Towards Visual Anomaly Detection in Domains with Limited Amount of Labeled Data

Dejan Štepec
XLAB Research
XLAB d.o.o.
Ljubljana, Slovenia
dejan.stepec@xlab.si

Danijel Skočaj
Faculty of Computer and Information Science
University of Ljubljana
Ljubljana, Slovenia
danijel.skocaj@fri.uni-lj.si

***Abstract*—Anomaly detection in visual data refers to the problem of differentiating abnormal appearances from normal cases. Supervised approaches have been successfully applied to different domains, but require abundance of labeled data. Due to the nature of how anomalies occur and their underlying generating processes, it is hard to characterize and label them. Recent advances in deep generative based models have sparked interest towards applying such methods for unsupervised anomaly detection and have shown promising results in medical and industrial inspection domains.**

***Index Terms*—anomaly detection, unsupervised, deep-learning, autoencoders, generative adversarial networks**

I. Introduction

Anomaly detection represents an important process of determining instances that stand out from the rest of the data. Detecting such occurrences in different data modalities is widely applicable in different domains such as fraud detection, cyber-intrusion, industrial inspection and medical imaging [1]. Detecting anomalies in high-dimensional data (e.g. images) is a particularly challenging problem that has recently seen a particular rise of interest, due to prevalence of deep-learning based methods.

Success of current deep-learning based methods has mostly relied on abundance of available data. Anomalies generally occur rarely, in different shapes and forms and are thus extremely hard or even impossible to label. Supervised deep-learning approaches have seen great success in different domains, including in anomaly detection [2]–[4]. Success of such methods is the most evident in the domains with well known characterization of the anomalies and abundance of labeled data. Specific to the visual anomaly detection domain, we usually also want to localize the actual anomalous region in the image. Obtaining such detailed labels to learn supervised models is a costly process and in many cases also impossible. Weakly-supervised approaches address such problems by requiring only image-level labels and are thus able to infer anomalous regions solely from weakly labeled data [5]–[7]. In an unsupervised setting, only normal samples are available, which are usually available in abundance. Such methods represent the most general case and are the most widely applicable. Deep generative methods have been recently applied to the problem of unsupervised anomaly detection (UAD) and have shown promising results [8], [9]. Current methods are usually developed for a particular domain or on synthetic datasets which limits their generality, as well applicability to real-world applications. They are also not really unsupervised, requiring only normal samples, with significant drops in performance with the presence of small amount of contaminated training data [10], [11].

In this work we focus on anomaly detection from images, which was just briefly mentioned in one of the most significant papers on general anomaly detection [1]. This clearly shows the state of this domain before the era of deep-learning. There have been a lot of advancements in recent years in the visual anomaly detection domain, but there is no survey work that clearly summarizes them. Most of the existing survey papers are addressing the wider scope of anomaly detection problem, lacking the focus on visual anomaly detection and its recent advancements [1]. Some survey papers are addressing recently popular deep-learning based anomaly detection approaches [2], but are describing applications to the broader field of anomaly detection and are not focusing on particular methods and application domains related to visual anomaly detection. Similarly there are survey papers that are focusing on a particular set of methods [12]. Our work is addressing some of this limitations, by providing a general overview and at the same time limiting the focus to a few application domains and representative state-of-the-art methods. We also explore and present open research problems, from methodological point of view, as well as novel challenging application domains, untapped by existing UAD methods.

II. Taxonomy of Learning Approaches

A. General Anomaly Detection

The general problem of anomaly detection, as well as domain specific applications have been a topic of a number of surveys and review articles [1], [2], [12]. In this work we emphasize survey paper [1], which provides an extensive overview, spanning multiple research areas and application domains. This survey paper is particularly interesting as it captures all the relevant research and application domains before the era of deep-learning and clearly shows the state of research interest towards visual anomaly detection.

Anomaly detection refers to the problem of differentiating abnormal appearances from normal cases. These abnormal appearances are in the literature known as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants, depending on the application domain [1]. Applications of anomaly detection can be found in fraud detection systems for credit cards and insurances, intrusion detection systems for cyber-security, industrial inspection and medical imaging. According to [1], anomalies can be categorized as point anomalies, contextual anomalies and collective anomalies. Point anomaly represents an individual data instance, that deviates from the rest the data and represents the simplest type of anomaly and is also the focus of research on anomaly detection. Point anomaly can be represented as a contextual anomaly, if it is not conforming to the expected behavior in a specific context (e.g. a low temperature in summer). Collective anomalies, on the other hand, represent a set of data points, which together represent a deviation from a normal behaviour.

Anomaly detection methods can also be used for novelty detection, as they offer capabilities to detect unseen patterns in data, that could translate to new actionable insights. The difference is that the novel patterns are usually used alongside the previously known patterns, after being detected.

B. Availability of Labeled Data

Availability of large scale datasets [13] with labeled data and proliferation of deep-learning based methods has brought tremendous improvements particularly to computer vision domain [14]. Obtaining labeled data is often very expensive, as it is usually done manually by a human expert. Obtaining labeled data for anomaly detection is even harder, or even impossible, due to the nature of anomaly occurrences and unknown underlying processes, that generate them. Important factor is also the level of details, that are provided with the labels. This is especially important for visual anomaly detection, where labels can be on the image level (i.e. contains anomaly or not) or at the pixel level, delineating location and the extent of anomalies.

Anomaly detection methods are categorized to the bellow presented modes, based on the extent, to which the labels are available [1].

1) Supervised anomaly detection: Methods trained in a supervised fashion require labeled data for normal, as well as anomalous cases. There is usually much more labeled data for normal instances, which makes this an extremely imbalanced classification problem. Generalization performance of such methods is usually worse, due to limited availability and fixed vocabulary of representative labels and can also vary by the application domain.

2) Semi-supervised anomaly detection: According to [1], semi-supervised anomaly detection methods require labeled data only for normal instances and are as such more widely applicable. Such classification is often interchangeably also used for current unsupervised anomaly detection approaches, where normal instances are usually implicitly labeled as normal. A more common and widely used semi-supervised setting is when there is a combination of large set of unlabeled samples and a small pool of labeled ones [15].

3) Weakly-supervised anomaly detection: Weakly-supervised anomaly detection has not been considered in [1], mostly due to recent advancements and applications of such methods in the domain of visual anomaly detection [5]–[7]. In the context of industrial inspection or anomaly detection in medical imagery, we want to detect the anomaly, as well as localize it. Detailed ground-truth localized annotations are expensive or impossible to obtain in many cases. Weakly-supervised anomaly detection approaches utilize only image-level labels (i.e. contains anomaly or not) and are able to localize anomalous regions, without pixel-level annotations in the case of visual anomaly detection.

4) Unsupervised anomaly detection: Unsupervised methods do not require any labeled data and are as such the most widely applicable. They are usually trained with normal samples only in order to learn the distribution and are later on capable of capturing out-of-distribution samples. This methods run on the assumption that normal samples are far more frequent than anomalous ones. With recently presented deep generative based methods, accurate detection and localization of anomalous regions is possible, without any supervision [8], [16]. In the literature most of these methods are treated as unsupervised, despite weak implicit supervision, which is introduced by selecting only normal samples for training. In real-life scenarios one should expect that there will be some small percentage of contaminated data in training samples [10], [11].

III. Visual Anomaly Detection

Visual anomaly detection is dealing with detecting and localizing anomalous regions in imagery data. We have seen great success in computer vision domain since the introduction of deep-learning based methods and consequently, visual anomaly detection has also seen increasing interest and success [2]. The primary benefit of deep-learning based methods is the data driven approach, which eliminates the need for the expert-level feature engineering, which has shown sub-optimal performance [14], [17]. Despite the proliferation of deep-learning based methods, there is relatively small amount of methods that are truly addressing anomaly detection problem, especially in a real-world setting.

In the next sections we discuss this recent improvements in the context of industrial inspection and medical domain. We first briefly present a few application domains and associated data, with the focus on two recently presented large-scale datasets. Later on we categorize the methods based on the availability of data and present the main representatives.

A. Datasets

Large-scale labeled datasets have been one of the main contributing factors to the recent success of deep-learning based methods [13], [14]. Due to the nature and frequency of anomaly occurrence in real-world application domains, it

is difficult to obtain such large-scale datasets for anomaly detection.

Most of the initial work on visual anomaly detection has been performed on existing classification datasets [13], [18], [19], by considering a subset of the existing classes as anomalous samples and the rest of them as normal. With this approach one gets access to large-scale datasets to develop anomaly detection methods, but the anomalous samples differ significantly from the normal ones and are as such not representing real-world conditions. Manufacturing defects in industrial inspection domain or lesions in medical imagery are usually hard to detect and do not alter the resultant image, to differ significantly from normal samples. Equally important in visual anomaly detection domain is also the ability to segment anomalous regions, which is especially vital for industrial inspection and medical domain.

Real-world datasets for anomaly detection are rare, due to difficulties to create them, as well as due to confidential and privacy concerns. Industrial inspection is performed with industrial grade cameras [4] and specialized devices, such as X-Ray CT scans [20], which can reveal the details of the manufacturing process. Similarly, medical imagery can contain personal information and needs to be reviewed by medical boards and in some cases, patients consents are needed [3]. Despite confidential and privacy concerns, there are some datasets, that have been made publicly available and two representative datasets, that will be used in our research work are presented next.

1) MVTec AD - A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection: The dataset presented in [21] represents the first comprehensive, multi-object, multi-defect dataset for anomaly detection in a real world scenario of an industrial inspection. In comparison with other works [22]–[24], that evaluate anomaly detection methods on existing classification datasets, this represents a much more realistic scenario, with anomalies manifesting in less significant differences from the training data. MVTec Anomaly Detection dataset consists out of 15 categories of different objects and textures. 3629 images are provided for training and validation and another 1725 for testing. The training data does not contain defects. Altogether, 73 different types of defects are encountered, with provided pixel-wise annotations for all the test examples. Example data, together with anomalies and pixel-wise annotations are presented in figure 1. The captured dataset represents close-to real-world conditions, with some of the objects being rigid, while others deformable or with natural variations. Some of the objects are captured in aligned poses and some in random rotations. All images were acquired using 2048 x 2048 industrial grade RGB camera and resultant images were cropped to different resolutions between 700 x 700 and 1024 x 1024 pixels. Some of the images were intentionally provided in gray-scale and under different (uncontrolled) illumination conditions, to increase variability.

A thorough evaluation of multiple state-of-the-art unsupervised anomaly detection methods was performed. AnoGAN [16] method, based on generative adversarial networks (GANs), as well as a method based on auto-encoders and structural similarity [25] were evaluated. Both of the methods are described in section IV-C. Additionally, they evaluate a classical Convolutional Neural Network (CNN) feature extraction approach [26], as well as traditional non deep-learning methods based on Gaussian Mixture Model (GMM) [27] and a simple variational model approach [28].

Results were reported based on the classification, as well as anomaly segmentation performance, across different object and texture categories. None of the evaluated methods performed consistently across different object and texture classes. Object categories were best classified using autoencoders [25], with L2 loss. Similarly this method performed the best on the segmentation task, but with the structured similarity (SSIM) [25] as the reconstruction loss. This benchmark also nicely represents the level of the generalization performance, especially with the AnoGAN method [16] and its non-competitive results in comparison with the state-of-the-art results in medical domain.

2) Detection of Lymph Node Metastases in Women with Breast Cancer: Advances in tissue digitalization and in slide scanning technology have opened the possibilities for computer-aided diagnostics to detect cancerous metastases in stained tissue sections. Digital pathology is a new emerging field, utilizing computerized analysis of histopathological images. Breast cancer is just one of the many cancers, where the extent of it is measured by histopathological analysis. Detecting metastases in such gigapixel imagery is prone to error and a time consuming process, where pathologists would benefit greatly by recent advances in computer vision domain.

A competition was organized in 2015 [3] in order to evaluate the machine learning based methods against pathologists. In the challenge setting, some deep-learning based methods achieved better results than a panel of 11 pathologists. 399 whole-slide images were collected from 399 patients at 2 hospitals in the Netherlands. All metastases in the slides where annotated by trained pathologists on the slide level (contains metastases or not - for the slide level classification), as well as separate metastases (for segmentation purposes). The set of images was randomly divided into train (n = 270) and test set (n = 129). All the data is publicly available on the competition websites[1][2]. The first task was designed to evaluate the detection of separate metastases and evaluate the performance against the reference annotations, provided by the pathologists. Free-response Receiver Operator Characteristics curve (FROC) was used to evaluate the performance, at 6 predefined false positive rates. FROC curve shows the true-positive fraction vs. the mean number of false-positive detections in metastasis-free slides only. The goal of the second task was to evaluate the discrimination performance on the whole-slide level. Area Under Curve (AUC) was used for evaluation against pathologists, with and without any time-constraint.

[1]https://camelyon16.grand-challenge.org/https://camelyon16.grand-challenge.org/

[2]https://camelyon17.grand-challenge.org/https://camelyon17.grand-challenge.org/

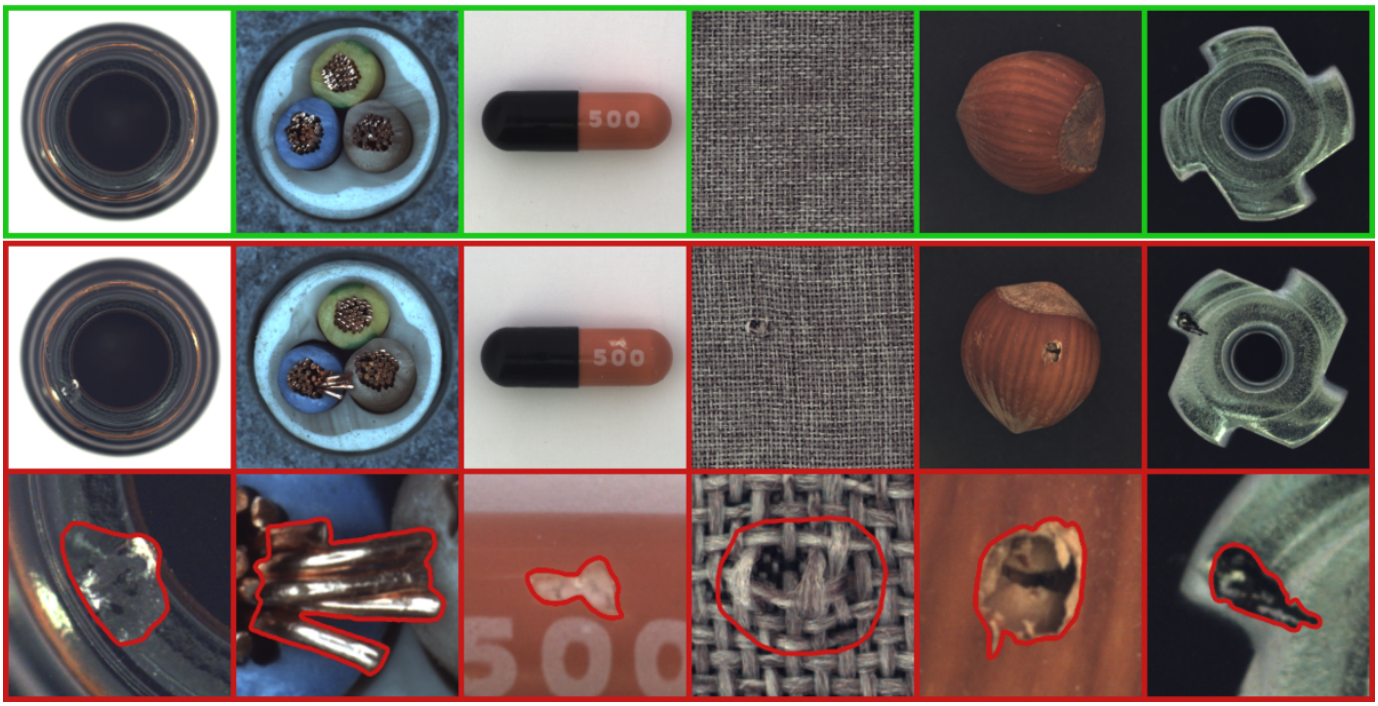


Fig. 1: Samples from MVTec AD dataset [21]. First row represents normal samples from the training set (bottle, cable, capsule, carpet, hazelnut, metal nut), second row the same objects with various defects and the third row presents pixel-wise annotations for defective samples. Image adapted from [21].

Out of 32 submitted methods, 25 used deep convolutional neural networks (CNNs) and overall performed significantly better compared to traditional approaches. However, preprocessing (e.g. standardizing stain variations, different sampling strategies - class imbalance problem) and augmentation procedures proved to play an important role, compared to the selection of the CNN architecture. After the competition, another approach was presented [29], which improves the competition results significantly and is presented in detail in the next section. All the solutions approached the problem in a supervised fashion, as a patch classification problem. We will approach this problem as an unsupervised visual anomaly detection problem instead. Same data was already utilized in a weakly-supervised fashion using Multiple Instance Learning (MIL) approach [6], utilizing only whole-slide level annotations, presented in section IV-B.

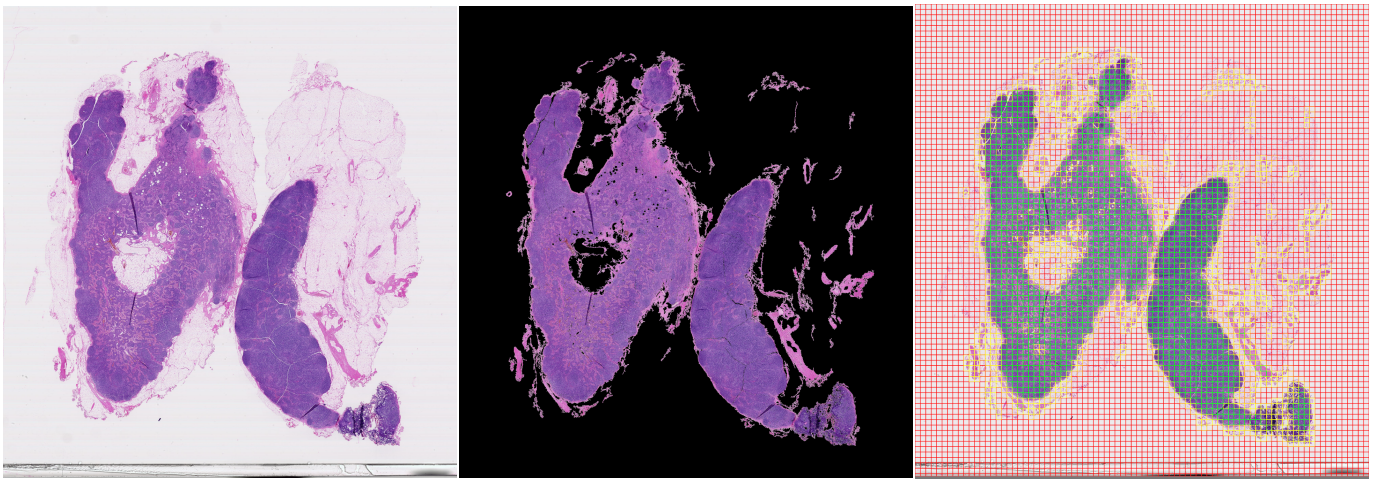
IV. Methods

In this section we review representative methods for anomaly detection, based on availability of the data. We focus on the medical domain, specifically on metastases detection from histopathological images. This particular domain represents a challenging task, that has not been considered directly as an anomaly detection problem. This particular problem has been considered in a supervised setting, as well as recently in a weakly-supervised fashion. Unsupervised approaches have not been considered yet. We present these existing approaches, as well as describe current state-of-the-art UAD approaches and present initial results, that demonstrate feasibility to apply them to the domain of metastases detection. The same methods are also applicable to other domains, especially industrial domain, which was presented in this paper; and representative methods for UAD described in this section have also been applied to that domain.

A. Supervised Anomaly Detection Methods

Winners of the Camelyon Grand Challenge 2016 [3] on detection of lymph node metastases presented their winning supervised based approach in a technical report [30]. Majority of digitized Whole Slide Image (WSI) consists of background white space, which needs to be segmented, to reduce computational time. Winners first utilized Otsu's algorithm [31] in HSV color space in order to generate segmentation masks. A simple filtering based on the green channel value can also be used, due to the purple and pink tones, resulting from H&E staining. Morphological operators are also applied to remove small objects and artifacts. Results of tissue filtering and patch extraction are presented in figure 2. We color-coded extracted patches based on the tissue percentage, in order to extract only the patches with sufficient amount of tissue. Metastasis detection framework was then proposed, consisting of patch-based classification part, which produces heatmaps, that are later on processed to obtain WSI-level and lesion-level labels.

Authors utilized GoogLeNet [32] as their best performing CNN architecture. Positive and negative 256 x 256 pixel patches were extracted, according to the provided lesion level labels and used to train the binary classification model. An additional model was learned on hard-negative examples, based



(a) Original WSI (b) Filtered WSI (c) Patches from WSI

Fig. 2: Preprocessing of original WSI presented in a) consists out of filtering tissue sections b) and extracting patches c), based on tissue percentage (green $\geq$ 90%, red $\leq$ 10% and yellow in-between). Best viewed in digital version with zoom.

on the initial model. The best results were obtained with the highest 40x WSI magnification. Learned models were applied in a sliding window fashion (overlapping patches), to obtain tumour probability maps. For lesion based detection, connected components were identified using the first model, which results were later averaged with the model learned on hard negative examples. For slide-level classification, 28 different geometrical and morphological features were extracted from heatmaps (e.g. percentage of tumour region over whole tissue region). Random Forest classifier was used to discriminate the WSIs with metastases from negative examples. Authors obtained an AUC score of 0.925 for WSI classification and an average FROC score of 0.705. These results showed that close to pathologist-level performance (AUC of 0.966 and FROC of 0.733) can be achieved with supervised deep-learning based models.

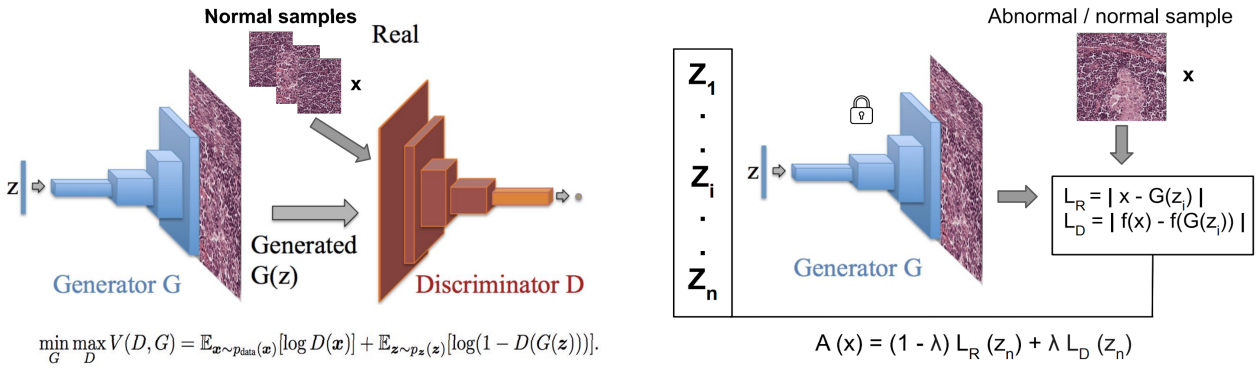
Above presented winning solution of the Camelyon 2016 challenge was later further improved by Google [29], by utilizing newer Inception architecture [33], careful image patch sampling and extensive image augmentations. They improved FROC sensitivity score for lesion based detection to 0.885 and AUC score for slide level classification to 0.986, though the evaluation protocol seems not to be exactly the same. They also show that statistically the same slide level classification performance can be achieved solely by using maximum value from the heatmap, instead of handcrafted features and Random Forest classifier.

B. Weakly-supervised Anomaly Detection Methods

Supervised approaches require abundance of labeled data, which is particularly severe in digital pathology, where digitization of glass slides is expensive, and pixel-level manual labels are time-consuming to obtain, due to gigapixel large pathology imagery. In [6] the authors present a weakly supervised approach, that only utilizes image level reported diagnosis as labels for training, omitting the need for expert pixel-wise annotations. Such a procedure can capture a much wider variance of clinical samples that is not captured in small supervised datasets. They collect large-scale pathology imagery (WSIs) from 1) prostate cancer (prostatic carcinoma), 2) skin cancer (basal cell carcinoma) and 3) breast cancer (axillary lymph nodes), together with slide-level diagnosis, obtained from electronic health records.

With negative slide-level diagnosis, one can be sure, that all the tiles within a negative WSI are negative, not containing the metastases or tumor. On the other hand, with a positive slide-level diagnosis, we know, that at least one tile is positive. This kind of classification problem is a classical formulation of Multiple Instance Learning (MIL), where training instances are arranged in sets, called bags, and a label is provided for the entire bag [34]. Solving MIL task induces the learning of a tile-level representation that can linearly separate the discriminative tiles in positive slides from all other tiles [6]. This is implemented on a tile-level using standard CNN based architectures (e.g. Resnet34) and probability is obtained for each of the tiles of being positive. The top ranked tile (or K top ranked) are selected and compared with slide-level ground truth labels, used in cross-entropy loss. In this way, weakly supervised tile-level classifier is learned, that is applied in a similar fashion as in [3]. They used handcrafted features from the obtained heatmaps and learned a Random Forest classifier for slide-level classification, similarly as in [30]. Additionally, they noticed the drawback of such handcrafted aggregation methods for slide-level classification and proposed a new Recurrent Neural Network (RNN) based model that uses features, learned during tile-level classification training.

The performance of the proposed weakly supervised method was evaluated on in-house data, that is not publicly available. They also compared the method with fully supervised approach on Camelyon 2016 challenge data [3]. They imple-



(a) AnoGAN GAN training

(b) AnoGAN anomaly detection

Fig. 3: AnoGAN method [16] consisting of DCGAN training a) and iterative optimization procedure b) to find an optimal latent vector for anomaly detection. Image adapted from [16] for digital pathology.

mented a modified supervised winning approach from [30], trained on Camelyon data and evaluated the approach on their in-house data, to evaluate the generalization performance. They noticed a 20% drop in AUC score (from test results on Camelyon data). In comparison, they evaluated their proposed weakly supervised MIL-RCNN method, trained on large-scale in-house data, on Camelyon test set and noticed only 7% drop in AUC score (from test results on in-house data). Unfortunately they do not report the results of their proposed method, when trained only on Camelyon data.

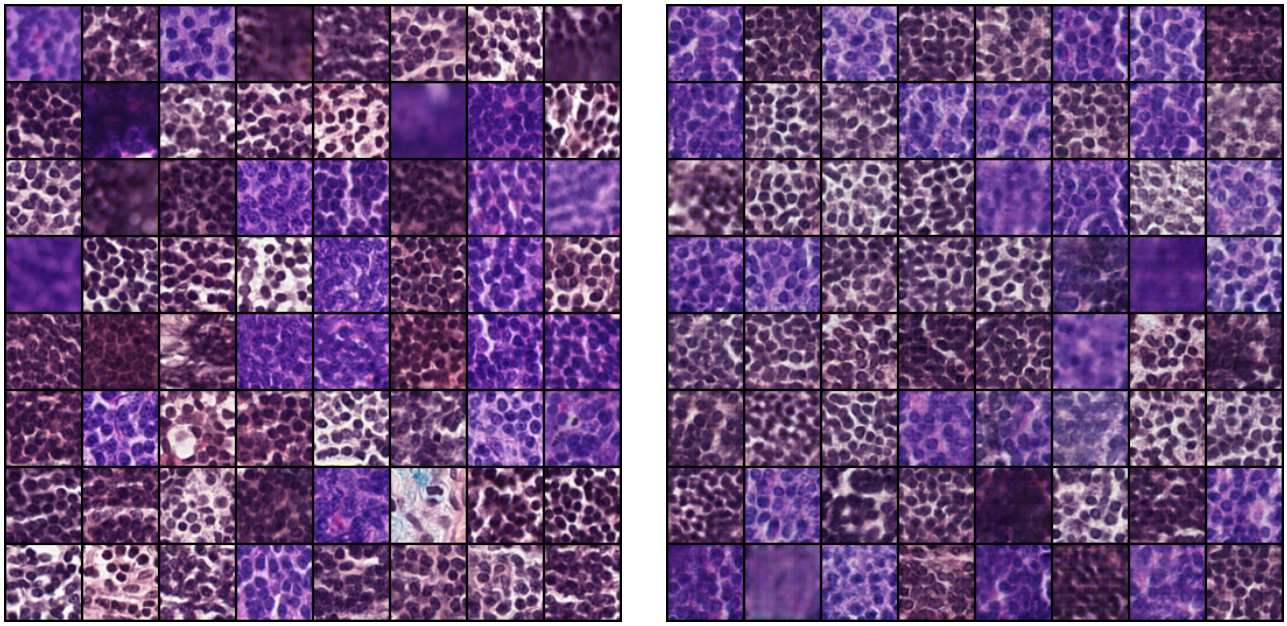
C. Unsupervised Anomaly Detection Methods

In comparison with supervised and weakly-supervised approaches, unsupervised approaches omit the need for expertly labeled data. UAD is a relatively new domain and has seen particular improvements and rise of interest with the introduction of deep generative methods. In this section we introduce two main approaches, one based on GANs [36] and the other one based on autoencoders [37]. None of the approaches has been applied to challenging digital pathology imagery. Besides introduction to the methods, we also present preliminary results, that demonstrate feasibility to apply presented methods for detection of cancerous regions in histology imagery.

1) GAN based UAD methods: AnoGAN method [16], presented in figure 3, represents the first work, where GANs are used for anomaly detection in medical domain. A rich generative model is constructed on healthy examples of optical coherence tomography images of the retina and a methodology is presented for image mapping into the latent space, to generate the closest example to the presented query image, to be able to detect and segment the anomalies in an unsupervised fashion. Given a set o healthy images, smaller patches were extracted and used to train a generative model, based on the DCGAN [35] architecture, in order to learn the manifold of healthy examples. In this way, the model captures the variability of the training examples in an unsupervised fashion. Labels are only given during the testing, to evaluate the detection performance.

GANs consists of generator (G) and discriminator part (D). The generator G learns a mapping $G(z)$, where z represents a sampled 1D vector from the uniformly distributed input noise, sampled from the latent space - consisting of healthy examples. Discriminator on the other hand, maps an input 2D image to a scalar value, representing the probability of the input being a real image, sampled from the training data, or a generated one - produced by $G(z)$. G and D are trained in an alternating fashion, using a two-player minimax game. The discriminator D is trained to maximize the probability to discriminate the real image, from the generated one. Generator (G) is on the other hand trained to fool the discriminator. After the adversarial training is completed, the generator learns how to generate realistically looking healthy examples, captured in the training set. When query image x is presented, to detect the anomaly, the goal is to find the closest point z in the latent manifold of healthy examples. This is done in an iterative fashion from a randomly sampled initial latent vector z_1, which is updated back using backpropagation in $i = 1, 2, ...n$ steps, via residual (L_R) and discrimination loss (L_D), to obtain the optimal latent vector z_n (only the coefficients of z are modified, G and D parameters are kept fixed). Residual loss captures similarity of the query image to the generated one $G(z_i)$, while discrimination loss ensures that the generated image $G(z_i)$ lies on the learned manifold of healthy training examples. The mapping and corresponding losses are inspired by the work of semantic image inpainting using GANs [38], which poses a similar problem setup. The combined residual and discrimination loss for z_n can be directly used as an anomaly score $A(x)$ and the resultant residual image between $G(z_n)$ and query image, for pixel-wise anomalous region segmentation. The whole process of training AnoGAN method [16] is visually presented in figure 3

Iterative optimization approach to find the optimal latent vector is time-consuming and not applicable for real-time anomaly detection. Recently presented f-AnoGAN method [8] greatly improves inference times, at a similar performance rate, by replacing iterative optimization approach with a trained



(a) Original samples

(b) Generated samples

Fig. 4: a) Original patches, extracted from histology image and b) generated artificial patches from DCGAN [35] based GAN network, as used in AnoGAN method [16].

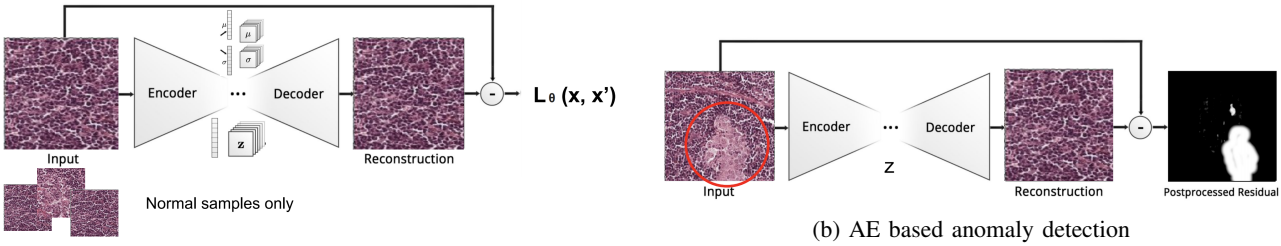
encoder mapping from images to corresponding location in the learned latent space. Besides, training GANs can be a very unstable process and mostly smaller resolution images are used. AnoGAN and f-AnoGAN methods utilize baseline DCGAN [35] and Wasserstein GAN (WGAN) [39] architectures and do not consider recent works, that are able to generate higher resolution images in a more stable way [40], [41]. Capability to generate realistically looking histology imagery is crucial, in order to generate accurate cellular structure. We present baseline results of the DCGAN [35] architecture in figure 4. These initial results with a baseline method, that was also used in AnoGAN method, demonstrate the applicability of such methods to digital pathology domain.

2) AE based UAD methods: Above presented UAD methods are modelling normal samples with GANs. Autoencoder (AE) based methods are one of the simplest and first approaches, that are also used for visual anomaly detection, by learning how to reconstruct the input image through a bottleneck, via encoder (E) and decoder (D) networks. Generative AEs (i.e. variational autoencoders [37]) were also introduced and used in a recent UAD work for lesion detection in brain MR images [9]. AE based method are trained in a self-supervised way, such that they learn how to reconstruct input training images. This is achieved by mapping an input to a bottleneck, which can in fact be a distribution or a direct mapping. When introduced with normal samples only, they learn how to reconstruct such normal samples and in the case of VAE, they are also able to generate them, similarly to GANs. When we introduce anomalous sample, the method is able to reconstruct it, the way that the normal sample should look like. We are then able to threshold the reconstruction error, in order to detect the anomaly, as well to segment them, by computing a residual image. This process is visually presented in figure 5, the way, that the method would be used in digital pathology setting.

GANs are known to produce very sharp images, due to adversarial training, but are having issues with stable training and mode collapses, which results in learning to generate just a few examples [35]. GAN and VAE concepts have been recently combined into VAEGAN framework [42], combining the best of the two approaches. Adding an adversarial loss and discriminator to the AE/VAE framework forces the decoder to generate a better reconstructions, that will fool a discriminator. In [9] they also used spatial VAEs, replacing the mapping to dense 1D bottleneck z with a fully convolutional encoder-decoder network, resulting in a higher dimensional spatial bottleneck z, omitting the loss of the spatial information in the bottleneck encoder function. The presented AnoVAEGAN UAD method was compared against AnoGAN [16] method and variations of AE/VAE arhitectures with different types of bottlenecks (i.e. dense vs. spatial) and their dimensions. Similar to GAN approaches, no AE based approach has been utilized for anomaly detection in gigapixel histology imagery.

V. Conclusion

Visual anomaly detection is an important process in many domains and recent advancements in deep generative based methods have shown promising results towards applying them in an unsupervised fashion. This has sparked research in many domains, that did not benefit much from traditional supervised deep-learning based approaches.



(a) AE based UAD method training

(b) AE based anomaly detection

Fig. 5: Basic architecture of AE based method for anomaly detection, consisting of AE training a) and AE inference b), resulting to residual image, used for anomaly detection and segmentation. Image adapted from [9] for digital pathology.

Most of the existing methods are applied to medical domain, where vast amount of imagery data is available, but without any detailed labels to learn state-of-the-art supervised models. All the appearances of anomalies in real-world applications are usually also not known in advance and are as such impossible to label. Benefits of such methods have recently also been recognized in industrial inspection domain, where the need for rapid product development is making the existing supervised approaches inappropriate to use, due to time constraints to collect anomalous samples, as well as wide-range of potential anomalies that can occur and are unknown in advance. The presented UAD methods have been developed and evaluated on particular limited real-world domains or even on existing classification datasets and significant performance drops are visible when applied to other domains. This has been seen through several presented works, that evaluated the existing UAD methods, along with the newly presented ones, on new application domains. Another important issue is the robustness of existing UAD methods to contaminated training data. Existing UAD methods are not really unsupervised due to the requirement that completely anomaly-free data is available for training the methods, therefore implicitly implying the need for weak labelling.

UAD in visual data is a relatively new domain, that has seen particular improvements with the introduction of generative based methods. Unprecedented amount of visual data that is captured every day in different domains represents an untapped potential for unsupervised based methods, that will be able to leverage this data as it is. Addressing the issues of current UAD approaches will enable their wider usage, especially in data-heavy domains. Dual-use of UAD approaches that enables novelty detection can also represent a major diagnostic tool for early cancer detection and rare disease detection, thereby support the development and evaluation of personalized medicine, and thus address a much wider societal challenge.

ACKNOWLEDGMENT

This work was partially supported by the European Commission through the Horizon 2020 research and innovation program under grant 826121 (iPC).

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

[2] R. Chalapathy and S. Chawla, "Deep Learning for Anomaly Detection: A Survey," *ArXiv*, vol. abs/1901.03407, 2019.

[3] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, , and the CAMELYON16 Consortium, "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, 12 2017. [Online]. Available: https://doi.org/10.1001/jama.2017.14585

[4] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-Based Deep-Learning Approach for Surface-Defect Detection," *Journal of Intelligent Manufacturing*, May 2019. [Online]. Available: https://doi.org/10.1007/s10845-019-01476-x

[5] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang, "Learning Fixed Points in Generative Adversarial Networks: From Image-to-Image Translation to Disease Detection and Localization," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[7] P. Courtiol, E. W. Tramel, M. Sanselme, and G. Wainrib, "Classification and disease localization in histopathology using only global labels: A weakly-supervised approach," *arXiv preprint arXiv:1802.02212*, 2018.

[8] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks," *Medical Image Analysis*, vol. 54, pp. 30 – 44, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841518302640

[9] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham: Springer International Publishing, 2019, pp. 161–169.

[10] L. Beggel, M. Pfeiffer, and B. Bischl, "Robust Anomaly Detection in Images using Adversarial Autoencoders," *ArXiv*, vol. abs/1901.06355, 2019.

[11] A. Berg, J. Ahlberg, and M. Felsberg, "Unsupervised Learning of Anomaly Detection from Contaminated Image Data using Simultaneous Encoder Training," *ArXiv*, vol. abs/1905.11034, 2019.

[12] F. D. Mattia, P. Galeone, M. D. Simoni, and E. Ghelfi, "A Survey on GANs for Anomaly Detection," 2019.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges,

L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[15] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep Semi-Supervised Anomaly Detection," 2019.

[16] 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 *Information Processing in Medical Imaging*, M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, and D. Shen, Eds. Cham: Springer International Publishing, 2017, pp. 146–157.

[17] Emeršič, D. Štepec, V. Štruc, and P. Peer, "Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 987–994.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *University of Toronto*, 05 2012.

[20] E. A. Donahue, T.-T. Quach, K. Potter, C. Martinez, M. Smith, and C. D. Turner, "Deep learning for automated defect detection in high-reliability electronic parts," in *Applications of Machine Learning*, M. E. Zelinski, T. M. Taha, J. Howe, A. A. S. Awwal, and K. M. Iftekharuddin, Eds., vol. 11139, International Society for Optics and Photonics. SPIE, 2019, pp. 30 – 40. [Online]. Available: https://doi.org/10.1117/12.2529584

[21] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD - A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," in *CVPR*, 2019.

[22] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, no. 1, 2015.

[23] R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," *arXiv preprint arXiv:1802.06360*, 2018.

[24] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International Conference on Machine Learning*, 2018, pp. 4393–4402.

[25] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *arXiv preprint arXiv:1807.02011*, 2018.

[26] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by cnn-based self-similarity," *Sensors*, vol. 18, no. 1, p. 209, 2018.

[27] T. Böttger and M. Ulrich, "Real-time texture error detection on textured surfaces with compressed sensing," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 88–94, 2016.

[28] C. Steger, M. Ulrich, and C. Wiedemann, *Machine vision algorithms and applications*. John Wiley & Sons, 2018.

[29] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

[30] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[31] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan 1979.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[34] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[36] 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.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.

[38] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493.

[39] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 214–223.

[40] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[41] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[42] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.