

# ST-MetaDiagnosis: Meta learning with Spatial Transform for rare skin disease Diagnosis

Delong Zhang

College of Computer Science and  
Engineering, Key Laboratory of  
Intelligent Computing in Medical  
Image, Ministry of Education  
Northeastern University Shenyang,  
China  
20184683@stu.neu.edu.cn

Mengqun Jin

College of Computer Science and  
Engineering, Key Laboratory of  
Intelligent Computing in Medical  
Image, Ministry of Education  
Northeastern University Shenyang,  
China  
20184528@stu.neu.edu.cn

Peng Cao\*

College of Computer Science and  
Engineering, Key Laboratory of  
Intelligent Computing in Medical  
Image, Ministry of Education  
Northeastern University Shenyang,  
China  
caopeng@cse.neu.edu.cn

**Abstract**—Skin conditions affect 1.9 billion people. Because of a shortage of dermatologists, most cases are seen instead by general practitioners with lower diagnostic accuracy. Current skin disease researches adopt the auto-classification system for improving the accuracy rate of skin disease classification. It is therefore an important task to develop Computer Aided Detection (CAD) systems that can aid/enhance dermatologists workflow and improve the classification performances. However, the long-tailed class distribution in the database and the limitation of ability to achieve a spatially invariant features make this problem challenging. We propose a ST-MetaDiagnosis, which utilizes meta-learning and spatial transform learning to facilitate quick adaptation and generalization of deep neural networks trained on the common diseases data for identification of rare diseases with much less annotated data. In particular, in order to predict the target risk where there are limited data samples, we train a meta-learner with spatial transforming from a set of related risk prediction tasks which learns how a good predictor is learned. The meta-learned can be directly used in target risk prediction, and the limited available samples can be used for further fine-tuning the model performance. Experiments on the recent ISIC 2018 skin lesion classification dataset show that our ST-MetaDiagnosis obtains 64.6% (accuracy) and 64.4% (F1-score) on the diagnosis of actinic keratosis, vascular lesion and dermatofibroma, demonstrating that ST-MetaDiagnosis can improve performance for predicting target risk with low resources comparing with the predictor trained on the limited samples available for this risk.

**Keywords**—few-shot learning, skin disease, computer aided diagnosis, meta learning, spatial transforming

## I. INTRODUCTION

Skin diseases remain a major cause of disability worldwide and contribute approximately 1.79% of the global burden of disease measured in disability-adjusted life years. It is the fourth leading cause of nonfatal disease burden globally, affecting 30-70% of individuals and prevalent in all geographies and age groups<sup>[1,2]</sup>. Furthermore, skin diseases may be cancerous, inflammatory or infectious and affect people of all ages, especially the elderly and young children. There are severe consequences of skin diseases such as death (in the case of melanoma), impairment of daily activities, loss of relationships, and damage to internal organs. Moreover, they also pose a real threat of mental illness leading to isolation, depression and even suicide. To decrease the associated consequences, cost, mortality and morbidity rate, skin diseases should be treated in their initial stages. However, dermatologists are consistently in short supply, particularly in rural areas, and consultation costs are rising.

Computer-aided diagnosis is crucial to improve treatment strategies for the skin diseases<sup>[3,4]</sup>. Many researchers have done investigations to broaden the availability of dermatology expertise with machine learning techniques. Recent advances in machine learning have facilitated the development of computer-aided diagnosis tools to assist in diagnosing skin disorders from images and obtain comparable performance with experienced dermatologists<sup>[5,6]</sup>. However, due to the intrinsic visual similarity between different types of skin lesions, it is difficult to distinguish different types of skin lesions even for the dermatologists. The general pipeline for existing computer-aided diagnosis methods follows three steps: preprocessing, feature extraction, and classification. The image features play a key role in the skin lesion classification task, and many conventional methods with hand-crafted features (colors, textures, shapes, etc.) as inputs have been proposed. Unfortunately, hand-crafted features have limited discriminative power, and they perform poorly when dealing with complex problems. Deep learning has proven to be successful in a multitude of computer vision tasks ranging from object recognition and detection to semantic segmentation<sup>[7-10]</sup>. Motivated by these successes, more recently, deep learning has been increasingly used in medical applications. With the availability of data, DL can lead to the assistance and automation of preliminary diagnoses which are of tremendous significance in the medical community. Despite the current research achievement, skin lesion classification is still a challenging task due to the following reasons:

- (1) Deep networks require a large amount of accurately annotated corresponding training data. Annotated images for diagnosis of rare or novel diseases are likely to remain scarce due to small affected patient population and limited clinical expertise to annotate images. Further, in case of the frequently occurring long-tailed class distributions in skin lesion and other disease classification datasets, conventional training approaches lead to poor generalization on classes at the tail end of the distribution due to biased class priors.
- (2) The automated skin diseases diagnosis is challenging due to the high variance in appearance and shape of the targeting lesions. The current deep learning methods, such as Convolutional Neural Networks, are still limited by the lack of ability to be spatially invariant to the input image in a computationally and parameter efficient manner. Moreover, we are also interested in the spatial localization without any annotation about the artifact positions of lesions.

These issues create an opportunity for incorporating machine learning systems into the doctor's workflow, aiding them in sieving through possible skin conditions. In order to solve the issues above, we propose a spatial transforming meta

Disease Diagnosis in this paper, we formulate the problem of disease identification from skin lesion images in low-data regimes as a few-shot learning problem by utilizing recent meta learning techniques. Hence, to facilitate learning from small amounts of annotated data, meta-learning<sup>[11,12]</sup> techniques have emerged. These techniques imbibe the system with the capability to rapidly adapt to new tasks and environments with very few training examples. On the other hand, spatial transformer networks (STNs)<sup>[13]</sup> are incorporated into the meta learning framework to enable convolutional neural networks (CNNs) to learn invariance to image transformations. The component of STNs were originally proposed to transform CNN feature maps as well as input images. This enables the use of more complex features when predicting transformation parameters. To the best of our knowledge, we are the first to embed the module of STNs into the procedure of meta learning. Our ST-MetaDiagnosis aims to improve the classification performance by meta-learning how to localization of attentional regions. Extensive experiments on ISIC 2018 Skin Lesion Dataset<sup>[14]</sup>, showing the effectiveness of the proposed framework. The experimental results demonstrate the potential of our MetaDiagnosis for real clinical practice.

The structure of the paper is as follows: in Section 2 we introduce the proposed method ST-MetaDiagnosis. In Section 3, the experimental results on a real datasets are given, which show excellent performance of the proposed method in comparison with other competitive methods. Finally, Section 3 concludes the paper.

## II. PROPOSED APPOROCH

We aim to train models that achieve fast adaptation with meta-train data(common disease), and the network model can quickly be adapted via a few steps of gradient descent with only few meta-test data(rare disease) to handle the new tasks.

### A. Problem Setting

The objective of the paper is to identify diseases from skin lesion images. Fig. 1. shows class distribution in the skin lesion datasets of ISIC 2018. The distribution is generally heavy-tailed with some classes having very few samples. The classes towards head of the class distribution (common-diseases), shown in red, are taken as train classes and classes at the tail of the distribution (new / rare disease), shown in blue color, are chosen as predicted classes.

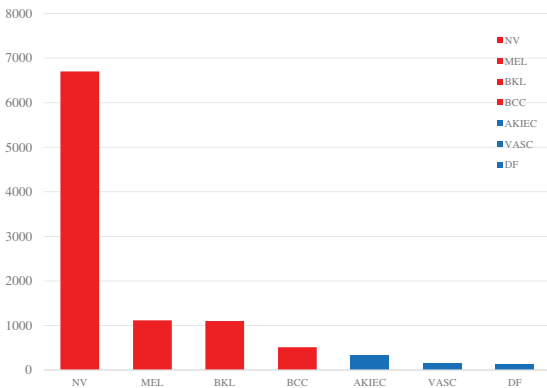


Fig. 1. The class distribution in skin lesion datasets

Generally, it is difficult to collect data for these rare diseases and obtain annotations from experienced physicians. This is due to the fact that new/rare diseases are being

discovered everyday which are difficult to annotate because of limited expertise. This limits the number of annotated samples for new / rare diseases as compared to those of common diseases. By low-resource, we mean that only limited skin disease can be used for the target prediction task, which is insufficient to train a good predictor by seen samples of the task themselves. Therefore, we formulate the problem of disease identification from skin lesion images in low-resource as a few-shot learning problem.

### B. Model-Angnostic Meta-learning Framework

Recently, transfer learning<sup>[15]</sup> has been demonstrated as an effective mechanism to achieve good performance in learning with limited samples in medical problem<sup>[16,17]</sup>. The transfer learning approaches requires the network to be pre-trained on a large amount of labelled data on a related domain and subsequently, fine-tuned on target domain data. However, these methods are successful only when sufficient labeled data is available in the target domain and do not guarantee optimal network initialization parameters that can quickly adapt to new target domains.

Meta-learning<sup>[18]</sup> is a recent trend in machine learning aiming at learning to learn from this experience, i.e., meta-data, to learn new tasks much faster. In meta-learning, the goal of the trained model is to quickly learn a new task from a small amount of new data, and the model is trained by the meta-learner to be able to learn on a large number of different tasks. The key idea underlying our method is to train the model's initial parameters such that the model has maximal performance on a new task after the parameters have been updated through one or more gradient steps computed with a small amount of data from that new task.

Formally, we consider a model represented by a parametrized function  $\delta_\theta$  with parameters  $\theta$ . When facing a new task, the model's parameter  $\theta$  is updated to  $\theta'_i$ . In this process, the updated parameter vector  $\theta'_i$  is computed using one or more gradient descent updates on task  $T_i$ . For example, when using one gradient update,

$$\theta'_i = \theta - \alpha \nabla_\theta L_{T_i}(\delta_\theta) \quad (1)$$

The step size  $\alpha$  may be fixed as a hyperparameter. For simplicity of notation, we will consider one gradient update for the rest of this section, but using multiple gradient updates is a straightforward extension. The model parameters are trained by optimizing for the performance of  $\delta_{\theta'_i}$ , with respect to  $\theta$  across tasks sampled from  $p(T)$ . More concretely, the meta-objective is as follows:

$$\min_{\theta} \sum_{T_i \sim p(T)} L_{T_i}(\delta_{\theta'_i}) = \sum_{T_i \sim p(T)} L_{T_i}(\delta_{\theta - \alpha \nabla_\theta L_{T_i}(\delta_\theta)}) \quad (2)$$

Note that the meta-optimization is performed over the model parameters  $\theta$ , where as the objective is computed using the updated model parameters  $\theta_0$ . In effect, our proposed method aims to optimize the model parameters such that one or a small number of gradient steps on a new task will produce maximally effective behavior on that task. The meta-optimization across tasks is performed via adaptive moment estimation, such that the model parameters  $\theta$  are updated as follows:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} L_{T_i}(\delta_{\theta'_i}) \quad (3)$$

where  $\beta$  is the meta step size. The full algorithm, in the general case, is outlined in Algorithm 1.

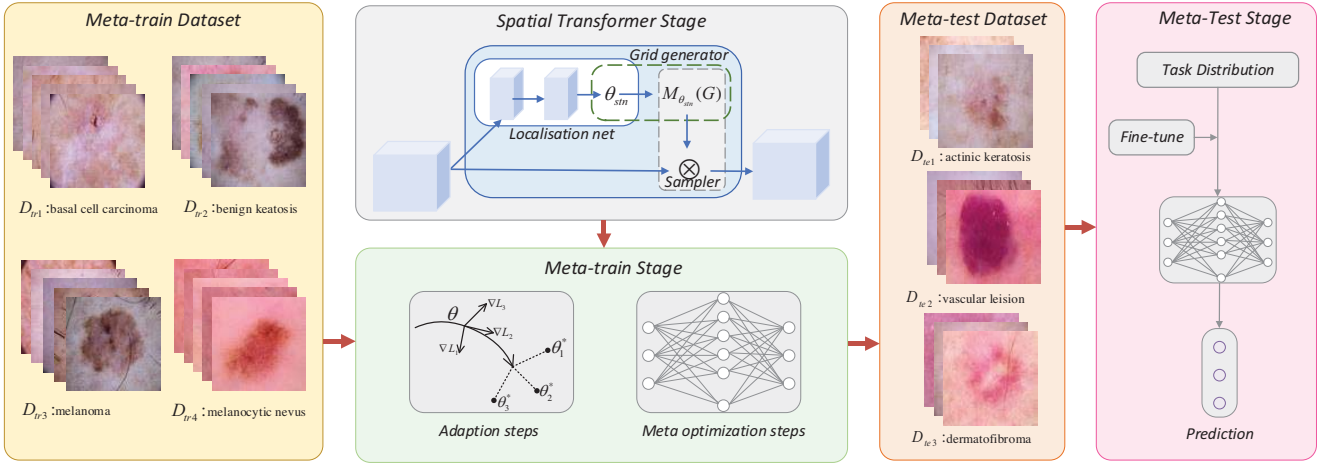


Fig. 2. The architecture of our proposed ST-MetaDiagnosis framework. A spatial transformer is composed of the localisation network and sampling mechanism. It can be flexibly inserted into every layer of the MAML network. We train meta-classifier on the meta-train dataset, feeding only a few samples. The representation  $\theta$  can adapt to new tasks rapidly, via only a few steps of gradient descent, and shows high accuracy.

#### Algorithm1 MAML

**Require:**  $p(T)$ : distribution over tasks  
**Require:**  $\alpha, \beta$ : step size hyperparameters  
1: randomly initialize  $\omega$   
2: **while** not done **do**  
3: Sample batch of tasks  $T_i \sim p(T)$   
4: **for all**  $T_i$  **do**  
5: Evaluate  $\nabla_{\theta} L_{T_i}(\delta_{\theta})$  with respect to  $k$  examples  
6: Compute adapted parameters with gradient descent:  $\theta'_i = \theta - \alpha \nabla_{\theta} L_{T_i}(\delta_{\theta})$   
7: **end for**  
8: Update  $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i \sim p(T)} L_{T_i}(\delta_{\theta'_i})$   
9: **end while**

Similarly, for discrete classification tasks with a cross entropy loss, the loss  $L_{T_i}(\delta_{\theta})$  takes the form:

$$\sum_{x_j^m, y_j^m \sim T_i} y_j^m \log \delta_{\theta}(x_j^m) + (1 - y_j^m) \log (1 - \delta_{\theta}(x_j^m)) \quad (4)$$

The MAML meta-gradient updates involves a gradient through a gradient. This requires an additional backward pass through  $\delta$  to compute Hessian-vector products, which is supported by standard deep learning libraries.

#### C. Spatial Transform Network

In this section we describe the formulation of a spatial transformer. It is a differentiable module that applies a learnable affine transformation to an input image, or more generally, to a feature map based on the input itself. It is composed of three parts, shown in Fig. 3.

First a localization network takes the input feature map  $U \in \mathbb{R}^{H \times W \times C}$  with width  $W$ , height  $H$  and  $C$  channels and outputs  $\theta_{stn}$ , the parameters of the transformation  $M_{\theta_{stn}}$  to be applied to the feature map:  $\theta_{stn} = f_{loc}(U)$ . It can take any form, such as fully-connected network or a convolutional network, but should include a final regression layer to produce the transformation parameters  $\theta_{stn}$ . The size of  $\theta_{stn}$  can vary depending on the transformation type that is parameterized. For an affine transformation  $\theta_{stn}$  is 6-dimensional.

Then, a grid generator selects a set of points of the input map, to produce the transformed output, using the predicted transformation parameters. To perform a warping of the input feature map, each output pixel is computed by applying a

sampling kernel centered at a particular location in the input feature map. In the affine case, the pointwise transformation is:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = M_{\theta_{stn}}(G_i) = A_{\theta_{stn}} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (5)$$

where  $(x_i^s, y_i^s)$  are the source coordinates in the input feature map that define the sample points,  $(x_i^t, y_i^t)$  are the target coordinates of the regular grid in the output feature map, and  $A_{\theta_{stn}}$  is the affine transformation matrix.

Finally, the feature map and the sampling grid are taken as inputs to the sampler, producing the output map sampled from the input at the grid points. The combination of these three components forms a spatial transformer. Placing spatial transformers within a CNN allows the network to learn how to actively transform the feature maps to help minimize the overall cost function of the network during training.

#### D. ST-MetaDiagnosis

For this scenario, we develop a model agnostic gradient descent framework with spatial transforming to train a meta-learner on a set of prediction tasks where the target skin disease prediction tasks are highly relevant. Fig. 2. shows the overall framework of the proposed ST-MetaDiagnosis.

We are interested in learning a mapping  $\phi: X \rightarrow Y$ , which given an image predicts the associated disease label. We model as the composition of two functions  $\phi = \phi_{cnn} \circ \phi_{stn}$  where  $\phi_{stn}: X \rightarrow \hat{X}$  estimates an affine transformation and applies it to the input image  $X$ , and  $\phi_{cnn}: \hat{X} \rightarrow Y$  assigns the prediction label to the transformed image. Intuitively, ST learns to localize regions of interest most salient in the input image. Formally, the ST layer extracts feature of an attentional region, from the feature maps of the whole input image. The procedure of image transform of ST module is illustrated in Fig. 3. The meta-training stage consists of a meta-learner for training the neural network to solve a large number of few-shot image classification tasks created from a set of training classes comprising of common diseases, with the classes being sampled from the head of the distribution, and finding effective network initialization parameters for the model. The initialization parameters involve two parts:

$\theta_{cnn}$  and  $\theta_{stn}$ . Both parameters are randomly initialized.

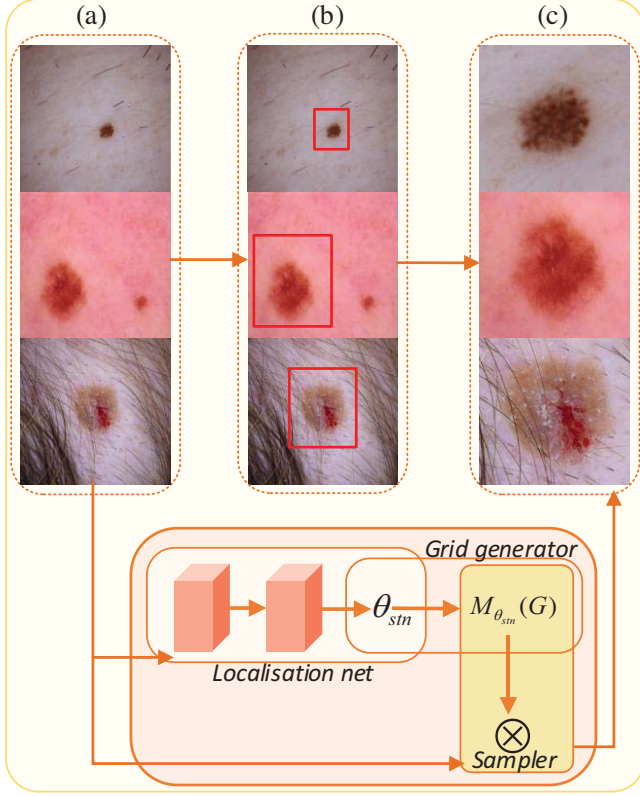


Fig. 3. An overview of using a spatial transformer as the first layer of MAML trained for Melanoma Detection Dataset classification. (a) The input to the STN is a skin disease image that is original without attention mechanism. (b) The localisation network of the ST predicts a transformation to apply to the input image. (c) The output of the ST, after applying the transformation, is feed into the MAML as an input.

---

#### Algorithm2 ST-MetaDiagnosis Training

---

**Require:**  $p(T)$ : distribution over tasks from common disease datasets  
**Require:**  $\alpha, \beta$ : step size hyperparameters  
1: randomly initialize  $\theta$ , including  $\theta_{cnn}$  and  $\theta_{stn}$   
2: **while** Outer-Loop not done **do**  
3: Sample batch of tasks  $T_i \sim p(T)$   
4: **while** Inner-Loop not done **do**  
5: Compute  $\nabla_{\theta} L_{T_i}(\delta_{\theta_{cnn}}, M_{\theta_{stn}}(G))$  with respect to  $k$  examples  
6: Parameters fast adaption with gradient descent:  $\theta'_i = \theta - \alpha \nabla_{\theta} L_{T_i}(\delta_{\theta_{cnn}}, M_{\theta_{stn}}(G))$   
7: **end for**  
8: Update  $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i \sim p(T)} L_{T_i}(\delta_{\theta'_i})$   
9: **end while**

---



---

#### Algorithm3 ST-MetaDiagnosis Testing

---

**Require:**  $p(T)$ : distribution over tasks from rare disease datasets  
**Require:**  $\theta$ : learned parameter  
1: Compute  $\nabla_{\theta} L_{T_i}(\delta_{\theta_{cnn}}, M_{\theta_{stn}}(G))$  with respect to  $k$  examples  
2: Parameters fast adaption with gradient descent:  $\theta'_i = \theta - \alpha \nabla_{\theta} L_{T_i}(\delta_{\theta_{cnn}}, M_{\theta_{stn}}(G))$   
3: Evaluate predicted results of Learner( $\{X_{T_i}, Y_{T_i}\}; \theta'_i$ )  
4: **end**

---

The training and testing of ST-MetaDiagnosis algorithm is shown below in Algorithm 2 and 3. In the meta-testing stage, the model is adapted to perform classification on a new set of unseen rare classes with very few examples.

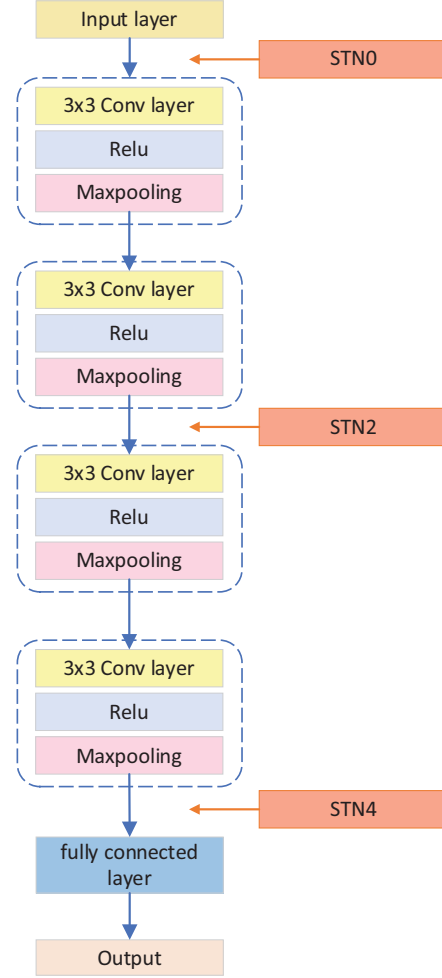


Fig. 4. ST-MetaDiagnosis network with multiple ST modules incorporated into different layers.

### III. EXPERIMENT

In this section, we introduce the dataset and conduct sufficient experiments to evaluate the quantitative performance improvement on both the multi-class and binary-class classification for the rare skin disease.

#### A. Dataset

We applied our method on the ISIC 2018 Skin Lesion Analysis Towards Melanoma Detection Dataset, with a total of 10,015 skin lesion images from seven skin diseases, including melanocytic nevus (6705), melanoma (1113), benign keratosis (1099), basal cell carcinoma (514), actinic keratosis (327), vascular lesion (142) and dermatofibroma (115). We utilize the 4 classes with largest amount of samples as common diseases (i.e., meta-train dataset  $D_{tr}$ ) and the left 3 classes as the rare diseases (i.e., meta-test dataset  $D_{te}$ ) to simulate the problem. Task instance  $T_i$  is randomly sampled from distribution over tasks  $p(T)$  and  $D_{tr}, D_{te} \in p(T)$ . During meta-train stage, learning task  $T_i$  are multi-class classification tasks and each task consists of 4 random classes with  $k$  samples per class in  $D_{tr}$ . Learning task  $T_i$  are multi-class classification tasks. It is noted that during meta-train stage, each task consists of 4 random classes with  $k$  samples

per class in  $D_{tr}$ , but sampling from  $D_{te}$ , each test task instance consists of 3 random classes with  $k$  samples per class.

### B. Comparison on the multi-class classification

In this section we display the baseline meta training details and explore the spatial transformer networks on MAML. Each skin image was processed into a size of  $84 \times 84$ . We employed 4 conv blocks as the backbone architecture and used Adam optimizer with a meta-learning rate of 0.001 and divide by 2 for every 10 epochs without increasing of validation accuracy. The STN and CNN are jointly trained. During training stage, we totally prepared 10000 iterations and stop training if meta-lr lower than  $1e-6$ . The batch size is 4, consisting of 4 tasks sampled from meta-train dataset. Each task consists of randomly  $k$  samples from 4 classes. We query 15 images from each of 4 classes to adapt parameters for  $T_i$ . During meta-test stage, the inference is performed by randomly sampling  $k$  samples from 3 classes from meta-test dataset, *i.e.*,  $D_{te}$ . The final report results is the accuracy, macro recall and macro f1-score over 5 runs. In our experiment, we incorporate STN module to different Conv Blocks to validate performance of ST-MetaDiagnosis.

TABLE I. EXPERIMENT RESULTS OF 3WAYS1SHOT

settings	network structure	ACC	Recall	F1
1shot	Finetune + Aug	0.4089	0.4015	0.4015
	MAML	0.4436	0.4462	0.4462
	ST-MetaDiagnosis-0	0.4806	0.4551	0.4551
	<b>ST-MetaDiagnosis-2</b>	<b>0.4810</b>	0.4562	0.4561
	<b>ST-MetaDiagnosis-4</b>	0.4788	<b>0.4597</b>	<b>0.4595</b>
	ST-MetaDiagnosis-02	0.4415	0.4412	0.4410
	ST-MetaDiagnosis-34	0.4453	0.4452	0.4452
	ST-MetaDiagnosis-024	0.4402	0.444	0.4445

TABLE II. EXPERIMENT RESULTS OF 3WAYS3SHOT

settings	network structure	ACC	Recall	F1
3shot	Finetune + Aug	0.5321	0.5324	0.5325
	MAML	0.5768	0.5757	0.5758
	ST-MetaDiagnosis-0	0.5832	0.5821	0.5821
	<b>ST-MetaDiagnosis-2</b>	<b>0.5979</b>	<b>0.5957</b>	<b>0.5955</b>
	ST-MetaDiagnosis-4	0.5887	0.5901	0.5904
	ST-MetaDiagnosis-02	0.5601	0.5603	0.5603
	ST-MetaDiagnosis-34	0.5672	0.567	0.5669
	Finetune + Aug	0.5321	0.5324	0.5325

TABLE III. EXPERIMENT RESULTS OF 3WAYS5SHOT

settings	network structure	ACC	Recall	F1
5shot	Finetune + Aug	0.5832	0.5831	0.5831
	MAML	0.6038	0.6016	0.6017
	ST-MetaDiagnosis-0	0.6220	0.6228	0.6215
	ST-MetaDiagnosis-2	0.6152	0.6128	0.6129
	<b>ST-MetaDiagnosis-4</b>	<b>0.6459</b>	<b>0.6438</b>	<b>0.6438</b>
	ST-MetaDiagnosis-02	0.6039	0.6010	0.6010
	ST-MetaDiagnosis-34	0.6203	0.6207	0.6207
	ST-MetaDiagnosis-024	0.6303	0.6293	0.6291

To demonstrate the overall performance of the proposed ST-MetaDiagnosis framework on the extremely low-data resource, we first compare our method with some strong

baselines, *i.e.*, fine-tuning, traditional MAML. We incorporated STN to the input layer, the fourth convolution layers and the fully-connected layer, respectively. The results of these experiments are shown in TABLE I, TABLE II, TABLE III. Fig. 5 shows the convergence of the comparable methods on the same validation dataset. For more detailed comparative purposes, we enclose the corresponding confusion matrix in Fig. 6, as well.

It can be seen that the proposed ST-MetaDiagnosis with only one ST module consistently achieved better or comparable performance than the fine-tuning method and the traditional MAML method, suggesting that effectiveness of ST-MetaDiagnosis for the few-shot learning on the skin disease diagnosis. Specifically, ST-MetaDiagnosis-2 achieves the best classification performance with an accuracy of 48.1%, 59.79 % and 64.59%, resulting in a 3.74%, 2.11% and 1.14 % increase in accuracy compared with the previous MAML model on the 3ways1shot, 3ways3shot and 3ways5shot, respectively. ST-MetaDiagnosis-4 achieves the best classification performance with an accuracy of 47.88 %, 58.87 % and 64.59 %, resulting in a 3.52%, 1.19% and 4.21 % increase in accuracy compared with the previous MAML model on the 3ways1shot, 3ways1shot and 3ways1shot, respectively.

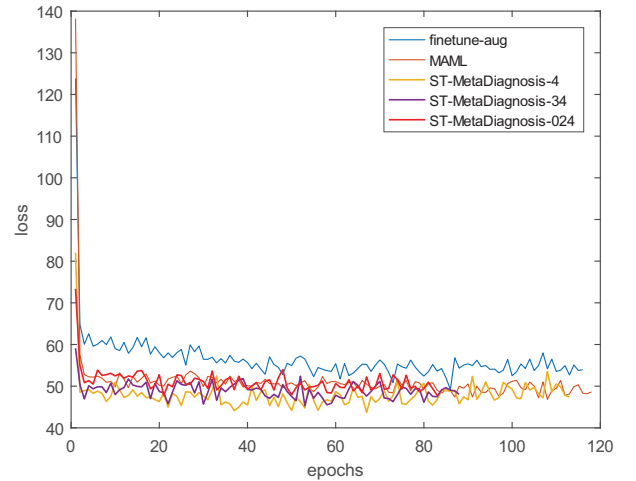


Fig. 5. The convergence of baseline methods and our ST-MetaDiagnosis network with different ST module.

An interesting observation is that the more ST modules do not achieves a better performance for few-shot learning. The reason might be that more ST layers lead to more parameters to be meta-learned. However, for the MAML model, each task is typically modeled by a low complexity base learner (such as a shallow neural network) and being unable to use deeper and more powerful architectures. More parameters result in overfitting. Although more ST layers do not further improve the performance of ST-MetaDiagnosis with only one ST layer, it still shows an increase in performance on the 3ways5shot compared with MAML. Furthermore, we also explore how the location of STN incorporated into different convolution layers affects the model performance. We embed STN module into different locations of Conv Blocks and find that our method can markedly improve the overall performance of the few-shot learning. The experiment demonstrates that STN incorporated higher Conv Blocks layer achieve a better performance. The result indicates that

the spatial transformer is appropriate for the higher-level feature map with larger receptive fields.

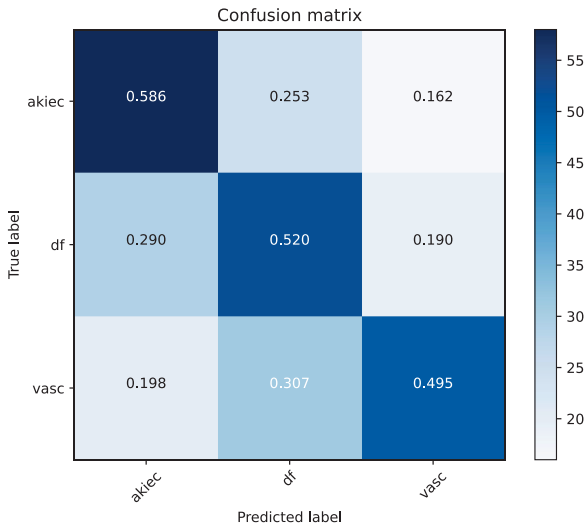


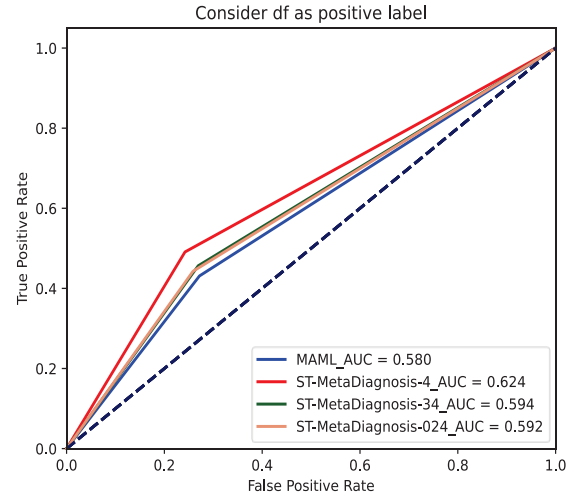
Fig. 6. The confusion matrix of ST-MetaDiagnosis for predicting rare skin disaster in multi-classification tasks.

### C. Comparison on the binary-class classification

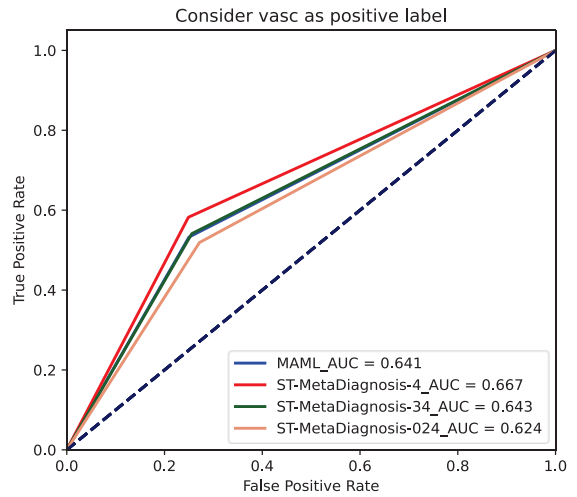
In addition to multi-classification tasks of the rare diseases, we also evaluate the performance on the binary-class classification by one-vs-the-rest scheme. We query 15 images from the meta-train dataset for each of the classes in a task during the meta-training stage. During meta-testing,  $k$  images are sampled from the training split of each class involved in the meta-test task. The value of  $k$  in our experiments is 1, 3, and 5 indicating 1-shot, 3-shot, and 5-shot, respectively. These images are used for fine-tuning the model obtained as a result of meta-training. The final inference is performed on the entire testing split of the classes in the meta-test task to compute the accuracy and AUC values. TABLE IV shows the comparable performance of the traditional MAML and our ST-MetaDiagnosis. We also show the ROC analysis of the various approaches in Fig. 7. Experimental results in TABLE IV and Fig. 7 demonstrate that our method can consistently and substantially outperform the traditional MAML method, confirming the importance of ST module in meta learning for rare diseases diagnosis.

TABLE IV. EXPERIMENT RESULTS 3WAYS1SHOT

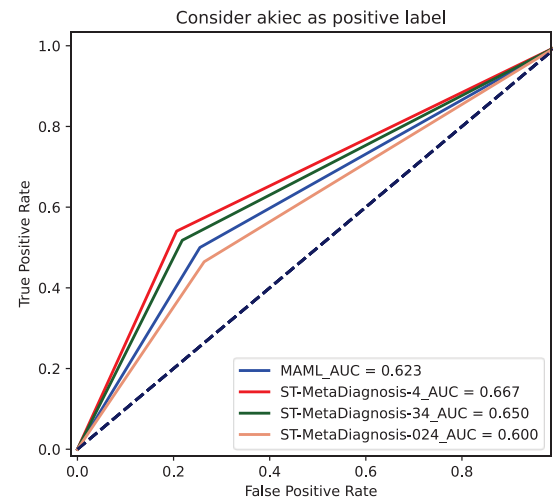
settings	network structure	Avg.AUC	Avg.ACC
1-shot	MAML	0.6174	0.6016
	<b>ST-MetaDiagnosis-4</b>	<b>0.6527</b>	<b>0.6578</b>
	ST-MetaDiagnosis-34	0.6291	0.6407
	ST-MetaDiagnosis-024	0.5843	0.6193
3-shot	MAML	0.7338	0.7456
	<b>ST-MetaDiagnosis-4</b>	<b>0.7632</b>	<b>0.7638</b>
	ST-MetaDiagnosis-34	0.7403	0.7407
	ST-MetaDiagnosis-024	0.7203	0.7393
5-shot	MAML	0.7829	0.7916
	<b>ST-MetaDiagnosis-4</b>	<b>0.8059</b>	<b>0.8138</b>
	ST-MetaDiagnosis-34	0.7933	0.7907
	ST-MetaDiagnosis-024	0.7903	0.7993



(a) one shot learning for comparable methods



(b) 3-shot learning for comparable methods



(c) 5-shot learning for comparable methods

Fig. 7. The ROC curves of MAML and ST-MetaDiagnosis

#### D. Comparison with the multi-instance learning

Due to the intrinsic visual similarity between different types of skin lesions, it is difficult to distinguish different types of skin lesions from the whole image even for the dermatologists. Therefore, it is inappropriate for learn the whole-image level features by regarding an image as a whole instance without considering local structures within the images. As an alternative solution, it is noteworthy that the appearance of lesion is often locally different. Different from ST-MetaDiagnosis, a local feature learning method for skin lesion classification task is proposed in this experiment for comparison. To solve the problem of supervised learning in the diagnosis of skin disease, we formulate multi-class skin disease diagnosis as a multi-class multi-instance problem where each image (bag) is labeled as one of three disease labels and consists of unlabeled candidate lesion regions (instances). On the other hand, we incorporate an attention mechanism<sup>[19]</sup> into a deep learning-based MIL network to identify the highly suspicious instances and improve the overall classification performance. The proposed attentional multi-instance learning for the skin disease diagnosis is named AMILDiagnosis. The framework is illustrated in Fig. 8. In the AMILDiagnosis framework, the origin images are segmented into multiple patches. Given training dataset  $\{(X_1, y_1), (X_2, y_2), \dots, (X_{N_t}, y_{N_t})\}$ , where  $X_i$  is a bag,  $y_i$  is the bag-level label, and  $N_t$  means the total number of training samples. A bag is composed of multiple instances, namely  $X_i = \{x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n_i)}\}$ , where  $n_i$  is the number of instances of  $X_i$ , and each instance has no label. Fig. 9 shows the network structure of multi-instance multi-classification and attention mechanism in AMILDiagnosis. With the embedding  $h_i^{(j)}$  of each remained instance in  $X_i$  learned by instance discriminator, the attention weight of each instance is calculated as follow:

$$a_i^{(j)} = \frac{\exp\left\{\omega^T \left( \tanh\left(v(h_i^{(j)})^T\right) \odot \text{sigm}\left(u(h_i^{(j)})^T\right) \right)\right\}}{\sum_{k=1}^{n_i} \exp\left\{\omega^T \left( \tanh\left(v(h_i^{(k)})^T\right) \odot \text{sigm}\left(u(h_i^{(k)})^T\right) \right)\right\}} \quad (6)$$

where  $\omega \in \mathbb{R}^{L \times 1}$ ,  $U \in \mathbb{R}^{L \times M}$ ,  $V \in \mathbb{R}^{L \times M}$  are parameters,  $\odot$  is an element-wise multiplication and  $\text{sigm}(\cdot)$  is the sigmoid non-linearity. Since  $\tanh(\cdot)$  non-linearity may not be effective in learning complex relationships, it is proposed to use the gating mechanism<sup>[20]</sup> together with the  $\tanh(\cdot)$  non-linearity to eliminate the troublesome linearity in  $\tanh(\cdot)$ . With the attention weight, the bag-level mapping relationship is composed of weighted instances, expressed by:

$$z_i = [a_i^{(1)} h_i^{(1)} \quad a_i^{(2)} h_i^{(2)} \quad a_i^{(3)} h_i^{(3)} \dots a_i^{(n_i)} h_i^{(n_i)}] \quad (7)$$

Let  $N_{max} = \max_{i=1 \dots N_t} n_i$  be the largest number of all the training bags. The weighted 2D instance-level is expanded to a tensor bag-level representation  $z_i \in \mathbb{R}^{N_{max} \times L \times P}$  by stacking multiple instances embedding, where  $P$  is the dimensionality of instance embedding. Finally, the bag-level prediction of the  $L \times 1$  dimension is obtained with tensor bag-level representation  $z_i$  by a FC layer, combined with a softmax activation function.

A comparison of skin disease diagnosis results between the ST-MetaDiagnosis and AMILDiagnosis is shown in Table V. It can be seen that the proposed MIL model achieves a better classification performance in terms of ACC, AUC and F1. This result suggests that the local representation learning is critical for the few shot learning with the complicated disease diagnosis. In the future, we will incorporate the multi-instance learning into the our ST-MetaDiagnosis to further improve the classification performance of few shot learning.

TABLE V. EXPERIMENT RESULTS OF COMPARISON WITH MIL

	ACC	AUC	Recall	F1
<b>ST-MetaDiagnosis</b>	0.6459	0.8059	<b>0.6438</b>	0.6438
<b>AMILDiagnosis</b>	<b>0.7361</b>	<b>0.8512</b>	0.6367	<b>0.6450</b>

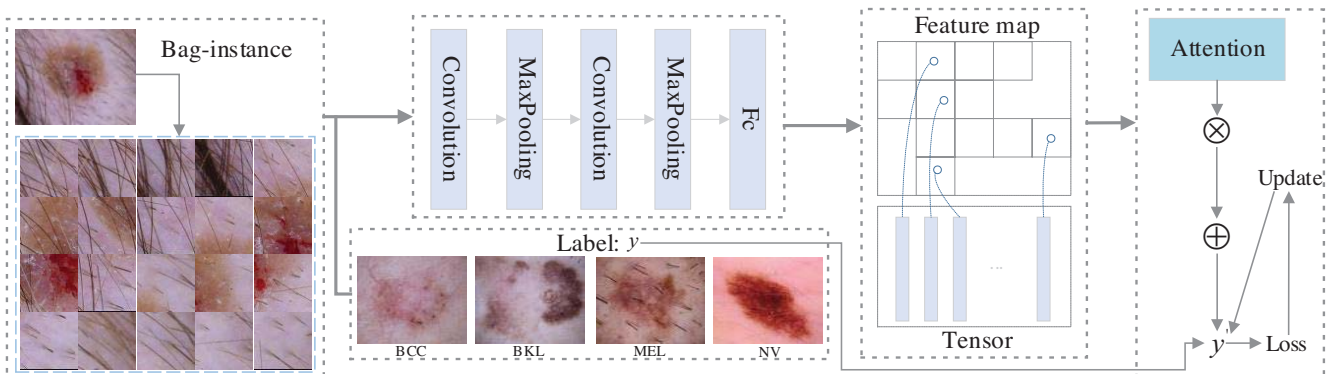


Fig. 8. An overview of the AMIL framework for skin disease

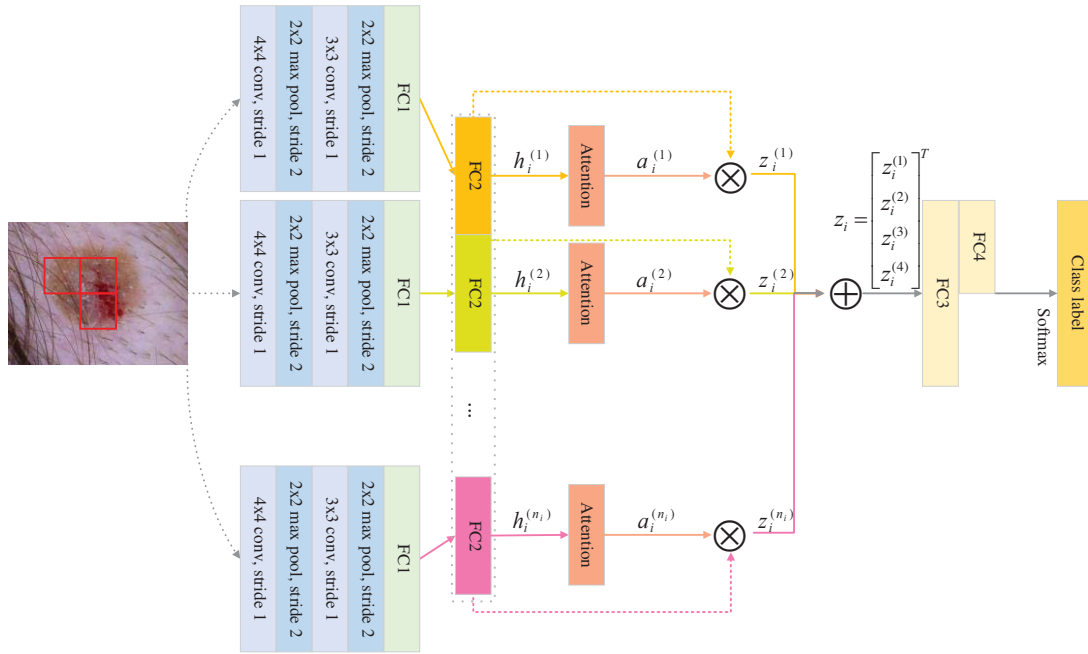


Fig. 9. The network structure of AMIL framework

#### ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (No.62076059) and the Fundamental Research Funds for the Central Universities (No. N2016001).

#### CONCLUSION

Skin disease is one of the most common human illnesses that affects 30% to 70% of individuals, with even higher rates in at-risk subpopulations where access to care is scarce. We consider the problem of clinical image classification for the purpose of aiding doctors in dermatological disease diagnosis. Diagnosis of dermatological conditions from images poses two major challenges for standard off-the-shelf techniques: First, the distribution of real-world dermatological datasets is typically long-tailed. Second, the lack of ability to learn a spatially invariant feature. Our network leverages Spatial Transformers Network (STN) and meta-learning to improve the skin diseases diagnosis performance. Extensive experiments demonstrated the advantage our method. We also claim that the proposed meta-learning based disease identification system can also be applied on other medical imaging datasets in future work.

#### REFERENCES

- [1] Roderick James Hay DM FRCP, and Lucinda Claire Fuller BM FRCP. The assessment of dermatological needs in resource-poor regions. *International Journal of Dermatology*. 2011.
- [2] Hay, R. J. , et al. "The global challenge for skin health." *British Journal of Dermatology* 172.6(2015):1469–1472.
- [3] Skin, Diagnosing, and Using, Diseases. "Diagnosing skin diseases using an artificial neural network." *International Conference on Adaptive Science & Technology* IEEE, 2010.
- [4] Chen, Min , et al. "AI-Skin : Skin Disease Recognition based on Self-learning and Wide Data Collection through a Closed Loop Framework." (2019).
- [5] Esteva, Andre , et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639(2017):115–118.
- [6] Hameed, Nazia , et al. "Multi-Class Multi-Level Classification Algorithm for Skin Lesions Classification using Machine Learning Techniques." *Expert Systems with Applications* 141(2019):112961.
- [7] Rodrigues, Douglas De A. , et al. "A new approach for classification skin lesion based on transfer learning, deep learning, and IoT system." *Pattern Recognition Letters* (2020).
- [8] Thomsen, Kenneth, et al. "Deep Learning for Diagnostic Binary Classification of Multiple-Lesion Skin Diseases." *Frontiers in Medicine* 7 (2020): 604.
- [9] Harangi, Balazs , A. Baran , and A. Hajdu . "Assisted deep learning framework for multi-class skin lesion classification considering a binary classification support." *Biomedical Signal Processing and Control* 62(2020):102041.
- [10] Milton, Md Ashraful Alam. "Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge." *arXiv preprint arXiv:1901.10802* (2019).
- [11] Hospedales, Timothy , et al. "Meta-Learning in Neural Networks: A Survey." *arXiv* (2020).
- [12] Munkhdalai, Tsendsuren, and Hong Yu. "Meta networks." *Proceedings of machine learning research* 70 (2017): 2554.
- [13] Jaderberg, Max , et al. "Spatial Transformer Networks." (2015).
- [14] Codella, Noel CF, et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)." 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 2018.
- [15] Pan, Sinno Jialin , and Q. Yang . "A Survey on Transfer Learning." *IEEE Transactions on Knowledge & Data Engineering* 22.10(2010):1345–1359.
- [16] Jia, Jinmeng , et al. "RDAD: A Machine Learning System to Support Phenotype-Based Rare Disease Diagnosis." *Frontiers in Genetics* 9(2018).
- [17] Prabhu, Viraj Uday. "Few-shot learning for dermatological disease diagnosis." PhD diss., Georgia Institute of Technology, 2019.
- [18] Finn, Chelsea , P. Abbeel , and S. Levine . "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks." (2017)
- [19] Ruoxian Song, Peng Cao, Jinzhu Yang, Dazhe Zhao, Osmar R. Zaiane, A Domain Adaptation Multi-instance Learning for Diabetic Retinopathy Grading on Retinal Images, IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2020), December 16-19, 2020
- [20] Dauphin, Yann N , et al. "Language Modeling with Gated Convolutional Networks." (2016).