Interpretability-Guided Content-Based Medical Image Retrieval

Wilson Silva[1,2(✉)], Alexander Poellinger[3], Jaime S. Cardoso[1,2], and Mauricio Reyes[4,5]

[1] INESC TEC, Porto, Portugal
wilson.j.silva@inesctec.pt
[2] Faculty of Engineering, University of Porto, Porto, Portugal
[3] Department of Diagnostic, Interventional and Pediatric Radiology, Inselspital, Bern University Hospital, Bern, Switzerland
[4] Insel Data Science Center, Inselspital, Bern University Hospital, Bern, Switzerland
[5] ARTORG Center for Biomedical Research, University of Bern, Bern, Switzerland

Abstract. When encountering a dubious diagnostic case, radiologists typically search in public or internal databases for similar cases that would help them in their decision-making process. This search represents a massive burden to their workflow, as it considerably reduces their time to diagnose new cases. It is, therefore, of utter importance to replace this manual intensive search with an automatic content-based image retrieval system. However, general content-based image retrieval systems are often not helpful in the context of medical imaging since they do not consider the fact that relevant information in medical images is typically spatially constricted. In this work, we explore the use of interpretability methods to localize relevant regions of images, leading to more focused feature representations, and, therefore, to improved medical image retrieval. As a proof-of-concept, experiments were conducted using a publicly available Chest X-ray dataset, with results showing that the proposed interpretability-guided image retrieval translates better the similarity measure of an experienced radiologist than state-of-the-art image retrieval methods. Furthermore, it also improves the class-consistency of top retrieved results, and enhances the interpretability of the whole system, by accompanying the retrieval with visual explanations.

Keywords: Medical image retrieval · Interpretability · Chest X-ray

1 Introduction

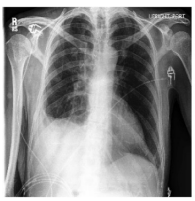
Accessibility to medical imaging technologies has considerably increased over the last decade, leading to an increase in the number of images that need to

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-59710-8_30) contains supplementary material, which is available to authorized users.

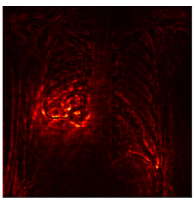
A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12261, pp. 305–314, 2020.
https://doi.org/10.1007/978-3-030-59710-8_30

be analyzed by radiologists in their daily workflow [11]. As the ratio of diagnostic demand to the number of radiologists is increasing, the effective available time per diagnostic has been decreasing and became a critical issue when diagnostic needs to be supported by confirmatory evidence of a potential suspected diagnosis. Currently, when in doubt for a suspected condition, radiologists turn to public or internal image databases where similar disease-matching images can be searched and compared against. Such a task is time-consuming and ineffective since it requires several iterations to find the right matching image supporting a final diagnosis. Therefore, it is of great value to develop disease-targeted content-based image retrieval (CBIR) systems that automatically present disease-matching similar images to the one being analyzed. CBIR systems mainly consist of two tasks: feature representation, and feature indexing and search [10]. For feature representation, one seeks to find a low-dimensional description of the image that is suitable for characterizing it well enough. In contrast, in feature indexing and search, the objective is more related to the efficiency of the retrieval process.

The focus of this work is on the feature representation task. To date, feature representation is mainly performed in one of three different ways: based on statistical measures, hand-crafted features, or through learned features. As pointed out by Li *et al.* [10], one successful approach to do feature representation is the use of a pre-trained Convolutional Neural Network (CNN), with a following fine-tuning phase using the medical dataset related to the task, as done in several state-of-the-art works [7,14,18]. Li *et al.* also mentioned that there are other possibilities, such as training from scratch in the medical dataset, or combining extracted deep features with hand-crafted features. Furthermore, in the absence of a sizeable labelled dataset, unsupervised approaches have also been proposed to perform feature extraction [3]. In terms of computing similarity among feature representations, it was referred by Ghorbani *et al.* [6] and demonstrated by Zhang *et al.* [20] that the Euclidean distance (L2 distance) measured in the activation space of final layers is an effective perceptual similarity metric.



Chest X-ray Image



Saliency Map

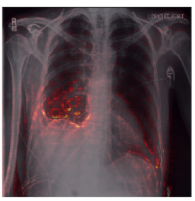


Image with Saliency

Fig. 1. Chest X-ray image and corresponding disease-related saliency map. In the saliency map, brighter colors mean higher relevance. (Color figure online)

The state-of-the-art approaches fail to give particular attention to the regions that are determinant for the medical condition, performing an overall image to image comparison, equally weighting anatomical and pathophysiological related image information. It is then of great importance to find a way to direct the focus of the image retrieval methods to the clinically relevant regions of a medical image, ideally, in an unsupervised manner. In that sense, interpretability methods [15], namely those which produce visual explanations in the form of saliency maps [2,12,13,16] appear as a suitable solution to find these relevant regions without supervision. In Fig. 1, we illustrate the motivation behind our central idea, by presenting an example of a Chest X-ray image and its corresponding saliency map, which points out to the disease-related image regions.

In this work, we explore the use of interpretability saliency maps as an attention mechanism to focus the feature representations to image regions that characterize the class to which they belong. As a proof-of-concept, experiments were conducted for the pleural effusion condition in Chest X-Ray images. The evaluation of the retrieval quality of the proposed method is based on its ranking capabilities and class-consistency.

2 Materials and Methods

2.1 Data

For the experiments, we used the publicly available CheXpert dataset [9], which consists of 224,316 chest radiographs from 65,240 patients collected from the Stanford Hospital. Each case was labelled for the presence of 14 different observations, with training set labels being automatically generated from the associated radiology reports, while both validation (200 chest radiographs) and test (500 chest radiographs) sets were labelled by board-certified radiologists. Currently, the test set is not publicly available since a competition is running[1]. Thus, the validation set was used for the evaluation of the proposed approach. From the validation set, we create two different sets: a test set with the cases to be analyzed, and a catalogue, with well-curated cases to be retrieved. For the sake of this work, we focused on the Pleural Effusion condition. Multiple reasons justify this decision, namely, most of the images of the validation set having been acquired in the anteroposterior position, and the data being highly imbalanced for certain medical conditions. Our work was also supported by an experienced board-certified radiologist, who provided us a ranking ground-truth for the catalogue cases, and the localization of the condition.

2.2 Method

State-of-the-art image retrieval methods analyze the image as a whole, producing a feature representation that characterizes the image in its entirety. Our proposed method aims to refine this feature computation process, enforcing the focus to relevant regions, and consequently improving medical image retrieval.

[1] https://stanfordmlgroup.github.io/competitions/chexpert/.

As the focus mechanism used here is based on interpretability saliency maps, we named our approach as Interpretability-guided CBIR (IG-CBIR). The method[2] is presented in Fig. 2 and described in the following paragraphs.

Training: The training process can be divided into two different steps. In **Step 1**, we train a CNN model to classify images into Pleural Effusion or Non Pleural Effusion. The CNN model used was the well-known DenseNet-121 [8], which was initialized with the pre-trained weights from ImageNet [4], and afterwards was fine-tuned (all weights) with the CheXpert training set. To accommodate for the use of the pre-trained network, grayscale images were replicated (by concatenation), so that they became three-channel RGB-like images. Furthermore, image resolution (224×224) and pre-processing were the same as for the ImageNet pre-training process. The model was trained for 10 epochs, using the Adadelta optimizer [19], a batch size of 32, and the binary cross-entropy loss. The number of epochs was optimized by splitting training set into train and validation. In **Step 2**, the goal is to enforce the focus of the network in clinically relevant regions. To do so, we generate saliency maps, using one of the standard interpretability methods provided by the iNNvestigate toolbox [1] (Deep Taylor [12] was the one used in this work). Afterwards, these training saliency maps are used to fine-tune the previously trained CNN. This fine-tuning stage follows the same procedures as before, with the only difference being that the inputs now are the saliency maps, instead of the original images. In short, this training phase results

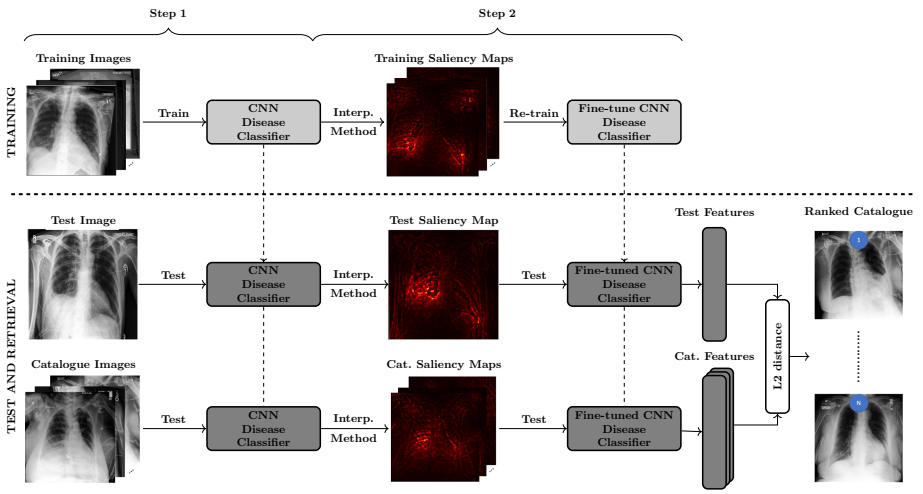


Fig. 2. Overview of the proposed approach. Blocks in light gray (■) mean CNNs are being trained (i.e., weights are being updated), whereas blocks in dark gray (■) represent trained CNNs (i.e., weights are fixed). In the saliency maps, brighter colors mean higher relevance. Blue circles indicate ranking positions. (Color figure online)

[2] Code available at https://github.com/wjsilva19/ig_cbir.

in two trained models: a first model to make predictions and generate saliency maps; and, a fine-tuned version of it to compute deep feature representations.

Test and Retrieval: The next step consists of using test and catalogue images as inputs to the first trained model (Fig. 2, lower part). With this, label predictions and saliency maps (for test and catalogue images) become available. Afterwards, these saliency maps are the input to the second trained model. During this stage, the features computed in the previous to last layer of the model are saved. The final step consists of calculating the Euclidean distance (L2) between the feature representations obtained for test and catalogue images, and rank the catalogue in terms of similarity to the test image (from most to least similar).

2.3 Evaluation

Baselines: We considered two types of baselines: a statistically-based, and a CNN-based. For the statistically-based baseline, we considered a well-known statistical measure of similarity, the structural similarity index (SSIM) [17]. The SSIM was computed directly between test and catalogue images, using its default values. Since high values of SSIM mean high similarity, the top retrieved images with this method are the ones with the highest similarity index. The second type of baseline is, like the proposed approach, based on deep learned features. As detailed in Li *et al.* [10], a current successful technique to learn feature representations of medical images is to use a CNN, pre-trained with natural images (e.g., ImageNet [4]), and fine-tune it in the application dataset (e.g., CheXpert). Afterwards, one can use the features computed in the last layers of the network to measure similarity. In practical terms, this CNN-based baseline consists of using the CNN disease classifier (from step 1 of Fig. 2) and saving the feature representations computed from the previous to last layer, which means that this baseline also works as ablation to assess the value of Step 2. Finally, the ranking of the catalogue images is performed in the same way as for the proposed approach, by computing and sorting the Euclidean distances between the input image and each catalogue image.

Assessing the Quality of the Retrieval: We considered two types of metrics, one to measure the quality of the ranking, and a second one to evaluate the class-consistency of the top retrieved images.

To measure the quality of the ranking, we used a standard metric in learning to rank tasks [5], the normalized Discounted Cumulative Gain (nDCG) - Eq. (1). The nDCG is the normalized version of the Discounted Cumulative Gain (DCG) - Eq. (2), being it normalized by the ideal/maximum possible value of DCG (IDCG). In both Eq. (1) and Eq. (2) the subscript p represents the number of retrieved images considered. Relevance values (rel_i) were assigned from 1 to 5.5, being 1 the least similar image according to the radiologist, and 5.5 the most similar one (i.e., the relevance of two contiguous positions differs by 0.5). This

$$\text{nDCG}_p = \frac{DCG_p}{IDCG_p} \quad (1) \qquad \text{DCG}_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (2)$$

was done with the goal of giving more importance to the first positions of the catalogue.

As images with high similarity ideally belong to the same class, we also measure the class-consistency of each method. For that, we considered a traditional retrieval evaluation measure, namely, precision - Eq. (3). In this context, relevant images are the ones that belong to the class of the test image.

$$\text{precision} = \frac{|\{\text{relevant images}\} \cap \{\text{retrieved images}\}|}{|\{\text{retrieved images}\}|} \quad (3)$$

3 Results

We performed five initial experiments, corresponding to five different sets of images, that resulted from the use of different seeds when doing the split of the validation data into test and catalogue (keeping the proportion of the classes). Due to time limitations of the radiologist, we considered catalogues of 10 images of size. In Fig. 3, we present the results in terms of nDCG for the top-4, top-7, and top-10 retrieved images[3].

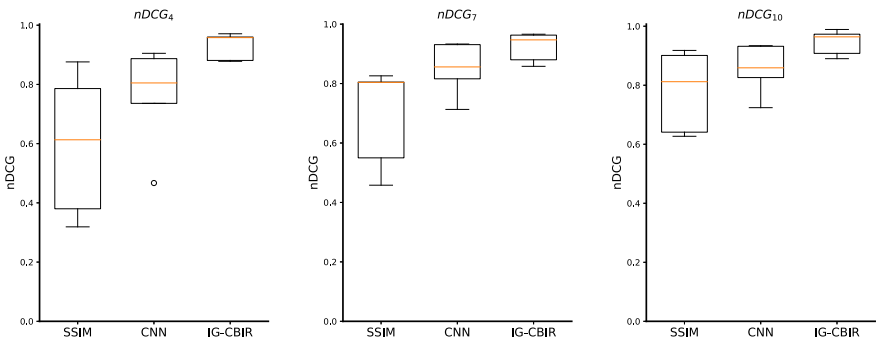


Fig. 3. Box-and-whisker plots regarding the nDCG results for Top-4, Top-7, and Top-10 retrieved images, respectively. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, and IG-CBIR is the proposed interpretability-based approach.

To evaluate the class-consistency, we only needed dataset images and labels, and no expert-based ranking, hence we considered larger catalogues, computing

[3] Detailed results are provided in Table 1 of the Supplementary Material.

precision results for three different settings: top-4 images retrieved when catalogue has 10 images; top-7 images retrieved when catalogue has 20 images; and, top-11 images retrieved when catalogue has 30 images. Class-consistency results

Table 1. Precision results. Top-X means X retrieved images (X: 4, 7, 11). Cat-Y means catalogue of size Y (Y: 10, 20, 30). X is also the number of relevant images in catalogue Y. Results are presented as average [min, max].

Method	Top-4 (Cat-10)	Top-7 (Cat-20)	Top-11 (Cat-30)
SSIM-based	0.55 [0.25, 0.75]	0.40 [0.29, 0.57]	0.42 [0.27, 0.55]
CNN-based	0.85 [0.50, 1.00]	0.60 [0.29, 0.86]	0.69 [0.55, 0.82]
IG-CBIR (Proposed)	**0.95** [0.75, 1.00]	**0.77** [0.71, 0.86]	**0.80** [0.73, 0.82]

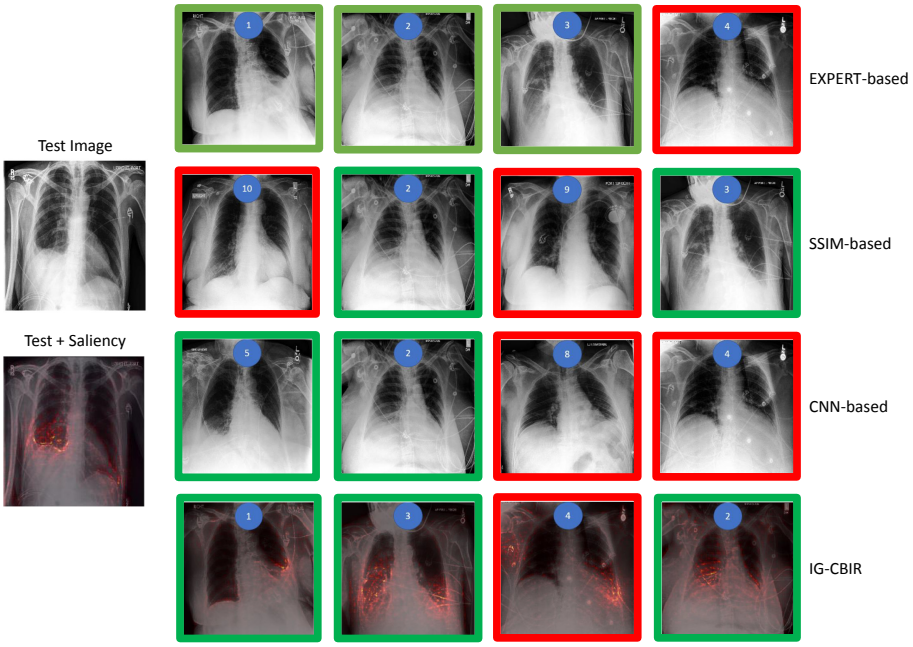


Fig. 4. Retrieved catalogue images for one example test image. From left to right: test image (and test image with saliency map superimposed) and most similar images sorted according to each method. From top to bottom: ground-truth defined by the radiologist, SSIM baseline results, CNN baseline results, and IG-CBIR results. IG-CBIR results are presented with image and saliency superimposed. Green boxes mean agreement with test image label whereas red boxes mean disagreement. Numbers on top of the images represent ranking position in the ground-truth based on expert rating. (Color figure online)

are presented in Table 1[4] in terms of average, minimum, and maximum values obtained by each method.

In Fig. 4, we present an example of a test image and the top-4 retrieved images given by each of the methods for that same test image. We show the test image and the test image with the corresponding saliency map superimposed (Fig. 4, left). It is important to point out that this saliency map is in agreement with the report of the radiologist: "Bilateral pleural effusion, stronger on the right side but also present, to a lesser extent, in the left side". On the right side, one can see the top-4 of the most similar catalogue images to the test one according to the expert and to each of the considered methods.

4 Discussion and Conclusion

We proposed to improve medical image retrieval by using interpretability saliency maps to focus the image retrieval system in the clinically relevant regions of the medical images. The proposed interpretability-based approach leads to an improvement in medical image retrieval, with the most significant improvement being related to the ranking quality of the retrieval. As demonstrated in Fig. 3, and illustrated in Fig. 4, the proposed approach resembles better the ranking order given by the radiologist than state-of-the-art image retrieval methods. It is also important to mention that the method's training process is expert-agnostic. We only use label information during training, and the labels that we use are the ones already provided by the CheXpert dataset. Furthermore, the method also improves the results in terms of class-consistency, as shown in Table 1. As we considered different sizes of catalogues for the class-consistency evaluation, we observed that the method seems to be robust to the catalogue size.

For both our proposed approach and the CNN-based baseline (or even any CNN-based approach), the quality of the retrieval will, of course, be limited by the classification performance of the model. Indeed, that was the reason for us not to consider conditions such as Atelectasis in this proof-of-concept. Nonetheless, as sizes of the databases grow, the classification performance for more diseases is expected to be in a suitable range for CNNs to be used in medical image retrieval applications. Classification performances obtained with our CNN model were in line with those reported by the CheXpert team [9].

In this work, the experiments were conducted with the Deep Taylor interpretability method. It can be noted that results may change according to the interpretability method to be used, since they produce very different saliency maps. Nonetheless, we are confident that the use of other interpretability saliency maps will also help in the refinement of the retrieval. As future work, we intend to explore different interpretability methods to generate the saliency maps, and also different ways of combining them to perform the fine-tuning stage, with the goal of improving the robustness of the method.

In conclusion, we have introduced a novel approach based on interpretability saliency maps to refine the quality of medical image retrieval achieved by deep

[4] Detailed results are provided in Table 2 of the Supplementary Material.

CNNs. As shown, this approach leads to a better similarity measure between medical images of the same condition, and, therefore, to a better image retrieval than that obtained using state-of-the-art approaches. Moreover, it also enhances the interpretability of the computer aided-diagnosis system, as it accompanies the retrieval with visual explanations.

Acknowledgements. This work was partially supported by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e Tecnologia within project POCI-01-0145-FEDER-028857, and also by Fundação para a Ciência e Tecnologia within PhD grant number SFRH/BD/139468/2018.

References

1. Alber, M., et al.: Innvestigate neural networks. J. Mach. Learn. Res. **20**(93), 1–8 (2019)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
3. Bengio, Y., Courville, A.C., Vincent, P.: Unsupervised feature learning and deep learning: a review and new perspectives. CoRR, abs/1206.5538, vol. 1, 2012 (2012)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR 2009 (2009)
5. Fernandes, K., Cardoso, J.S.: Hypothesis transfer learning based on structural model similarity. Neural Comput. Appl. **31**(8), 3417–3430 (2017). https://doi.org/10.1007/s00521-017-3281-4
6. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Advances in Neural Information Processing Systems, pp. 9273–9282 (2019)
7. Hofmanninger, J., Langs, G.: Mapping visual features to semantic profiles for retrieval in medical imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 457–465 (2015)
8. 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)
9. Irvin, J., et al.: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031 (2019)
10. Li, Z., Zhang, X., Müller, H., Zhang, S.: Large-scale retrieval for medical image analytics: a comprehensive review. Med. Image Anal. **43**, 66–84 (2018)
11. McDonald, R.J., et al.: The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad. Radiol. **22**(9), 1191–1198 (2015)
12. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. **65**, 211–222 (2017)
13. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

14. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
15. Silva, W., Fernandes, K., Cardoso, J.S.: How to produce complementary explanations using an ensemble model. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
16. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
17. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
18. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 589–596. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_72
19. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
20. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)