Human Crowd Detection for Drone Flight Safety Using Convolutional Neural Networks

Maria Tzelepi and Anastasios Tefas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
Email: mtzelepi@csd.auth.gr, tefas@aiia.csd.auth.gr

Abstract—**In this paper a novel human crowd detection method, that utilizes deep Convolutional Neural Networks (CNN), for drone flight safety purposes is proposed. The aim of our work is to provide light architectures, as imposed by the computational restrictions of the application, that can effectively distinguish between crowded and non-crowded scenes, captured from drones, and provide crowd heatmaps that can be used to semantically enhance the flight maps by defining no-fly zones. To this end, we first propose to adapt a pre-trained CNN on our task, by totally discarding the fully-connected layers and attaching an additional convolutional one, transforming it to a fast fully-convolutional network that is able to produce crowd heatmaps. Second, we propose a two-loss-training model, which aims to enhance the separability of the crowd and non-crowd classes. The experimental validation is performed on a new drone dataset that has been created for the specific task, and indicates the effectiveness of the proposed detector.**

Index Terms—**Crowd detection, Drones, Safety, Convolutional Neural Networks, Deep Learning.**

1. Introduction

The recent advent of Drones in a wide range of applications such as visual surveillance, rescue, and entertainment, is accompanied by the demand of safety. Apart from the robustness, a primary step to settle the issue of safety constitutes in defining no-fly zones for crowd avoidance, since a drone may operate close to crowds, and is potentially exposed to environmental hazards or unpredictable errors that render emergency landing inevitable. Furthermore, Drone flight regulations in several Countries' national legislation, especially in European Countries, request that the drones should not fly over crowds, and even more several laws define the minimum distance the drone can fly near a crowd. Thus, it is of utmost importance for the drone to be able to detect crowds in order to define no-fly zones and proceed to re-planning during flying operations. This feature will also allow for pushing on relaxing the restrictions for autonomous flying of drones keeping safe individual persons and crowds in the flying area. To this aim, in this paper we address the problem of crowd detection from drones, towards crowd avoidance utilizing the state-of-the-art deep CNNs, [1], [2].

Over the last few years, deep CNNs have been established as one of the most promising avenues of research in the computer vision area, providing outstanding results in a series of vision recognition tasks, such as image classification [3], [4], face recognition [5], digit recognition [6], [7], pose estimation [8], object and pedestrian detection [9], [10], and content based image retrieval [11], [12]. It has also been demonstrated that features extracted from the activation of a CNN trained in a fully supervised fashion on a large, fixed set of object recognition tasks can be re-purposed to novel generic recognition tasks, [13].

CNNs comprise of a number of convolutional and subsampling layers with non-linear neural activations, followed by fully connected layers (an overview of the utilized network is provided in Fig. 1). That is, the input image is introduced to the neural network as a three dimensional tensor with dimensions (i.e., width and height) equal to the dimensions of the image and depth equal to the number of color channels (usually three in RGB images). Three dimensional filters are learned and applied in each layer where convolution is performed and the output is passed to the neurons of the next layer for non-linear transformation using appropriate activation functions. After multiple convolution layers and subsampling the structure of the deep architecture changes to fully connected layers and single dimensional signals. These activations are usually used as deep representations for classification, clustering or retrieval. The size of the input image is fixed and usually scaling is performed before feeding the image to the network whenever there are fully-connected layers in the CNN. In order to allow for arbitrary image dimensions the CNN should be fully convolutional.

In this work, we propose a crowd detection method for drone flight safety, using fully convolutional deep CNNs. Our focus is to provide a lightweight CNN model, which, satisfying the computational and memory limitations of our application, can distinguish between crowded and non-crowded scenes, captured from drones and provide semantic heatmaps that can be used to semantically enrich the flying zones. First, we utilize a pre-trained CNN model, we adapt it by discarding the fully-connected layers, we add a final

ISBN 978-0-9928626-7-1 © EURASIP 2017

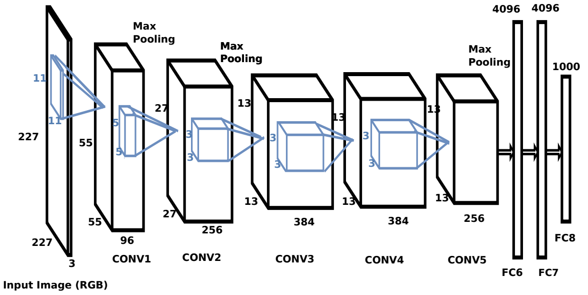


Figure 1. Overview of the CaffeNet Architecture

convolutional layer, and subsequently we retrain all the convolutional layers on the utilized dataset. Second, we propose a novel two-loss-training procedure, which aims to enhance the separability of the crowd and non-crowd classes. Finally, we created a new drone crowd dataset to evaluate the proposed approach since there is no such dataset publicly available.

The fully-convolutional nature of the proposed model is very important in handling input images with arbitrary dimension, and estimating a heatmap for the crowded areas. This will allow for semantically annotating the corresponding maps and defining no-fly zones. Additionally, the proposed fully-convolutional model will allow for handling low computational and memory resources on the drone whenever there is another process (*e.g.*, replanning, SLAM, etc.) that takes place, and only low dimensional images can be processed on the fly for crowd avoidance.

The remainder of the manuscript is structured as follows. The proposed method is described in detail in Section 2. The dataset used for the evaluation is presented in Section 3. Experimental results are provided in Section 4. Finally, conclusions are drawn in Section 5.

2. Proposed Method

In this paper we propose a human crowd detection method for drone flight safety, that uses deep CNNs.

We utilize the BVLC Reference CaffeNet model, which is an implementation of the AlexNet model trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 to classify 1.3 million images to 1000 ImageNet classes, [3]. The model consists of eight trained neural network layers; the first five are convolutional and the remaining three are fully connected. Max-pooling layers follow the first, second and fifth convolutional layers, while the ReLU non-linearity ($f(x) = max(0, x)$) is applied to every convolutional and fully connected layer, except the last fully connected layer (denoted as FC8). The output of the FC8 layer is a distribution over 1000 ImageNet classes. The softmax loss is used during the training. An overview of the CaffeNet architecture is provided in Fig. 1.

A common practice in classification problems is to utilize a pre-trained CNN model, trained on large datasets such as Imagenet, replace the classification layer with a new one that represents the labels of a specific dataset and is initialized randomly, and retrain the network on the specific dataset, using backpropagation. The basic reason underlying this practice is the insufficient amount of training data. Furthermore, a CNN model, trained on large dataset, have learned in the earlier layers more generic features, that could be useful in other tasks.

Thus, we first replace the last layer with a new classification layer that represents the two classes, *Crowd* and *Non-Crowd*, of our problem, following the aforementioned practice. This approach serves as baseline against the proposed method.

2.1. Modifying a pretained model

Despite the effectiveness of the CaffeNet model, consisting of 61M parameters, the computational limitations of our application render it prohibitory to use it on the fly, even if the training procedure has been performed offline. Towards this end, since the fully-connected layers of the CaffeNet, as in most CNN models, occupy the most of the parameters (59M parameters out of a total of 61M parameters), we propose a new model by discarding the fully-connected portion of the network, and by attaching an extra convolutional layer, denoted as CONV6. The new convolutional layer, with receptive field equal to the whole input, is initialized randomly, while all the convolutional layers up to the CONV5 are initialized on CaffeNet's weights. The softmax loss is used during training. We denote this model by *One-Loss Convolutional*.

This modification drastically reduces the amount of the model parameters, and consequently the computational cost is restricted. Additionally, this also allows arbitrary-sized input images, since the fixed-length input requirement concerns the fully-connected layers. Consequently, this allows for using low-resolution images, which can be very useful in our application, since it can further restrict the computational cost.

2.2. Two-Loss Convolutional model

Inspired by the Linear Discriminant Analysis (LDA) [14] method, which aims at best separating samples of different classes, by projecting them into a new low-dimensional space, which maximizes the between-class separability while minimizing their within-class variability, we also propose a new model architecture. The new model, apart from the softmax loss layer which preserves the between class separability, includes an extra loss layer which aims at bringing the samples of the same class closer to each other.

To achieve this goal, considering a labeled representation $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is the image representation and y_i is the corresponding image label, we adapt the CNN model, aiming to minimize the squared distance between $\mathbf{x}_i$ and its m relevant representations. Here, we define as relevant an image belonging to the same class to another.

Let $\mathcal{I} = \{\mathbf{I}_i, i = 1, \ldots, N\}$ be the set of N images of the training set, and $\mathbf{x} = MAC_5(\mathbf{I})$ the output of the so-called

ISBN 978-0-9928626-7-1 © EURASIP 2017

MAC_5 layer of the CaffeNet model on an input image $\mathbf{I}$. The MAC_5 layer, is an extra pooling layer which implements the Maximum Activations of Convolutions (MAC) [15], over the width and height of the output volume, for each of 256 feature maps of the CONV5 layer. Then we denote by $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^{256\times1}, i = 1, \ldots, N\}$ the set of N feature representations emerged in the MAC_5 layer, and by $\mathcal{R}^i = \{\mathbf{r}_k \in \mathbb{R}^{256\times1}, k = 1, \ldots, K^i\}$ the set of K^i relevant representations of the i-th image. We compute the mean vector of the m representations of R^i to the certain image representation $\mathbf{x}_i$, and we denote it by $\boldsymbol{\mu}_+^i \in \mathbb{R}^{256\times1}$.

Then, the new target representations for the images of $\mathcal{I}$ can be determined by solving the following optimization problem:

$$\min_{\mathbf{x}_i\in\mathcal{X}} \mathcal{J}^+ = \min_{\mathbf{x}_i\in\mathcal{X}} \sum_{i=1}^{N} \|\mathbf{x}_i - \boldsymbol{\mu}_+^i\|_2^2, \tag{1}$$

We solve the above optimization problem using gradient descent. The first-order gradient of the objective function $\mathcal{J}^+$ is given by:

$$\begin{aligned}\frac{\partial \mathcal{J}^+}{\partial \mathbf{x}_i} &= \frac{\partial}{\partial \mathbf{x}_i}\left(\sum_{i=1}^{N} \|\mathbf{x}_i - \boldsymbol{\mu}_+^i\|_2^2\right) \\ &= \frac{\partial}{\partial \mathbf{x}_i}\left((\mathbf{x}_i - \boldsymbol{\mu}_+^i)^\intercal(\mathbf{x}_i - \boldsymbol{\mu}_+^i)\right) \\ &= 2(\mathbf{x}_i - \boldsymbol{\mu}_+^i),\end{aligned} \tag{2}$$

The update rules for the n-th iteration can be formulated as:

$$\mathbf{x}_i^{(n+1)} = \mathbf{x}_i^{(n)} - 2\zeta(\mathbf{x}_i^{(n)} - \boldsymbol{\mu}_+^i), \quad \mathbf{x}_i \in \mathcal{X}, \tag{3}$$

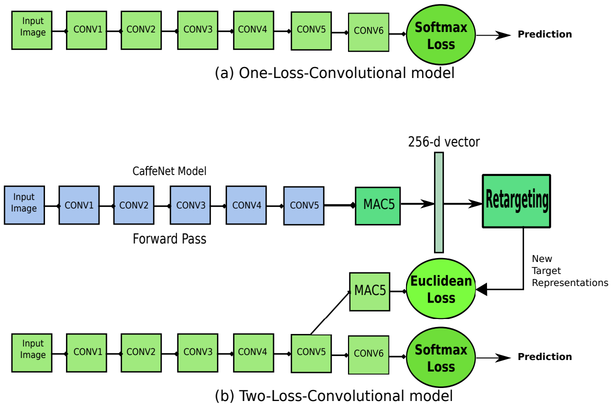
where the parameter $\zeta, \in [0, 0.5]$ controls the desired distance from the relevant representations.

Thus, using the above representations as targets in the CONV5 layer, we formulate an additional regression task for the neural network. The Euclidean loss is used during training for the regression task. The network is initialized on the CaffeNet's weights up to CONV5 layer and the two-loss-training is performed using back-propagation. This model is denoted as *Two-Loss Convolutional*. The two proposed models are illustrated in Fig. 2.

The proposed Two-Loss Convolutional model can be considered as having an extra regularization layer that exploits information from the data samples that are relevant to the input image. The additional cost for retargeting is performed once during the training and does not affect the complexity of the model during deployment and testing.

3. Dataset

In order to assess the performance of the proposed method, since there is no publicly available crowd dataset of drone videos/images, we constructed our own dataset by querying specific keywords to the Youtube[1] video search engine. We selected 60 drone videos with keywords describing crowded events (*e.g.* parade, festival, marathon, protests, political rally, etc). Non-crowded videos have been also gathered by searching for unspecified drone videos. Non-crowded images (e.g., bikes, cars, buildings, etc.) also randomly selected from the UAV123[2], and senseFly-Example-drone[3] datasets. Sample frames from the gathered video sequences are provided in Fig.3.



(a) One-Loss-Convolutional model

(b) Two-Loss-Convolutional model

Figure 2. One-Loss Convolutional and Two-Loss Convolutional models.

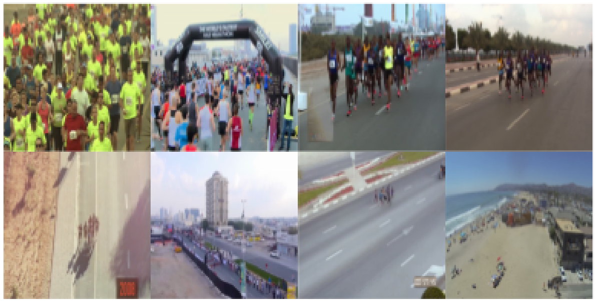


Figure 3. Sample images of the *Crowd-Drone* dataset.

In order to validate the performance of the proposed method, we left out of the training entire video sequences, and from their corresponding extracted frames we formulated the test set.

Thus, the train and test image datasets are described below:

	Train	Test
Crowd	2184	727
Non Crowd	1914	429
Total	4098	1156

TABLE 1. DATASET INFORMATION

4. Experimental Results

In this section we present the experiments conducted in order to evaluate the proposed method.

[1] http://www.youtube.com/

[2] https://ivul.kaust.edu.sa/Pages/Dataset-UAV123.aspx

[3] https://www.sensefly.com/drones/example-datasets.html

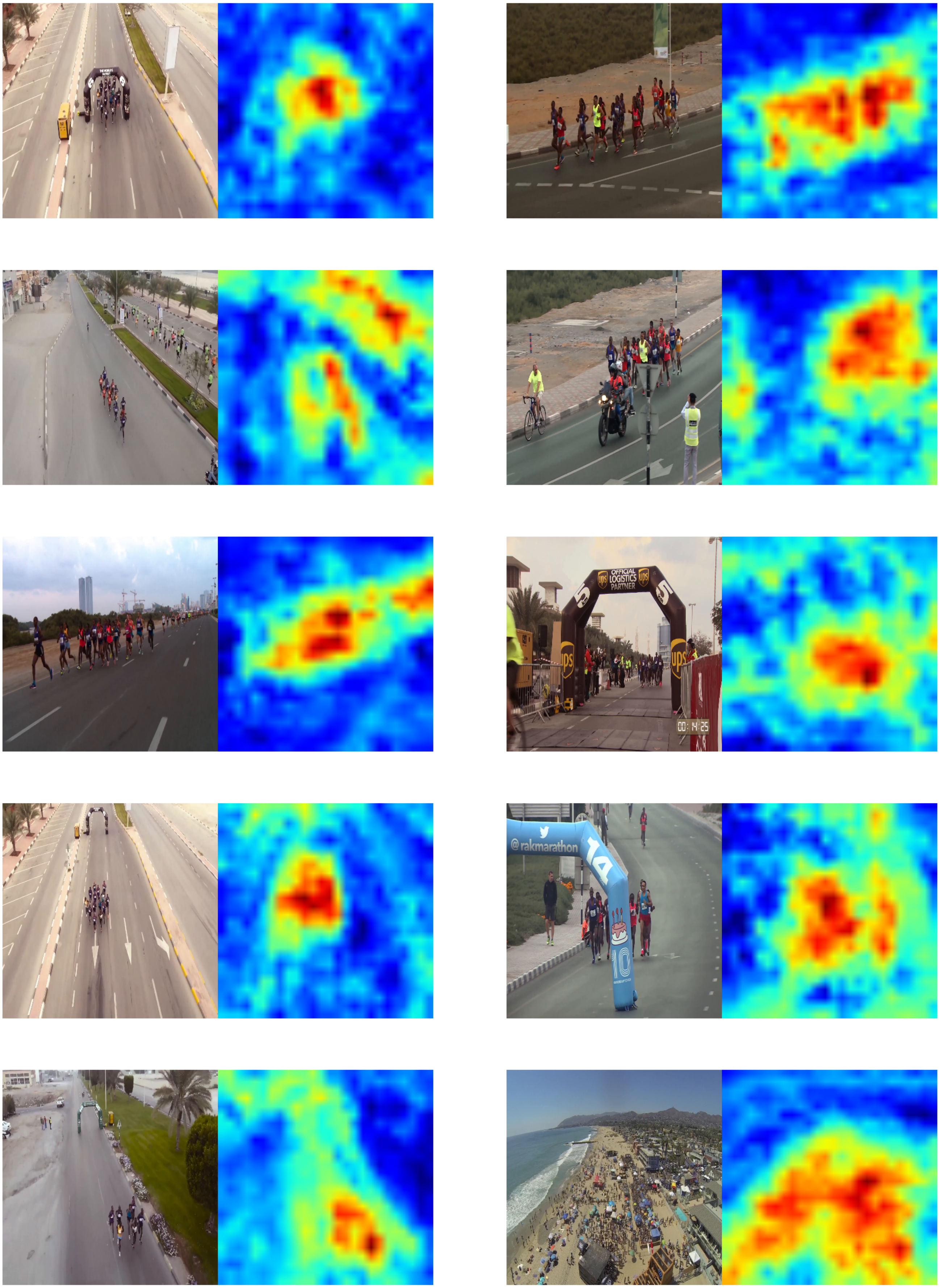


Figure 4. Heatmaps: The left part of each of ten pairs of images shows the original image, and the right one the corresponding heatmap.

We implemented the proposed method using the Caffe Deep Learning framework, [16]. We use the adaptive moment estimation algorithm (Adam) [17], instead of the simple gradient descent for the network optimization, with the default parameters. The parameter ζ in (3) is set to 0.4.

Table 2 illustrates the experimental results of the proposed method against the baseline CNN with fully connected layers and the one-loss fully convolutional CNN. Performance is measured in terms of Classification Accuracy. The best result is printed in bold.

Training Approach	Parameters	Layers	Accuracy
CaffeNet	61M	8	0.9299
One-Loss Convolutional	2.3M	6	0.91
Two-Loss Convolutional	2.3M	6	**0.9532**

TABLE 2. CLASSIFICATION ACCURACY

From the provided results, we can observe that the One-Loss-Convolutional model performs slightly worse than the refined CaffeNet model since the drastic reduction of the model parameters affects also the performance, however the model parameters reduction is very important and can be considered that compensates for the slightly reduced accuracy. Finally, we see that the proposed Two-Loss-Convolutional training procedure, achieves considerably improved performance against the baselines, with a significantly lighter architecture (2.3M parameters compared to 61M of the baseline CNN).

In Fig.4 we provide the heatmaps for the class *Crowd* of the proposed classifier. That is, ten test images of size 1024×1024 are fed to the network, and we compute the output of the network at the layer CONV6, for the label *Crowd* which is the desired heatmap.

5. Conclusions

In this paper we proposed a novel human crowd detection method, for drone flight safety purposes, utilizing fully convolutional deep CNNs. The first approach, includes the CaffeNet model adaptation in order to comply with the computational requirements of the specific application, and also benefit from the fully-convolutional networks properties. Second motivated by the LDA method, we also proposed a two-loss-training procedure, which optimized the lightweight model to distinguish between crowded and non-crowded images and concurrently enhance the two classes separability. Experimental evaluation on the constructed Crowd-Drone dataset indicates the effectiveness of the proposed method, outperforming the CaffeNet's baseline.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE). This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

References

[1] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems 2*. Morgan Kaufmann Publishers Inc., 1990, pp. 396–404.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1701–1708.

[6] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

[7] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard *et al.*, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural networks: the statistical mechanics perspective*, vol. 261, p. 276, 1995.

[8] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3626–3633.

[11] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 584–599.

[12] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 157–166.

[13] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[14] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[15] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *CoRR*, vol. abs/1511.05879, 2015.

[16] 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 *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.