Concept-based Explanation for Fine-grained Images and Its Application in Infectious Keratitis Classification

Zhengqing Fang
Zhejiang University
zq_fang@zju.edu.cn

Kun Kuang*
Zhejiang University
kunkuang@zju.edu.cn

Yuxiao Lin
Zhejiang University
yuxiaolinling@zju.edu.cn

Fei Wu*
Zhejiang University
wufei@cs.zju.edu.cn

Yu-Feng Yao*
Zhejiang University
yaoyf@zju.edu.cn

ABSTRACT

Interpretability has become an essential topic as deep learning is widely applied in professional fields (e.g., medical image processing) where high level of accountability is required. Existing methods for explanation mainly focus on computing the importance of low-level pixels or segments, rather than the high-level concepts. Concepts are of paramount importance for human to understand and make decisions, especially for those fine-grained tasks. In this paper, we focus on the real application problem of classification of infectious keratitis and propose a visual concept mining (VCM) method to explain the fine-grained infectious keratitis images. Based on our discovered explainable visual concepts, we further propose a visual concept enhanced framework for infectious keratitis classification. Extensive empirical experiments demonstrate that (i) our discovered visual concepts are highly coherent with the physicians' understanding and interpretation, and (ii) our visual concept enhanced model achieves significant improvement on the performance of infectious keratitis classification.

CCS CONCEPTS

• **Applied computing** → **Graphics recognition and interpretation**; *Imaging*; • **Computing methodologies** → *Supervised learning by classification*; *Cluster analysis*; *Anomaly detection*.

KEYWORDS

interpretability, visual concept, deep learning, keratitis classification

ACM Reference Format:
Zhengqing Fang, Kun Kuang, Yuxiao Lin, Fei Wu, and Yu-Feng Yao. 2020. Concept-based Explanation for Fine-grained Images and Its Application in Infectious Keratitis Classification. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413557

*Corresponding authors.

Multimedia '20, October 12–16, 2020, Seattle, United States

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413557

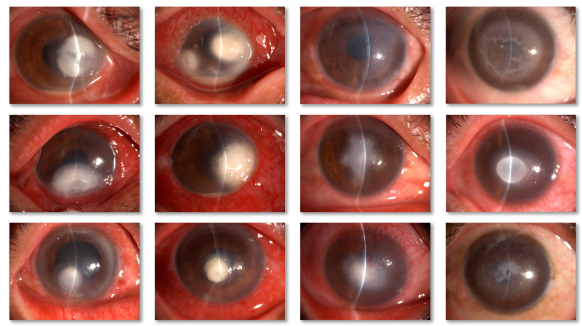


Figure 1: Examples of four categories of corneal diseases, include bacterial keratitis (BK), fungal keratitis (FK), herpes simplex viral stromal keratitis (HSK), and others referring to the corneal diseases except aforementioned three types of corneal infectious diseases (OTHERS), among which the manifestations of the diseases are subtle for identification by non-professionals.

1 INTRODUCTION

Artificial intelligence, especially deep learning has demonstrated remarkable performances in medical image analysis. However, this increasing performance comes as a cost of increasing model complexity and opacity. As a result, most of those models are used in a black-box way without being able to explain model decisions. However, in the medical field, methods require high level of accountability and transparency, which means one need to explain machine decisions, predictions and justify their reliability [24].

Many machine learning explanation methods have been proposed to bring understanding on black-box learning models, such as LIME [17], SHAP [30], GradCAM [21] and Guided-Backpropagation [22]. These methods give explanation for a model by computing or approximating the importance of each individual feature or low-level pixel. However, they are found to be algorithm-centric with few human-subject tests to verify their contributions for human interpretability [24] and lack of discussion about the relationship between per-sample saliency and corresponding category. Moreover, [14] showed that these methods do not increase human understanding and trust of the model.

Recently, a line of research has focused on providing explanations around deep learning models in the form of human "concepts" levels, including TCAV [14] and ACE [10]. Instead of computing the importance of each individual feature or pixel, these methods output the important concepts that are coherent with the human understanding. For example, "black and white stripes" is the main concept for detecting Zebra and "neckline" is the concept to identify the Shirt as shown in Figure 2a. Concepts are meaningful, coherent and important visual patterns that could provide great explanations to increase human understanding of deep models [15]. However, these concepts based methods come with their own drawbacks. [14] and [10] concentrated on concepts of each certain class, neglecting the fact that there might be many common concepts among different categories. For example, the "black and white strips" could be a common concept for *zebra* and the *black and white shirt* as shown in Figure 2a. Hence, [14] and [10] would lead to misunderstanding or confusing on those concepts across categories and hurt their importance for classification, especially in fine-grained tasks.

In this paper, we focus on the concept based explanation for a real application problem of infectious keratitis classification, which is a fine-grained task in medical field. As shown in Figure 1, three most common keratitis are bacterial keratitis (BK), fungal keratitis (FK) and herpes simplex viral stromal keratitis (HSK). We define those corneal disease entities other than aforementioned three categories of infectious keratitis as OTHERS. During diagnosis, physicians identify the subtle clinical manifestations/concepts on the cornea lesion area as criterions.

To provide the concept based explanation in fine-grained task, we propose a novel visual concept mining (VCM) algorithm, which consists of two main components: potential concept generator and visual concept extractor. The potential concept generator is designed for catching the subtle concepts by automatically searching and grouping important pixels via saliency map calculation, and producing salient patches which contain accountable manifestations as find-grained potential concepts. To address the challenges from common concepts, we propose a visual concept extractor which learns the concept similarity and diversity among different classes with Deepcluster [3] techniques, and quantifies their correlation and unique contribution to each class. Figure 2b(ii) demonstrates our discovered visual concepts for a case of fungal keratitis (FK). In this case, there are 4 kinds of concepts indexed by $1, 2, 3, 4$. The concept 1 is a common concept of classes bacterial keratitis and fungal keratitis, and concept 2 is a common concept of classes fungal keratitis and herpes simplex viral keratitis. These common concepts would not be discovered by previous methods. However, with considering the correlation between concepts, our algorithm demonstrates that the combination of concepts 1 and 2 is a great explanation for the class of FK. Moreover, the explanation is exactly coherent with the physician understanding as we demonstrated in Figure 2b(ii).

What's more, we propose a visual concept enhanced framework to joint our discovered visual concepts with the features extracted by traditional deep model for infectious keratitis classification. Experimental results show that our discovered visual concepts can significantly improve the performance of the base deep model.

Our Contribution. To summarize, our contributions are listed as follows:

- We investigate the concept based explanation problem on the real medical application of infectious keratitis classification, which is a fine-grained task.
- We propose a novel visual concept mining algorithm, consisting of potential concept generator and visual concept extractor, to automatically generate explainable visual concepts for fine-grained infectious keratitis classification.
- We propose a visual concept enhanced framework to strengthen the performance of traditional deep model via incorporating the features of discovered visual concepts.
- Extensive experiments demonstrate that our discovered visual concepts are (i) meaningful and explainable: they are coherent with the physician understanding and interpretation; and (ii) important: they can be used to improve the performance of infectious keratitis classification.

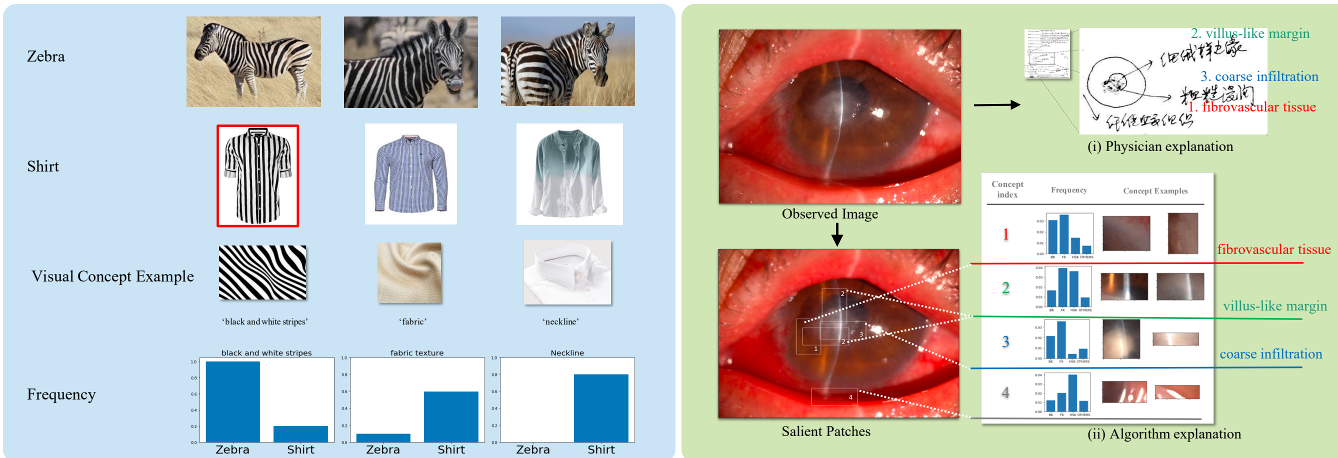
2 RELATED WORK

Deep learning in Medical Field. Deep learning methods perform better as models become wider [23] and deeper [12] [13], and are widely used in medical image analysis [18]. These CNN-based AI algorithms can perform anatomical structure segmentation on CT images [33], classify normal or abnormal findings of chest radiographs [7], perform screening for lung [1] or breast cancer [25], detect critical findings in head CT scans [5], classify liver lesion [9] and detect lymph node metastases in pathology images [2] [8]. All above show remarkable results, however, their practicality remains to be clinically examined.

Another research direction is computer aided medicine, in which deep learning also plays an important role. In segmentation tasks, Unet [19] proposed an architecture consisting of a contracting path to capture context and a symmetric expanding path to enable precise localization, which achieved a high performance on cell tracking challenge. Followed by many frameworks such as Unet++ [34], Unet leads a tendency of combining segmentation and classification [26] and has been applied in retinal vessel [27] and Nuclei [29] segmentation.

Explainable Artificial Intelligence. Explainable artificial intelligence has become a hotspot in machine learning research community that intends to figure out why and what is accountable if things go wrong, or how to leverage them further [24]. A well-known interpretation method is Class Activation Mapping(CAM) [31], which produces saliency-maps which correspond to different categories. Grad-CAM [21] uses the gradients of any target class feeding into the final convolutional layer to produce a coarse localization, which is applicable to a wide variety of CNN model-familiars. As the combination with Guided Back-propagation [22], Guided Grad-CAM achieves a fine-grained visualization and could quantify the contribution of each individual pixel. LIME [17] explains the predictions of any classifier by learning an interpretable model locally around the prediction. Shapely values(SHAP [30])'s interpretation contains the most (and least) important segments of input images.

Recent researches [32] [14] [4] [10] have focused on providing explanations in the form of high-level human "concepts". IBD [32] decomposes the prediction of one image into human-interpretable conceptual components. TCAV [14] produces estimation of how important a concept is for the prediction. ProtoPNet [4] is trained to learn visual prototype vector and calculate similarity for prediction. ACE [10] proposed a method to automatically extract visual concept from certain class's images.



(a) Visual Concepts in General Classification

(b) Concept-based Explanation in Fine-grained Images

Figure 2: Examples of visual concepts in general classification and our fine-grained infectious keratitis classification. (a) shows three common visual concepts and compares their frequency in "zebra" and "shirt". We can explain why a zebra image is correctly classified or how wrong classification take place in shirt image once we find the concept "black and white strips", because human have prior knowledge that nearly all zebra have this visual concept, while few shirts do. (b) illustrates our concept based explanation for a case of fungal keratitis (FK). Our visual concept mining algorithm extracts 4 kinds of concepts, indexed with 1, 2, 3, and 4, to explain the observed medical image as FK category. Sub-figure (i) illustrates the explanations of physician and the frequency in sub-figure (ii) demonstrates the distribution of each concept in different keratitis categories. Our discovered visual concepts shows high coherency with interpretation of physician.

All the above-mentioned methods neglect the fact that there are many common visual concepts in different categories and their interpretation are limited in certain class. In other words, they just answer the question that how the image is correctly predicted but cannot answer the question of what is the distinguishment of one class from others.

Visual Data Mining. Visual data mining, or unsupervised object discovery, aims to find image fragments with same semantic meaning from a large image dataset automatically. As the popularity of deep learning grows, many self-supervised deep-learning-based representation learning methods have emerged. RotNet [6] proposed a self-supervised task of rotation recognition to learn image feature representation from unlabeled image dataset. DeepCluster [3] is a clustering method that jointly learns a network generating image representation and the cluster assignments of the resulting features. It iteratively uses k-means to group image features, and uses the subsequent assignment to train the network. BowNet [11] learns perturbation-invariant and context-aware image features by training a model to predict bags-of-visual-words representation of original images given perturbed images as input. We propose a method based on DeepCluster to automatically learn representations of collected salient patches and to extract visual concepts from them.

3 VISUAL CONCEPTS IN INFECTIOUS KERATITIS INTERPRETATION

In this section, we first introduce the background of diagnosing infectious keratitis and its necessity to be interpretable. Then we enumerate some difficulties of applying traditional computer vision methods to detect related clinical manifestations. Finally, we give an explanation of why our automatically mined visual concepts are suitable for representing clinical manifestations.

Background. Infectious keratitis are the most common entities of corneal diseases, in which pathogen grows in the cornea leading to inflammation and destruction of the corneal tissues. Microorganisms that causes corneal infection involve bacteria, viruses, fungi and protozoa. Triage and diagnosis of diseases are carried out by physicians through observation based upon experience and knowledge constructed by individuals so ophthalmologists can only achieve 49.27±11.5% diagnostic accuracy according to [28]. Though the deep learning method proposed by [28] could achieve 80.00% diagnostic accuracy, poor interpretation limits its practicality. For junior physicians, the difficulty of diagnosing keratitis is the lack of experience to distinguish subtle manifestations. Thus, interpretation based on manifestations became a necessity for deep learning to be reliable and practical.

Representative Clinical Manifestations. The uttermost feature of infectious keratitis is the pathogen growth in the cornea leading to focal mass cloudiness and the cornea roughness, ineluctably bringing out unique characteristics of each pathogenic microorganism for its growth in the tissue [28]. Experienced ophthalmologists usually describe them using medical terms subjectively, e.g. "infiltrate", "lesion", "edema", "cloudiness", "opacity", "stroma thinning", "dense scarring", etc. However, manifestations are of indistinct edges and uncertain amounts in keratitis images. After plenty of surveys, we conclude that traditional methods under supervised condition are not suitable for detecting manifestations, such as multitask learning, object detection or instance segmentation, because the collecting of manifestations notations is challenging. Unlike regular tasks, expertise is highly needed when labeling manifestations, and moreover, lacking of standards making it more difficult to perform.

Superiority of Visual Concepts. In this paper, we focus on mining visual concepts for detecting manifestations without any prior knowledge of the manifestations annotations. Ideally, the discovered visual concepts should be coherent with the manifestations that have 1) Meaningfulness: an example of a concept should be meaningful/understandable to human; 2) Coherency, Examples of a concept should be similar to each other while being different from examples of other concepts; 3) Importance, a concept should be important features for prediction or diagnosis.

4 METHOD

In this section, we present the visual concept mining (VCM) framework, which consists of two main components, potential concept generator and visual concept extractor, as shown in Figure 3. Potential concept generator is designed for automatically searching salient patches that contain clinical manifestations to distinguish different keratitis. Those salient patches are with preliminary interpretability for classification, but with large number, hence, we treat them as potential concepts. Then, the visual concept extractor is designed for mining meaningful, coherent concepts with a clustering based method to explain the keratitis.

Next, we will introduce the details of each component[1].

4.1 Potential Concept Generator

To approximately locate the representative clinical manifestations of keratitis in the condition that labeling is challenge, we designed Potential Concept Generator which employs Guided Grad-CAM, a widely-adopted interpretation method calculating pixel-level importance, combined Unet to estimate saliency map and produce salient patches containing most of the accountable manifestations. Three main procedures are necessary to construct a Potential Concept Generator: 1) Classification and segmentation model pretraining. 2) Saliency map calculating. 3) Candidate anchors screening.

Pretrain. Considering the gap between slit-lamp microscopic image datasets and other datasets like ImageNet [20], we trained a backbone using Densenet121 from scratch, aiming to minimize the NLL loss function(the negative log likelihood loss) $loss = \sum_{n=0}^{N}(nll(W(I_n), y_n))$ given a training set I and ground truth label y, where $W(.)$ denotes the trainable weights of Densenet121. An optional Unet can be trained using a few pixel-level labels with objective function $\min_{\Phi}(\sum_{n=0}^{N}\sum_{i,j} BCE(\Phi(I_n)_{i,j}, mask_{i,j}))$, where N denotes the number of samples, $BCE(.)$, $\Phi(.)$denotes Binary Cross Entropy loss function and Unet parameters correspondingly.

Calculate Saliency Map. For each sample in the training set, we applied Guided Back-propagation and Grad-CAM to visualize salient pixels. As shown in Eq.1, Guided Back-propagation save all the positive gradient that we can quantify contribution of every pixel.

$$\Omega_{n\,i,j} = relu(\frac{\partial y_n}{\partial I_{n\,i,j}}) \tag{1}$$

Grad-CAM was applied for calculating salient regions. Through weighted combination of forward activation maps, we obtained a coarse heatmap of 7×7 size, as shown in Eq.2, where $w_k^{y_n} = \frac{1}{Z}\sum_i\sum_j \frac{\partial y_n}{\partial A_{ij}^k}$, $R(.)$ is linear interpolation operation for scaling to the same size as input.

$$L_{GradCAM}^{y_n} = R(relu(\sum_k w_k^{y_n} A^k)) \tag{2}$$

So far we have got segments from Unet as location constraint, heatmap from Grad-CAM as activation constraint and saliency score for each pixel as saliency constraint. As shown if Eq.3, we define the saliency map formula for keratitis.

$$S_n = \alpha_1\Phi(I_n) + \alpha_2\Omega_n + \alpha_3 L_{GradCAM}^{y_n} \tag{3}$$

where α_1,α_2,α_3 are hyperparameters. S_n is the saliency map corresponds to sample I_n.

Candidate anchors screening. We applied the same anchor generating strategy as FasterRCNN [16], using 3 scales and 3 aspect ratios, yielding k = 9 anchors at each sliding position. For a saliency map of a size $W \times H$(typically $224 * 224$ in our application), there are $W \times H \times k$ anchors in total. With so many candidate anchors, we design a two-stage screening strategy based on saliency distribution and similarity.

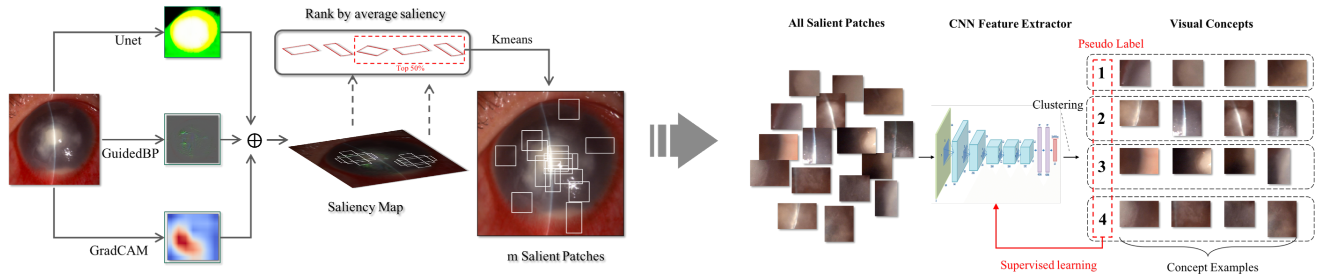
Screening by saliency. For each anchor p in sample n, we could calculate corresponding average saliency value $\bar{s} = \frac{1}{A_p}\sum_{i,j\subset p} S_{n\,i,j}$, and saliency variance $\hat{s} = \frac{1}{A_p}\sum_{i,j\subset p}(S_{n\,i,j} - \bar{s})^2$, where A_p denotes total pixel number in p. A candidate patch would be selected if both of corresponding average saliency $\bar{s}$ and saliency variance $\hat{s}$ rank in top 50%.

Screening by similarity. In order to remove redundancy, we use Kmeans clustering to select the most representative m salient patches ($m = 10$, in our task). Each candidate patch p cropped from original image was encoded to 1024-dimensional feature vector with pretrained Densenet weight $W(.)$. Here, we obtained m cluster centroids $C \in \mathbb{R}^{m\times 1024}$, and the nearest patch to each centroid is chosen. The objective of Kmeans training is $\min_{C\in\mathbb{R}^{m\times 1024}}\sum_P \|W(p) - C_p\|^2$, where C_p denotes the nearest centroid to patch p. As shown in Fig3, we obtain K distinctive salient patches which are high-resolution and vital to prediction.

4.2 Visual Concept Extractor

Since we have N training samples and m salient patches for each sample, a new dataset P with $N_P = m \times N$ unlabeled salient patches could be constructed, representing the most typical manifestations of keratitis. To figure out the correlation of all salient patches and their medical explanation for clinical diagnosis, we propose to learn the pattern similarity and diversity of samples in P in an unsupervised manner. Deepcluster [3], an inspiring self-supervised representing learning method, is suitable for our task. Given a set P, DeepCluster iteratively learns the features $\Theta(P) \in \mathbb{R}^{N_P\times d}$ and groups them into

[1]Implementation available at https://github.com/createrfang/VisualConceptMining.git



Stage One: Potential Concept Generator

Stage Two: Visual Concept Extractor

Figure 3: Overview of the VCM framework. The Potential Concept Generator uses correctly predicted images as input, generates and selects K salient patches for each image based on the result of Unet, GuidedBP and GradCAM, then crops the corresponding patches from original high-resolution image. The Visual Concept Extractor uses DeepCluster to group salient patches acquired from previous step, results in k different visual concepts.

K clusters. The training process, precisely, learns a $K \times d$ centroid matrix C and the cluster assignments y_p of each salient patch p by solving the following problem:

$$\min_{C \in \mathbb{R}^{K \times d}, \Theta} \sum_{P} \|\Theta(p) - y_p\|^2 \tag{4}$$

where $\Theta(.)$ denotes parameters of AlexNet. In our implementation details $K = 32$, $d = 256$ after PCA Dimensionality Reduction from $\Theta(p) \in \mathbb{R}^{4096}$.

The cluster assignments y_p of each salient patch can be viewed as pseudo labels, which are used to update Θ, after calculating cross entropy loss and backward gradient. We repeat these procedures until the loss converges and labels become stable. According to ACE [10], the patches in final clusters satisfy the three properties: meaningfulness, coherency and importance. So we can claim that each cluster represents an individual visual concept, and the patches in this cluster are examples of corresponding visual concept. Because it is too professional and too subjective to name each visual concept properly, we used the cluster index $k \in \{1, 2, \ldots, 31\}$in terms of corresponding visual concept.

4.3 Statistical Analysis on Visual Concepts

So far, we have got K=32 visual concepts which are highly related to 4 infectious keratitis categories: several of them are general in all images, some of them are common in certain two categories and some of them are unique for certain category. To estimate the correlation of concepts and classes, we construct a 32×4 metric M, where M_{ij} denotes the number of patches belong to i-th visual concept and cropped from j-th class's images. Then the frequency of i-th visual concept in j-th category is $P(k = i|C = j) = \frac{M_{ij}}{\sum_{k=1}^{32} M_{kj}}$, as shown in Figure 4.

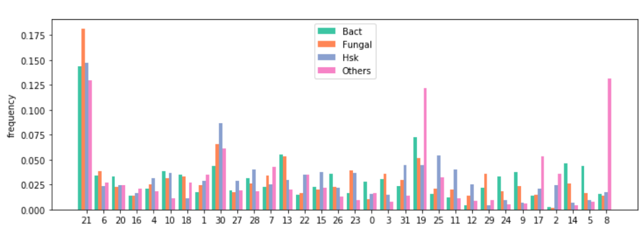


Figure 4: A histogram presents visual concepts' frequency in four categories, sorted by specificity score in ascending order from left to right.

According to the frequency in 4 categories, we can easily determine which class a visual concept appears most, marked as c_{max}. In some ways, if we find a visual concept in an unknown image, we tend to guess it belongs to class c_{max}. To imitate this process, we design a value function to quantify the relationship between visual concept i and its c_{max}, shown as Eq.5.

$$S_i = \frac{P(k = i|C = c_{max})}{\sum_{c_j \in C - c_{max}} P(k = i|C = c_j)} \tag{5}$$

where S_i denotes the specificity score. We sort visual concepts by specificity score in ascending order and present the result in Figure 4.

Interpretation Framework. The clustering model and its weights, concepts' centroids with samples and their statistical results mentioned above are saved as parts of our framework. When interpreting new samples, We follow the procedures shown below:

(1) Fed test samples into Potential Concept Generator and obtain several high-resolution patches that model interested in.
(2) Extract patches' features via saved model and weights, and find the nearest centroid for each patch.
(3) Visualize prior knowledge of corresponding visual concepts. We not only provide statistical results, but also present similar visual concept examples for the convenience of analog learning.

5 VISUAL CONCEPT ENHANCED CLASSIFICATION

Figure 4 suggests that some visual concepts are of high specificity that are worthy to focus on, for example, the concept 8 is almost unique for the OTHERS class, while several are so confusing that

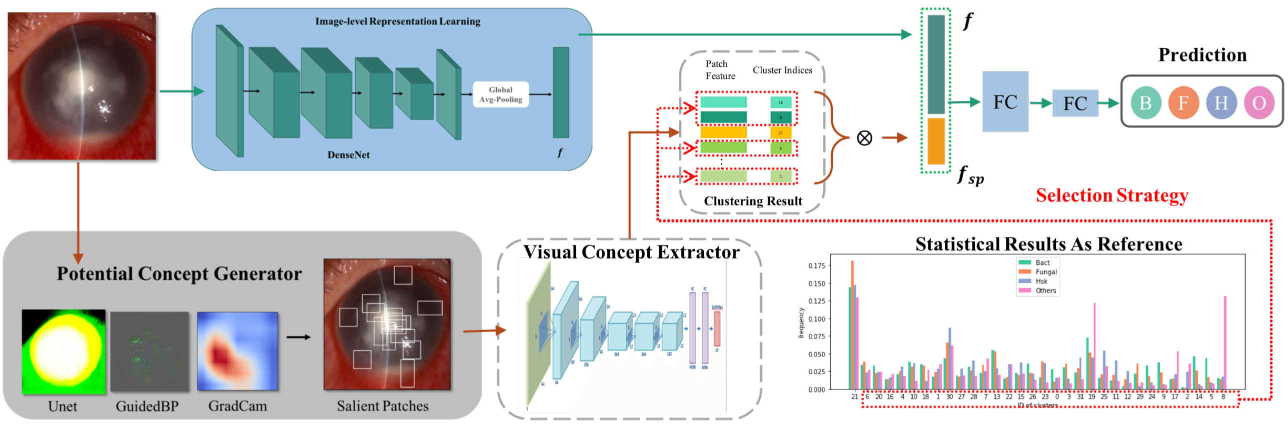


Figure 5: The overall architecture of Visual Concept Enhanced Classification. The framework incorporates image feature f produced by DenseNet and high-resolution patches' feature f_{sp} produced by Potential Concept Generator and Visual Concept Extractor, guided by a selection strategy that refers to prior statistical results.

we want the algorithm to ignore, for example, concepts 21 is a common concept for all classes. To achieve this goal, we creatively designed a Visual Concept Enhanced Classification model with addition of Selective Concept Branch. The overview of architecture is introduced as Figure 5.

Overall Architecture. It's an end-to-end model incorporating discovered visual concepts and pixel-level features of image for classification. We put input into backbone(Densenet121 in our experiment) and get feature f containing global information, meanwhile we obtain local feature f_{sp} from Selective Concept Branch, representing information of high-resolution visual concepts' patches. After concatenating f and f_{sp}, we apply a two-layer fully connected(FC) network for classification. The trainable part are backbone network(Denset121) and FC layers.

Selective Concept Branch. The branch's workflow is based on interpretation framework, joint with a selecting procedure. The choice strategy is determined by statistical results and objectives: to remove confusing concepts, to pick specify concepts and to reinforce certain category for example. Finally, we merge the feature of selected concepts' patches with linear addition and obtain f_{sp}.

6 EXPERIMENT

In this section, we evaluate the meaningfulness and coherency of our discovered visual concepts based on the understanding and interpretation of physician, also check their importance via visual concept enhanced approach for infectious keratitis classification.

6.1 Dataset descriptions

Microorganisms causing corneal infection involve bacteria, viruses, fungi and protozoa, and have different manifestations due to different pathogens. Figure1 presents the representative slit-lamp microscopic images of bacterial keratitis (BK), fungal keratitis (FK), herpes simplex viral stromal keratitis (HSK), and the OTHERS represents those corneal disease entities rather than aforementioned three categories of the corneal diseases. These images are selected from a high-quality infectious keratitis dataset proposed by Xu et al. [28], in which images are taken from patients with corneal infection at the active stage, including bacterial keratitis, fungal keratitis and herpes simplex viral stromal keratitis.

The dataset we used involved 2,284 images from 867 patients. The training set consists of randomly selected 387 images of bacterial keratitis, 519 images of fungal keratitis, 488 images of HSV stromal keratitis, and 528 images of other corneal diseases, from 747 patients. The testing set consists of randomly selected 86 images of bacterial keratitis, 97 images of fungal keratitis, 51 images of HSV stromal keratitis, and 128 images of other diagnosis, from 120 patients.

6.2 Baseline

In our paper, we have two main tasks, one is interpretation on keratitis images, the other is infectious keratitis classification.

In the task of interpretation, we evaluate the performance of concept based interpretation of our VCM algorithm, comparing with Grad-CAM [21] for interpretation on pixel or patch level, and ACE [10] for interpretation on concept level. We also compare these interpretations with the physician interpretations.

In the task of classification, we applied traditional deep model, such as DenseNet121 [13], ResNet [12] and VGG16 [23] for baselines. To comprehensively demonstrate the importance of our discovered visual concept, we first implement a simple method, named VCSP, by directly using visual concepts feature for classification, then, we implement our visual concept enhanced classification (VCEC) based on the DenseNet121 model by fusing features from both DenseNet121 and our discovered visual concepts for classification.

6.3 Results on Interpretation

In this section, we evaluate the visual interpretation performances of our visual concept mining approach for concept based interpretation. Figure 6 compares the interpretation of baselines (Grad-CAM [21]

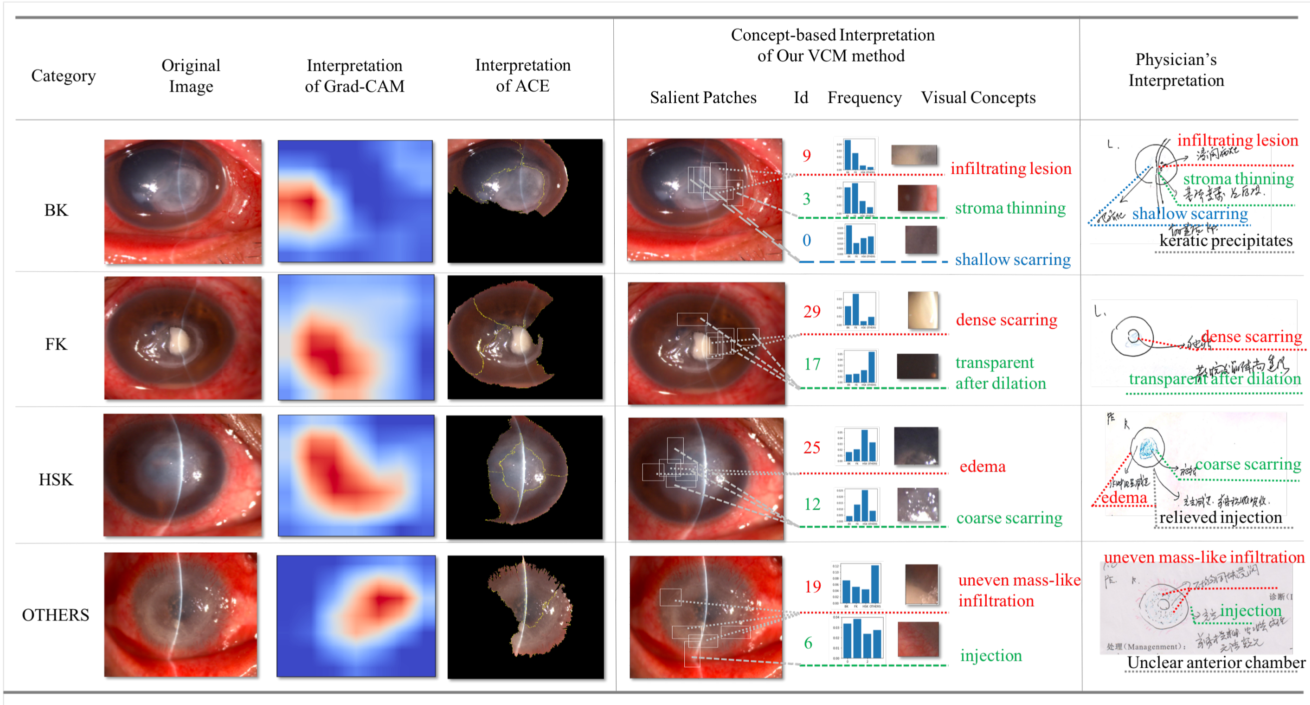


Figure 6: The comparison of different interpretations for keratitis images from each category. Our VCM interpretation contains salient patches, related visual concepts with extra examples for better comprehension, and corresponding frequency plots in different classes(BK, FK, HSK and OTHERS). Coherent clinical manifestations and visual concepts are shown in the same color.

and ACE [10]) and our model with the gold-standard physician's interpretation on four kinds of keratitis images. From figure 6, we have following observations and analyses:

(1) Grad-CAM only highlight the position of important pixels for its interpretation, which can be considered emphasize the salient region for the given class. From the results, the emphasized salient region overlaps the lesion intuitively, but it is hard for us/physician to distinguish the subtle difference between different keratitis.
(2) The discovered concepts by ACE is too coarse to explain the keratitis images. Although these concepts can cover the physician interpretation, for example, the discovered concept in FK class contains the part of "dense scarring", and concept in HSK class includes the part of "coarse scarring", they are hardly to provide concept explanation in our fine-grained keratitis image classification.
(3) Our discovered visual concepts present significant coherency with the physician's understanding and interpretation, for example, the discovered concept 9 in BK class is exactly the "lesion of infiltrate" in the physician's interpretation, and the meaning and position of concept 25 in HSK class is highly coherent with the physician's interpretation. Moreover, the representative concepts are different across classes, bringing a more meaningful and human-friendly explanation for each kind of keratitis.

6.4 Results on Classification

In this section, we evaluate the importance of our established visual concepts by visual concept enhanced framework for infectious keratitis classification.

Experimental Settings. All reported results are the average of the last epoch in an 100-epoch training, with a 10-step schedule decreasing learning rate beginning from 0.1 on a single 10 Gbs Titan V GPU. In DenseNet121 and ResNet50, images are scaled to 224×224 with 32 batch size while the size is 299×299 in VGG16.

Evaluation Metrics. In this paper, we focus on the problem of infectious keratitis classification. Hence, we use the accuracy($Acc = \frac{TP+TN}{P+N}$), $F_1 = \frac{2TP}{2TP+FN+FP}$ as evaluation metrics, where P and N denote the numbers of positive and negative samples, and TP, TN, FP and FN denote the numbers of true positive, true negative, false positive and false negative samples in prediction correspondingly. With considering on the data imbalance among different keratitis classes as we demonstrated in the data descriptions. We also employed Macro-F_1 ($MF_1 = \frac{1}{n}\sum_{i=1}^{n} F_1^i$) as an important evaluation metric, where n refers to the number of classes. In our problem, $n = 4$.

Experimental Results. We report the results on classification in Table 1, where we also demonstrate the accuracy of SOS model in [28]. From the results, we have following observations and analyses:

(1) Among the three deep model baselines, DenseNet121, ResNet and VGG16, the DenseNet121 achieved the best performance.

Table 1: Results of infectious keratitis classification.

Algorithm	*Acc*	F_1 Score				MF_1
		BK	FK	HSK	Others	
DenseNet	78.56	0.431	0.872	0.651	0.790	0.686
VGG	65.18	0.254	0.764	0.548	0.745	0.578
ResNet	69.10	0.275	0.810	0.566	0.750	0.601
SOS [28]	80.20	-	-	-	-	-
Human [28]	49.3±11.5	-	-	-	-	-
VCSP	59.30	0.257	0.743	0.434	0.529	0.490
VCEC	80.52	0.418	0.886	0.651	0.837	0.698
VCEC-P_1	82.26	0.454	0.890	0.670	0.856	0.717
VCEC-P_4	83.35	0.487	0.891	0.682	0.872	0.733
VCEC-P_7	80.73	0.452	0.868	0.655	0.865	0.710
VCEC-P_{10}	81.50	0.470	0.881	0.664	0.834	0.721
VCEC-P_{12}	**84.78**	**0.559**	0.893	**0.705**	**0.885**	**0.760**
VCEC-P_{15}	81.61	0.503	0.890	0.682	0.872	0.723
VCEC-D_1	82.37	0.492	**0.894**	0.686	0.842	0.728
VCEC-D_6	82.92	0.488	0.883	0.701	0.872	0.736
VCEC-D_{11}	80.84	0.466	0.887	0.656	0.828	0.709

That's why we choose DenseNet121 as backbone in our VCEC algorithm.

(2) By utilizing the sequential relation among different image patches, the SOS model revives a great accuracy with 80.2%. Here, we directly use the results in [16], since SOS need hand-labeled or predefined patches sequence as input.
(3) Our naive model, VCSP, achieved 0.49 on F1 score and 59.30% on accuracy, which is still better than the average performance from human. This demonstrates our discovered visual concepts are informative for infectious keratitis classification.
(4) By roughly incorporating all discovered visual concepts in our VCEC framework, our method VCEC achieved the best performance comparing with all the baselines. But the improvement is puny, since there are some common or confusing visual concepts incorporated.

To deeply demonstrate the importance of our discovered visual concepts, we propose following two strategies to select (or delete) the most informative (or confusing) visual concepts based on our statistical analysis in section 4.3:

- Pick-strategy: we pick the top-k specific visual concepts to enhance the base model for classification, our VCEC framework with this strategy mark as VCEC-P_k.
- Drop-strategy: we drop the top-k confusing visual concepts to enhance the base model for classification, our VCEC framework with this strategy mark as VCEC-D_k.

From Table 1, we observed (i) by picking the first specific visual concept (concept 8 as shown in Figure 4), our model VCEC-P_1 can significantly improve accuracy and MF_1 from our rough model VCEC; (ii) by dropping the most confusing visual concept (concept 21 as shown in Figure 4), our model VCEC-D_1 can also improve the performance of classification; (iii) By picking the top-12 specific visual concepts, our model VCEC-P_{12} achieved the best performance with 84.78% on the accuracy and 0.76 on the MF_1.

Overall, the visual concepts discovered by our VCM model and picking/dropping strategies are important, and can indeed enhance the deep model on the problem of infectious keratitis classification.

7 DISCUSSION

7.1 Quality of Visual Concepts

In this section, we discuss how to improve quality of automatically learned visual concepts. In our method: (1) Unet segmentation limits patches in cornea and lesion area and reinforces their meaningfulness. (2) Guided Grad-CAM adds probability in active area while generating salient patches and guarantees their importance. (3) Deep-clustering aggregates salient patches which contain same patterns as visual concepts' examples, guaranteeing coherency.

Consequently, in our framework, quality of visual concepts is determined by segmentation task, saliency evaluating task and clustering task correspondingly. There are various developing methods to solving these three tasks nowadays, and they could be applied to take place the methods we present individually and freely. We admit that there is room for further investigation, which remain open for future work.

7.2 Contributions of Visual Concepts

In this section, we analyze how visual concepts play a role in interpretation and classification. For interpretation, visual concepts provide a semantic way to express the information in each sample, and the distribution of visual concepts also helps a lot.

We have done quantity of experiments that outperforms baselines. A reasonable guess about our performance is the increasing parameters and extra information from high-resolution patch. The comparison of experiments VCEC and DenseNet denied it. We utilized all salient patches and retrain the whole network and got 1.96% promotion while the best performance got 6.22% with 0.76 F_1 score achieved by picking-12 strategy. Experiment results in Table1 demonstrate our visual concept enhanced model achieves significant improvement on the problem of infectious keratitis classification.

8 CONCLUSION

In this work, we developed the VCM framework for interpreting CNN for fine-grained tasks. The framework includes a Potential Concept Generator which produces salient patches containing most accountable features, and a Visual Concept Extractor which clusters salient patches into several groups as visual concepts. We also developed the VCEC framework which utilizes the interpretation result to improve the performance of the model.

In our experiment, we applied our framework on infectious keratitis classification task. The result indicated that, although without detailed clinical manifestations annotations, the discovered visual concepts are coherent with the physicians' understanding. The classification result using only the visual concept is on par with the average performance of ophthalmologists. By enhancing the base model using the discovered visual concept, our method significantly improved the performance of the base model, and beat the previous state-of-the-art method on this task.

ACKNOWLEDGMENTS

This research is supported by National Key Research and Development Program of China (No. 2018AAA0101900), the Fundamental Research Funds for the Central Universities, the Health Commission of Zhejiang Province (WKJ-ZJ-1905 and 2018ZD007) and the National Natural Science Foundation of China (61625107, 61751209). We thank Wenjia Xie and Yesheng Xu for medical consultation, MinQin Zhu for experiment assistance, Ming Kong and Prof. Qiang Zhu for insightful discussions and comments.

REFERENCES

[1] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* 25, 6 (2019), 954–961. https://doi.org/10.1038/s41591-019-0447-x

[2] Zachi I. Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M. McKie, Dorothy J. Ladewig, Gaurav Satam, Patricia A. Pellikka, Maurice Enriquez-Sarano, Peter A. Noseworthy, Thomas M. Munger, Samuel J. Asirvatham, Christopher G. Scott, Rickey E. Carter, and Paul A. Friedman. 2019. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine* 25, 1 (2019), 70–74. https://doi.org/10.1038/s41591-018-0240-2

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11218 LNCS (2018), 139–156. https://doi.org/10.1007/978-3-030-01264-9_9 arXiv:1807.05520

[4] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2018. This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS (2018), 1–12. arXiv:1806.10574 http://arxiv.org/abs/1806.10574

[5] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. 2018. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet* 392, 10162 (2018), 2388–2396. https://doi.org/10.1016/S0140-6736(18)31645-3

[6] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, Jimeng Sun, and Sutter Health. 2017. RotNet unsupervised learning. *Iclr2018* 2016 (2017), 1–14. https://doi.org/10.1145/3097983.3098126 arXiv:arXiv:1611.07012v1

[7] Jared A. Dunnmon, Darvin Yi, Curtis P. Langlotz, Christopher Ré, Daniel L. Rubin, and Matthew P. Lungren. 2019. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology* 290, 3 (2019), 537–544. https://doi.org/10.1148/radiol.2018181422

[8] et al Teeple E, Collins J, Shrestha S, Dennerlein J. 2018. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Physiology & behavior* 176, 1 (2018), 139–148. https://doi.org/10.1016/j.physbeh.2017.03.040

[9] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* 321 (2018), 321–331. https://doi.org/10.1016/j.neucom.2018.09.013 arXiv:1803.01229

[10] Amirata Ghorbani, James Zou, James Wexler, and Been Kim. 2019. Towards Automatic Concept-based Explanations.

[11] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. 2020. Learning Representations by Predicting Bags of Visual Words. (2020). arXiv:2002.12247 http://arxiv.org/abs/2002.12247

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90 arXiv:1512.03385

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua, July (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243 arXiv:1608.06993

[14] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *35th International Conference on Machine Learning, ICML 2018* 6 (2018), 4186–4195. arXiv:1711.11279

[15] Yunhe Pan. 2020. Multiple Knowledge Representation of Artificial Intelligence. *Engineering* 6, 3 (2020), 216–217. https://doi.org/10.1016/j.eng.2019.12.011

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031 arXiv:1506.01497

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Augu (2016), 1135–1144. https://doi.org/10.1145/2939672.2939778 arXiv:arXiv:1602.04938v3

[18] Guoguang Rong, Arnaldo Mendez, Elie Bou Assi, Bo Zhao, and Mohamad Sawan. 2020. Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering* 6, 3 (2020), 291–301. https://doi.org/10.1016/j.eng.2019.08.015

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9351 (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28 arXiv:arXiv:1505.04597v1

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y arXiv:1409.0575

[21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (2020), 336–359. https://doi.org/10.1007/s11263-019-01228-7 arXiv:1610.02391

[22] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2015. Striving for simplicity: The all convolutional net. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings* (2015), 1–14. arXiv:1412.6806

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June (2015), 1–9. https://doi.org/10.1109/CVPR.2015.7298594 arXiv:1409.4842

[24] Erico Tjoa and Cuntai Guan. 2019. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. 1 (2019). arXiv:1907.07374 http://arxiv.org/abs/1907.07374

[25] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, Thibault Fevry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. 2020. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging* 39, 4 (2020), 1184–1194. https://doi.org/10.1109/TMI.2019.2945514 arXiv:1903.08297

[26] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Chao-Wei Zhao, and Ming-Ming Cheng. 2020. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. (2020), 1–11. arXiv:2004.07054 http://arxiv.org/abs/2004.07054

[27] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. 2018. Weighted Res-UNet for High-Quality Retina Vessel Segmentation. *Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018* (2018), 327–331. https://doi.org/10.1109/ITME.2018.00080

[28] Yesheng Xu, Ming Kong, Wenjia Xie, Runping Duan, Zhengqing Fang, Yuxiao Lin, Qiang Zhu, Siliang Tang, Fei Wu, and Yu-Feng Yao. 2020. Deep Sequential Feature Learning in Clinical Image Classification of Infectious Keratitis. *Engineering* xxxx (2020). https://doi.org/10.1016/j.eng.2020.04.012 arXiv:2006.02666

[29] Zitao Zeng, Weihao Xie, Yunzhe Zhang, and Yao Lu. 2019. RIC-Unet: An Improved Neural Network Based on Unet for Nuclei Segmentation in Histology Images. *IEEE Access* 7 (2019), 21420–21428. https://doi.org/10.1109/ACCESS.2019.2896920

[30] Kai-gang Zhang, Yue-dong Zhang, and Mei Wang. 2012. A Unified Approach to Interpreting Model Predictions Scott. *Nips* 16, 3 (2012), 426–430. arXiv:1705.07874

[31] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), 2921–2929. https://doi.org/10.1109/CVPR.2016.319 arXiv:1512.04150

[32] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11212 LNCS (2018), 122–138. https://doi.org/10.1007/978-3-030-01237-3_8

[33] Xiangrong Zhou, Ryosuke Takayama, Song Wang, Takeshi Hara, and Hiroshi Fujita. 2017. Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. *Medical Physics* 44, 10 (2017), 5221–5233. https://doi.org/10.1002/mp.12480

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11045 LNCS (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1 arXiv:arXiv:1807.10165v1