IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Anomaly Detection in Medical Imaging with Deep Perceptual Autoencoders

NINA SHVETSOVA[1,5], BART BAKKER[2], IRINA FEDULOVA[1], HEINRICH SCHULZ[3], AND DMITRY V. DYLOV[4] (Member, IEEE)

[1]Philips Research, Moscow, Russia
[2]Philips Research, Eindhoven, Netherlands
[3]Philips Research, Hamburg, Germany
[4]Skolkovo Institute of Science and Technology, Moscow, Russia
[5]Goethe University Frankfurt, Frankfurt, Germany

Corresponding author: Dmitry V. Dylov (e-mail: d.dylov@skoltech.ru).

ABSTRACT

Anomaly detection is the problem of recognizing abnormal inputs based on the seen examples of normal data. Despite recent advances of deep learning in recognizing image anomalies, these methods still prove incapable of handling complex images, such as those encountered in the medical domain. Barely visible abnormalities in chest X-rays or metastases in lymph nodes on the scans of the pathology slides resemble normal images and are very difficult to detect. To address this problem, we introduce a new powerful method of image anomaly detection. It relies on the classical autoencoder approach with a redesigned training pipeline to handle high-resolution, complex images, and a robust way of computing an image abnormality score. We revisit the very problem statement of fully unsupervised anomaly detection, where no abnormal examples are provided during the model setup. We propose to relax this unrealistic assumption by using a very small number of anomalies of confined variability merely to initiate the search of hyperparameters of the model. We evaluate our solution on two medical datasets containing radiology and digital pathology images, where the state-of-the-art anomaly detection models, originally devised for natural image benchmarks, fail to perform sufficiently well. The proposed approach suggests a new baseline for anomaly detection in medical image analysis tasks[a].

[a]The source code is available at https://github.com/ninatu/anomaly_detection/

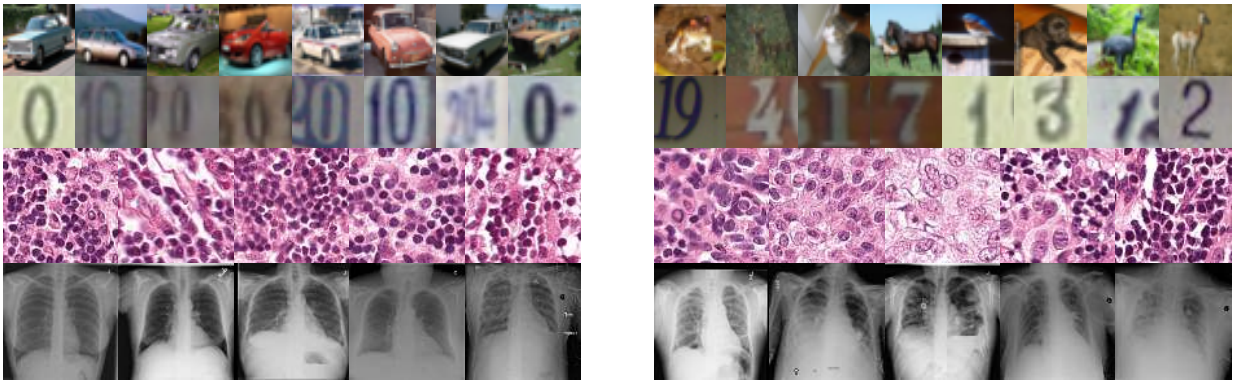
INDEX TERMS Anomaly Detection, Autoencoders, Chest X-Rays, Radiology, Digital Pathology

I. INTRODUCTION

ANOMALY detection is a crucial task in the deployment of machine learning models, where knowing the "normal" data samples should help spot the "abnormal" ones [5], [6]. If an input deviates from the training data substantially (*e.g.*, the input belongs to a class not represented in the training data), it is usually impossible to predict how the model will behave [7], [8]. This trait is especially important in high-consequence applications, such as medical decision support systems, where it is especially vital to know how to recognize the anomalous data. Identification of rare occurrences is another important application where anomaly detection is useful. For example, in pathology, where labeling diverse microscopy datasets is both time-consuming and expensive, the rare types of cells and tissues require specialized expertise from the annotator [9], [10]. Forthright anomaly classification and segmentation algorithms are typically prone to mistakes either because of the lack of sufficient annotation (thousands of labeled examples needed for supervised models) or because of the lack of representative data altogether (*e.g.*, the case of some rare pathologies). Moreover, these algorithms are affected by the need to deal with very unbalanced and *apriori* noisy data, frequently leading to inaccurate results (*e.g.*, the findings on chest x-rays can be so subtle that they can lead to disagreement in the interpretation [11], [12]). Because the normal cases greatly prevail over the abnormal ones, the anomaly detection could alleviate the annotation burden by automatically pointing to the rare samples.

In recent years, deep learning techniques achieved important advances in image anomaly detection [13]–[21].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107163, IEEE Access



Normal Abnormal

FIGURE 1: Examples of normal vs. abnormal images of considered datasets. Natural images: (first row) cars vs other classes of CIFAR10 dataset [1], (second row) digits "0" vs digits "1" – "9" of SVHN dataset [2]. Medical images: (third row) healthy tissue vs. tissue with metastases in H&E-stained lymph nodes images from Camelyon16 challenge [3], (fourth row) normal chest X-rays vs. chest X-rays with abnormal findings from NIH dataset [4].

However, these efforts were primarily focused on artificial problems with distinct anomalies in natural images (e.g., outliers in images of "cars" in the CIFAR10 dataset [1], see Figure 1). The medical anomalies, however, differ from those in the natural images [15], [19], [22]. Contrary to the natural images, the anomalies in the medical domain tend to strongly resemble the normal data. For example, detection of obscure neoplasms in chest X-rays [4] and of metastases in H&E-stained lymph node images [3] manifest a blatant challenge at hand, with the anomalous tissues being barely different from the normal ones (see Figure 1). Although deep learning has proved useful for a variety of biomedical tasks [23]–[26], only recently, a few groups started dedicating their effort to the anomaly detection problem [15], [19], [27]. However, to the best of our knowledge, a thorough comparison of the state-of-the-art (SOTA) solutions in the medical domain is still missing despite the pressing demand and the prospective clinical value.

In our paper, we evaluate and compare the strongest SOTA approaches ([15], [18], and [19]) on the two aforementioned medical imaging tasks. We find these methods either to struggle detecting such types of abnormalities, or to require a lot of time and resources for training. Moreover, the SOTA approaches lack a robust way of setting up model hyperparameters on *new* datasets, which complicates their use in the medical domain. Thus, we revisit the problem of image anomaly detection and introduce a new powerful approach, capable of tackling these challenges. The proposed method leverages the efficacy of *autoencoders* for anomaly detection [28], the expressiveness of *perceptual loss* [29] for understanding the content in the images, and the capabilities of the *progressive growth* [30] to approach training on high-dimensional image data.

Recent related studies [29], [31], [32] showed the effectiveness of deep features as a perceptual metric between images (the perceptual loss), and as a score of anomaly [18]. Also, the use of the perceptual loss for training autoencoders has been very popular in a variety of tasks [18], [29], [32]–[36] except – inexplicably– in the task of image anomaly detection where it has been somewhat dismissed so far. Trained only on normal data, the autoencoders tend to produce a high reconstruction error between the input and the output when the input is an abnormal sample. That property has been used intensively for anomaly detection [13], [14], [16], [37], [38]. Yet, we propose to compel the autoencoder to reconstruct *perceptive* or *content* information of the normal images, by using *only* the perceptual loss during autoencoder training. As such, the reconstructed image may not be an image altogether, but a tensor that stores the "content" of the original image. The main idea behind it is not to force the network to reconstruct a realistically looking image, but to let it be flexible in understanding the content of the normal data. Section III-A covers the details.

To further improve the expressiveness of the autoencoder and to allow it to capture even the fine details in the data, we propose to train the model using the progressive growing technique [30], [39], starting from a low-resolution network and adding new layers to gradually introduce additional details during the training. In particular, we present how to achieve a smooth growth of perceptual information in the loss function, and show that this greatly improves the quality of anomaly detection in the high-resolution medical data. We will describe it in Section III-B.

Lastly, we propose a new approach to the basic setup of anomaly detection model. Most approaches [13], [16]–[18], [20] prescribe not to use any anomaly examples during the model setup, dismissing the questions of optimization and of hyperparameter selection for such models. However, in reality, some types of abnormalities to detect are actually known (for example, the most frequent pathologies on the chest X-rays). Therefore, we consider the *weakly-supervised* scenario where a low number of anomalies with confined variability are available for use in optimal model hyperparameter selection (Section III-C). We believe this scenario

reflects the real tasks encountered in practice, provides a clear pipeline for setting up the model on new data, and helps to obtain reproducible results.

To summarize our main results quantitatively, the proposed solution achieves 93.4 ROC AUC[1] in the detection of metastases in H&E stained images of lymph nodes on Camelyon16 dataset [3], and 92.6 in the detection of abnormal chest X-rays on the subset of NIH dataset [4], which outperforms SOTA methods (by 2.8% and 5.2% in absolute value, respectively).

A. CONTRIBUTIONS

1) We compare the three strongest SOTA anomaly detection methods (the hyperparameters of which we fine-tuned to their optima) in two challenging medical tasks: Chest X-rays and H&E-stained histological images. To the best of our knowledge, this is the first candid comparison of anomaly detection models in digital pathology and in one of the largest Chest X-ray datasets available in the community[2] [4]. We also disclose the source code of all our experiments to facilitate the development of anomaly detection in medical imaging[3].
2) We introduce a new anomaly detection approach that utilizes the autoencoder with the perceptual loss. The proposed model is very easy to implement and train on new data, and it provides a strong anomaly detection baseline. We further extend the proposed method with progressive growing training (in particular, we introduce how to gradually grow the perceptual information in the loss function), allowing us to adapt the anomaly detection to the *high-resolution* medical data. The proposed solution outperforms SOTA methods on both datasets.
3) We revisit the training setup of the anomaly detection problem statement. To address the problem of choosing the model hyperparameters in the absence of the validation dataset, we propose to relax the unrealistic assumption that no abnormal examples are available during training. We show that even a small number of abnormal images (e.g., 0.5% of the training dataset) is enough to select the hyperparameters (outcome within 2% of the optimum). We believe that such a simple solution will standardize tuning of hyperparameters of different models, eliminate the ground for misunderstanding, and improve reproducibility.

II. RELATED WORK

Anomaly detection has been extensively studied in a wide range of domains, including but not being limited to fraud detection [40], cyber-intrusion detection [41], anomalies in videos [42], financial analytics [43], and the Internet of Things [44]. An extensive survey is out of the scope of our manuscript and can be found in [5], [6]. Herein, we will focus on anomaly detection in images.

Distribution-based methods. Conceptually, abnormal examples lie in low probability density areas of the "normal" data distribution; samples with a lower probability are thus more likely to be an anomaly. Distribution-based methods try to predict if the new example lies in the high-probability area or not. Kernel density estimation (KDE) [45] or Gaussian mixture models (GMM) [46] aims to model data distribution directly. One-class SVM [47], Isolation Forest [48], SVDD [49] methods create a boundary around normal examples. The latest methods extend classical solutions by using deep data representation. For example, Deep IF [19] successfully employed Isolation Forest on features extracted from a deep pre-trained network. DAGMM [50] proposed to use GMM on learned data representation. Deep SVDD [20] trains a network representation to minimize the volume of a hypersphere of the normal samples. However, the most critical part of such approaches is given in learning discriminative data representation. As shown in [19] anomaly detection performance may drop if there is a domain shift between the source dataset (for training data representation) and the target task.

Reconstruction-based methods. PCA and autoencoder-based [51] methods rely on the fact that the model trained only on normal data can not accurately reconstruct anomalies. Methods that use Generative Adversarial Networks (GANs), such as AnoGAN [15], use a similar idea: the generator, trained only on normal data, cannot generate abnormal images. The reconstruction error, thus, indicates the abnormalities. The latest methods broadly extend this idea by employing different combinations of autoencoders and adversarial losses of GAN's (OCGAN [16], GANomaly [52], ALOCC [53], DAOL [22] PIAD [18]), variational or robust autoencoders [37], energy-based models (DSEBM [13]), probabilistic interpretation of the latent space [54], [55], bi-directional GANs [56], memory blocks [14], etc. The main difficulties of such approaches are: choosing an effective dissimilarity metric and searching for the right degree of compression (the size of the bottleneck). The work [32] shows the extraordinary effectiveness of deep features as a perceptual dissimilarity metric; however, the very perceptual loss term was not considered in the anomaly detection problem. To the best of our knowledge, only [18] demonstrated a successful use of a perceptual metric for the anomaly detection task. We believe that a powerful dissimilarity measure is the key component of reconstruction-based methods. Below, we show that a simple, yet effective, combination of a deep autoencoder with the perceptual loss yields a suitably accurate anomaly detection baseline.

A recent model Deep GEO [17] employed a new method of image anomaly detection based on the idea of the self-supervised learning. The authors proposed to create a self-labeled dataset by applying different geometric transformations to images, with each geometric transformation (90°

[1] Area Under the Curve of Receiver Operating Characteristic in %.
[2] The only prior work [22] considered a portion of dataset [4].
[3] https://github.com/ninatu/anomaly_detection/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107163, IEEE Access

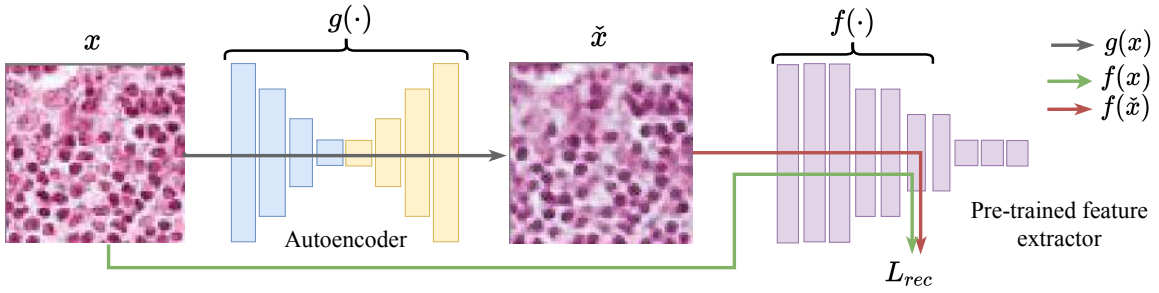


FIGURE 2: The proposed Deep Perceptual Autoencoder for image anomaly detection: g denotes the autoencoder network, f denotes a feature extractor, x is an image, and $\check{x} = g(x)$ is a reconstructed "image" (a *content* tensor). Reconstruction loss L_{rec} calculates difference between deep features $f(x)$ and $f(\check{x})$.

rotation, 180° rotation, etc.) being a new class in the dataset. After training a classifier on such a self-labeled dataset, the authors proposed to connect the abnormality of a new input to the average quality of the classification. Powerful in the natural image domain, this method, obviously, is sub-optimal for the homogeneously looking biomedical images (the rotation of which is pointless).

Despite a large number of anomaly detection methods that appeared in the recent years, only several papers [15], [19], [22], [57] included medical images in their experiments, still dismissing a detailed comparison with the latest strongest models. An accompanying problem here is the absence of standardized benchmark for the medical anomaly detection challenge. Herein, we fill these gaps by implementing major SOTA methods and by comparing their performance on two popular medical datasets with different types of abnormalities.

III. METHOD

A. DEEP PERCEPTUAL AUTOENCODER

Autoencoder-based approaches rely on the fact that autoencoders can learn shared patterns of the normal images and, then, restore them correctly. The key idea of our method is to simplify the learning of these common factors inherent to the data, by providing a loss function that measures "content dissimilarity" of the input and the output (the dissimilarity between the overall spatial structures present in the two images, without comparing the exact pixel values), which we called Deep Perceptual Autoencoder (DPA) (see Figure 2). It was shown that the perceptual loss – which computes a distance between the deep features obtained from an object classification neural network pre-trained on a large diverse dataset – can capture the content dissimilarity of the images [29], [31]. We further propose to use *nothing but* the perceptual loss to train the autoencoder and to compute the restoration error during the evaluation, without worrying about reconstructing the *whole* input information in the image by some other loss terms (e.g., no need for the output "image" to look realistic). We demonstrate that such a loss makes the autoencoder more flexible in gaining meaningful cues of the "normality" of the data, ultimately leading to much better anomaly detection results.

Let g be the autoencoder network, and x be an image. During the training, the autoencoder minimizes the difference between x and the reconstructed "image" $\check{x} = g(x)$, being called the reconstruction loss $L_{rec}(x, \check{x})$. To compute the perceptual loss as the reconstruction loss between x and $\check{x}$, we compute the difference between the deep features of these images ($f(x)$ and $f(\check{x})$, respectively). We adopt relative-perceptual-L1 loss from Ref. [18] as it is robust to noise and to changes in the image contrast:

$$L_{rec}(x, \check{x}) = \frac{\|\hat{f}(x) - \hat{f}(\check{x})\|_1}{\|\hat{f}(x)\|_1}, \tag{1}$$

where

$$\hat{f}(x) = \frac{f(x) - \mu}{\sigma} \tag{2}$$

are the normalized features, with the mean μ and the standard deviation σ of the filter responses of the layer being pre-calculated on a sufficiently large dataset. During the evaluation, the same $L_{rec}(x, g(x))$ is used to predict the abnormality for the new input x.

Remark. Deep Perceptual Autoencoder stems from the idea of PIAD [18] in that it relies on a strong image similarity metric. However, PIAD employs GANs for mapping the image distribution to the latent distribution and for the inverse mapping, which entails the following disadvantages: 1) GANs are hard to train (training is unstable and highly sensitive to hyperparameters); 2) GANs are time-consuming and resource-hungry (requiring alternative optimization of the generator and the discriminator); 3) GANs require additional hyperparameter tuning due to multi-objective loss function (the adversarial and the reconstruction loss terms). On the contrary, we propose the use of autoencoders that are much simpler to set up and train, while combining it with perceptual loss was shown to be more powerful than GANs utilizing the perceptual loss.

PIAD compels the latent vectors to be normally distributed and the reconstructed images to replicate the input images, while Deep Perceptual Autoencoders purposely discard these restrictions, so that the reconstructed image is a tensor that actually stores the "content" of the input image. We believe that this makes the model more flexible in capturing key fea-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107163, IEEE Access

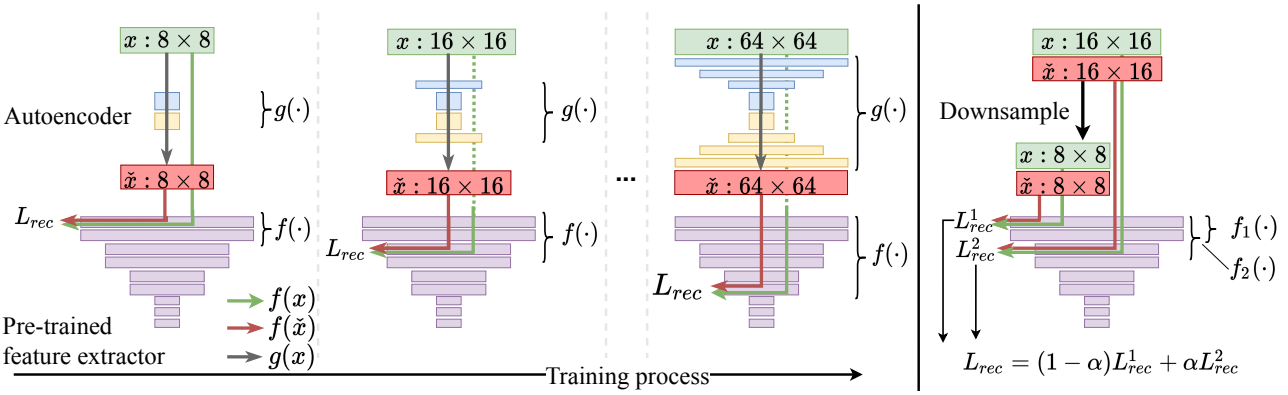


FIGURE 3: Progressive training process for high-resolution images. (Left) The layers are incrementally faded to the autoencoder g, and the depth of the features f increases synchronously. (Right) The corresponding gradual increase of the "resolution" of the perceptual loss L_{rec}.

tures responsible for determining the degree of "normality" of the data.

B. PROGRESSIVE GROWING

To improve the expressive power of the autoencoder, we propose to train it by harnessing the methodology of progressive growth [30]. Illustrated in Figure 3, the suggested pipeline gradually *grows* the "level" of the "perceptual" information in the loss function. In the beginning of the training, the loss function computes the dissimilarity between the low-resolution images using the features from the coarse layers of the network, but as the training advances, the "level" of abstraction is increased through including deeper features. It seems intuitively essential because the "content" information is absent in the low-resolution images, with only the main color and the high-level structure being stored there. The novelty that we propose in our solution, therefore, is to *synchronize* addition of the new layers to the autoencoder with the gradual increase of the depth of the features entailed in the calculation of the perceptual loss (see Figure 3 (Right)).

Both the autoencoder g and the perceptual loss L_{rec} have a low "resolution" in the beginning (Figure 3 (Left)). For example, the input and the output of the autoencoder are 8×8-pixel images x and $\check{x}$, and the loss L_{rec} computes the distance between the features $f(x)$ and $f(\check{x})$ of the coarse layer f (the pre-trained feature extractor network). As the training advances, the layers are incrementally added to the autoencoder g, and the depth of the features f is increased.

While doubling the resolution of the autoencoder, for example, from 8×8 to 16×16, the new layers are introduced smoothly, with the parameter α linearly increasing from 0 to 1. As it was proposed in [30], [39], during this process, both the input x and the output $\check{x}$ are the mixtures of the new high-resolution 16×16 image and the previous low-resolution 8×8 image, upsampled by a factor of two (not shown in Figure). In a similar manner, we smoothly increase the "level" of information supplied to L_{rec} from the features f_1 to the features f_2:

$$L_{rec} = \alpha * L_{rec}\big(f_2(x), f_2(\check{x})\big) + (1-\alpha) * L_{rec}\big(f_1(\text{down}(x)), f_1(\text{down}(\check{x}))\big), \quad (3)$$

where down$(\cdot)$ performs downsampling by a factor of two.

Thus, the training process consists of alternating the two routines: fixed-resolution training and resolution doubling. The training starts with a small autoencoder (with most of the layers excluded) on low-resolution images with the perceptual loss over the coarse features. Then we perform the "resolution doubling": we smoothly add new layers, scale up the resolution, and increase the depth of the features in the perceptual loss. Such an alternation is then repeated until the target image resolution is reached. Further details can be found in the released source code[3].

C. WEAKLY-SUPERVISED PARADIGM

Any anomaly detection model has many hyperparameters, the tuning of which is essential for the quality of the detection (in our method, these are the number of convolutions in the autoencoder, the size of the bottleneck, etc.). The majority of the anomaly detection papers declare no need to see the abnormal examples to set up their models, remaining vague with regard to how to choose the hyperparameters and how to deal with those cases when some new data needs to be analyzed by the same model. Some works mention tuning hyperparameters based on an unsupervised metric, like the value of the restoration error in the reconstruction-based methods [16], [19]. However, lower reconstruction loss does not mean better anomaly detection quality. For example, better reconstruction due to a larger bottleneck can cause the autoencoder to reconstruct anomalous data accurately as well.

In practice, however, one can have access to some labeled anomalies during the model setup. The number of such examples may be small, and they may not represent all possible abnormalities in the data, so it is typically tricky to use them in training. In our work, we propose a new *weakly-supervised* training paradigm where a low number of labeled anomalous

examples of a limited variation (i.e., a confined number of the types of anomalies) is available during the model setup as a "validation" or an "optimization" set.

This small set serves a single purpose – select the model's hyperparameters during its setup. Unlike works [19], [22] where a small subset of *all* anomalous data aims to improve the target performance, we propose to use a small subset of *limited* types of anomalies merely for the initiation of the model. This is a key difference because, in practice, it is difficult to cover all types of anomalies, even if just several examples of each. Below, we report extensive evaluation of how many types and examples of abnormalities are needed to validate the model. We show that even 20 abnormal images (which is less than 0.5% of the training set) are enough to select the model within 2% of the optimal one. The proposed paradigm reflects real-world clinical scenarios, allows candid comparison with standardized design of experiments, and provides a framework to ensure that the results are reproducible.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION PROTOCOL

We evaluated all approaches and SOTA baselines in the problem statement of *novelty detection*, where the training data are assumed to be free of anomalies.

1) Medical Images

To perform an extensive evaluation of anomaly detection methods in the medical domain, we examined two challenging medical problems with different image characteristics and the appearance of abnormalities.

a: Metastases Detection in Digital Pathology

Detecting metastases of lymph nodes is an extremely important variable in the diagnosis of breast cancer. However, the examination process is time-consuming and challenging. Figure 1 shows examples of the tumor and normal tissues. Tissues exhibiting metastasis may differ from healthy types only by texture, spatial structure, or distribution of nuclei, and can be easily confused with normal tissue. We considered the task of detecting metastases in H&E stained images of lymph nodes in the Camelyon16 challenge [3]. We trained anomaly detection models only on healthy tissue aiming to identify tissue with metastases. The training dataset of Camelyon16 consists of 110 tumour-containing whole slide images (WSIs) and 160 normal WSIs, while the testing dataset has 80 regular WSIs and 50 tumorigenic WSIs.

For all slides, we performed the following preprocessing. Firstly, we divided tissue from the background by applying Otsu's thresholding [58]. Then we randomly sampled 768x768 tiles (maximum 50 from one slide without overlapping) of healthy tissue (from entirely normal images) and tumor tissue (from slides with metastases) and performed color normalization [59]. For the hyperparameter search we sampled tiles only from 4 out of 110 train tumor slides (validation set of confined variability). We obtained 7612 normal training images, 200 tumor images for validation, and 4000 (normal) + 817 (tumor) images for the test. During training, we randomly sampled 256x256 crops from 768x768 normalized tiles, and to the test, we used only a central 256x256 crop (to reduce border effect during normalization). The original WSIs were done with 40x magnification of tissue, but during the hyperparameter search, we also considered x10 and x20 magnification by bilinear downsampling images (256x256 to 128x128, and 64x64).

b: Anomaly Detection on Chest X-Rays

Chest X-ray is one of the most common examinations for diagnosing various lung diseases. We considered the task of the recognition of fourteen findings, such as Atelectasis or Cardiomegaly, on the chest X-rays in the NIH dataset (ChestX-ray14 dataset) [4] (Figure 1). Searching abnormalities on a chest x-ray is challenging even for an experienced radiologist since abnormality may occupy only a small region of lungs, or be almost invisible. The dataset consists of 112,120 frontal-view images of 30,805 unique patients: 86523 for training, 25595 for evaluation. We split the dataset into two sub-datasets having only posteroanterior (PA) or anteroposterior (AP) projections, because organs on them look differently. We tried different preprocessing during the hyperparameter search: rescaling to 256x256, 128x128, and 64x64 and histogram equalization, central crop (3/4 of the image size) to delete "noisy" borders. We considered images without any disease marker as "normal" and used them for training. Abnormal images for hyperparameter searching comprised of the training images of the most frequent disease ('Infiltration') out of fourteen possibilities. We also evaluated model on subset containing "clearer" normal/abnormal cases (provided by [22]). This subset consists of 4261 normal images for training, 849 normal and 857 abnormal images for validation, and 677 normal and 677 abnormal images for testing.

2) Natural Images

We also evaluate the methods on two natural image benchmarks CIFAR10 [1] and SVHN [2]. Both datasets provide an official train-test split and consist of 10 classes. Following previous works [13], [16]–[20], we used a one-vs-all evaluation protocol: we design 10 different experiments, where only one class is alternately considered as normal, while others treated as abnormal. In all experiments, images were rescaled to 32x32 resolution. During the hyperparameter search, we tried different preprocessing: 1) using the original RGB-images and 2) converting the images to grayscale. We randomly sampled one abnormal class of the train set as a validation set with abnormal images (that has only one type of abnormalities out of nine). These conditions were fixed in all methods compared beneath.

3) Metrics

Following the convention in the field of anomaly detection [16]–[20], [22], [52], we used the Area Under the Curve of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107163, IEEE Access

	AnoGAN	GANomaly	DAGMM	DSEBM	DeepSVDD	OCGAN	DeepGEO	PIAD	Deep IF	Ours (w/o p. g.)
CIFAR10	57.6/-	58.1/-	57.5/-	58.8/-	64.8/-	65.7/-	86.6/86.5	78.8/81.3	87.2/**87.3**	83.9
SVHN	53.3/-	-	51.8/-	57.1/-	57.3/-	-	93.3/**93.5**	77.0/76.3	59.0/62.4	80.3

TABLE 1: ROC AUC in % for CIFAR10 and SVHN datasets averaged over all ten experiments in the dataset (see Section IV-A2) and over three different runs per experiment (each experiment we repeated three times with different model initialization). For methods results are reported in two options: ROC AUC obtained with authors' default hyperparameters (left), ROC AUC obtained with hyperparameters found by cross-validation in weakly-supervised paradigm (right).

Receiver Operating Characteristic (ROC AUC) as an evaluation metric that integrates the classification performance (normal *vs.* abnormal) for all decision thresholds. This precludes from the need to choose a threshold for the predicted abnormality scores, allowing for assessing the performance of the models "probabilistically" and without a bias.

B. BASELINES.

We considered the following strongest SOTA baselines of different paradigms: Deep GEO [17] Deep IF [19] and PIAD [18]. On natural images we also competed against AnoGAN [15], GANomaly [52], DAGMM [50], DSEBM [13], DeepSVDD [20], and OCGAN [16] methods. On the NIH dataset, we also compared our results to DAOL framework [22], [60], purposely developed for detecting anomalies in chest X-rays.

C. IMPLEMENTATION DETAILS.

We implemented Deep IF and PIAD approaches using extensive descriptions provided by the authors. For GANomaly and Deep GEO, we adapted the official code for our experiment setups. Results of DAOL and OCGAN methods were obtained in the corresponding papers. For other approaches, we used results as reported in [19]. For the strongest baselines, we also perform a hyperparameter tuning precisely as it is proposed herein (for a standardized comparison). For the Deep GEO approach, we searched for an optimal number of the classifier's training epochs (we find the method to be sensitive to this parameter). For Deep IF, we searched for the best feature representation – the best layer of the feature extractor network. For PIAD, we searched for the optimal size of the latent vector, the best feature layer in the relative-perceptual-L1 loss, and the most suitable number of training epochs. Also, for all algorithms, we searched for the best image preprocessing to yield the highest scores.

For the proposed approach, in all experiments, we used autoencoders with pre-activation residual blocks. For the computation of the relative-perceptual-L1 loss, we used the VGG19 network that was pre-trained on *ImageNet*. We trained the autoencoder until after the loss on the hold-out set of normal images stops decreasing. During the hyperparameter selection, we search for the best size of the autoencoder bottleneck and for the best feature layer of relative-perceptual-L1 loss. Further details are covered in the released algorithm code[3].

Hyperparameter search was performed by maximizing average ROC AUC over 3 "folds". Namely, we split training normal data into 3 subsets (as a 3-fold cross-validation) and did the same for validating the abnormal data. In each run, we left one normal subset and one abnormal subset for testing, and used the other two normal subsets for training. Only in the experiments with the NIH subset, Ref. [22], we did not perform cross-validation but ran the experiment thrice on the same train-validation split to precisely replicate the DAOL experimental settings.

D. RESULTS

1) Natural Images

As mentioned above, for CIFAR10 and SVHN datasets, we conducted ten experiments, where each class alternatively was considered normal. In such experiments, an anomaly is an image of an object of a different class. Therefore, abnormal images are very different from normal data (compared to anomalies on medical images), but normal data also have high variability. The average results over all experiments in a dataset are reported in Table 1. Notice, while testing on these datasets, we do not use progressive growth in our method because the image resolution is only 32×32.

The approaches that we called the strongest baselines (Deep GEO, PIAD, Deep IF) and our method significantly outperform other methods, with margin 20% (except for Deep IF on SVHN dataset). The Deep GEO approach, which classifies the geometric transformations of images, excels in distinguishing digits from each other (SVHN dataset). The reason for that is that digits have a simple geometrical structure, and their geometric transformations are easily distinguishable. Our approach shows the second-best result. However, Deep IF fails – features obtained by the *ImageNet*-pre-trained neural network turned out to be not discriminative for this task.

Images of CIFAR10 dataset have a more challenging geometrical structure than SVHN ones, hence Deep GEO shows lower performance. However, since the domain shift between *ImageNet* and CIFAR10 dataset is smaller, Deep IF also shows good results. Our proposed method closely approaches the reported performance of the leaders Deep GEO and Deep IF. We noticed that our approach in both datasets outperformed PIAD by $\sim$3%.

We report that reconstruction-based approaches, like ours and PIAD, are inferior to Deep GEO for the natural image tasks. We hypothesize that this stems from the high variability of the normal images, and the autoencoder overgeneralizes on the anomaly data. Indeed, during hyperparameter tuning, we search for the optimal autoencoder capacity, where

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107163, IEEE Access

	DAOL	Deep GEO	PIAD	Deep IF	Ours (w/o p. g.)	Ours (with p. g.)
Camelyon16	-	$52.4 \pm 11.1/45.9 \pm 2.1$	$85.4 \pm 2.0/89.5 \pm 0.6$	$87.6 \pm 1.5/90.6 \pm 0.3$	92.7 ± 0.4	$\mathbf{93.4 \pm 0.3}$
NIH (a subset)	$-/80.5 \pm 2.1$	$85.8 \pm 0.6/ 85.3 \pm 1.0$	$88.0 \pm 1.1/87.3 \pm 0.9$	$76.6 \pm 2.7 /85.3 \pm 0.4$	92.0 ± 0.2	$\mathbf{92.6 \pm 0.2}$
NIH (PA proj.)	-	$60.2 \pm 2.6/63.6 \pm 0.6$	$68.0 \pm 0.2/68.7 \pm 0.5$	$52.2 \pm 0.5/47.2 \pm 0.4$	70.3 ± 0.2	$\mathbf{70.8 \pm 0.1}$
NIH (AP proj.)	-	$53.1 \pm 0.3/54.4 \pm 0.6$	$57.4 \pm 0.4/\mathbf{58.6 \pm 0.3}$	$54.3 \pm 0.5/56.1 \pm 0.2$	$\mathbf{58.6 \pm 0.1}$	58.5 ± 0.0

TABLE 2: ROC AUC in % with standard deviation (over 3 runs). For baselines results are reported in two options: ROC AUC obtained with authors' default hyperparameters (left), ROC AUC obtained with hyperparameters found by cross-validation in weakly-supervised paradigm (right). For our method, results are showed with and without progressive growing regime of training.

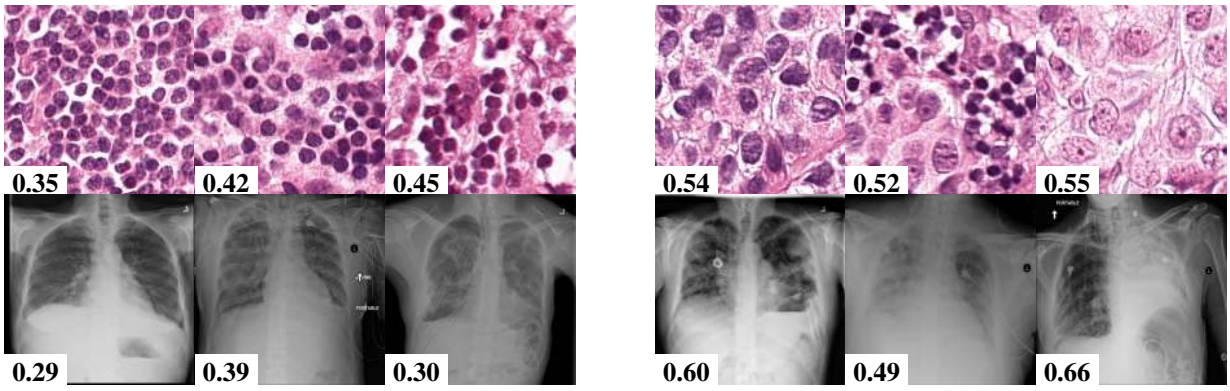


FIGURE 4: Examples of normal (left) and anomaly (right) images of H&E-stained lymph node of Camelyon16 challenge [3] (top) and chest X-rays of NIH dataset [4] (bottom). We also showed the predicted anomaly score by the proposed method. The higher the score, the more likely to be an anomaly. Notice how the proposed method spots even the borderline cases.

	PIAD	Ours (w/o p. g.)	Ours (with p. g.)
Camyleon16	84	105	160
NIH (a sub.)	287	59	177
NIH (PA proj.)	151	89	159
NIH (AP proj.)	275	70	107

TABLE 3: Average training time (minutes). Experiments were run on GeForce GTX 1080 Ti with Pytorch 1.4.0.

the autoencoder reconstructs the normal data well but does not generalize on the other data. However, when the training data is highly variable, the autoencoder generalizes better on the unseen classes.

2) Medical Images

Given Deep GEO, PIAD, and Deep IF are superior to the other methods on natural image datasets, we chose them for evaluation on the medical images too, where such a diverse variability is absent (see Table 2). For our method, we report results obtained with and without the progressive growing training regime.

Remarkably, our approach significantly outperforms Deep GEO and Deep IF in both medical datasets. The Deep GEO shows poor performance on the digital pathology data (only 52.4% with the default hyperparameters proposed by the authors), where the images are invariant to geometric transformations. Indeed, digital pathology scans do not have space orientation; rotations and translations of them are not distinguishable. We can tell that the Deep GEO is not applicable to such data. For NIH PA and AP projections, ROC AUC's are also very low. Our hypothesis is that if abnormality occupy a small region of the image, the classifier still distinguishes the geometric transformations well, so the quality clasification hardly indicates such abnormalities. For NIH (a subset) with more "obvious" abnormalities (the abnormal region is larger), the Deep GEO approach shows better results.

Deep IF displayed inferior performance on the NIH (PA proj.) and NIH (AP proj.) datasets, achieving ROC AUC scores of 52.2% and 54.3% respectively (Table 2, default hyperparameters proposed by the authors). This might probably be due to the domain shift between ImageNet and X-ray images, whereby the features obtained from the pre-trained network turned out to be not discriminative for this task. However, for the Camelyon16 and NIH (a subset) experiments, the ROC AUC is quite high. We postulate that if the feature extractor network was pre-trained on images which closely-resembled those of the X-ray image analysis task, Deep IF would demonstrate a significantly-improved performance on the NIH (PA proj.) and the NIH (AP proj.) datasets as well. We conclude that the weakest side of this approach is feature representation. Currently, there exists no explicit algorithm for obtaining discriminative feature spaces for differing experimental setups.

It is worth noting that our method also uses the features pre-trained on *ImageNet*, but for image comparison (rather than image representation), since the learnt features (though not being discriminative due to the domain shift),) might still prove useful in determining the key differences between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107163, IEEE Access

	Camelyon16	NIH (a subset)	NIH (PA proj.)	NIH (AP proj.)
(1) L1 + unsupervised	21.1 ± 1.4	70.8 ± 0.6	66.5 ± 0.1	52.4 ± 0.1
(2) PL + unsupervised	87.9 ± 0.6	89.3 ± 0.2	68.9 ± 0.1	56.4 ± 0.2
(3) PL + weakly-supervised	92.7 ± 0.4	92.0 ± 0.2	70.3 ± 0.2	$\mathbf{58.6 \pm 0.1}$
(4) PL + 1· adv + weakly-supervised	79.4 ± 4.0	64.4 ± 7.8	52.3 ± 3.3	51.5 ± 3.4
(5) PL + 0.1· adv + weakly-supervised	90.8 ± 0.7	82.2 ± 2.6	59.2 ± 1.4	55.4 ± 0.9
(6) PL + 1· L1 + weakly-supervised	75.3 ± 1.6	91.7 ± 0.4	70.7 ± 0.2	57.3 ± 0.1
(7) PL + 0.1· L1 + weakly-supervised	93.0 ± 0.3	92.0 ± 0.1	70.6 ± 0.2	58.5 ± 0.1
(8) PL + 1· L1 + 1· adv + weakly-supervised	57.5 ± 6.3	59.3 ± 5.0	50.1 ± 2.0	51.7 ± 0.8
(9) PL + 0.1· L1 + 0.1· adv + weakly-supervised	90.6 ± 1.0	78.2 ± 1.0	60.8 ± 1.8	55.5 ± 0.4
(10) PL + weakly-supervised + progressive growing	$\mathbf{93.4 \pm 0.3}$	$\mathbf{92.6 \pm 0.2}$	$\mathbf{70.8 \pm 0.1}$	58.5 ± 0.0

TABLE 4: Ablation study. ROC AUC in % with standard deviation (over 3 runs).

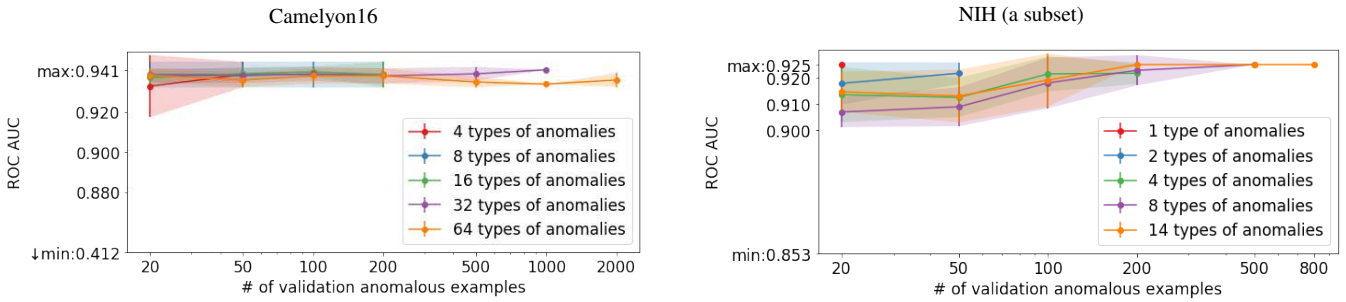


FIGURE 5: Dependence of the quality of anomaly detection (of our approach) on the number of anomaly examples (the x-axis) and their variability (the different lines) in the validation set. The highest (max) and the lowest (min) performance achievable on these hyperparameter spaces are shown on the plots. For Camelyon16, we consider metastases tiles from one slide as abnormality of one type, for NIH dataset, type of abrnomality is unique finding. We used the same 3-fold cross-validation split and hyperparameter space as in previous experiments. We sampled a validation set for each configuration (# anomaly types, # anomaly examples) seven times. For each sample of the validation set, we selected the best hyperparameters on the cross-validation split. Then we evaluated the quality of the model trained on all training images with chosen hyperparameters on test split. Here we showed mean and std of test ROC AUC's (computed over three samples of the validation set for this configuration)

images (via the perceptual loss).

With a smaller margin, our algorithm (in both options: with and without progressive growing) is also ahead of the PIAD method. We would like to highlight that our method is generally easier and faster to train while being the least resource-hungry amongst the models considered herein. The average time of training of the final models is provided in Table 3. Moreover, the use of the progressive growing training approach gained us an additional 1% in the image quality (as depicted by the ROC AUC scores in Table 2).

We illustrate the predictions of our model in Figure 4.

E. HYPERPARAMETERS TUNING ANALYSIS

While proposing the use of a small and restricted set of anomalies during the model setup, we naturally asked exactly how many anomaly types and how many anomaly examples would be required. In Figure 5, we demonstrate the dependence of the quality of the anomaly detection on the number of anomaly examples and their variability in the validation set. The experiment shows that even a small number of abnormal samples (for example, 20) of *one type* of anomaly is enough, to "reject" inferior hyperparameter configurations. In the two experiments considered, having 20 abnormal examples of the same type of abnormality (which is less than 0.5% of the normal examples used in training) proved sufficient to select the hyperparameters within the 2% margin of the optimal configuration.

F. ABLATION STUDY

To stress the importance of every component proposed herein, we performed an extensive ablation study. Table 4 considers ten ablation scenarios.

(1): Autoencoder (AE) training with the L1 loss and the hyperparameter optimization using *unsupervised* criteria (the reconstruction loss).

(2): The same, but with L1 replaced by perceptual loss (PL).

(3): The previous one with the hyperparameters corresponding to the best validation ROC AUC (*weakly-supervised* scenario).

(4)–(9): Here, we added the adversarial loss (with weights 1 and 0.1) or L1 norm, or both of them to the loss function during the training (to force the reconstructed image to have a realistic look or to restore the whole input image). To compute adversarial loss we trained the discriminator jointly with autoencoder using Wasserstein GAN with a Gradient Penalty objective [61].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3107163, IEEE Access

TABLE 5: Results on natural images. ROC AUC in % with std in each experimental configuration on the CIFAR 10 and SVHN datasets (each experiment was repeated three times with different model initialization). We used a one-vs-all evaluation protocol: 10 different experiments for each dataset were designed, where only one class (a column name) is alternately considered as normal, while the others are treated as abnormal. For these methods, the results are reported for two options: ROC AUC obtained with the default hyperparameters proposed by the authors of corresponding works (default), and ROC AUC obtained with the hyperparameters found by cross-validation in weakly-supervised paradigm (weakly s.))

		CIFAR10									
	hyperparams	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
Deep GEO	default	75.4±0.9	96.0±0.2	79.9±1.9	73.6±0.2	87.4±0.4	**87.6±0.7**	85.2±0.9	95.1±0.1	94.3±0.0	91.3±0.3
	weakly-s.	75.7±1.0	96.0±0.2	**80.4±1.1**	72.9±0.9	88.0±0.2	86.3±0.9	84.6±0.5	**95.4±0.0**	**94.3±0.2**	91.4±0.5
PIAD	default	81.8±0.1	87.1±0.3	74.9±0.3	60.7±0.2	78.1±0.5	70.6±1.4	81.7±0.8	84.4±0.4	86.3±0.4	82.3±0.6
	weakly-s.	84.3±0.2	86.7±1.1	74.4±0.9	59.6±2.1	85.0±1.1	73.6±1.1	83.8±1.2	87.0±1.1	88.8±0.2	89.4±0.7
Deep IF	default	85.2±1.2	94.3±0.4	72.5±4.0	**76.8±1.2**	**89.9±0.7**	86.1±1.0	90.3±1.7	89.1±1.0	92.0±1.0	95.6±0.1
	weakly-s.	**87.1±0.9**	**97.0±0.3**	75.2±2.9	73.7±1.8	88.9±1.0	85.0±2.6	**90.5±0.9**	86.3±1.7	93.4±0.3	**95.7±0.3**
Ours	weakly-s.	86.5±0.2	92.2±0.3	76.8±0.6	58.7±1.2	85.1±0.4	77.7±0.9	88.9±0.1	89.1±0.2	91.4±0.5	92.2±0.4

		SVHN									
		0	1	2	3	4	5	6	7	8	9
Deep GEO	default	89.0±0.6	84.1±1.1	96.9±0.1	**91.3±0.3**	97.3±0.0	96.2±0.3	96.0±0.2	98.2±0.1	**86.4±0.3**	97.4±0.1
	weakly-s.	**90.6±0.6**	**84.8±0.6**	**97.2±0.2**	91.1±0.1	**97.5±0.1**	**96.3±0.0**	**96.2±0.2**	**98.4±0.0**	85.6±0.9	**97.6±0.2**
PIAD	default	85.6±0.4	79.2±1.1	74.8±0.4	69.2±0.0	77.3±0.8	74.6±0.9	76.3±0.7	77.5±0.4	78.0±0.2	77.4±0.3
	weakly-s.	86.3±0.9	80.2±0.9	76.2±0.8	71.4±1.1	77.0±0.5	71.9±0.9	70.6±0.5	78.4±0.2	79.6±0.7	71.9±0.8
Deep IF	default	65.3±1.1	68.7±1.8	51.9±0.8	57.1±1.5	56.7±2.1	64.9±1.5	50.9±0.8	56.2±1.9	63.7±0.9	54.3±1.1
	weakly-s.	75.0±1.2	70.5±1.5	51.1±0.8	59.0±0.9	57.7±1.4	68.4±0.6	54.5±0.2	58.7±0.7	69.6±1.6	59.1±1.5
Ours	weakly-s.	88.4±0.2	82.7±0.8	80.0±0.8	72.9±0.1	79.1±0.7	77.4±0.7	78.0±0.8	79.0±0.2	83.5±0.2	82.1±0.3

(10): The last training scenario finally considers the progressive growing.

Remarkably, the use of the perceptual loss **(2)** outperforms the mere L1 norm **(1)** with a large margin. We also observed that the method of selecting the hyperparameters by revealing a subset of anomalies of confined variability **(3)** noticeably benefits the anomaly detection performance (when compared to the unsupervised criteria **(2)**). We also note the advantage of our approach over the autoencoder that encourages fully restored or realistically looking images **(4)–(9)** (using additional adversarial or L1 norm loss). We thus confirm our hypothesis that the use of the perceptual loss alone provides the autoencoder with more flexibility to gain a meaningful interpretation of the "normality". Our experiments demonstrate that the additional losses only deteriorate the performance. Finally, the proposed progressive growing technique **(10)** allow us to gain further improvements, solidifying the entire model as the new medical image anomaly detection baseline.

G. LIMITATIONS

The proposed method excelled on the medical datasets but was unable to outperform the assayed SOTA on natural image baselines. We, thus, hypothesize that the method may be better suited for data having low to medium variability (such as the medical images from the same image acquisition modality) rather than high variability datasets (such as CIFAR10 or SVHN). The high diversity present in the natural data may lead to overgeneralization of the autoencoder: in this case, the autoencoder may produce low reconstruction error even for the abnormal data. A controlled study on the dependence of the performance of all methods on data diversity requires specifically prepared datasets (perhaps, with synthetic anomalies and a strict measure of 'data diversity') and will be the subject of future work. Still, one may argue that a very sensitive anomaly detection algorithm, capable of discriminating against similarly-looking normal data and tuned for a given imaging modality, would actually be preferred by individual imaging niches over more generalized but less sensitive methods.

V. CONCLUSIONS

In this manuscript, we evaluated a range of state-of-the-art image anomaly detection methods, the performance of which we found to be sub-optimal in the challenging medical problems. We proposed a new method that uses an autoencoder to understand the representation of the normal data, with optimization being performed with regard to perceptual loss in the regime of progressive growing training. To overcome the problem of setting up the model on the new data, we propose to use a small set of anomalous examples of a limited variation simply for selecting the model's hyperparameters. We believe that this realization reflects real-world clinical scenarios, allows consistent structuring of the experiments, and enables the generation of reproducible results in the future. The proposed approach achieved 0.934 ROC AUC in the detection of metastases and 0.926 in the detection of abnormal chest X-rays. Our work, thus, establishes a new strong baseline for anomaly detection in medical imaging.

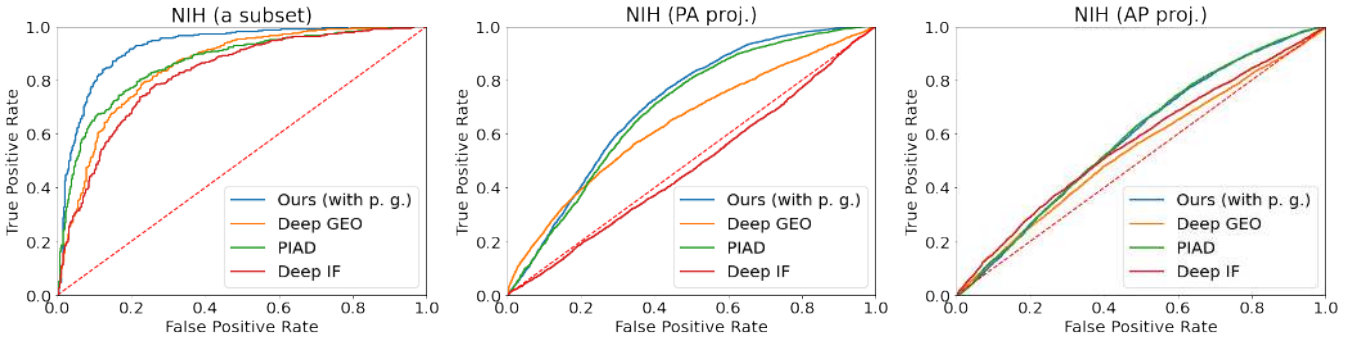


FIGURE 6: ROC curves of Deep GEO, PIAD, Deep IF, and the proposed model on the NIH (a subset), NIH (PA proj.), and NIH (AP proj.) datasets.

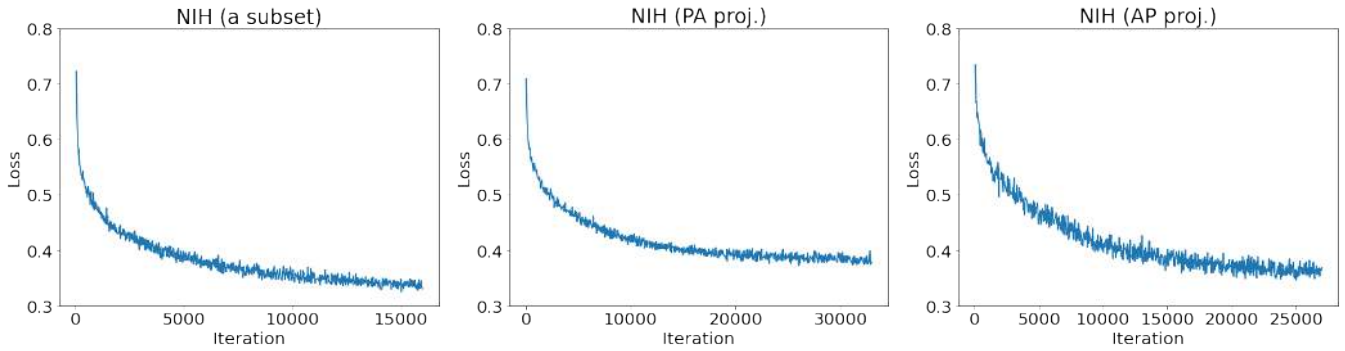


FIGURE 7: Training performance curves of the proposed method on the NIH (a subset), NIH (PA proj.), and NIH (AP proj.) datasets.

.

APPENDIX A

Table 5 reports additional details on average ROC AUC values with standard deviation calculated over 3 runs for every experimental configuration on the CIFAR10 and the SVHN datasets. We considered ten different configurations for each dataset, where one label is designated as normal while the others are considered abnormal.

We also show ROC curves for the considered methods on the NIH (a subset), NIH (PA proj.), and NIH (AP proj.) datasets in Figure 6 and the training performance graph (loss *vs.* iteration) of the proposed method in Figure 7.

REFERENCES

[1] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[2] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[3] 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 *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097–2106.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[6] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[7] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 427–436.

[8] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[9] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images." *Journal of the American Medical Informatics Association: JAMIA*, vol. 20, no. 6, p. 1099–1108, 2013.

[10] A. Chowdhury, D. V. Dylov, Q. Li, M. MacDonald, D. E. Meyer, M. Marino, and A. Santamaria-Pang, "Blood vessel characterization using virtual 3d models and convolutional neural networks in fluorescence microscopy," *IEEE ISBI 2017*, pp. 629–632, April 2017.

[11] T. Olatunji, L. Yao, B. Covington, and A. Upton, "Caveats in generating medical imaging labels from radiology reports with natural language processing," in *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*, 2019.

[12] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest x-ray analysis: A survey," *Medical Image Analysis*, p. 102125, 2021.

[13] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," *arXiv preprint arXiv:1605.07717*, 2016.

[14] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 1705–1714.

[15] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 146–157.

[16] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2898–2906.

[17] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Advances in Neural Information Processing Systems*, 2018, pp. 9758–9769.

[18] N. Tuluptceva, B. Bakker, I. Fedulova, and A. Konushin, "Perceptual image anomaly detection," in *Pattern Recognition*, S. Palaiahnakote, G. Sanniti di Baja, L. Wang, and W. Q. Yan, Eds. Cham: Springer International Publishing, 2020, pp. 164–178.

[19] K. Ouardini, H. Yang, B. Unnikrishnan, M. Romain, C. Garcin, H. Zenati, J. P. Campbell, M. F. Chiang, J. Kalpathy-Cramer, V. Chandrasekhar *et al.*, "Towards practical unsupervised anomaly detection on retinal images," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 2019, pp. 225–234.

[20] L. Ruff, N. Görnitz, L. Deecke, S. A. Siddiqui, R. Vandermeulen, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International Conference on Machine Learning*, 2018, pp. 4390–4399.

[21] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study," *Medical Image Analysis*, p. 101952, 2021.

[22] Y.-X. Tang, Y.-B. Tang, M. Han, J. Xiao, and R. M. Summers, "Deep adversarial one-class learning for normal and abnormal chest radiograph classification," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950. International Society for Optics and Photonics, 2019, p. 1095018.

[23] B. P. Nguyen *et al.*, "Prediction of fmn binding sites in electron transport chains based on 2-d cnn and pssm profiles," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[25] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[26] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H.-j. Bae, and N. Kim, "Deep learning in medical imaging," *Neurospine*, vol. 16, no. 4, p. 657, 2019.

[27] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Scale-space autoencoders for unsupervised anomaly segmentation in brain mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 552–561.

[28] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[29] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[30] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[31] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[33] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5933–5942.

[34] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the ECCV*, 2018, pp. 172–189.

[35] G. Grund Pihlgren, F. Sandin, and M. Liwicki, "Improving image autoencoder embeddings with perceptual loss," in *International Joint Conference on Neural Networks*, 2020.

[36] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[37] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 665–674.

[38] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International Symposium on Neural Networks*. Springer, 2017, pp. 189–196.

[39] A. Heljakka, A. Solin, and J. Kannala, "Pioneer networks: Progressively growing generative autoencoder," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 22–38.

[40] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.

[41] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, pp. 1–13, 2017.

[42] B. Kiran, D. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[43] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems*, vol. 55, pp. 278–288, 2016.

[44] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for iot big data and streaming analytics: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2923–2960, 2018.

[45] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[46] G. J. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.

[47] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class svm for learning in image retrieval." in *ICIP (1)*. Citeseer, 2001, pp. 34–37.

[48] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[49] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[50] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations*, 2018.

[51] G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu, "A comparative study of rnn for outlier detection in data mining," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE, 2002, pp. 709–712.

[52] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*. Springer, 2018, pp. 622–637.

[53] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3379–3388.

[54] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 481–490.

[55] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, pp. 1–18, 2015.

[56] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially learned anomaly detection," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 727–736.

[57] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders," *arXiv preprint arXiv:1806.04972*, 2018.

[58] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[59] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[60] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[61] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.