Discriminative Deep Neural Network for Predicting Knee OsteoArthritis in Early Stage

Yassine Nasser[1](✉), Mohammed El Hassouni[1], and Rachid Jennane[2]

[1] FLSH, Mohammed V University in Rabat, Rabat, Morocco
{yassine_nasser,Mohamed.Elhassouni}@um5.ac.ma
[2] Institut Denis Poisson, University of Orleans, 45100 Orleans, France
Rachid.Jennane@univ-orleans.fr

Abstract. Knee osteoarthritis (OA) is a degenerative joint disease that causes physical disability worldwide and has a significant impact on public health. The diagnosis of OA is often made from X-ray images, however, this diagnosis suffers from subjectivity as it is achieved visually by evaluating symptoms according to the radiologist experience/expertise. In this article, we introduce a new deep convolutional neural network based on the standard DenseNet model to automatically score early knee OA from X-ray images. Our method consists of two main ideas: improving network texture analysis to better identify early signs of OA, and combining prediction loss with a novel discriminative loss to address the problem of the high similarity shown between knee joint radiographs of OA and non-OA subjects. Comprehensive experimental results over two large public databases demonstrate the potential of the proposed network.

Keywords: Convolutional Neural Network · Discriminative loss · Knee osteoarthritis · Plain radiography

1 Introduction

Osteoarthritis (OA) is a degenerative joint disease caused by the breakdown of the cartilage located at the end of the bone. Generally, OA is characterized by stiffness, swelling, pain and a grating sensation on movement which lead to a decrease in quality of life. Knee OA is the most common type of osteoarthritis. Due to their safety, availability and accessibility, the standard imaging modality for knee OA diagnosis is radiography (X-ray). The major hallmarks features of knee OA such as joint space narrowing, osteophytes formation, and subchondral bone changes could be visualized using X-ray images. Based on these pathological features, The Kellgren and Lawrence (KL) grading system [1] splits knee OA severity into five grades from grade 0 to grade 4. Grade 0 indicates the definite absence of OA and grade 2 early presence of OA. However, X-ray image patterns at early stage of knee OA present differentiation challenges and often result in

I. Rekik et al. (Eds.): PRIME 2022, LNCS 13564, pp. 126–136, 2022.
https://doi.org/10.1007/978-3-031-16919-9_12

high inter-reader variability across radiologists. Hence, the KL grading system is semi-quantitative, which introduces subjectivity/ambiguity into decision making and makes knee OA diagnosis more challenging.

Recently, a significant body of literature has been proposed on the application of deep learning networks to X-ray images for knee OA detection and prediction. In [2,3], Anthony *et al.* applied deep Convolutional Neural Networks (CNN) to automatically detect knee joint regions and classify the different stages of knee OA severity. In [4], Tiulpin *et al.* proposed an approach based on Deep Siamese CNN, which reduces the number of learnable parameters compared to standard CNNs. In their paper, the authors use an independent test set for evaluating its obtained results. In [5], Chen *et al.* applied a custom YOLOv2 model to detect the knee joint and fine-tuned a CNN model with a novel ordinal loss to classify knee OA severity.

All aforementioned deep learning based studies used Convolutional Neural Networks. However, classical CNNs rely mainly on the global shape information extracted from the last layers and ignore the texture information that characterizes bone architecture changes due to OA.

In [6], Nasser *et al.* introduced a Discriminative Regularized Auto-Encoder (DRAE) for early knee OA prediction using X-ray images. The proposed DRAE was based on Auto-Encoders (AE) with a combination between the standard AE training criterion and a novel discriminative loss. The mean goal was to maximize the class separability and learn the most useful discriminative features into the classifier. The limitation of this study that it was focused only on texture changes and neglected the overall deformation of the knee shape.

In this study, we propose to use a deep CNN model to predict knee OA in early stage from plain radiographs. Inspired by previous research in texture CNN [10,11], and the recently proposed discriminative regularization [6], we propose a new network to consider both shape and texture changes and maximize the class separability between OA and non-OA subjects.

The remainder of this paper is organized as follows. We report in Sect. 2 a detailed description of the proposed method. Section 3 presents the experimental settings. The results of a comparative evaluation with effective alternative solutions are discussed in Sect. 4. Finally, we give some concluding remarks and perspectives in Sect. 5.

2 Proposed Method

2.1 Overview

Conventional CNN architectures usually lead to extract complex correlations in upper layers corresponding to shape information and neglecting fine properties that contain the texture information [10,11]. However, knee osteoarthritis diagnosis depends on shape and texture properties across the entire distal knee joint. Thus, it is important to consider both features to create the training model. Nevertheless, early diagnosis of OA remains a challenging task, due to the high degree of similarity between non-OA and OA cases. Moreover, several studies

[7–9] have shown that in case of strong inter-class similarities or strong intra-class variations, and using only softmax loss, features learned with conventional CNNs of the same class are often scattered, and those learned from different classes overlap. Therefore, the discriminative aspect of the OA diagnostic model should also be improved.

To address these issue, we propose a new method based on the standard DenseNet [12]. The method combines texture information extracted from the mid-level layers with deep features in the top layer to better identify early signs of OA from inputs images (see Fig. 2). Moreover, we propose to add a novel discriminative loss function to the standard softmax in order to maximize the distance between non-OA and OA subjects.

2.2 DenseNet Learning Model

Our proposed network is derived from the classical DenseNet architecture [12], which is a densely connected convolutional network pre-trained on ImageNet [14]. In this section, a brief review of its architecture is given.

Let x_l be the output of the l^{th} layer. In conventional CNNs, x_l is computed by applying a nonlinear transformation H_l to previous layer's output x_{l-1}:

$$x_l = H_l(x_{l-1}) \tag{1}$$

During consecutive convolutions, activation function and pooling operation, the network obtains robust semantic features in the top layers. However, fine image details related texture tend to disappear in the top layers of the network.

Inspired by the main idea of the ResNet learning model [13], which introduces a residual block that sums the identity mapping of the input to the output of a layer, and in order to improve the information flow between layers, DenseNet proposes a direct connection from any layer to all subsequent layers. Consequently, the l^{th} layer receives the feature maps from all preceding layers as inputs. Thus, it is possible to define the output of the l^{th} layer as:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{2}$$

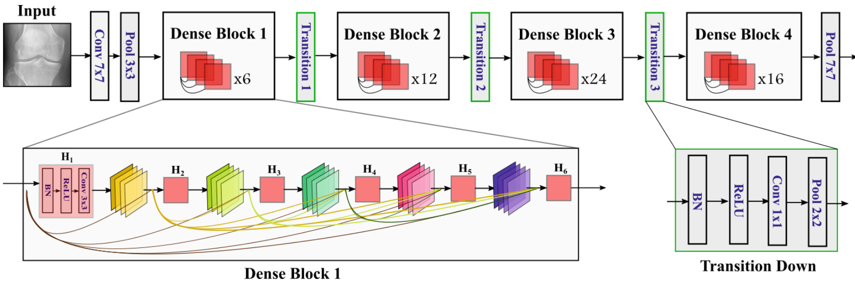

Fig. 1. Architecture of the DenseNet-121 learning model.

where [...] represents the concatenation operation, $H_l(.)$ is a composite function of the following consecutive operations: Batch Normalization (BN), Rectified Linear Units (ReLU), and a 3×3 Convolution (Conv). We denote such composite function as one layer.

DenseNet-121 used in our experiments consists of four dense blocks, each of which has 6, 12, 24 and 16 layers. In order to reduce the number of feature-maps, DenseNet introduces a transition down block between each two contiguous dense blocks. A transition down layer consists of a batch of normalization followed by a ReLU function, and a 1×1 convolutional layer followed by a 2×2 max-pooling layer. Figure 1 provides an illustrative overview of the architecture of DenseNet and the composition of each block.

2.3 Proposed Discriminative Shape-Texture DenseNet

In order to tackle the high similarity between OA and non-OA knee X-ray images at the early stages and to better detect the early signs of OA, we force the proposed network to : (i) learn a deep discriminative representation and (ii) consider both texture and shape information at the different layers of the model.

Learning a Deep Discriminative Representation. To learn deep discriminative features, a penalty term is imposed on the mid-level representations of the DenseNet (see Fig. 2). Apart from minimizing the standard classification loss, the objective is to improve the discriminative power of the network by forcing the representations of the different classes to be mapped faraway from each other. More specifically, we incorporate an additional discriminative term to the original classification cost function. The new objective function, $\mathcal{L}_T$ consists of two terms including the softmax cross-entropy loss and the discriminative penalty one:

$$\mathcal{L}_T = \mathcal{L}_C + \lambda \mathcal{L}_D \tag{3}$$

where λ is a trade-off parameter which controls the relative contribution of these two terms.

$\mathcal{L}_C$ is the softmax cross-entropy loss, which is the traditional cost function of the DenseNet model. It aims at minimizing the classification error for each

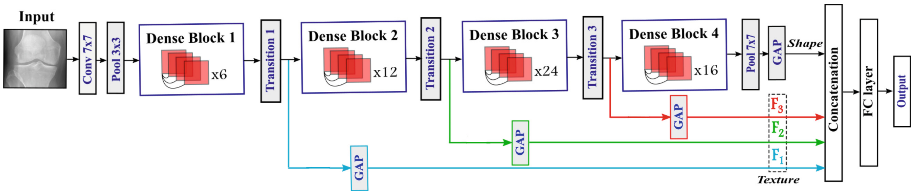

Fig. 2. Overview of the proposed method. Combination of texture and shape information to improve the prediction of OA in early stage. F_l is the global average pooling of the output of the l^{th} transition layer.

given training sample. Over a batch X of multiple samples of size N, the binary CE loss is defined as:

$$J_{Cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{4}$$

$\mathcal{L}_D$ represents the discriminative loss used to enforce the discriminative ability of the proposed model. $\mathcal{L}_D$ attempts to bring "similar" inputs close to each other and "dissimilar" inputs apart. To compute $\mathcal{L}_D$, we first feed the set of training samples X to the network and compute the outputs (feature maps) in each layer for each training sample, $x_i \in X$. Then, we compute $F_l(x_i)$, the Global Average Pooling (GAP) of the output feature maps of each transition layer l. Finally, the total discriminative loss $\mathcal{L}_D$ is defined as follows:

$$\mathcal{L}_D = \sum_{l=1}^{L} E_l \tag{5}$$

where E_l is the discriminative loss at a transition layer l. In the current study, we test two loss functions, the online Triplet Hard and SemiHard losses [21] and the Ω_{disc} one used in [6].

The Triplet loss [21], aims to ensure that the image x_i^a (anchor) is closer to all images x_i^p (positive) belonging to the same class, and is as far as possible from the images x_i^n (negative) belonging to an other class. Hence, when using a triple loss, E_L can be defined as

$$E_l = \sum_{i=1}^{N} \max(d(F_l(x_i^a), F_l(x_i^p)) - d(F_l(x_i^a), F_l(x_i^n)) + \epsilon) \tag{6}$$

where d is a distance metric, ϵ is a margin that is enforced between positive and negative pairs.

The Ω_{disc} loss [6], attempts to encourage classes separability, at each transition layer l, by maximizing the distance between the means μ_l^p and μ_l^n of the learned feature sets ($F_l(x_i^p)$ and $F_l(x_i^n)$) of each class and minimizing their variances v_l^p and v_l^n. The discriminative loss E_l which will be minimized in the use case of Ω_{disc} is defined then

$$E_l = \frac{v_l^p + v_l^n}{|\mu_l^p - \mu_l^n|^2} \tag{7}$$

Combining Shape and Texture. As mentioned above, several studies have shown that the first layers of CNNs are designed to learn low-level features, such as edges and curves which characterize the texture information, while the deeper layers are learned to capture more complex and high-level patterns, such as the overall shape information [17,18]. Moreover, CNN layers are highly related to filter banks methods widely used in texture analysis, with the key advantages that the CNN filters learn directly from the data rather than from handcrafted features. CNNs have also an architecture of learning which increases the abstraction level of the representation with depth [10,11,19].

Based on these studies and especially on the main idea of the texture and shape CNN (T-CNN) learning model [10], we propose a simple and efficient modification to the DenseNet architecture to improve its ability to consider both texture and shape.

Figure 2 illustrates the proposed architecture for combining texture information of the mid-level layers with the shape information of the top layer. First, using a specific concatenation layer, we fuse into a single vector the selected $\{F_l|l = 1, .., L\}$ which contain meaningful information about texture with the features of the last network layer that represent shape information. Then, we feed this vector to the final classification layer (i.e. the Fully Connected (FC) layer). Consequently, the network can learn texture information as well as the overall shape from the input image. This combination of features at different hierarchical layers enables to describe the input image at different scales.

3 Experimental Setup

3.1 Data Description

Knee X-ray images used to train and evaluate the proposed model were obtained from two public datasets: The Multicenter Osteoarthritis Study (MOST) [16] and the OsteoArthritis Initiative (OAI) [15]. The entire MOST database (3026 subjects) is used for the training, and the OAI baseline database (4796 subjects) is used for validation and test. The model was trained with regions of interest (ROI) corresponding to the distal area of the knee extracted from right knees and horizontally flipped left ones. Each ROI was associated with its KL grade. The objective of this study is to distinguish between the definite absence (KL-G0) and the definite presence of OA (KL-G2), which is the most important and challenging task, due to the high degree of similarity between their corresponding X-ray images, as shown in Fig. 3. KL-G1, is a doubtful one and was not considered in the current study. Table 1 summarizes the number of training, validation and testing samples.

3.2 Implementation Details

Our experiments were conducted using Python with the framework Tensorflow on Nvidia GeForce GTX 1050 Ti with 4 GB memory. The proposed approach was evaluated quantitatively using four metrics: Accuracy (Acc); Precision (Pr); Recall (Re) and F1-score (F1).

Dataset Preparation. As shown in Table 1, data are imbalanced. To overcome this issue during the training stage, data were balanced using the oversampling technique. To do so, different random linear transformations were applied to

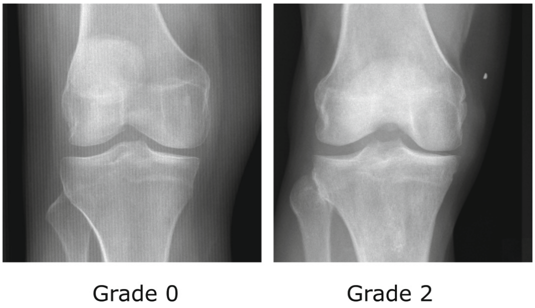

Fig. 3. Knee joint X-ray samples showing the high similarity between KL grades 0 and 2.

Table 1. Dataset description and distribution

Group	Dataset	KL-0	KL-2
Train	MOST	6008	3045
Validation	OAI	1116	806
Test	OAI	2313	1545

the samples, including: (i) random rotations using a random angle varying from -15^0 to 15^0, (ii) color jittering with random contrast and random brightness with a factor of 0.3, and (iii) a gamma correction.

Training Phase. As mentioned previously, DenseNet [12] pre-trained on ImageNet [14] was retained as our basic network structure (section II). The input size of the ROIs is 224 × 224, which is the standard size used in the literature. The proposed model was trained and optimized end-to-end using Adam optimizer with an initial learning rate of 0.0001. Hyper-parameters (λ, batch size, size of the fully connected layer, ration of dropout) were tuned using grid search on the validation set.

4 Experimental Results

In this section, the performance of our proposed method is evaluated for early knee OA detection. Firstly, two discriminative loss functions are tested. Then, the proposed network is compared to the deep learning pre-trained models, including the standard DenseNet [12], ResNet [13] as well as to Inception-V3 [20]. Finally, a visualisation analysis using t-SNE scatter plots is performed.

We test Triplet Hard and SemiHard losses with three distance metrics: l2-norm, squared l2-norm and the cosine similarity distance. We test also the discriminative loss Ω_{disc} proposed in [6]. The results are reported in Table 2. As can be seen, the best overall classification performance is obtained using the Ω_{disc}

discriminative loss with an accuracy rate of 87.69%. In term of the F1-score, the highest value (87.06%) is also reached using the Ω_{disc} discriminative loss, which corresponds to a precision rate of 87.48% and recall rate of 86.72%. We notice that Triplet SemiHard loss with l2-norm distance achieves competitive performance with Ω_{disc} loss. These results show that Ω_{disc} discriminative loss, leads generally to better performance compared to other tested losses. Hence, it is retained for the following experiments.

Table 2. Classification Performance of the proposed method using different discriminative loss functions

Discriminative loss	Distance metric	Acc (%)	Pr (%)	Re (%)	F1 (%)
Triplet hard	l2-norm	86.21	85.51	86.31	85.82
	squared l2-norm	86.50	85.94	85.93	85.94
	cosine similarity	86.39	85.76	86.02	85.88
Triplet SemiHard	l2-norm	87.48	**87.88**	85.94	86.66
	squared l2-norm	86.91	86.74	85.82	86.21
	cosine similarity	85.82	85.16	85.49	85.31
Ω_{disc} used in [6]	x	**87.69**	87.48	**86.72**	**87.06**

The proposed method is compared to some deep learning pre-trained networks, that are the standard DenseNet [12], ResNet [13] as well as Inception-V3 [20]. Results are reported in Table 3. As can be seen, the proposed method achieved the highest prediction performance compared to the other networks. In terms of accuracy, our proposed method obtains a score of 87.69% compared to 85.07%, 86.49% and 84.03% achieved by ResNet-101, DenseNet-169 and Inception-V3, respectively. The highest F1-score (87.06%) is obtained also by our proposed model. Even though DenseNet-169 achieved a high precision compared to other networks, it still has a low recall (75.08%). Therefore, with the exception of the precision values of DenseNet-169, our approach outperforms all other networks for all four metrics. In particular, a significant improvement in terms of F1-score is observed, as our model increases results by 5.14% from the 81.92% achieved by the standard DenseNet to 87.06% for the proposed method.

In addition to the quantitative evaluation, we check whether our model is able to increase the segregation of classes. To this end, we display the 2D scatter plots using t-distributed Stochastic Neighbor Embedding (t-SNE) [22] on each features levels $\{F_1, F_2, F_3\}$. Results are illustrated in Fig. 4. The first column shows the feature vector F_1 extracted from the first transition layer. As can be seen, the two classes significantly overlap. This may be due to common textual features shared between classes, such as edges and contours that form the overall joint shape. The second column shows the learned feature vectors F_2 obtained from the second transition layer. In this case, the network improves the separation between the two classes but not enough. The last column shows the learned

Table 3. Comparison of the proposed method to the deep learning pre-trained networks

Methods		Acc (%)	Pr (%)	Re (%)	F1 (%)
ResNet	ResNet-50	83.23	88.41	74.49	80.85
	ResNet-101	85.07	83.56	80.04	81.76
	ResNet-152	84.86	75.99	84.64	80.08
DenseNet	DenseNet-121	85.66	82.76	81.10	81.92
	DenseNet-169	86.49	**89.50**	75.08	81.66
	DenseNet-201	84.76	86.22	73.72	79.48
Inception	Inception-V3	84.03	83.39	75.08	79.02
Proposed method		**87.69**	87.48	**86.72**	**87.06**

features vector F_3 obtained from the third transition layer. Results show that by going deeper, our proposed model learned two discriminant representations. Thus, it leads to a better classes discrimination and thus a good prediction of knee OA at an early stage.

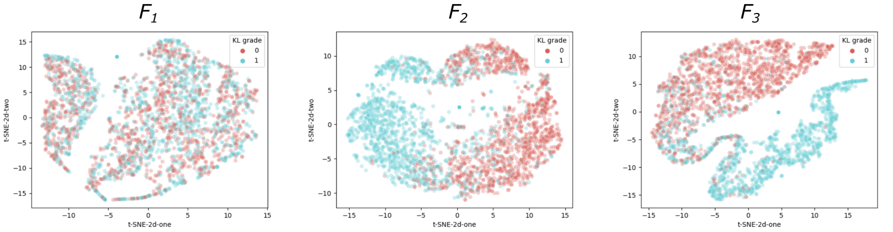

Fig. 4. Obtained t-SNE scatter plots for each feature levels using our proposed network.

5 Conclusion

In this paper, we proposed a novel deep learning method based on CNNs architecture with two distinct ideas: (i) combining the learned shape and texture features, (ii) enhancing the discriminative power to improve the challenging classification task, where a high similarity exists between early knee OA cases and healthy subjects. We tested the performance of our method using two discriminative losses with several distance metrics. The experimental results show that the proposed method surpasses the most influencial deep learning pre-trained networks. The results are promising and a further extension in a context of multi-classification with more KL grades and other loss functions will be considered in a future work.

References

1. Kellgren, J.H., Lawrence, J.: Radiological assessment of osteo-arthrosis. Ann. Rheum. Dis. **16**(4), 494 (1957)
2. Antony, J., McGuinness, K., O'Connor, N.E., Moran, K.: Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 1195–1200. IEEE (2016)
3. Antony, J., McGuinness, K., Moran, K., O'Connor, N.E.: Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In: Perner, P. (ed.) MLDM 2017. LNCS (LNAI), vol. 10358, pp. 376–390. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62416-7_27
4. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., Saarakkala, S.: Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based (2018)
5. Chen, P., Gao, L., Shi, X., Allen, K., Yang, L.: Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. Comput. Med. Imaging Graph. **75**, 84–92 (2019)
6. Nasser, Y., Jennane, R., Chetouani, A., Lespessailles, E., El Hassouni, M.: Discriminative regularized auto-encoder for early detection of knee osteoarthritis: data from the osteoarthritis initiative. IEEE Trans. Med. Imaging **39**(9), 2976–2984 (2020)
7. Wen, Y., Zhang, K., Li, Z., Qiao, Yu.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
8. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 302–309. IEEE (2018)
9. Cheng, G., Yang, C., Yao, X., Guo, L., Han, J.: When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. IEEE Trans. Geosci. Remote Sens. **56**(5), 2811–2821 (2018)
10. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3828–3836 (2015)
11. Andrearczyk, V., Whelan, P.F.: Using filter banks in convolutional neural networks for texture classification. Pattern Recogn. Lett. **84**, 63–69 (2016)
12. Huang, G., Liu, Z., van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
13. 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)
14. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
15. The Osteoarthritis Initiative (2020). https://nda.nih.gov/oai/
16. Multicenter Osteoarthritis Study (MOST) Public Data Sharing (2020). https://most.ucsf.edu/

17. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
18. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
19. Liu, L., Chen, J., Fieguth, P., Zhao, G., Chellappa, R., Pietikäinen, M.: From BoW to CNN: two decades of texture representation for texture classification. Int. J. Comput. Vision **127**(1), 74–109 (2019). https://doi.org/10.1007/s11263-018-1125-z
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
21. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
22. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)