IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2020, accepted August 13, 2020, date of publication September 1, 2020, date of current version September 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3020802

Utilizing Knowledge Distillation in Deep Learning for Classification of Chest X-Ray Abnormalities

THI KIEU KHANH HO[1] AND JEONGHWAN GWAK[2]

[1]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea
[2]Department of Software, Korea National University of Transportation, Chungju 27469, South Korea

Corresponding author: Jeonghwan Gwak (james.han.gwak@gmail.com)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2020R1I1A3074141, in part by the Brain Research Program through the NRF funded by the Ministry of Science, ICT and Future Planning under Grant NRF-2019M3C7A1020406, and in part by the Korea National University of Transportation in 2020.

ABSTRACT Automatic screening and diagnosis of lung abnormalities from chest X-ray images has been recently drawing attention from the computer vision and medical imaging communities. Previous studies of deep neural networks have predominantly demonstrated the effectiveness of lung disease binary classification procedures. However, large numbers of medical images—which can be labeled with a variety of existing or suspected pathologies—are required to be interpreted and reported upon daily by an individual radiologist; this poses a challenge in maintaining a consistently high diagnosis accuracy. In this paper, we present a competitive study of knowledge distillation (KD) in deep learning for classification of abnormalities in chest X-ray images. This method aims to either distill knowledge from cumbersome teacher models into lightweight student models or to self-train these student models, to generate weakly supervised multi-label lung disease classifications. Our approach was based on multi-task deep learning architectures that, in addition to multi-class classification, supported the visualizations utilized in saliency maps of the pathological regions where an abnormality was located. A self-training KD framework, in which the model learned from itself, was shown to outperform both the well-established baseline training procedure and the normal KD, achieving the AUC improvements of up to 6.39% and 3.89%, respectively. Through application to the publicly available ChestX-ray14 dataset, we demonstrated that our approach efficiently overcame the interdependency of 14 weakly annotated thorax diseases and facilitated the state-of-the-art classification compared with the current deep learning baselines.

INDEX TERMS ChestX-ray14, deep neural networks, knowledge distillation (KD), multi-class classification, saliency maps, self-training KD.

I. INTRODUCTION

With the potential to escalate simple thoracic ailments into cancers, lung diseases are one of the leading causes of death worldwide. Chest radiography is the most common medical imaging technique used to diagnose them, owing to its efficiency in the identification and detection of cardiothoracic, pulmonary, and interstitial diseases; it currently occupies a significant role in lung disease treatment practices [1]. Accurate analysis of the large quantities of patient health information represents a major challenge for radiologists because the timely reporting of potential findings is necessary for effective treatment. The overlapping of tissue structures in the X-ray images or the low contrast resolutions with which they need to distinguish the lesion and surrounding tissues greatly increase the complexity of interpretation. This results in a certain number of missed detections and diagnoses. The wide applicability and interpretational difficulties of chest X-ray images have led to the introduction of computer-aided detection (CAD) systems into medical imaging practices. CAD systems are predominantly divided into four steps: preprocessing, Region of Interest (ROI) segmentation, ROI feature extraction, and disease identification; however, there is a vital need for them to not only automatically process large numbers of medical images, but also enhance the certainty of accurate disease prediction.

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

The extensive and successful developments of artificial intelligence (AI), along with the accumulation of large numbers of medical images, have opened up the promising possibility of building a CAD system integrated AI techniques. In particular, deep learning-based methods have achieved remarkable performances in various image-recognition tasks, including image classification [2]–[5] and semantic segmentation [6]–[9]; these methods have been proposed for re-application on anatomical and pathological medical imaging domains. Advanced deep learning, in combination with the construction of large medical databases, has recently enabled these algorithms to surpass the performances of conventionally medical techniques, particularly in tasks such as pulmonary nodule detection [10], detection of lymph node metastases in breast cancer [11], cerebral micro-bleeding detection [12], [13], skin cancer classification [14], [15], pneumonia diagnoses from radiographs [16], diabetic retinopathy detection [17], cardiologist-level arrhythmia classification [18], and cerebral micro-bleeding identification [19].

Whilst the general trend for deep learning models such as convolutional neural networks (CNNs), which learn features in an end-to-end manner with respect to millions of parameters, is towards deeper, wider, and more complex architectures, the expensive computation costs limit the capability of a deep learning solution in real-world applications. That is, the weights are transferred from a pre-trained model to a new network with a need of matching the network architecture in case of transfer learning. This means that the new network should be as sophisticated as the old one. Hence, it is arduous to deploy a cumbersome model to many applications. For instance, self-driving vehicles and mobile robots have limited memory and power resources. Even when these are in abundant supply, for example when a data system is hosted in a network cloud, effective deep networks serving clients at a lower cost are still necessary. However, data privileges or privacy issues can restrict access to the source data domain in real transfer learning problems. Therefore, it is essential to transfer the knowledge of a network trained on the data by accessing only the training data of the target domain.

Therefore, a recent study [20] proposed a knowledge distillation (KD) procedure to capture and transfer the knowledge of a trained teacher model to a student model. Typically, a teacher network exhibits a greater learning capacity and higher performance and can be used to teach a lower-capacity student network by providing soft-targets. Dark knowledge describing the similarity of privileged information from different classes can be transferred from these soft-targets, to enhance the performance of the student model. This process guides the training of a student network, and further uses an additional distillation loss to encourage the student model to mimic some aspects of the teacher model. Originally motivated by resource-efficient neural network compression tasks [20], KD procedures have found a variety of applications in such areas as adversarial defense [21], privileged learning [22], and learning with noisy data [23]. To extend this idea of mimicking the softened class scores provided by the teacher model, Fit-Nets [24] added hints to guide the intermediate layers' training. Liu *et al.* [25] introduced a supervisory signal for KD in the form of spatial attention, by computing the sum of squared activations along the channel dimension; this intuitively encouraged the student model to produce similar normalized spatial attention maps to the teacher model. As expected, recent works have expanded the scope of KD, for example by using semi-supervised adaptive distillation for a learning-efficient detector [54], knowledge adaptation for segmenting sematic regions [55], and a teacher assistant (an intermediary between teachers and students) for KD improvements [56]. With these new findings from the deep learning community, it is of great interest and importance to find ways of exploiting KD performances in medical imaging fields.

Meanwhile, the manual marking of the pathologically abnormal areas of X-ray images, performed by expert radiologists, typically requires more effort than simply labeling them. In other words, the bounding boxes for disease localization tasks are much more descriptive and informative than a single class label. As a consequence, chest X-ray datasets such as ChestX-ray8 and ChestX-ray14 [26] have recently been published. These provide comprehensive disease labels along with a small subset of abnormal region annotations, which are suitable for weakly supervised learning problems. Therefore, designing models for such small numbers of annotated masks is a crucial step toward clinical applications. Many attention-based mechanisms have recently been developed and have demonstrated the feasibility of the localization and recognition of multiple objects, in spite of using only simple class labels during training [27]. In addition, identifying regions containing unexpected and unique abnormalities within an image is of critical importance. Saliency mapping techniques [28], [29] identified such regions as being distinctive by using primitive signature features such as texture, shape, and color. Rendered as a heat-map in which hot regions correspond to a considerable impact on the model's final decision, saliency maps represent an important step toward understanding chest X-ray images and further improving models' classification performances.

For the first time in thorax multi-class classification (to the best of our knowledge), we address this problem by using the promising performances of KD approaches to support the automatic classification of 14 abnormalities appearing in chest X-rays, along with saliency map visualizations to ensure the accurate identification of abnormal regions. The main contributions of this paper are summarized as follows:

- We utilized a variety of saliency mapping techniques, including vanilla/guided back-propagation, smooth gradients, and SmoothGrad integrated gradients to better understand our deep learning model's decision-making process.
- We proposed different KD training approaches, including original basic training, standard KD (deeper teachers teach lower-cost students), reversed KD

(lower-cost students teach deeper teachers), defective KD (teachers trained over the first 50 iterations teach lower-cost students), and self-training KD (models teach themselves); we then compared their respective classification performances.

The remainder of the paper is organized as follows. Section 2 describes the relevant recent works on chest X-ray lesion classification. In Section 3, we describe our proposed approaches for saliency map visualization and different KD training methods for thorax multi-classification. In Section 4, we introduce the ChestX-ray14 dataset and summarize our obtained results. The paper concludes in Sections 5 and 6 with a discussion and suggestions for future works, respectively.

II. RELATED WORKS

Asides from the various screening methods applied to detect suspected lung diseases, the promising results obtained from implementing deep learning techniques in chest X-ray image analysis tasks have recently attracted much attention [30], [31]. Several open-access datasets of chest X-ray images have allowed scientists to train, verify, fine-tune, and evaluate their new deep learning algorithms; these datasets have included chest X-rays with and without lung cancer nodules from the Japanese Society of Radiological Technology [32], frontal and lateral chest radiographs of disease annotations from the Indiana dataset [33], two databases (from Montgomery County and Shenzhen Hospital) to improve the CAD of pulmonary diseases [34], normal versus tuberculosis cases from the Royal Tropical Institute [35], and ChestX-ray14—the largest publicly available database currently available, containing annotations of 14 different lung diseases [26]. The TUNA-Net framework was proposed by [36] for pneumonia recognition on two public chest X-ray datasets; this model adapted the labeled adult chest X-rays in the source domain such that they appeared as though they had been taken from pediatric X-rays in an unlabeled target domain. TUNA-Net achieved a 96.3% AUC (the area under the receiver operating characteristic (ROC) curve) value in binary pediatric pneumonia classification. Salehinejad *et al.* [37] employed deep convolutional generative adversarial networks (DCGAN) to generate artificial images from five common pathological classes, then applied it to chest X-rays. The authors reported that data augmentation using these synthesized images increased the diversity of the training data, substantially improving the generation performance and classification of unseen data.

There have been many deep learning models proposed to achieve outstanding classification results on the ChestX-ray14 dataset. Rajpurkar *et al.* [30] proposed CheXNet - a 121-layer convolutional neural network for pneumonia classification. A 14-disease classification task was also attempted and competitive results were obtained under their proposed method. They also compared the performances of four radiologists on a subset of 420 annotated images against the CheXNet model and found that CheXNet exceeded the average radiologist performance, as measured by the F1 metric. A unified weakly supervised multi-label image classification and localization framework was introduced by Wang *et al.* [26] to evaluate the ChestX-ray8 dataset. After implementing a variety of pre-trained deep models and excluding the fully connected and soft-max layers, a transition layer, global pooling layer, prediction layer, and loss layer were all inserted. This approach facilitated the identification of plausible spatial regions due to the combination of activations from the transition layer and weights from the prediction inner-product layer. Their initially quantitative classification and localization results were promising, despite the procedure remaining too computationally strenuous for full implementation as an automated high-precision CAD system.

More recently, a variety of deep learning-based techniques have sought to approach the ChestX-ray14 problem. ChestNet [38] contained two main branches: a classification branch, which served as a unified network with a pre-trained ResNet-152 model to manage the complexities of handling local handcrafted features; and an attention branch, which explored the correlations between different disease labels and allowed for the localization of abnormal regions. In its performance comparison, it was shown to outperform three state-of-the-art deep learning models employing official patient-wise splits without extra training data. TieNet [39] was introduced to first classify ChestX-ray14 images by extracting distinctive X-ray images and embedded texts from corresponding reports; it was later converted into a chest X-ray-reporting system in a simulation, to output disease classifications with a preliminary report. It achieved an average AUC of over 90%, which was an improvement of 6 % compared to the baseline on an unseen and hand-labeled OpenI dataset. A multi-level attention model, implemented as an end-to-end trainable CNN-recurrent neural network (RNN) to highlight the meaningful regions, was also built in this study.

A fully convolutional recognition network [40] improved AUC scores in classifications of most diseases compared to the reference models, as well as remarkable prediction scores of disease localizations. Wang *et al.* [57] introduced ThoraxNet, which contained two branches for 14-label prediction and abnormality localization. The classification branch used a pre-trained ResNet-152 and the attention branch was equipped with several convolutional layers and the gradient-weighted class activation mapping (Grad-CAM) module. This procedure yielded AUC scores of 0.788 and 0.896 by using the patient-wise official split and image-wise random split, respectively. It obtained higher AUCs compared to other deep models training with no external data. Ho and Gwak [41] proposed a pre-trained DenseNet-121 model to localize pathologically abnormal areas and a handcrafted, deep feature integration approach to classifying 14 disease classes. The authors demonstrated that their proposed methods could efficiently manage interdependencies between class annotations and achieved superior classifications to the then-current reference baseline on the ChestX-ray14 dataset.

From the existing reports on ChestX-ray14, the transferal of features extracted from pre-trained models is seen to be preferable. However, the trends in model compression—in which a larger pre-trained model is built to allow the smaller model to learn complex features whilst minimizing the computation and memory costs—has not yet been investigated for X-ray datasets. In particular, a large and complex network or an ensemble model is first trained and extracts important feature information from the given data, thereby producing targeted predictions. A small network is then trained with the help of this more cumbersome model. The small model is able to produce comparable results or replicate the cumbersome model's results. Therefore, we propose different KD training strategies for 14-disease classification, as well as a variety of saliency mapping techniques for abnormal feature visualizations in X-ray images.

III. PROPOSED APPROACHES

We conducted extensive experiments to examine both the dominant features visualized by saliency maps and the common features of dark knowledge in KD.

A. SALIENCY MAPS

As the most common technique for interpreting deep neural networks (DNNs), saliency maps [42], [43] represent the gradient of the output class with respect to the input, based on a score function. They note how the changes in the output correspond to changes in input image pixels. The output value is increased under small changes in the pixels or exclusively positive values in the gradients. Thus, visualizing these gradients provides an intuitive measure of attention. In our design, using an input vector $x \in R^d$ and a model with the function $S : R^d \rightarrow R^{14}$ results in an explanation map of $S : R^d \rightarrow R^d$, which maps inputs to particular objects of the same shape. Each dimension is then associated with the relevance or importance of the final output's dimension.

1) GRADIENT [44]

The gradient of the scalar logit for a specific class for the input is expressed as

$$E_G(x) = \frac{\partial S}{\partial x}. \tag{1}$$

2) GUIDED BACK-PROPAGATION (GBP) [45]

GBP indicates the change in how the back-propagated gradient varies with ReLU. Using $\{f^l, f^{l-1}, \ldots, f^0\}$ as the feature maps derived during the forward pass of a DNN and $\{R^l, R^{l-1}, \ldots, R^0\}$ as the intermediate representation obtained during the backward pass (more concisely, $f^l = relu\left(f^{l-1}\right) = max(f^{l-1}, 0)$ and $R^{l+1} = \frac{\partial f_{out}}{\partial f^{l+1}}$), GBP aims to achieve zero outputs for all negative gradients; the mask is then computed as

$$R^l = 1_{R^{l+1}>0} 1_{f^l>0} R^{l+1}, \tag{2}$$

where $1_{R^{l+1}>0}$ retains only positive gradients and $1_{f^l>0}$ retains only positive activations.

3) INTEGRATED GRADIENTS (IG) [46]

The gradient saturation is addressed by summing over-scaled values of the input. IG for an input x is defined as:

$$E_{IG}(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x} + \alpha(x - \bar{x}))}{\partial x} d\alpha, \tag{3}$$

where $\bar{x}$ is typically set to zero and is the baseline input representing an absence of features in the input sample x_i.

4) SMOOTHGRAD (SG) [47]

SG seeks to alleviate noise and visual diffusion by averaging over all explanations of noisy versions of an input. Given an explanation E and a sample x, the SG explanation E_{SG} is defined as

$$E_{SG}(x) = \frac{1}{N} \sum_{i=1}^{N} E(x + g_i), \tag{4}$$

where the noise vectors $g_i \sim N(0, \sigma^2)$ are drawn independently and identically distributed from the normal distribution.

B. KNOWLEDGE DISTILLATION (KD)

In the standard KD model [20], knowledge is encoded and transferred based on the forms of the softened class scores. The total loss of the student model's training is given by

$$L = (1 - \alpha) L_{CE}(y, \sigma(z_s)) + 2\alpha T^2 L_{CE}(\sigma\left(\frac{z_s}{T}\right), \sigma(\frac{Z_T}{T})), \tag{5}$$

where $L_{CE}(.,.)$ represents the cross-entropy; y represents the one-hot vector of ground truths; σ is the soft-max function; z_s and z_T are the output logits of student and teacher models, respectively; α is a balancing hyper-parameter; and T is the temperature hyper-parameter. In (5), the first term denotes the cross-entropy loss using ground truth labels whilst the second term encourages the student model to mimic the softened class scores from the teacher model.

As shown in the standard KD from Fig. 1, the student model was trained using the predictions of the teacher model along with the ground truth hard labels. A variant of the soft-max function including a temperature parameter T was used to produce the soft labels as

$$sofmax(I_L, T) = \frac{\exp(I/T)}{\sum_i \exp(I_i/T)}, \tag{6}$$

where I is the input logits to the soft-max layer, and a higher value of T produces a smoother probability distribution over the 14 classes.

Thus, the total loss function L is a combination of the KD loss (soft loss) L_{soft}, the cross-entropy loss between the soft predictions of the teacher and students, and the hard loss L_{hard}, given as:

$$\begin{aligned} L_{soft} &= H(softmax(I_T, T), softmax(I_S, T)), \\ L_{hard} &= H(Y_S, Y_{GT}) = H(softmax(I_S, 1), Y_{GT}), \\ L &= L_{soft} + L_{hard}. \end{aligned} \tag{7}$$

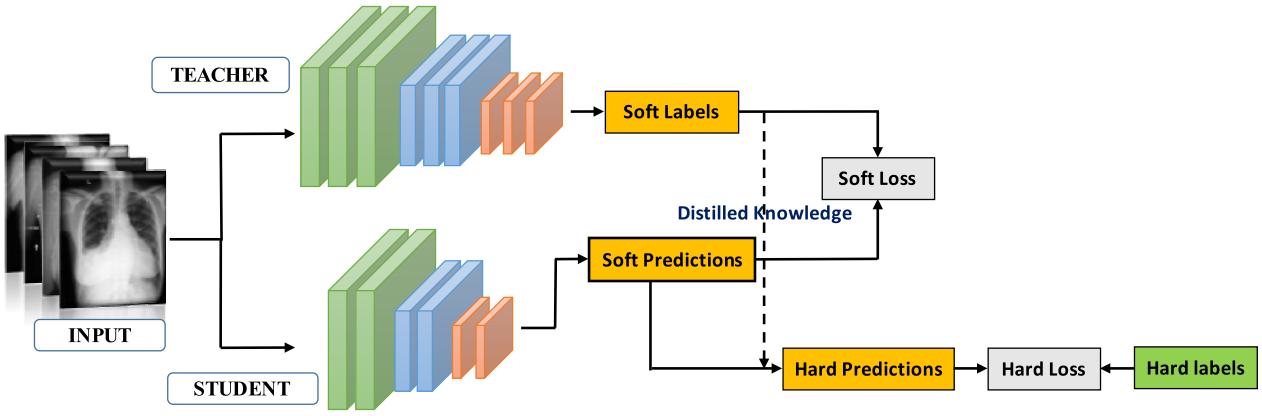

FIGURE 1. Standard KD – a student model learns from the teacher model's guidance (soft loss) and ground-truth labels (hard loss).

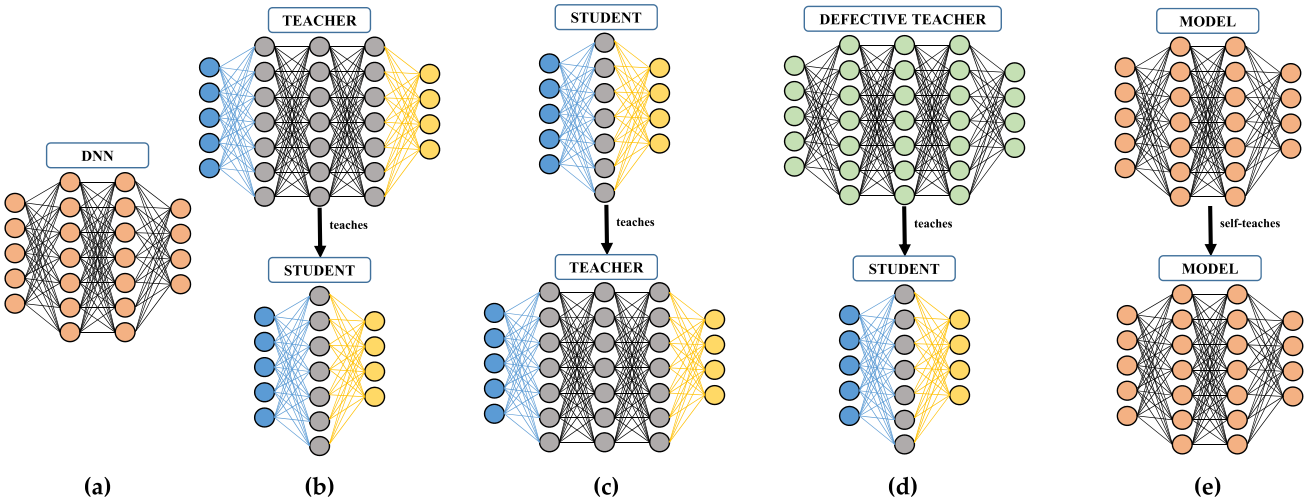

FIGURE 2. The proposed training strategies on the ChestX-ray14 dataset: (a) - Base training, (b) - Standard KD, (c) - Reversed KD, (d) - Defective KD, and (e) - Self-training KD.

However, it is commonly understood that if we reverse the KD operation, the teacher will not be significantly improved because the student model is too weak to learn and transfer useful knowledge. Also, using a poorly-trained teacher model that has been trained on 50 first epochs may yield worse performances than normal KD or reverse KD procedures. Finally, if we self-train the model, it may achieve better results compared to all of the above strategies. For example, the model would learn from its softened-class targets with a 10% error when being trained from itself with a 90% accuracy criterion. To address these concerns, we followed the KD procedure illustrated in Fig. 1 and conducted all experiments pertaining to the five main training strategies: base training—simply training normal DNNs in an end-to-end manner; standard KD—training a teacher model to teach a student model; reversed KD—training a student model to teach a teacher model; defective KD—poorly training a teacher over the first 50 epochs to then teach a student model; and self-training KD—training a model to teach itself (Fig. 2). To feasibly conduct all KD training approaches, we selected six types of DNN models—with identical input sizes—to examine our proposed training methods, including MobileNet-v2 [2], VGG-19 [3], ResNet-32, ResNet-50, and Resnet-152 [4], and DenseNet-121 [49]. The first four models (MobileNet-v2, VGG-19, ResNet-32, ResNet-50) were used as student models; they are all relatively small and simple models, though sufficiently powerful to either learn X-ray features from both themselves and more complex teacher models (ResNet-152, DenseNet-121) or to transfer the distilled knowledge to deeper networks.

IV. EXPERIMENTAL RESULTS

A. CHESTX-RAY14 DATASET

We evaluated our proposed approaches on the publicly available, recently published ChestX-ray14 dataset [26]; this is considered to be the largest collection of up-to-date front-view chest radiographs, containing a total of 112,120 X-ray images acquired from 30,805 unique patients. Each image is marked with a single or multiple pathological labels denoting 14 diseases, based on radiology reports with over 90% accuracy. In addition, there were 984 annotated images provided by board-certified radiologists.

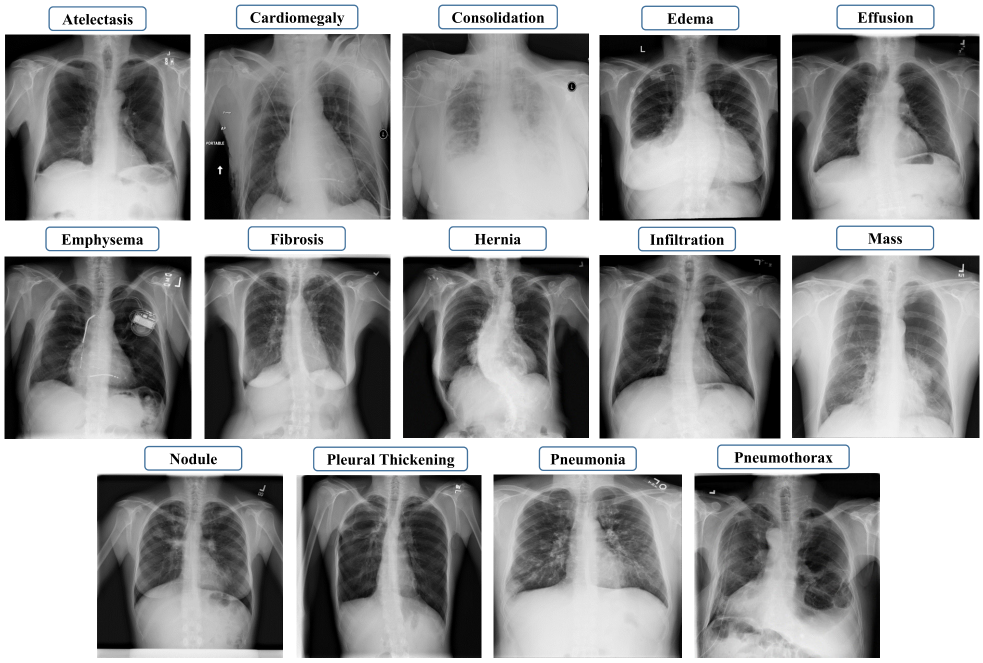

FIGURE 3. 14 random samples associated with 14 thorax diseases in the Chestx-ray14 dataset.

Fig. 3 shows 14 examples of common thorax diseases. The original 1024 × 1024 PNG images were downscaled to 224 × 224 PNG images for all KD experiments and 229 × 229 PNG images for InceptionV3 [48] saliency map visualizations—these were normalized into the range [-1,1] based on the mean and standard deviation.

It is greatly important to consider the data division step for proper evaluation of our proposed methods. On the patient-wise official split considered, all images from the same patient are only present in one of the training, validation, and testing subsets. Meanwhile, the image-wise random split would randomly divide all X-ray images into three subsets without considering on which subject an X-ray image was acquired. In other words, there is an average of 3.6 images per patient. The radiographs from the same subject are likely to appear in both training, validation and testing sets simultaneously leading to achieve much better performance than using the patient-level split. However, this should not be accepted in pattern classification tasks since it is burdensome to establish consistent benchmarks or sometimes known as ''cheating'' if patient samples from testing sets appear in the training data. Plus, because of the impact of randomness, it is required to conduct experiments multiple times to average the AUC scores. Concerning these sorts of problems, we thereby solely utilized the patient-level split which formulates more proper criteria to evaluate any models in thorax disease prediction.

Using the patient-wise official split, we divided the data from 30,805 unique patients into 70% for training, 10% for validation, and 20% for testing. We also augmented the training and validation datasets using randomized horizontal flipping procedures. Python 3.6.10 with Tensorflow 2.1.0, CUDA 9, and cuDNN 7.5 deep learning dependencies were used for implementing both (i) the visualizations from the different saliency mapping techniques and (ii) the 14-category classifications based on the five KD training strategies. We conducted our experiments within a total computation time of one week, using an i7-4770K 4-core CPU, a GeForce GTX 1070 GPU, and 32G of memory.

B. SALIENCY MAP VISUALIZATION

In this section, we discuss three selected saliency mapping techniques, including GBP, SmoothGrad, and SmoothGrad integrated Gradients. We assessed their efficacy in visualizing distinguishable thorax diseases. The main purpose of this saliency mapping task was to attempt to visualize the measure of attention for abnormal regions that were not originally annotated in the 984 ground-truths from our dataset. The findings from our saliency mapping algorithms may significantly help radiologists make decisions concerning the locations of abnormal regions, despite the lack of prior annotations for the X-ray images. From our observation, the InceptionV3 model was seen to better visualize the hot attention areas on the ChestX-ray14 dataset (with higher AUC scores) than other deep models (both other students and teacher models). Fig. 4 shows four examples of the thorax abnormalities—without any ground-truth annotations—identified by the InceptionV3 model, with AUC scores of 0.854 for Effusion, 0.739 for Fibrosis, 0.762 for Hernia, and 0.768 for Nodule classes. Concerning the efficacy of the saliency mapping methods, the integration of SmoothGrad and Gradients outperformed others; it produced more easily obtainable and clearer images for further disease analysis.

Our knowledge of thorax symptoms, along with the observations from Panel (d), demonstrated that the generated pleural effusion images indicated a hot attention region

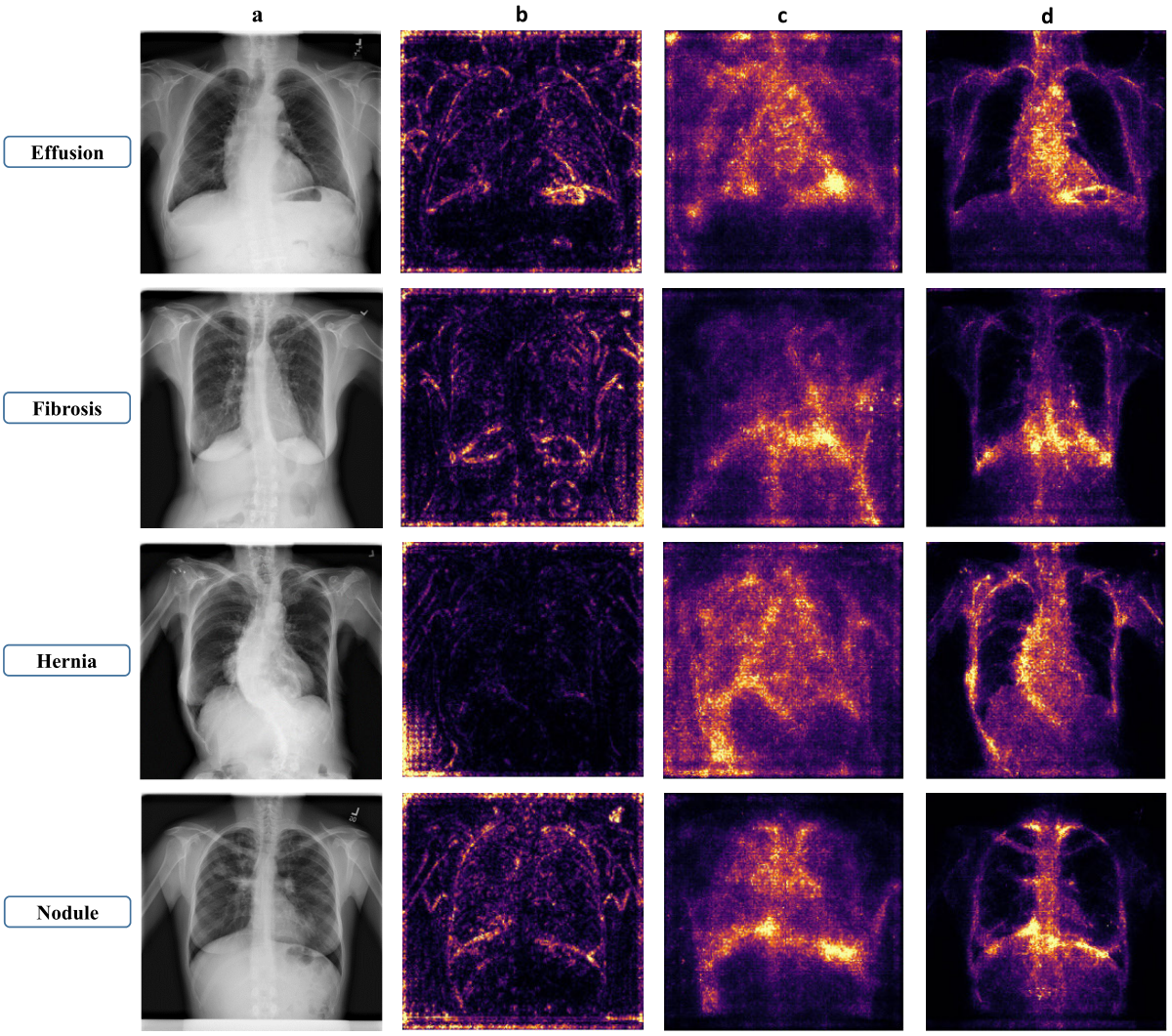

FIGURE 4. Several abnormal thorax images and their generated results from different saliency mapping techniques. Panel (a) shows the original chest X-ray images; Panels (b), (c), and (d) show the saliency maps generated by GBP, SmoothGrad, and SmoothGrad integrated Gradients, respectively.

localized in the lower right-hand side, where the hot region was seen to track along the lateral wall and the right costophrenic angle was obscured by a meniscus. This finding was correctly noted by experienced radiologists as the true screening analysis of the chest X-ray images. Similarly, pulmonary fibrosis (Row 2) exhibited an increase of subpleural reticular markings with lower lobe predominance across both lungs. In Row 3, there was a mild opacification of the bilateral right-hand lobes, with an air-fluid level consistent with a hiatal hernia. The upper left-hand lobe pulmonary nodule was slightly marked in Row 4 of Fig. 4. The foci of the saliency maps were indeed on abnormally affected regions, and the disturbance effects and noises were reducible. These findings from the SmoothGrad integrated Gradients saliency mapping method were remarkably conducive to the analysis and distinguishing of different thorax diseases, even when the absence of annotation labels was taken into account.

To validate the potential of saliency mapping techniques, SmoothGrad integrated Gradients, in particular, we compared its aptitude of localizing abnormal regions with our previous study [4] using class activation map (CAM), which extracted the weight activations from the last convolutional layers of the pre-trained DenseNet-121 model (see Fig. 5). The blue bounding boxes denoted the ground truths in a total of 984 available annotations from the ChestX-ray14 dataset. Although we formally verified the ability of CAM methods, inaccurately abnormal region localizations from several instances were inevitable. Those abnormal instances, nevertheless, could be situated with the relatively high certainty by our proposed saliency map technique. In specific, we showed eight samples that were wrongly located by the precedent CAM method (the red area represented the most indicative pathology region while the blue indicated normal regions), except for cardiomegaly sample while it was clearly able to see that all eight samples were precisely highlighted in the SmoothGrad integrated Gradients method when comparing with the ground truths. This first demonstrated the significant improvement of saliency maps compared to previously baseline disease visualization techniques and further could be extended in the extensive X-ray analysis.

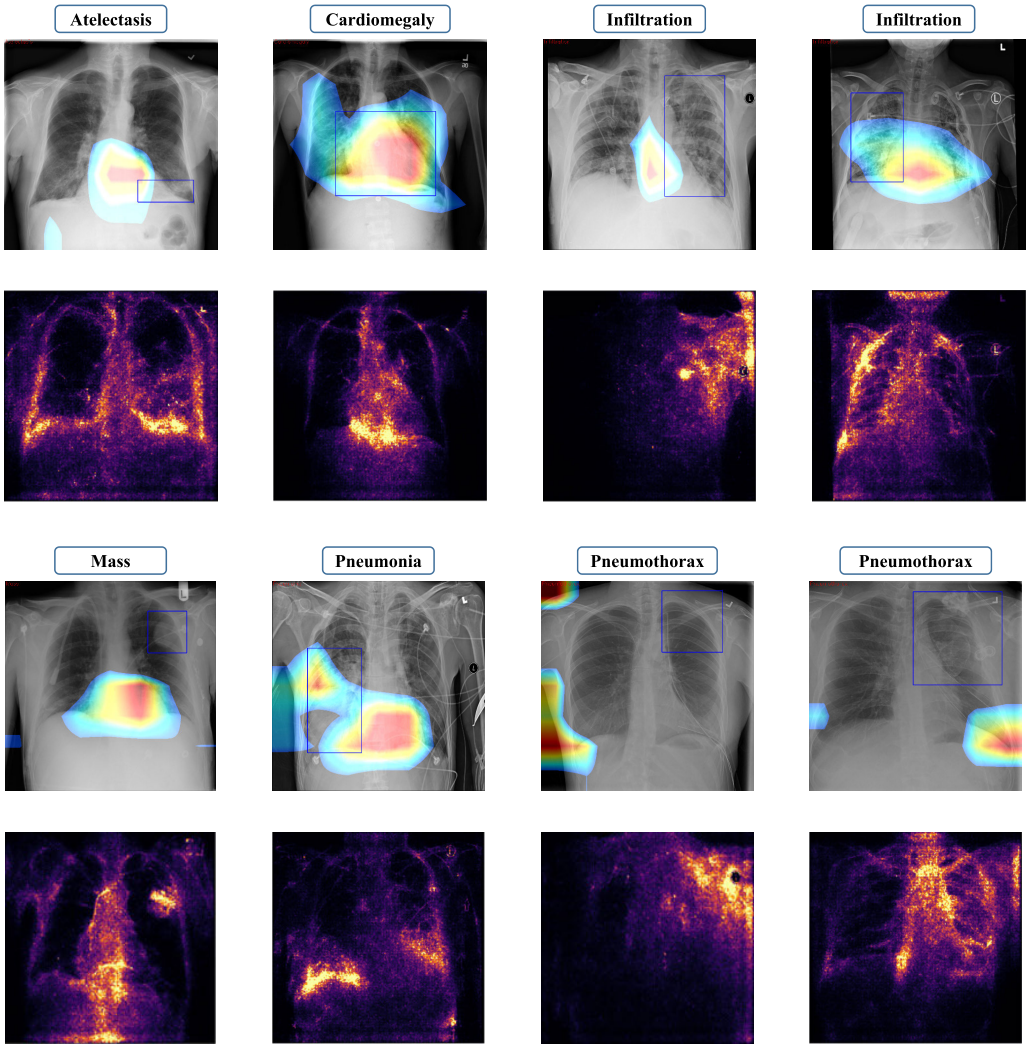

FIGURE 5. Eight abnormal thorax samples generated by CAM (first and third rows) and SmoothGrad integrated Gradients (second and fourth rows) methods. The blue bounding box represents the ground truths available from the ChestX-ray14 dataset.

C. THORAX MULTI-CLASS CLASSIFICATION RESULTS

In this section, we describe the extensive experimental results for the 14 thorax disease category classification using base training and three KD training strategies. In Table 1, results are shown for the six pre-trained deep models used for normal transfer learning (referred to as base training). DenseNet-121 obtained the best average AUC score with 80.97%, followed by ResNet-152 (79.01%), VGG-19 (76.17%), ResNet-50 (71.66%), MobileNet-v1 (67.10%), and ResNet-32 (66.05%). This suggests that the more complex and deeper models outperformed other smaller and simpler models when dealing with the challenging multi-class classification of chest X-ray images.

As expected, Standard KD at first outperformed the base training method. The student model was significantly improved by learning from the teacher. In particular, MobileNet-v1, VGG-19, ResNet-32, and ResNet-50 achieved 7.02%, 1.22%, 8.01%, and 3.86% AUC improvements, whereas a decrease of 0.21% was observed when DenseNet-121 was taught by ResNet-152. Similarly, when DenseNet-121 acted as a teacher model, it significantly improved upon all performances of the student models (7.84%, 2.5%, 10.36%, and 7.03% improvement accuracies with MobileNet-v1, VGG-19, ResNet-32, and ResNet-50, respectively); it even outperformed the ResNet-152 teacher model. This sheds light on the perspective that the weak student models could be significantly enhanced by superior teacher models.

Meanwhile, as aforementioned, we assumed that as the teacher model becomes more accurate, soft probabilities will extensively capture the underlying target class distribution and therefore deliver better supervision to the student model. That is, the smaller and less accurate models cannot be good teacher models. We, thereby, conducted Reversed KD experiments to settle this issue. The majority of experiments reported that the teacher models were not improved by Reversed KD training strategies (teachers were taught and trained by students), except in the case of Reversed ResNet-152/DenseNet-121. Therefore, we confidently confirm our hypothesis that the student models were

TABLE 1. The average classification AUC score using base training and KD training approaches.

Teacher - Original	Student – Original		Standard KD	Reversed KD	Defective KD
	Base Training	Self-training			
ResNet-152: 0.7901	MobileNet-v1: 0.671	0.7068	0.7412	0.6864	0.7018
	VGG-19: 0.7617	0.7715	0.7739	0.6989	0.7149
	ResNet-32: 0.6605	0.6934	0.7406	0.6725	0.6965
	ResNet-50: 0.7166	0.7438	0.7552	0.7205	0.7278
	DenseNet-121: 0.8097	**0.8256**	**0.8076**	**0.8097**	**0.7993**
DenseNet-121: 0.8097	MobileNet-v1: 0.671	0.7068	0.7494	0.6902	0.7218
	VGG-19: 0.7617	0.7715	0.7867	0.7076	0.7616
	ResNet-32: 0.6605	0.6934	0.7641	0.6752	0.7416
	ResNet-50: 0.7166	0.7438	0.7869	0.7512	0.7876
	DenseNet-121: 0.8097	**0.8256**		**0.8196**	**0.8021**

incapable of transferring effective knowledge to the teacher models. Moreover, we explored the Defective KD training strategies, in which the teacher model was trained over the first 50 iterations, with the defective knowledge transferred to student models; we observed that student models could be greatly improved even with distilled knowledge from poorly-trained teacher models. For instance, MobileNet-v2, ResNet-32, and ResNet-50 student models were improved (by 3.08%, 3.60%, and 1.12% AUC, respectively) with ResNet-152, whilst it also achieved 5.08%, 8.11%, and 7.1% AUC score improvements compared to the base training with DenseNet-121. Although the poorly-trained teacher model performed less accurately than the Standard KD (as expected), the capacity for transferring knowledge to lower-cost student models was evaluated as of a higher level compared to both base training and Reversed KD in most of the experiments. Defective KD could be used to generate the soft targets of the model, where these learned soft targets then guide the teacher model's regularization processes.

To better demonstrate the distillation approach, we considered updating the output distribution of the teacher model using information from itself or simpler models, this is the so-called Self-training KD framework, in which there is no teacher model. The self-training KD method was applicable to the cases in which either a teacher model is unavailable or limited computation resources are provided. Concretely, the model was first assigned to train in the normal way to obtain a pre-trained model, it was then used for self-training by transferring the soft-targets, as described in Eq. (5). Formally, we minimized the Kullback-Leibler (KL) divergence of the logits between model M and its pre-trained model M^t, using the loss function

$$L_{self-train} = (1-\alpha)\, H(q,p) + \alpha D_{KL}(p_\tau^t, p_\tau), \qquad (8)$$

where D_{KL} is the KL divergence; q is the ground-truth label; p and $p_\tau^t = softmax\left(z_k^t\right) = \frac{\exp(z_k^t/\tau)}{\sum_{i=1}^{K}\exp(z_i^t/\tau)}$ (z^t is the output logits of pre-trained models) are the output probabilities of M and M^t, respectively; τ is the temperature; and α is the weight parameter used to balance the two terms.

We trained five baseline models, including MobileNet-v2, VGG-19, ResNet-32, ResNet-50, and DenseNet-121. The baseline models were trained for 200 iterations with an initial learning rate of 0.1, an SGD optimizer (with a momentum of 0.9 and a weight decay of 5e-4), and a grid search for finding the optimum hyper-parameter values. Column 3 in Table 1 shows that Self-training KD consistently outperformed the base training approach. For example, MobileNet-v2, VGG-19, ResNet-32, ResNet-50, and DenseNet-121 increased their accuracy performances by 3.58%, 0.98%, 3.29%, and 3.35%, respectively. However, Standard KD outperformed the self-training methods because the weak models by themselves transferred inefficient knowledge, except in the case of DenseNet-121. From our observations, Self-training KD with DenseNet-121 obtained the highest average AUC (82.56%), followed by Reversed KD with ResNet-152/DenseNet-121 (80.97%), and Defective KD with DenseNet-121/DenseNet-121 (80.21%). Fig. 6 shows the stable and accurate performance of Self-training KD with DenseNet-121 via the training and validation accuracies, as well as the accuracy improvement compared with the base training method.

Although we acknowledge that KD frameworks (standard KD and defective KD) demonstrated significant improvements compared to stand-alone model (base models), it is insufficient to train a student model with rich input information to obtain a well-trained teacher model, as we observed results from the reversed KD approach. In addition, it is undoubtedly noted that the training performance of the Self-training framework accumulated a higher degree of time-consuming, computational costs, and resource burdens compared to any simply base trained methods. Table 2 denotes the execution time for each approach (base training and three types of KD approaches). In general, the training time of each KD approach using the teacher model (ResNet-152) is lesser than which of the teacher model (DenseNet-121). Moreover, among six base models, the base training approach

TABLE 2. The training time of each KD approach (unit: minute).

Base Model	Base Training	Standard KD		Reversed KD		Defective KD		Self-train KD
		ResNet-152	DenseNet-121	ResNet-152	DenseNet-121	ResNet-152	DenseNet-121	
MobileNet-v1	182	358	431	393	417	329	335	345
VGG-19	192	416	444	413	429	339	342	362
ResNet-32	168	388	414	391	397	315	320	319
ResNet-50	198	419	433	422	439	338	352	375
ResNet-152	246	--	--	--	486	--	--	--
DenseNet-121	258	486	508	498	508	397	413	508

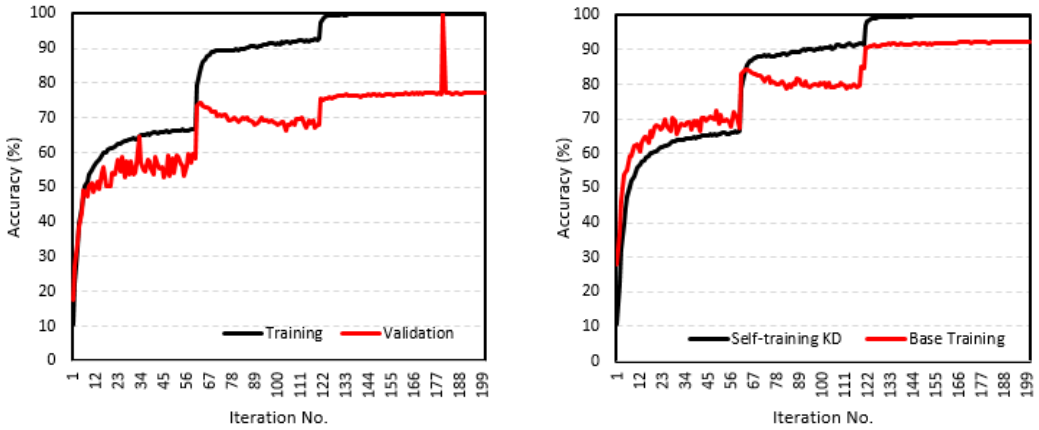

FIGURE 6. (left) Training and validation accuracies of Self-training KD and (right) the training accuracy comparison between Self-training KD and its base training from the DenseNet-121 model during 200 iterations.

using stand-alone models obtained the smallest amount of training time (approximately 258 minutes trained on the base DenseNet-121) while the Self-train KD (Self-train DenseNet-121) consumed the highest amount of training time (approximately 508 minutes). Despite showing the training accuracy of self-trained and base training methods in 200 iterations (Fig. 6 - right), the amount of time costs differently for each iteration of two frameworks. DenseNet-121-based self-trained KD consumed twice as the amount of training time as the base training, which sometimes even led our computational resources to be exhausted.

To justify the potential of the proposed methods, we compared our best results achieved by the Self-training DenseNet-121 model with five state-of-the-art deep learning frameworks on the ChestX-ray14 dataset by evaluating on the per-class AUC scores and the average AUC scores, as shown in Table 3. The highest AUC score was punctuated in boldface for each row. Although the works of Guendel *et al.* [58] and Wang *et al.* [57] yielded another exceptional classification result by utilizing the image-wise random split - without consideration on which subject a radiograph was acquired and the radiographs from the same subject thus could be appeared in both training and testing concurrently, we disregarded their phenomenal results in Table 3. Instead, we included results of studies using the patient-wise official data split deemed to be a fairer and more proper evaluation of CAD on the classification of thorax diseases. Note that the work of Guendel *et al.* [58] trained their model not only on the ChestX-ray14 dataset but also on an external set of 180,000 images from the PLCO dataset [59]. The diagnosis performances presented in Table 3 indicate that our proposed framework obtained very competitive results with the highest per-class AUC scores in seven disease classes and the highest average AUC score.

V. DISCUSSION AND FUTURE WORKS

We demonstrated the suitability of saliency mapping techniques for visualizing the abnormal regions of chest X-ray images, as well as the competitive distilling performance achieved by transferring knowledge both from the large, highly regularized models into smaller ones and from the model into itself, to classify 14 pathological thorax diseases. However, our work has several notable limitations. First, although we attempted to evaluate a comprehensive machine-human annotated chest X-ray dataset, simulating the practical clinical challenges of handling over 100,000 images, it was difficult to correctly visualize and discriminate the 14 classes by applying a deep learning framework when the database was unbalanced and weakly supervised. The appearance of a thorax disease is usually accompanied by other related diseases visible in chest X-ray images; for instance, pneumothorax is often associated with pneumonia. The low rate of agreement between multiple radiologists in this dataset revealed a large bias; and the diagnoses should be voted upon by the majority of these experts. Therefore, there is a need to utilize external training datasets and an independent validation, such as MIMIC-CXR [60] or PLCO [59] to verify the generalizability of our proposed frameworks.

Second, we analyzed the output of saliency maps, SnoothGrad integrated Gradients in particular, which offered a good visual representation of isolating the abnormal regions

TABLE 3. Per-class AUC of proposed and other five methods in the ChestX-ray14 literature using the patient-wise official split.

Pathology	Yao *et al.* [61]	Wang *et al.* [38]	Guendel *et al.* [58]	Ho *et al.* [41]	Wang *et al.* [57]	Proposed
Atelectasis	0.772	0.743	0.767	**0.795**	0.751	0.794
Cardiomegaly	**0.904**	0.875	0.883	0.887	0.871	0.896
Effusion	0.859	0.811	0.828	0.875	0.828	**0.882**
Infiltration	0.695	0.677	**0.709**	0.703	0.681	0.705
Mass	0.792	0.783	0.821	0.835	0.799	**0.844**
Nodule	0.717	0.698	**0.758**	0.716	0.715	0.752
Pneumonia	0.713	0.696	0.731	0.742	0.694	**0.763**
Pneumothorax	0.841	0.810	0.846	0.863	0.825	**0.878**
Consolidation	0.788	0.723	0.745	0.786	0.742	**0.798**
Edema	0.882	0.833	0.835	**0.892**	0.835	0.870
Emphysema	0.829	0.822	0.895	0.875	0.843	**0.918**
Fibrosis	0.767	0.804	**0.818**	0.756	0.804	0.803
Pleural Thickening	0.765	0.751	0.761	0.774	0.746	**0.779**
Hernia	**0.914**	0.899	0.896	0.836	0.902	0.876
Average	0.798	0.781	0.807	0.810	0.787	**0.826**

and could further assist the deep network in classification decisions. However, there was no certainty that saliency map results (Fig. 4) could be correctly localized abnormalities due to the lack of disease annotations from ChestX-ray14. This means we were not able to make a comparison between our outcomes with real ground truths. Plus, we solely evaluated its aptitude on very limited numbers of annotated X-ray images which might lead to generating a huge tendency of localization. That is, the integration of center bias and background information was not always helpful for cases in which the abnormal areas (saliency targets) were unidentifiable in the margins of the X-ray images, or in cases where there were multiple diseases in the X-ray image. Thus, it is critically important to design an attention visualization model not only to facilitate generalization but also to help diagnosis the models' failures by identifying biases or fair and bias-free outcomes from the datasets, as was done in [51].

Third, our extensive experiments demonstrated the potential of KD strategies in chest X-ray disease classifications. Although we demonstrated the outperformance of Self-training KD in terms of classification results compared with base-training and standard-training KD, the time-consuming and enormous costs of computation presented substantial shortcomings of the Self-training KD framework. Besides, our KD model independently extracted instance features as the distilled knowledge from specific layers of the teacher models, without considering the instance's relationship to the student models or the inference procedure. It is difficult for student models to directly fit all the layer outputs from teachers. Therefore, it is necessary to create new KD designs that can help reduce the intra-class variances and magnify inter-class differences in the feature space, as well as prevent the occurrence of significant performance drops when both teachers and students have different architectures, as seen in [52]. It might also be better to replace the process of mimicking the teacher's representation space with that of preserving the pairwise similarities in the student's own representation space [53].

Lastly, there are 60,412 normal images and 51,708 images with at least one or more labels that yield to the problem of interdependency among labels. For example, an image, which is indicated with the presence of edema, possibly includes the presence of both consolidation (air space opacification) and pleural effusion (the pleural space with the abnormal fluid). This generated much disturbance for our proposed models to be trained and produced lower AUC scores since the proposed method recognized the potential of these interdependencies and further predicted pathological outcomes across all thorax categories ineffectively. Therefore, an approach, which allows the distillation at different internal points across the teacher and entitles the student to learn and compress the abstraction in the hidden layers systematically, is necessarily required. With proper internal representations, the student may outperform its conventional approach on either ground-truth labels, soft-labels, or both. From our observation, the poorly-trained teacher could remarkably enhance the student itself (as results shown by Defective KD), it is justifiable to interpret KD as a regularization term and to scrutinize KD from the perspective of Label Smoothing Regularization (LSR) [62]. LSR can mitigate the over-confidence problem and improve model calibration by replacing the one-hot labels with smoothed labels. The smoothed label can be split into two parts: the first part is the ordinary cross-entropy (one-hot label) distribution and the output; the second part corresponds to the virtual teacher to provide soft-targets by a uniform distribution. This indeed furnishes efficient regularization for the student and feasibly overcome the issue of interdependency among thorax labels.

VI. CONCLUSION

In this work, we proposed KD training strategies along with three types of saliency mapping techniques, with the aim of

correctly classifying and visualizing 14 pathological thorax diseases from the public ChestX-ray14 datasets. Our experiments demonstrated the feasibility of implementing different KD training strategies, suggesting that the targeted models into which the distilled knowledge is transferred can be enhanced by the self-training KD method when difficulties arise in choosing superior teachers or when limited computation resources are available. Also, the results of the saliency mapping algorithms show promise in highlighting abnormal regions, despite featuring unbalanced and limited annotations of pathologies. Its capabilities can further represent a powerful tool with which clinicians or radiologists can review and interpret the decision-making processes of CAD algorithms in thorax disease diagnoses.

REFERENCES

[1] D. R. Brenner, J. R. McLaughlin, and R. J. Hung, "Previous lung diseases and lung cancer risk: A systematic review and meta-analysis," *PLoS ONE*, vol. 6, no. 3, Mar. 2011, Art. no. e17479.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: http://arxiv.org/abs/1412.7062

[10] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in CT images using deep convolutional neural networks," *Pattern Recognit.*, vol. 85, pp. 109–119, Jan. 2019.

[11] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, and O. Geessink, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *J. Amer. Med. Assoc.*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

[12] S. Wang, C. Tang, J. Sun, and Y. Zhang, "Cerebral micro-bleeding detection based on densely connected neural network," *Frontiers Neurosci.*, vol. 13, p. 422, May 2019.

[13] S. Wang, Y. Jiang, X. Hou, H. Cheng, and S. Du, "Cerebral micro-bleed detection based on the convolution neural network with rank based average pooling," *IEEE Access*, vol. 5, pp. 16576–16583, 2017.

[14] S. S. Chaturvedi, K. Gupta, and P. S. Prasad, "Skin lesion analyser: An efficient seven-way multi-class skin cancer classification using MobileNet," 2019, *arXiv:1907.03220*. [Online]. Available: http://arxiv.org/abs/1907.03220

[15] P. Xie, K. Zuo, Y. Zhang, F. Li, M. Yin, and K. Lu, "Interpretable classification from skin cancer histology slides using deep learning: A retrospective multicenter study," 2019, *arXiv:1904.06156*. [Online]. Available: http://arxiv.org/abs/1904.06156

[16] L. E, B. Zhao, Y. Guo, C. Zheng, M. Zhang, J. Lin, Y. Luo, Y. Cai, X. Song, and H. Liang, "Using deep-learning techniques for pulmonary-thoracic segmentations and improvement of pneumonia diagnosis in pediatric chest radiographs," *Pediatric Pulmonology*, vol. 54, no. 10, pp. 1617–1626, Oct. 2019.

[17] X. Zeng, H. Chen, Y. Luo, and W. Ye, "Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30744–30753, 2019.

[18] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, p. 65, 2019.

[19] S. Wang, J. Sun, I. Mehmood, C. Pan, Y. Chen, and Y. Zhang, "Cerebral micro-bleeding identification based on a nine-layer convolutional neural network with stochastic pooling," *Concurrency Comput., Pract. Exper.*, vol. 32, no. 1, Jan. 2020, Art. no. e5130.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[21] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[22] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," 2015, *arXiv:1511.03643*. [Online]. Available: http://arxiv.org/abs/1511.03643

[23] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1910–1918.

[24] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*. [Online]. Available: http://arxiv.org/abs/1412.6550

[25] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4176–4182.

[26] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[27] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," 2014, *arXiv:1412.7755*. [Online]. Available: http://arxiv.org/abs/1412.7755

[28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*. [Online]. Available: http://arxiv.org/abs/1312.6034

[29] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9505–9515.

[30] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-Rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[31] Y. Gordienko, P. Gang, J. Hui, W. Zeng, Y. Kochura, O. Alienin, and S. Stirenko, "Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer," in *Proc. Int. Conf. Comput. Sci., Eng. Edu. Appl.*, Jan. 2018, pp. 638–647.

[32] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-I. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of Radiologists' detection of pulmonary nodules," *Amer. J. Roentgenology*, vol. 174, no. 1, pp. 71–74, Jan. 2000.

[33] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.

[34] S. Jaeger, S. Candemir, S. Antani, Y. X. J. Wáng, P. X. Lu, and G. Thoma, "Two public chest X-Ray datasets for computer-aided screening of pulmonary diseases," *Quantum Imag. Med. Surg.*, vol. 4, no. 6, p. 475, 2014.

[35] S. Ryoo and H. J. Kim, "Activities of the Korean institute of tuberculosis," *Osong Public Health Res. Perspect.*, vol. 5, pp. 43–49, 2014.

[36] Y. Tang, Y. Tang, V. Sandfort, J. Xiao, and R. M. Summers, "TUNA-Net: Task-oriented UNsupervised adversarial network for disease recognition in cross-domain chest X-rays," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2019, pp. 431–440.

[37] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in X-Rays using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 990–994.

[38] H. Wang and Y. Xia, "ChestNet: A deep neural network for classification of thoracic diseases on chest radiography," 2018, *arXiv:1807.03058*. [Online]. Available: http://arxiv.org/abs/1807.03058

[39] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-Rays," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9049–9058.

[40] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8290–8299.

[41] T. K. Ho and J. Gwak, "Multiple feature integration for classification of thoracic disease in chest radiography," *Appl. Sci.*, vol. 9, no. 19, p. 4130, Oct. 2019.

[42] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*. [Online]. Available: http://arxiv.org/abs/1312.6034

[43] B. Kim, J. Seo, S. Jeon, J. Koo, J. Choe, and T. Jeon, "Why are saliency maps noisy? Cause of and solution to noisy saliency maps," 2019, *arXiv:1902.04893*. [Online]. Available: http://arxiv.org/abs/1902.04893

[44] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K. R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, Aug. 2010.

[45] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*. [Online]. Available: http://arxiv.org/abs/1412.6806

[46] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, pp. 3319–3328, 2017.

[47] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*. [Online]. Available: http://arxiv.org/abs/1706.03825

[48] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *Proc. 2nd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2017, pp. 783–787.

[49] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "DenseNet: Implementing efficient ConvNet descriptor pyramids," 2014, *arXiv:1404.1869*. [Online]. Available: http://arxiv.org/abs/1404.1869

[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[51] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[52] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7096–7104.

[53] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1365–1374.

[54] S. Tang, L. Feng, W. Shao, Z. Kuang, W. Zhang, and Y. Chen, "Learning efficient detector with semi-supervised adaptive distillation," 2019, *arXiv:1901.00366*. [Online]. Available: http://arxiv.org/abs/1901.00366

[55] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 578–587.

[56] S.-I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," 2019, *arXiv:1902.03393*. [Online]. Available: http://arxiv.org/abs/1902.03393

[57] H. Wang, H. Jia, L. Lu, and Y. Xia, "Thorax-net: An attention regularized deep neural network for classification of thoracic diseases on chest radiography," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 475–485, Feb. 2020.

[58] S. Guendel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, "Learning to recognize abnormalities in chest X-rays with location-aware dense networks," in *Proc. Iberoamer. Congr. Pattern Recognit.* Cham, Switzerland: Springer, Nov. 2018, pp. 757–765.

[59] J. K. Gohagan, P. C. Prorok, R. B. Hayes, and B.-S. Kramer, "The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the national cancer institute: History, organization, and status," *Controlled Clin. Trials*, vol. 21, no. 6, pp. 251S–272S, Dec. 2000.

[60] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*. [Online]. Available: http://arxiv.org/abs/1901.07042

[61] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*. [Online]. Available: http://arxiv.org/abs/1710.10501

[62] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4694–4703.

THI KIEU KHANH HO received the B.S. degree in biomedical engineering from International University - Vietnam National University, in 2017. She is currently pursuing the master's degree with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju. From 2017 to 2018, she worked with the Optimal Dementia Care Research Center, GIST, and the Biomedical Research Institute, Seoul National University Hospital, Seoul, South Korea, as a Research Assistant. Her research interests include brain–computer interface, biomedical signal/image processing, bioinformatics, machine learning, and deep learning-based medical applications and systems.

JEONGHWAN GWAK received the Ph.D. degree in machine learning and artificial intelligence from the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2014. From 2002 to 2007, he had worked for several companies and research institutes as a Researcher and a Chief Technician. From 2014 to 2016, he worked as a Postdoctoral Researcher with GIST, and from 2016 to 2017 as a Research Professor. From 2017 to 2019, he was a Research Professor with the Department of Radiology, Biomedical Research Institute, Seoul National University Hospital, Seoul, South Korea. In 2019, he joined the Korea National University of Transportation (KNUT) as an Assistant Professor and he is currently the Director of the Applied Machine Intelligence Laboratory. His current research interests include deep learning, computer vision, image and video processing, evolutionary computation, optimization, and relevant applications of medical and visual surveillance systems.