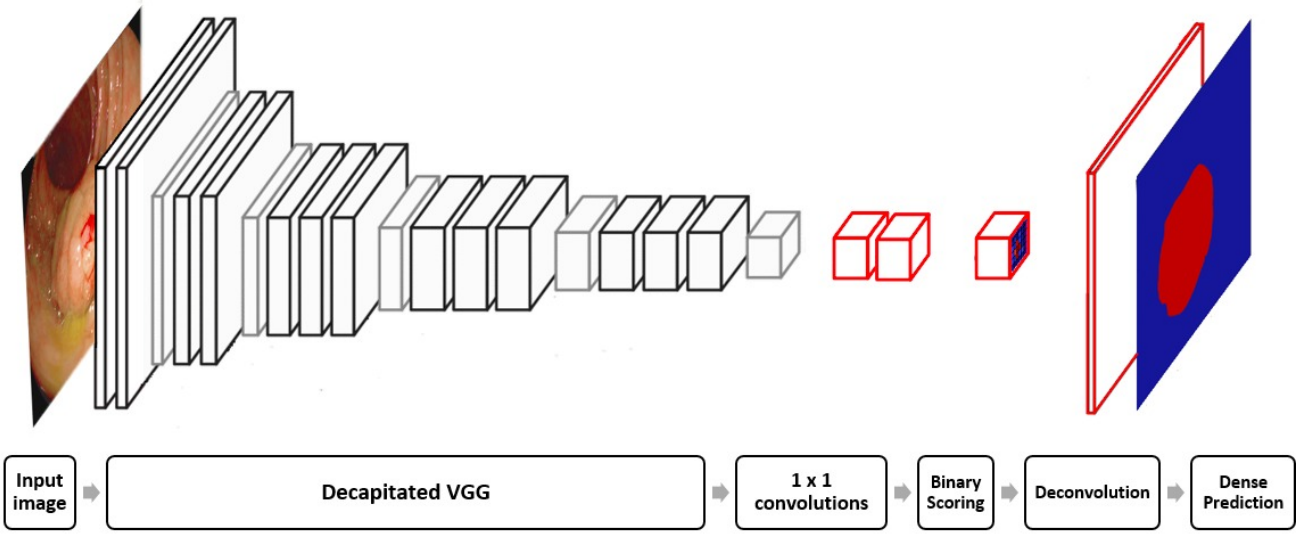

Figure 2: Illustration of a FCN-VGG for polyp detection. The decapitated VGG consists of the VGG network architecture[11] without the fully connected and scoring layers. The layers in red color are the ones that were added to form FCN-VGG. Image adapted from.[12]

convolution and activation, it is common to use pooling operations within the network, which down-sample input representation and significantly reduce computation cost.[10]

Traditionally, CNNs are trained using a loss function that measures the difference between the predicted results and the desired output. The parameters of the network are updated via stochastic gradient descend by back-propagating the gradients using the chain rule.[10]

In classification CNNs, networks end with one or more fully connected layers, which allows the production of non-spatial outputs. These layers can be viewed as a convolution where the kernel has the same dimensions as the input. By replacing these with convolutions, it is possible to convert traditional classification networks into FCNs. Even though the output maps can yield any size, these are typically reduced by subsampling within the network. To connect these coarse outputs to dense pixels, we use deconvolution as the interpolation strategy.[9]

When convolving an image with a stride S, and $S > 1$, convolution essentially performs a subsampling operation. Using this principle, by simply reversing forward and backward pass operations, it is possible to implement in-network up-sampling of factor S. This results in a very fast and effective up-sampling that achieved state-of-the-art in semantic segmentation of the PASCAL VOC dataset.[9]

We develop FCNs based on three well-known convolutional architectures, which achieved excellent results in different ILSVRC challenges: AlexNet,[13] GoogLeNet[14] and VVG.[11]

AlexNet has 5 convolutional layers alternated with max-pooling and three fully connected layers. In 2012, it was the first CNN to win the ILSVRC challenge[8] and it kickstarted the CNN revolution in computer vision.[13] GoogLeNet won the 2014 ILSVRC, and it differentiate itself from other architectures by replacing the traditional convolutional layers by "Inception modules". This module performs a series of parallel convolutions and pooling operations using different size filters. This allows to cover receptive fields with different size and thus extract more information.[14] Finally, VGG also achieved very good results in the 2014 ILSVRC by replacing the AlexNet's 11×11 filters by sequences of two or three 3×3 convolutional layers. Other characteristic is that as spatial dimensions are reduced by pooling, the number of filters is increased, based on the idea that with shrinking dimensionality the CNN should increase its depth.[11]

We converted these CNNs into FCNs by discarding every fully connected layer and replacing it with a 1×1 convolution layer with the same channel dimension. The last convolution is 2-dimensional to reflect the binary

nature of our problem. For GoogLeNet, the final average pool is discarded and only the final loss is used for fine-tuning. Every network is finalized with a deconvolution layer with 32 stride and 64×64 kernel, responsible for upsampling the coarse output to a dense scoring map with the same dimensions as the input. The proposed FCN-VGG is illustrated in Figure 2.

2.3 Experimental validation

In total, the three MICCAI-challenge datasets have 19514 frames. However, only 4664 of these corresponds to images with polyps. Due to this disparity, networks trained with the full dataset did not converge satisfactory and results were not clinically useful. We verified that by fine-tuning the FCNs with only polyp images, better results were achieved. Random mirroring was used during training.

All layers were fine-tuned by stochastic gradient descend with momentum and back-propagation. We chose the highest fixed learning rate that did not cause divergence. This corresponded to 10^{-9} for FCN-Alexnet and FCN-VGG and 10^{-11} for FCN-GoogLeNet. A momentum of 0.9 was used and the learning rate for bias was doubled. We zero initialize the scoring layer and increased its learning rate by a factor of 10. Convergence took 150K iterations for FCN-AlexNet and FCN-GoogLeNet and 50K for FCN-VGG. All models were trained and tested with Caffe[15] in a single NVIDIA Tesla K40 GPU.

3. RESULTS

The developed FCNs were used to segment polyps from the ETIS-Larib dataset of the MICCAI challenge. We report our results in terms of polyp segmentation and detection.

3.1 Segmentation

As far as we know, this is the first work to generate full dense polyp segmentations. Since there is no segmentation state of the art, we decided to evaluate our method using common segmentation evaluation metrics: mean pixel precision and mean pixel recall. These results are presented in the Table 1. Furthermore, examples of all CNN segmentations for three different polyp segmentations can viewed in Figure 3.

Table 1: Mean pixel precision and recall calculated with the segmentations obtained by the proposed FCNs in the ETIS-Larib dataset

Networks	Mean pixel precision (%)	Mean pixel recall (%)
FCN-AlexNet	27.87	35.54
FCN-GoogLeNet	25.83	37.82
FCN-VGG	70.23	54.20

Overall, all networks correctly identified at least 35% of the pixels with polyps. For example, in Figure 3, all networks achieved good segmentations for the first polyp. However, it is clear that FCN-VGG vastly outperform the other architectures, with a mean pixel recall of 54.20%. Even bigger discrepancies are demonstrated in the mean pixel precision, which means that FCN-VGG create far less false alarms. This can be seen in Figure 3, where FCN-Alexnet and FCN-GoogLeNet perform poorly in segmenting the polyp in the second image (FCN-GoogLeNet segments a greater area) and miss the polyp in the last case. On the other hand, FCN-VGG produces very accurate segmentation for all three polyps.

3.2 Detection

As stated before, this work was developed after the 2015 MICCAI challenge, so it is not one of the competing methods. However, for comparison purposes, we used the segmentations outputted by our method to calculate the polyp detection rate with the metrics advocated by the MICCA-challenge directives. For detection, we consider segmentation blobs that intercept with polyps ground truth as true positives and as a false positive otherwise. Polyps that do not overlap with any blobs are accounted as false negatives.

In Table 2 we can see that FCN-AlexNet and FCN-GoogLeNet achieved detection recalls comparable to the OUS method. However, these networks achieve it with a low precision, which mean they have a much higher

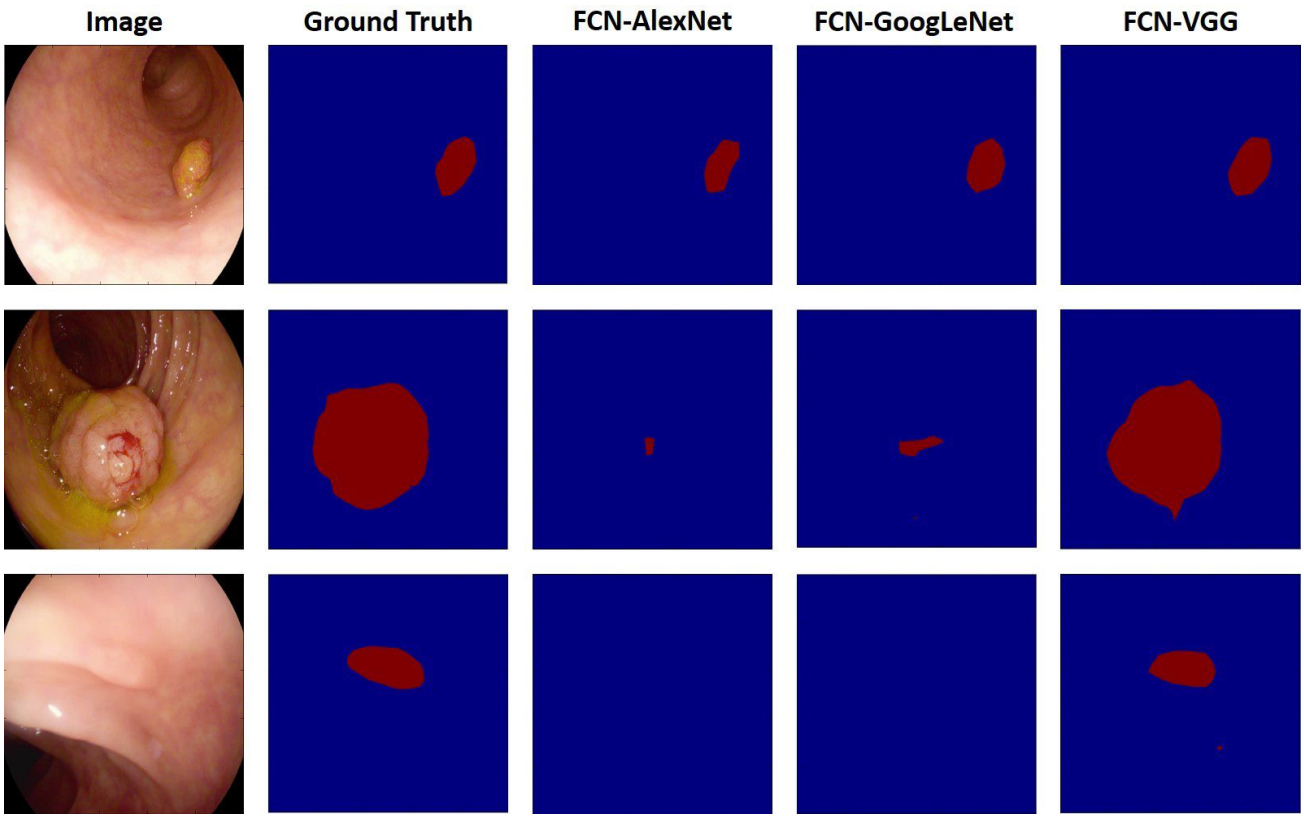

Figure 3: Example of three segmentation produced by three FCN networks.

Table 2: Detection precision and recall rates obtained by the proposed FCNs on the ETIS-Larib dataset. The two best 2015 MICCAI challenge methods are shown for comparison

Method	Precision (%)	Recall (%)
FCN-AlexNet	44.08	63.78
FCN-GoogLeNet	41.85	65.76
FCN-VGG	73.61	86.31
CUMED	72.30	69.20
OUS	69.50	63.00

rate of false positives. The relative small amount of parameters in FCN-AlexNet are not enough to successfully encode all the colon information with the available training data, and it has the worst performance of all networks. GoogLeNet has a deeper architecture which should, in theory, allow it to perform better than AlexNet. However, inception modules are notorious hard to train and, with the limited amount of data available, FCN-GoogLeNet performed similarly to FCN-AlexNet.

Just as for segmentation, Table 2 shows that FCN-VGG achieves the best performance. With a detection precision of 73.3%, FCN-VGG achieved a detection recall of 86.31%, greatly outperforming all other methods, including the best method in the 2015 polyp detection MICCAI challenge. The neurons of the AlexNet trained on the Imagenet dataset are well known for representing very general feature extractors, specially in the first couple of layers where it can clearly be seen variations of edge, color and texture filters. A lot of these filters proved themselves useful in the polyp detection problem because similar characteristics can be used to identify normal and abnormal tissue. Without this initialization, it would be hard to learn these kind of filters with the limited amount of medical images available. Moreover, the deeper and higher dimensionality of the FCN-VGG architecture allows it to better encode the complex colon scene than its shallower counterpart, the FCN-AlexNet.

Small false positives, like the one in the third row, are responsible for a reduction in detection precision of the FCN-VGG but could be easily removed with simple post-processing.

3.3 Computation speed

The inference speed of a network depends on the amount operation performed during the forward pass. Networks with a higher amount of learned parameters and layers take longer to produce an output.

Inference for a single 500×500 image requires, on average, 60 ms, 51 ms, and 295 ms for FCN-VGG, FCN-GoogLeNet and FCN-ALexNet, respectively. FCN-AlexNet is the fastest polyp segmentation architecture, requiring an average of 51 ms to process an image. As expected, the relative low amount of parameters that hinders the CNN accuracy, allows it to nearly run in real-time. The inception modules of FCN-GoogLeNet allow the network to keep a low amount of total parameters. This means that, despite having a deeper architecture, FCN-GoogLeNet is only slightly slower that FCN-AlexNet. Finally, although the large amount of parameters in FCN-VGG, facilitates high-accuracy, it also makes it approximately 5 times slower than other architectures.

4. CONCLUSION

In this paper we develop FCN architectures by extending traditional CNNs, for polyp detection in colonoscopy images. To our knowledge, this is the first work to develop and evaluate FCNs for polyp detection in colonoscopy images. Furthermore, this is the first method to produce full polyp segmentations. We surpassed the state-of-the-art in polyp detection and achieved high segmentation accuracy.

From the three architectures presented, FCN-VGG achieved the best overall results, with a polyp detection precision and recall of 73.61% and 86.31%, respectively. Our future work will be focused in improving precision accuracy by gradually incorporating the remaining background images available in the MICCAI challenge databases into the fine-tuning stage. Recent advances in deep learning, such as batch normalization and residual networks, might provide a good tool to further improve segmentation accuracy.

5. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Horizon 2020 Programme (H2020/2014- 2020) under Grant Agreement num. 688592 (EndoVESPA Project). Danail Stoyanov receives funding from the EPSRC (EP/N013220/1, EP/N022750/1, EP/N027078/1, NS/A000027/1), The Wellcome Trust (WT101957, 201080/Z/16/Z) and the EU-Horizon2020 project EndoVESPA (H2020-ICT-2015-688592).

REFERENCES

[1] Ciuti, G., Caliò, R., et al., "Frontiers of robotic endoscopic capsules: a review," *Journal of Micro-Bio Robotics* **11**(1), 1–18 (2016).
[2] Leufkens, A. M., Van Oijen, M. G., et al., "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy* **44**(5), 470–475 (2012).
[3] Tajbakhsh, N., Gurudu, S. R., et al., "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.* **35**(2), 630–644 (2016).
[4] Bernal, J., Sanchez, F. J., et al., "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imaging Graph.* **43**, 99–111 (2015).
[5] Tajbakhsh, N., Gurudu, S. R., et al., "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in [*Proc. IEEE ISBI*], 79–83 (2015).
[6] Brandao, P., Mazomenos, E., et al., "Validating convolution neural networks for automatic polyp detection in robotic colonoscopy," in [*Proc. CRAS*], 1–2 (2016).
[7] Bernal, J., Tajkbaksh, N., et al., "Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge," *(submitted) IEEE Trans. Med. Imag.* (2016).
[8] Russakovsky, O., Deng, J., et al., "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vision* **115**(3), 211–252 (2015).
[9] Long, J., Shelhamer, E., et al., "Fully convolutional networks for semantic segmentation," in [*Proc. IEEE CVPR*], 3431–3440 (2015).
[10] Dumoulin, V. and Visin, F., "A guide to convolution arithmetic for deep learning," 1–28 (2016).

[11] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," in [*Proc. ICLR*], 1–14 (2015).
[12] Noh, H., Hong, S., et al., "Learning Deconvolution Network for Semantic Segmentation," in [*Proc. IEEE ICCV*], 1520–1528 (2015).
[13] Krizhevsky, A., Sutskever, I., et al., "Imagenet classification with deep convolutional neural networks," in [*Adv. Neural Inf. Process. Syst.*], **25**, 1097–1105.
[14] Szegedy, C., Liu, W., et al., "Going deeper with convolutions," in [*Proc. IEEE CVPR*], 1–9 (2015).
[15] Jia, Y., Shelhamer, E., et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," *Proc. ACM MM* , 675–678 (2014).