# Fast Aircraft Detection Using End-to-End Fully Convolutional Network

Ting-bing Xu Guang-liang Cheng Jie Yang Cheng-Lin Liu
National Laboratory of Patter Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing 100190, China
Email: {tingbing.xu, guangliang.cheng, yangjie, liucl}@nlpr.ia.ac.cn

***Abstract*—Aircraft detection from remote sensing images of complex background is a challenging task. Existing aircraft detection methods usually consist of two separated stages: proposal generation and window classification, which may be suboptimal for the aircraft detection task. To overcome this shortcoming, a unified aircraft detection framework is proposed to simultaneously predict aircraft bounding boxes and class probabilities directly from an arbitrary-sized remote sensing image. Specifically, an end-to-end fully convolutional network (FCN) replaces the fully connected layers in traditional CNN model with convolutional layers, which can greatly reduce the model size while obtaining the comparable detection accuracy. To directly detect aircrafts under multiple scales and different aspect ratios, multiple referenced boxes are introduced. The overall framework can be optimized by minimizing a multi-task loss with end-to-end training. Extensive experiments on a common dataset have demonstrated that the proposed method yields much lower false alarm rates at different detection rates than the state-of-the-art methods, and its speed is more than 35 times faster than the compared methods.**

***Index Terms*—Aircraft Detection, Fully Convolutional Network, End-to-End.**

## I. Introduction

Aircraft detection from remote sensing images finds wide applications in military surveillance. It is a challenging task because aircrafts are not salient objects and have multi-size and multi-color characteristics. Although various methods [1]–[8] have been proposed to address the aircraft detection task, it remains far from being well solved.

Generally speaking, an object detection system consists of two stages: generating candidate regions (proposals) and verifying regions by window-based classification. This strategy has also been used in many aircraft detection systems [1]–[8]. Over the past decades, the sliding window approach [9], [10] has been widely used for object locating. This method tries to locate objects at arbitrary location and scale in the image. Though it can achieve satisfactory recall rate, it is highly time-consuming because there are millions of windows per image need to be processed. To alleviate this shortcoming, some approaches for generating region proposals [11]–[14] have been proposed in recent years. They can generate a moderate number of around one or several thousand proposals per image. Integrating such proposal generators can largely accelerate object detection algorithms [15]–[17]. However, the time-consuming of proposal generation is still considerable and this process is hardly realized through GPU. For further acceleration, some new proposal methods based on convolutional features, such as regional proposal network (RPN) [18], MultiBox [19] and DeepMask [20] are greatly suitable for implementation on GPU.

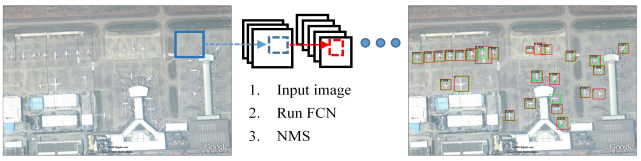


Fig. 1. **Aircraft Detection System.** (1) input image (of arbitrary size), (2) fully convolutional network on the image, (3) Non-maximum suppression (NMS) detection confidence thresholding.

As for aircraft detection, Hsieh et al. [1] introduced a method using invariant moments, wavelet and support vector machine (SVM). Yildiz et al. [2] integrated Gabor filters and SVM to detect stationary aircrafts. A key-points and sparse coding model was proposed by Liu et al. [3] for aircraft detection. Xu et al. [4] introduced an artificial bee colony algorithm with edge potential function to detect aircrafts. Li et al. [5] detect aircrafts using saliency computation and symmetry. These methods have a common shortcoming: they fail to detect tiny and blurred aircrafts under the complex backgrounds. To achieve better detection performance, Chen et al. [6], [7] used multiple thresholds for Localization and deep belief networks (DBNs) or deep CNN model were used to classify aircraft regions. Similarly, Wu et al. [8] combined BING technique and CNN model to fulfill aircraft detection.

The above detection methods still consist of two separated stages, thus they can not be optimized in end-to-end framework. To overcome the above shortcoming, the YOLO [22] and SSD [23] used a single deep neural network to perform localization task and detection task simultaneously. They can run fast and achieve a competitive performance. However, their input images need to be warped to a fixed size (such as $448 \times 448$ or $500 \times 500$), which may not be suitable for aircraft detection. To our knowledge, the size of remote sensing image tends to be large with complex backgrounds and small aircraft targets. The warped strategies may resize the large image to a small one, which increases the difficulty of aircraft detection.

To overcome the above shortcomings and effectively detect

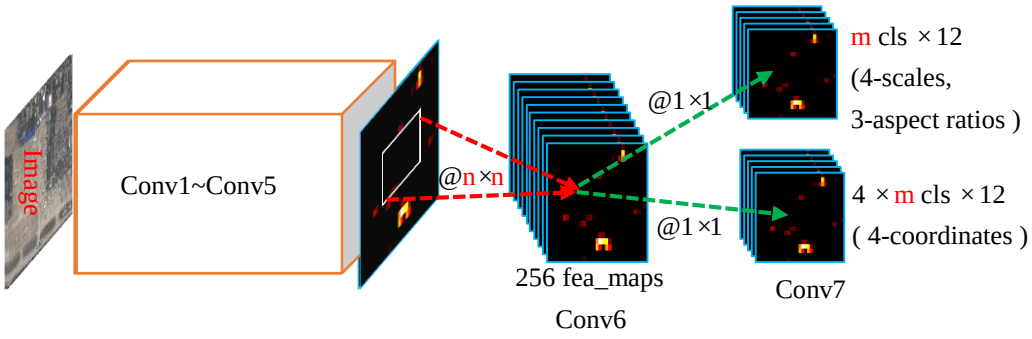


Fig. 2. **FCN Architecture.** Our aircraft detection network has 7 conv-layers, and the seventh layer includes two $1 \times 1$ conv-layers (box-classification layer and box-regression layer). Conv1∼Conv5 come from the Zeiler-Fergus (ZF)-Net [26].

aircrafts from remote sensing images, a unified framework based on single fully convolutional network is proposed. It can process arbitrary-sized images and achieves simultaneously low time complexity and high accuracy rate. The overview of the proposed method is shown in Fig. 1. The main highlights of the proposed method are summarized as follows:

1) A unified aircraft detection framework using an end-to-end fully convolutional network is proposed. It can simultaneously predict aircraft bounding boxes and class probabilities.
2) The FCN can greatly reduce the parameters by replacing CNN model's fully-connected layers with convolutional layers.
3) Our method achieves better performance than state-of-the-art methods, and is 35 times faster than the previous DBN-based method.
4) To detect multi-sized and different aspect-ratios aircrafts, a multi-classified reference boxes [18] is introduced.

The reminder of this paper is organized as follows. Section 2 presents model design. Section 3 shows experimental results and analysis. Concluding remarks will be drawn in Section 4.

## II. Model Design

Our network architecture is inspired by the Region Proposal Network (RPN) [18]. To reduce the number of candidate windows, region proposal methods yield a set of class-independent candidate regions, which tell the detector where to focus. Some typical proposal generation methods, such as Selective Search [12], BING [13] and EdgeBox [14], rely inexpensive features or inference schemes to achieve high speed. The key disadvantage of these methods is can not be realized through GPU. Recently, the methods like RPN [18], MultiBox [19] and DeepMask [20] extract rich convolutional features to improve the quality of proposals. We propose that such rich features not only facilitates proposal generation, but also enables direct detection instead of traditional two-stage approaches. Since fully connected layers are not suitable for object locating due to the spatial information loss, we will replace routine CNN model's fully connected layers with convolutional layers, forming a single fully convolutional network (FCN). The FCN is used to simultaneously predict multi-sized aircraft boxes and give the probabilities of aircraft class for those boxes. This unified architecture can be trained end-to-end and can be easily implemented by a GPU.

**Network Architecture:** Networks on Convolutional feature maps (NoC) [25] and Residual Network (ResNet) [24] showed that the recognition capability of convolutional (conv) layers is higher than fully connected (fc) layers. Our FCN architecture only uses two conv layers to replace fc layers in the traditional CNN model. The full network architecture is shown in Fig. 2. This network takes an image of arbitrary size as input and outputs a set of candidate aircraft regions, each with a confidence score. He et al. [16] indicate that the conv layers operate in a sliding-window manner and output feature maps that represent the spatial arrangement of the activations. Therefore, conv operations do not require a fixed input image size and can yield feature maps of arbitrary. As we can see from Fig. 2, a unit on conv6 feature map extracts a $n \times n$ spatial window over conv5 feature map, which corresponds to an effective receptive field ($171 \times 171$ for $n = 3$ and ZF-Net) of input image. Every spatial position (corresponding to a region of input image) on conv6 feature map obtains a 256-dimension feature vector, which is fed into the box-classification layer (cls) and box-regression layer (box). For the aircraft detection task, there are only two classes (i.e., aircraft and background), thus we set $m$ (in Fig. 2) as 2 in our experiments.

**Multi-scale Design:** In Fig. 2, a unit on conv6 only corresponds to a fixed-size receptive region of input image. Thus the system is not suitable for detecting various sized aircraft targets. The design of multi-classified referenced boxes [18] can friendly solve our multi-scale and multi-aspect ratios problem. Traditional methods use pyramids of input image to solve this problem, but have high time complexity. The spatial pyramid pooling network (SPP-Net) [16] uses pyramids of filters, and multi-sized convolutional kernels to extract features on conv5 feature map (variable $n$ in Fig. 2), but suffers from high implementation complexity. We select multi-sized predefined boxes that can be taken as pyramids of box-regression reference. In this way, a unit on a feature map of conv7 layer represents a certain scale and aspect ratio referenced box. We set 12 classified boxes with 4 scales and 3 aspect ratios, thus, a fixed receptive region on input image

TABLE I
SIZES OF 12 CLASSIFIED (4 SCALES AND 3 ASPECT RATIOS) REFERENCED BOXES.

| Set | $21^2$,6:5 | $38^2$,6:5 | $68^2$,6:5 | $118^2$,6:5 | $21^2$,1:1 | $38^2$,1:1 | $68^2$,1:1 | $118^2$,1:1 | $21^2$,4:5 | $38^2$,4:5 | $68^2$,4:5 | $118^2$,4:5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size | 24×19 | 43×33 | 76×59 | 133×103 | 21×21 | 38×38 | 68×68 | 118×118 | 20×24 | 36×43 | 63×76 | 111×133 |

will have 12 referenced boxes. The specific sizes are shown in Table 1. For locating, a box needs 4 coordinates, here we use the parameterizations of coordinates as below [18]:

$$\begin{aligned} b_x &= (x - x_r)/w_r, b_y = (y - y_r)/h_r, \\ b_w &= log(w/w_r), \quad b_h = log(h/h_r), \\ b_x^* &= (x^* - x_r)/w_r, b_y^* = (y^* - y_r)/h_r, \\ b_w^* &= log(w^*/w_r), \quad b_h^* = log(h^*/h_r), \end{aligned} \tag{1}$$

where $x$ and $y$ denote the center coordinates of a box, $w$ and $h$ are its width and height. $x$, $x_r$ and $x^*$ (similarly to $y$, $w$ and $h$) are the x-coordinate of the predicted box, referenced box and ground-truth box, respectively. $b$ denotes the 4 parameterized coordinates of predicted box from values of box-regression layer and $b^*$ denote the 4 parameterized coordinates of ground-truth box. Thus, we will obtain the coordinates $(x, y, w, h)$ of predicted box by combing the result $(b_x, b_y, b_w, b_h)$ of box-regression layer with consistent referenced box ($x_r$, $y_r$, $w_r$, $h_r$) during the test stage. Through this design, the whole aircraft detection system can efficiently predict aircraft targets at a wide range of scales and aspect ratios when only one scale input image is processed by the network.

**Training Scheme:** For training the single fully convolutional network, we assign labels to the units of box-classification layer and box-regression layer. Every feature map represents a certain scale and aspect ratio. We set a positive label (aircraft) for every unit on the feature map according to the following rules [15]: 1) The referenced box has the highest Intersection-over-Union (IoU) overlap with a ground-truth box, or 2) the referenced box has an IoU overlap that is higher than 0.75 with any ground-truth box. Thus, a single ground-truth box may be set as positive label to multiple units, which add positive samples in an image. Meanwhile, we set a negative label (background) for the predefined referenced box when its IoU overlap is lower than 0.3 with all the ground-truth boxes. Other referenced boxes do not serve as training samples. Finally, we minimize a multi-task loss function [17] including classification loss and box regression loss. For an referenced box $i$, its loss function is [18]:

$$L(p_i, b_i) = L_{cls}(p_i, p_i^*) + \lambda p_i^* L_{box}(b_i, b_i^*), \tag{2}$$

where the classification loss $L_{cls}$ is negative log loss and the regression loss $L_{box}$ is smooth $L_1$ loss function as defined in [17]. $p_i$ is the predicted probability of the referenced box $i$ being an aircraft, and $p_i^*$ is 1 only when the referenced box is defined as a positive sample. $b_i$ is the 4 parameterized coordinates of predicted box, and $b_i^*$ is the 4 parameterized coordinates of positive referenced box by combing the positive referenced box with the consistent ground-truth box. The second term $p_i^* L_{box}$ indicates that the box regression loss is activated only by positive referenced boxes. The hyperparameter $\lambda$ controls the balance between the two task losses. Ren et al. [18] showed the final results are insensitive to $\lambda$ when it is in a certain range. We find that satisfactory results can be obtained for aircraft detection when setting $\lambda$ = 10. Thus we keep this parameter fixed in the following experiments.

## III. EXPERIMENTS

We evaluated the performance of our method experimentally and compared with five other representative methods.

### A. Dataset and Implementation

The images used in our experiments are the ones used in [6]. They were collected from Google Earth around airports, including many international cities such as Los Angeles, Atlanta and Beijing. The size of image is around $1300 \times 900$. As used by the DBN-based method [6], we selected 18 images (including 505 aircrafts) for testing and 58 images (including 1019 aircrafts) for training. In the images, there are some small and blurred aircraft targets, and the backgrounds are diversified and complex. In order to overcome aircrafts' rotation problem, we augmented the number of training images with rotation ( $90°$, $180°$, $270°$) and horizontal flip strategies.

During the train stage, a mini-batch samples come from a single image containing a few positive and many negative referenced boxes. To avoid the bias towards negative samples, we randomly sampled 512 referenced boxes to compute loss function of a mini-batch, while maintaining ratio of positive and negative samples to 1:1. We fine-tuned the last two (or the all) layers of FCN model, while Conv1∼Conv5 are pretrained from the PASCAL VOC2007 detection task. Our whole model can be trained end-to-end by back propagation and stochastic gradient descent. The whole project is written in Matlab and C++ (Caffe [27]) based on CPU (Intel i5-3470) and GPU (GTX960). We need to set the detailed parameters for fine-tuning the whole FCN model. Specifically, the learning rates of two newly added layers and Conv1∼Conv5 layers start from 0.001 and 0.0001, respectively. Both of them are divided by 10 at every $6k$ iterations. Besides, the weight decay is 0.0005, the momentum is 0.9, all weights are initialized by gaussian with zero mean and std = 0.01.

### B. Results and Analysis

To evaluate the quantitative performance of the proposed method, we use the False Alarm Rate (FAR) and Detection Rate (DR) [6] as evaluation metrics, which are defined as:

$$\begin{aligned} \text{FAR} &= \frac{\text{number of False Alarms}}{\text{number of Aircrafts}} \times 100\%, \\ \text{DR} &= \frac{\text{number of Detected Aircrafts}}{\text{number of Aircrafts}} \times 100\%, \end{aligned} \tag{3}$$

TABLE II
FARS OF SIX METHODS AT VARIOUS DR LEVELS.

| Method | Detection Rate | | | | |
|---|---|---|---|---|---|
| | 85% | 80% | 75% | 70% | 65% |
| Wavelet+SVM [1] | 67.35% | 26.72% | 21.58% | 12.50% | 9.43% |
| Gabor+SVM [2] | 64.80% | 25.31% | 19.20% | 11.91% | 8.38% |
| HOG+SVM | 61.98% | 24.15% | 17.13% | 11.08% | 8.12% |
| DBNs [6] | 37.92% | 15.64% | 10.19% | 7.23% | 5.25% |
| **Proposed** | **19%** | **12.48%** | **7.72%** | **5.54%** | **4.75%** |
| SSD500 [23] | 6.34% | 4.16% | 2.97% | 1.98% | 1.78% |
| **Proposed(all-ft)** | **4.36%** | **2.38%** | **1.78%** | **1.39%** | **1.2%** |

**Table 2** shows the result of FAR at different DR levels of six methods, including the proposed method and five compared methods. Row 5 shows result of only fine-tuning the final two newly added layers of our model. Row 7 shows performance of fine-tuning the all layers of our model. At the same DR level, lower FAR indicates better detection performance. The results of Wavelet+SVM, Gabor+SVM, HOG+SVM and DBNs methods are provided by Chen et al. [6]. We can clearly see that our FCN model has better performance than some traditional methods when we only fine-tuned the two added layers. Particularly, when the DR is higher, the proposed method has much lower FAR. For examples, at $85\%$ DR, the DBN-based method has FAR $37.92\%$, while the FAR of our method is $19\%$, more than $50\%$ relatively lower than the previous best method [6] on our dataset.

In order to further verify the effectiveness of our FCN model, we introduced current state-of-the-art object detector SSD [23] on PASCAL VOC for our aircraft detection task. When we fine-tuned the whole layers of our FCN model, the detection results (Row 7 in Table 2) achieve comparable or even better performance than the SSD method [23] on our dataset. It has been demonstrated that our FCN model is really more suitable for aircraft detection from arbitrary-sized remote sensing images, even under the condition of large-sized complex backgrounds and small aircraft targets. The SSD method has slightly worse performance than ours because it needs to warp input image to a smaller fixed size ($500 \times 500$), which increases the difficulty of aircraft detection.

TABLE III
AVERAGE DETECTING TIME PER IMAGE (ABOUT $900 \times 1300$).

| Methods | localization -DBNs [6] (CPU) | **Proposed CPU** | -DBNs [6] GPU | **Proposed GPU** |
|---|---|---|---|---|
| Time(s) | $> 160^{\star}s$ | **4.5s** | $3s$ | **0.2s** |

**Table 3** shows the running time of testing an image. "$\star$" denotes the result is referenced from the literature [8]. For a fair and objective comparison, the two types of aircraft detection system are all tested under the condition of CPU. Compared to the DBN-based method [6], our method is more than 35 times faster. The DBN-based method is slow because it has two separated stages: First, it yields large number of region proposals by multiple thresholding algorithm. Then, each proposal is processed independently by DBN. Due to the localization task of DBN-based method can not be realized on GPU, so we only test time consumption of DBN on GPU. On the contrary, our approach is very fast because the FCN is an end-to-end procedure and can be easily performed on GPU.

TABLE IV
THE SIZE OF PARAMETERS IN VARIOUS MODELS

| Models | AlexNet [28] | ZF-Net [26] | **Proposed** |
|---|---|---|---|
| size | $244MB$ | $249MB$ | **17MB** |

**Table 4** shows the comparison of parameter size for various models. Here we use the number of bytes ("$MB$" denotes million bytes) that requires to store all of parameters in the trained model. It is worth noting that our FCN model is based on ZF-Net [26]. We replaced the fc layers with conv layers, but our model size is only about $1/14$ of the ZF-Net. It demonstrates that using the conv layer rather than the fc layer can greatly reduce the parameter size while obtaining comparable detection performance (seeing Table 2).

Fig. 3. **Example images with detection Results.** Practical aircraft detection result after using NMS operation (IoU = 0.3) and setting a score threshold as 0.89. The running time is about $0.2s$ per image on GTX960. The green boxes are ground-truth boxes and red boxes are predicted boxes.

**Fig. 3** shows example images with detected aircrafts boxed. It is shown that the proposed method can detected different scaled and multi-directional aircraft targets under complex background, and our method yields quite a few false alarms.

## IV. CONCLUSION

In this paper, we propose an end-to-end single fully convolutional network (FCN) for aircraft detection from remote sensing images. The architecture is designed to consider multi-scale and oriented objects, without separating into two stages and using pyramid of multi-scaled images. It allows the whole input image to be efficiently processed into a deep convolutional feature map and obtains directly detection results. Experimental results show that the proposed method is more than 35 times faster than the best previous method on our aircraft dataset while the false alarm rate is much lower at different DR levels. For our future work, we will extend the FCN framework for multi-class object detection.

## REFERENCES


[1] J. Hsieh, J. Chen, C. Chuang, K. Fan. Aircraft type recognition in satellite images. IEE Proceedings Vision, Image and Signal Processing, 152(3): 307-315, 2005.

[2] C. Yildiz, E. Polat, Detection of stationary aircrafts from satellite images, 19th IEEE Conference on Signal Processing and Communications Applications, pp. 515-521, 2011.

[3] L. Liu and Z. Shi, Airplane detection based on rotation invariant and sparse coding in remote sensing images, Optik - International Journal for Light and Electron Optics, 125(18): 5327-5333, May 2014.

[4] C. Xu and H. Duan, Artificial bee colony (ABC) optimized edge potential function (EPF) approach to target recognition for low-altitude aircraft, Pattern Recognit. Lett., 31(13): 1759-1772, Oct. 2010.

[5] W. Li, S. Xiang, H. Wang, C. Pan, Robust airplane detection in satellite images, Proc. ICIP, pp. 2877-2880, 2011.

[6] X. Chen, S. Xiang, C. Liu, C. Pan, Aircraft Detection by Deep Belief Nets. Asian Conference on Pattern Recognition (ACPR), 2013.

[7] X. Chen, S. Xiang, C. Liu, C. Pan, Aircraft Detection by Deep Convolutional Neural Networks. IPSJ Transactions on Computer Vision and Applications. Vol.7: 10-17, Jan. 2015.

[8] H. Wu, H. Zhang, J. Zhang, F. Xu, Fast Aircraft Detection In Satellite Images Based On Convolutional Neural Networks. In: ICIP, 2015.

[9] N. Dala, B. Trikks, Histograms of oriented gradients for human detection. In: CVPR, 2005.

[10] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan: Object detection with discrimina-tively trained part based models. IEEE Trans. Pattern Analysis and Machine Intelligence, 32(9): 1627-1645, 2010.

[11] B. Alexe, T. Deselaers, V. Ferrari: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. 34(11): 2189–2202, 2012.

[12] Jasper R. R. Uijlings and Koen E. A. van de Sande and Theo Gevers and Arnold W. M. Smeulders: Selective Search for Object Recognition. International Journal of Computer Vision. 104(2): 154–171, 2013.

[13] M. Cheng, Z. Zhang, W. Lin, P. Torr: BING: Binarized normed gradients for objectness estimation at 300fps. In: CVPR, 2014.

[14] C. Zitnick and P. Dollr, Edge boxes: Locating object proposals from edges. In: ECCV, 2014.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: ECCV, 2014.

[17] R. Girshick. Fast R-CNN. In: ICCV, 2015.

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS, 2015.

[19] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov, Scalable, high-quality object detection, arXiv:1412.1441 (v1), 2015.

[20] P. Pinheiro, R. Collobert, and P. Dollar, Learning to segment object candidates. In: NIPS, 2015.

[21] T. Ojala, M. Pietikainen, T. Maenpaa. Multiresolution Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7): 971-987, 2002.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In: NIPS, 2015.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu and A. Berg. SSD: Single Shot MultiBox Detector. arXiv:1512.02325 (v2), 2016.

[24] K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. In: CVPR, 2016.

[25] S. Ren, K. He, R. Girshick, X. Zhang, J. Sun. Object detection networks on convolutional feature maps. arXiv: 1504.06066, 2015.

[26] M. Zeiler, R. Fergus. Visualizing and understanding convolutional neural networks. In: ECCV, 2014.

[27] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In Proc. of the ACM International Conf. on Multimedia, 2014.

[28] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In: NIPS, 2012.