Attention Guided Slit Lamp Image Quality Assessment

Mingchao Li[1], Yerui Chen[1], Kun Huang[1], Wen Fang[2], and Qiang Chen[1](✉)

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
Chen2qiang@njust.edu.cn

[2] Department of Ophthalmology, The First Affiliated Hospital with Nanjing Medical University, Nanjing 210094, China

Abstract. Learning human visual attention into a deep convolutional network contributes to classification performance improvement. In this paper, we propose a novel attention-guided architecture for image quality assessment (IQA) of slit lamp images. Its characteristics are threefold: First, we build a two-branch classification network, where the input of one branch uses masked images to learning regional prior. Second, we use a Forward Grad-CAM (FG-CAM) to represent the attention of each branch and generate the saliency maps. Third, we further design an Attention Decision Module (ADM) to decide which part of the gradient flow of both two branch saliency maps will be updated. The experiments on 23,197 slit lamp images show that the proposed method allows the network closer to human visual attention compared with other state-of-the-art methods. Our method achieves 97.41%, 84.79%, 92.71% on AUC, F1-score and accuracy, respectively. The code is open accessible: https://github.com/nhoddJ/CSRA-module.

Keywords: Slit lamp images · Image quality assessment · Forward Grad-CAM · Attention Decision Module

1 Introduction

Convolutional neural network has been widely used in image fusion, object detection and image classification and has achieved widespread success. In the field of medical image analysis, it also achieves performance close to human experts and beyond [1–4]. Recent research shows that learning human visual attention into a convolutional network can help improve classification effect [5]. This is because the introduction of clinical prior knowledge (e.g., the shape and size of the Lesion area) allows the network to learn more and becomes more robust.

Learning human visual attention into a deep convolutional network contributes to classification performance improvement [6–10]. Huang et al. [8] utilized masks between the internal limiting membrane (ILM) layer and the retinal pigment epithelium (RPE) layer to guide macular disease diagnose. Wang et al. [9] utilized iris region masks to assist image quality assessment (IQA) in the iris region. He et al. [11] proposed a multi scale

H. Fu et al. (Eds.): OMIA 2021, LNCS 12970, pp. 179–188, 2021.
https://doi.org/10.1007/978-3-030-87000-3_19

feature extractor to get deep features of fovea region masks to assist diabetic macular edema (DME) grading. These researches show that the clinical semantic region attention mechanisms often lead performance improvement of the classification task.

In this work, we focus on the task of slit lamp image quality assessment. Slit lamp images are mainly used to observe ocular surface diseases, which is one of the common clinical ophthalmology diseases. The common clinical manifestations include dry eye disease (DED), blepharitis, seasonal allergic conjunctivitis, etc. For DED, its ocular surface irritation and ocular surface damage have a significant impact on the visual acuity between blinks. In severe cases, it can cause ocular surface inflammation, lacrimal glands inflammation and vision loss. The latest epidemiological data survey shows that DED and new cases that occur with environmental changes account for about 20% of the population [11]. As an important tool to judge ocular surface inflammations and elevated intraocular pressure [12], bulbar conjunctiva hyperemia grading needs high-quality image to analyze morphological features of blood vessels. Lesions analysis and feature quantification also need high-quality images. Therefore, it is necessary to evaluate the image quality of the slit lamp images to screen high-quality images.

In this work, we propose a novel attention-guided architecture for image quality assessment of slit lamp images. Our key insight is to let the attention of the classification network focus on the region marked by human experts, so that the network learns human visual attention. To this end, we build a two-branch classification network, where the input of one branch uses masked images to learning regional prior. Second, we use a Forward Grad-CAM (FG-CAM) to represent the attention of each branch and generate the saliency maps. Third, we further design an Attention Decision Module (ADM) to decide which part of the gradient flow of both two branch saliency maps will be updated.

This paper makes contributions as follows:

(1) We propose a novel attention-guided architecture for image quality assessment of slit lamp images. Experimental results show that it achieves visual attention closer to human experts than state-of-the-art baselines.
(2) We design a Forward Grad-CAM and an attention decision module. The FG-CAM is used to represent the network attention and can participate in network training, while ADM is used to update the branch gradients.

2 Dataset

The dataset we use contains 47095 slit lamp images taken from clinical purposes among several hospitals between 18/3/2015 and 05/10/2019. The dataset contains a variety of diseases, e.g., pterygium, trichiasis, pinguecula, hemorrhage, edema and cases of different degrees of conjunctival hyperemia. Further, the dataset also contains a variety of lighting conditions, e.g., Retro-illumination and indirect illumination, while the cases with ocular fluorescein staining are excluded in this analysis.

We select 11831, 2000 and 9367 images as training set, validation set and test set, respectively. Note that, the training set and test set are patient-independent. All the images are resized to 224 × 224. To evaluate the image quality of the bulbar conjunctiva area in these images, 9 trained graduate students annotated three types of labels that illumination (‘Good’, ‘Medium’, ‘Bad’), blur (‘Slight’, ‘Medium’, ‘Sever’) and image quality (‘Accept’ or ‘Refuse’) for the train and validation dataset, while only image quality (‘Accept’ or ‘Refuse’) for the test dataset. 2 experienced experts finally determine the category of image quality. To obtain the bulbar conjunctiva mask, the two experts performed pixel-level annotations on 1045 additional slit lamp images. A U-Net [13] model was trained to acquire bulbar conjunctival region masks for the above train, validation and test dataset. The final dataset, called SLIQA, contains slit lamp images, image quality labels, and bulbar conjunctival region masks.

3 Method

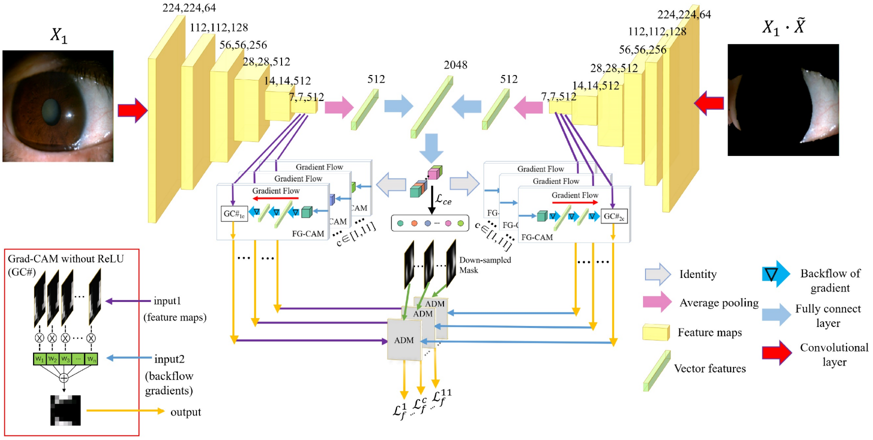

Fig. 1. The proposed attention-guided architecture for slit lamp image quality assessment.

Overview: Our proposed architecture as shown in Fig. 1 contains three parts: (1) Basic two-branch CNN. (2) Trainable Forward Grad-CAM. (3) Attention Decision Module. The two-branch CNN with different inputs is introduced in Sect. 3.1. The trainable Forward Grad-CAM (FG-CAM) used to obtain the saliency maps of two branches is introduced in Sect. 3.2. The Attention Decision Module (ADM) used to update the branch gradients is introduced in Sect. 3.3.

3.1 Multi-task Two-Branch Architecture

We denote the original slit lamp image as X and the bulbar conjunctiva region mask $\tilde{X}$. We firstly build a two-branch CNN, where the backbone we used is VGG [14], and the two branches are concatenated at the first length 512 fully connected layer. The inputs of two branches are X and $X \cdot \tilde{X}$ respectively, where $\cdot$ denotes pixel-wise multiplication. Then after two fully connected layers, the final fully connected layer output is length 11 category score vector relative to 4 classification tasks, including levels of illumination, blur, image quality and bulbar conjunctiva region area level.

The image quality classification is the main task while the others are auxiliary classification tasks. Level of the area L_{Area} is calculated as:

$$L_{Area} = \begin{cases} 0,\ A\left(\tilde{X}\right) < 0.15 \\ 1,\ 0.15 \le A\left(\tilde{X}\right) < 0.3 \\ 2,\ 0.3 \le A\left(\tilde{X}\right) \end{cases} \tag{1}$$

where $A(\bullet)$ indicates the area ratio of the bulbar conjunctiva area to the total area. This multi-tasking design is to extract more effective features while accelerating the convergence of the network.

3.2 Trainable Forward Grad-CAM

In this work, we seek a CAM that can participate in network training, not just for visualization. Inspired by [15], we use a trainable Forward Grad-CAM (FG-CAM) to describe the saliency of attention. It is expressed as:

$$A = conv(f, w) \tag{2}$$

where f is the feature map of the convolutional layer, and w represents the neuron importance weights obtained by the gradients flowing back through a global average pooling layer. Different from [15], we remove the ReLU operation, which is designed for visualization in the work of Selvaraju et al. [16].

3.3 Attention Decision Module

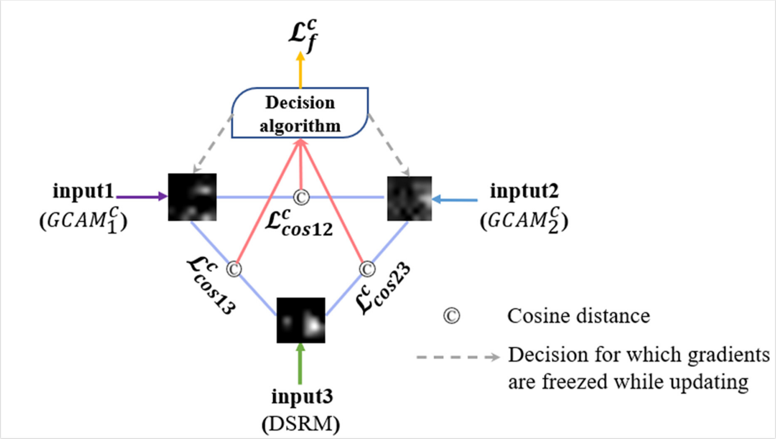

Fig. 2. Our proposed attention decision module. $GCAM_1^c$, $GCAM_2^c$ denote two outputs of FG-CAM modules in Fig. 1. DSRM denotes down-sampled semantic region mask. We compare the cosine distance of these three inputs one to one, and we make final decision which gradients flows will be frozen by Table 1.

We further propose an attention decision module (ADM) as shown in Fig. 2. For ADM with respect to class c, FG-CAM module outputs $GCAM_1^c$, $GCAM_2^c$ are obtained by Eq. (3) from two branch final convolutional layer features of our CNN respectively:

$$GCAM_i^c = \sum_k \frac{\partial \mathrm{y}^c}{\partial A^k} A^k \tag{3}$$

where A^k denotes k^{th} feature map at the final convolutional layer, y^c denotes score of class c, $\frac{\partial \mathbf{y}^c}{\partial A^k}$ denotes gradient matrix that contains derivative of function y^c with respect to A^k by forward propagation. The DSRM is calculated as:

$$DSRM = DownSample\left(\tilde{X}\right) - f\left(\tilde{X}\right) \tag{4}$$

where $DownSample(\bullet)$ denotes mean pooling module in this paper, and $f\left(\tilde{X}\right)$ is a scalar to adjust pixel value distribution of DSRM. Once three inputs are prepared, and then we calculate their cosine distances $\mathcal{L}_{cos13}^c$, $\mathcal{L}_{cos23}^c$, $\mathcal{L}_{cos12}^c$ with respect to the class c by:

$$\mathcal{L}_{cosij}^c = 1 - \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \tag{5}$$

where v_i, v_j denote two vectors to be calculated cosine distance, $\mathcal{L}_{cosij}^c \in [0, 2]$, v_1, v_2, v_3 are flattened by $GCAM_1^c$, $GCAM_2^c$, $DSRM$ respectively. After that we make a final decision of the output $\mathcal{L}_f^c$ by the following algorithm in Table 1:

Table 1. The decision algorithm of ADM with input1, input2, input3 in Fig. 2.

```
ADM Algorithm
1: Firstly, denote two thresholds $th_1 \in [0,1]$, $th_2 \in [0,1]$
2: If $\mathcal{L}_{cos12}^c > th_2$ and max{$\mathcal{L}_{cos13}^c$, $\mathcal{L}_{cos23}^c$} $> th_1$:
       $\mathcal{L}_f^c = \mathcal{L}_{cos12}$
       if $\mathcal{L}_{cos13}^c > \mathcal{L}_{cos23}^c$:
           freeze gradient flow of $GCAM_2^c$
       else:
           freeze gradient flow of $GCAM_1^c$
       end if
   else:
       $\mathcal{L}_f^c = 0$
   end if
```

th_1 is a threshold to describe the tolerability of disimilarity between $GCAM_i^c$ and DSRM. $\mathcal{L}_f^c$ will be set to 0 when $GCAM_i^c$ is similar to DSRM enough. $GCAM_i^c$ is expected to tend to be different from DSRM to some extend, because we believe that the weight distribution of the neural network attention region $GCAM_i^c$ is not necessarily similar to that of semantic region DSRM, and the specific extent is decided by the neural network itself. th_2 is a threshold to describe the tolerability of maximal angle between v_1 and v_2. $\mathcal{L}_f^c$ will be set t 0 when the angle is small enough. $GCAM_1^c$, $GCAM_2^c$ are expected to tend to focus on different regions, because we believe more information tends to mine when attention regions on the final convolutional layer features of two branches are different.

Combining Fig. 1, frozen the gradient flow of $GCAM_i^c$ denotes that in backpropagation the gradient backflow of $GCAM_i^c$ with respect to class c does not be optimized, which means the one that has a bigger difference with respect to down-sampled semantic region mask will learn to the other one but not learn with each other. Note that the last fully-connected layer parameters are shared between $GCAM_1^c$, $GCAM_2^c$, so the gradients of these parameters will not be frozen.

Overall, the total loss of our model is:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \beta \cdot \frac{1}{l} \cdot \sum_c^l \mathcal{L}_f^c \tag{6}$$

where l denotes the number of classes, and β is a coefficient to adjust the contribution between cross entropy loss and the ADM loss.

4 Experiments

4.1 Implementation Details

All Experiments in this paper obey the following rules: The Adam optimizer is adopted with the learning rate of 0.0001 firstly. When the average training accuracy of the multi-task classification is above 85%, the learning rate is set to 0.00001. It will be early stopped when the training accuracy of task image quality is above 98%, which is judged to be overfitting. The mini-batch size is set to 8 and all the experiments run on an NVIDIA GTX 1080Ti GPU.

4.2 Parameter Influence

The th_1 and th_2 in Table 1 will affect the tolerability of dissimilarity among Grad-CAM maps and the semantic region mask, and the β in Eq. (6) will affect the balance between cross-entropy loss and ADM loss. As shown in Table 2, we can see different th_1 and th_2 have little effect on AUC, F1, Accuracy, which shows our proposed method has good robustness. When β is set to 0.03, it will have an obvious performance deduction on the metrics. The reason is that our ADM loss accounts too small proportion to guide neural network attention to the goal region.

4.3 Comparison with Other Methods

Our method is compared with other similar methods as shown in Table 3. All the inputs of the compared methods are the original images. The AFN [1] and LACNN [5] are designed for lesion region mask attention, so there are not any improvement on our task compared with baseline [14]. AFN has long train time because extra structure is added on fully connected layer. The GAIN proposed by Li et al. [15] first utilized forward

Table 2. Parameter influence. Each experiment is repeated three times. The basic combination of the parameters is $\text{th}_1 = 0.8$, $\text{th}_2 = 0.4$, $\beta = 0.1$. We change one of these parameters, and the other two parameters remain unchanged for one experiment.

Parameter	AUC (%)	F1-score (%)	Accuracy (%)
$\text{th}_1 = 0.7$	97.30 ± 0.10	84.29 ± 0.39	92.67 ± 0.09
$\text{th}_1 = 0.8$	97.42 ± 0.09	84.81 ± 0.55	92.65 ± 0.16
$\text{th}_1 = 0.9$	97.32 ± 0.04	84.34 ± 0.55	92.36 ± 0.33
$\text{th}_2 = 0.3$	97.37 ± 0.07	84.60 ± 0.61	92.65 ± 0.17
$\text{th}_2 = 0.5$	97.29 ± 0.08	84.44 ± 0.79	92.76 ± 0.16
$\beta = 0.03$	97.18 ± 0.09	83.75 ± 0.77	92.20 ± 0.48
$\beta = 0.3$	97.37 ± 0.08	84.41 ± 0.28	92.52 ± 0.06

Grad-CAM to guide CNN's attention, and it has slightly improvement on AUC and F1-score compared with baseline, but its serial repeat feature extractors take big cost of time. The DFS proposed by Wang et al. [9] is designed for semantic region mask attention, and it has a little improvement compared with baseline, but its added segmentation head attention takes long time. Our proposed method has obvious improvement on AUC, F1-score and accuracy, and also takes short training time because our architecture has not any extra structures or serial repeat parts.

Table 3. Comparison with other similar methods, where each experiment is repeated six times. We denote the train time of the baseline as one unit time.

Method	AUC (%)	F1-score (%)	Accuracy (%)	Train time
Baseline [14]	96.99 ± 0.13	83.31 ± 0.41	92.10 ± 0.28	**1**
AFN [1]	96.97 ± 0.15	83.07 ± 0.48	91.87 ± 0.26	3.115
GAIN [15]	97.07 ± 0.20	83.46 ± 0.77	91.96 ± 0.33	3.067
LACNN [5]	96.99 ± 0.18	83.21 ± 0.52	91.92 ± 0.33	1.308
DFS [9]	97.16 ± 0.13	83.58 ± 0.48	92.00 ± 0.17	3.719
Ours	**97.41 ± 0.14**	**84.79 ± 0.42**	**92.71 ± 0.28**	1.966

The FG-CAM visualization results of each method is shown in Fig. 3. For column A, AFN has a deviation while other methods focus on the overexposure region. For column B, AFN and our proposed method focus on the left underexposure region in column B2 while baseline and LACNN focus on the error region of the cornea in column B1. For column C, all the methods focus on the overexposure region in column C2, but the semantic region masked image has many black holes in the overexposure region which will impede semantic comprehension. Only baseline and our proposed method focus on the whole overexposure region in column C1. For column D, baseline, LACNN and our proposed method focus on the pterygium region in D1, and our proposed method pays more attention to the bulbar conjunctiva region, while baseline and LACNN pay more attention to the angulus oculi medialis region which is out of the bulbar conjunctiva region. Overall, our proposed method not only focuses on the bulbar conjunctiva region, but also notices the specific abnormality regions. Moreover, our method also notices the whole abnormality region on the original image branch, while the semantic region masked image has obvious black holes that have a big influence on semantic comprehension.

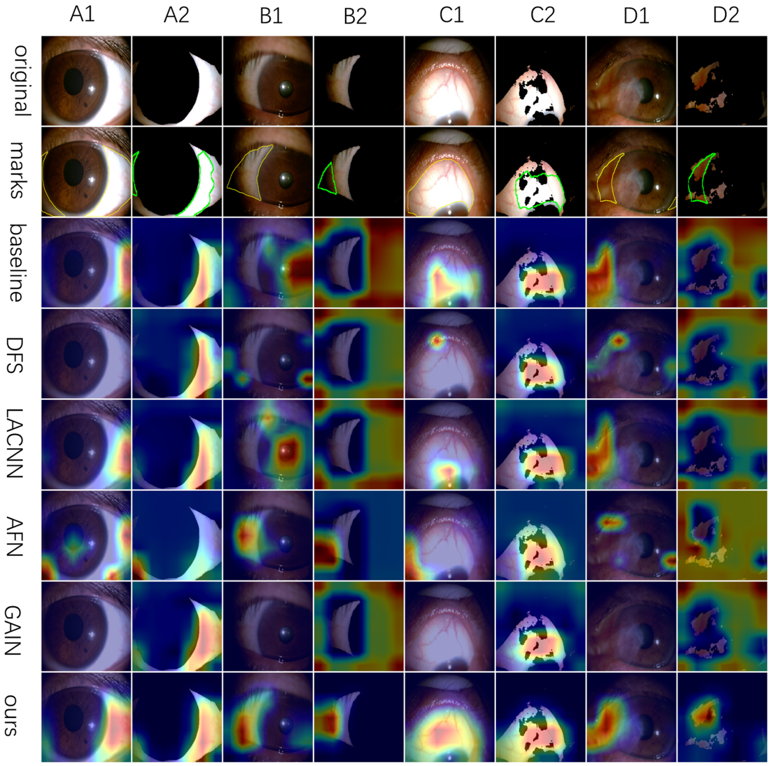

Fig. 3. Grad-CAM visualization results of each method. Column A1, B1, C1, D1 are four examples with different original inputs respectively. A2, B2, C2, D2 are bulbar conjunctiva regions masked image with respect to A1, B1, C1, D1 respectively. The first row contains the original images of four pairs of examples. The second row contains eight marked images, where yellow marks in the odd column show reference bulbar conjunctiva regions while green marks in even column show reference abnormality regions. The third row to the eighth row are the visualization of each method. (Color figure online)

5 Conclusion

We proposed an attention-guided architecture with two-branch CNN for the slit lamp image quality assessment. Experimental results show that it achieves visual attention close to human experts and thus improves classification performance. Compared with the-state-of-art methods, our proposed method has a better performance on AUC, F1-score, Accuracy metrics. Moreover, our method has the potential to migrate to other attention-dependent tasks.

References

1. Lin, Z., et al.: A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 74–82. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_9
2. Yang, Y., Li, T., Li, W., Wu, H., Fan, W., Zhang, W.: Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 533–540. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_61
3. Zhou, Y., et al.: Collaborative learning of semi-supervised segmentation and classification for medical images. In: CVPR, pp. 2074–2083 (2019)
4. Kermany, D.S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. In: Cell, pp. 1122–1131 (2018)
5. Fang, L., Wang, C., Li, S., et al.: Attention to lesion: lesion-aware convolutional neural network for retinal optical coherence tomography image classification. IEEE TMI **38**(8), 1959–1970 (2019)
6. Rasti, R., et al.: Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE TMI **37**(4), 1024–1034 (2019)
7. Chen, Q., et al.: Automated drusen segmentation and quantification in SD-OCT images. Med. Image Anal. **17**(8), 1058–1072 (2013)
8. Huang, L., et al.: Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network. IEEE Signal Process. Lett. **26**(7), 1026–1030 (2019)
9. Wang, L., Zhang, K., Ren, M., et al.: Recognition oriented iris image quality assessment in the feature space. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–9 (2020)
10. Wang, X., Ju, L., Zhao, X., Ge, Z.: Retinal abnormalities recognition using regional multitask learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 30–38. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_4
11. He, X., Zhou, Y., Wang, B., Cui, S., Shao, L.: DME-Net: diabetic macular edema grading by auxiliary task learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 788–796. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_87
12. Masumoto, H., et al.: Severity classification of conjunctival hyperaemia by deep neural network ensembles. J. Ophthalmol. **2019**, 1–10 (2019)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

14. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Li, K., et al.: Tell me where to look: guided attention inference network. In: CVPR, pp. 9215–9223 (2018)
16. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)