Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images

Yu Tian[1,3](✉), Guansong Pang[1], Fengbei Liu[1], Yuanhong Chen[1],
Seon Ho Shin[2], Johan W. Verjans[1,2,3], Rajvinder Singh[2],
and Gustavo Carneiro[1]

[1] Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia
yu.tian01@adelaide.edu.au
[2] Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, Australia
[3] South Australian Health and Medical Research Institute, Adelaide, Australia

Abstract. Unsupervised anomaly detection (UAD) learns one-class classifiers exclusively with normal (i.e., healthy) images to detect any abnormal (i.e., unhealthy) samples that do not conform to the expected normal patterns. UAD has two main advantages over its fully supervised counterpart. Firstly, it is able to directly leverage large datasets available from health screening programs that contain mostly normal image samples, avoiding the costly manual labelling of abnormal samples and the subsequent issues involved in training with extremely class-imbalanced data. Further, UAD approaches can potentially detect and localise any type of lesions that deviate from the normal patterns. One significant challenge faced by UAD methods is how to learn effective low-dimensional image representations to detect and localise subtle abnormalities, generally consisting of small lesions. To address this challenge, we propose a novel self-supervised representation learning method, called Constrained Contrastive Distribution learning for anomaly detection (CCD), which learns fine-grained feature representations by simultaneously predicting the distribution of augmented data and image contexts using contrastive learning with pretext constraints. The learned representations can be leveraged to train more anomaly-sensitive detection models. Extensive experiment results show that our method outperforms current state-of-the-art UAD approaches on three different colonoscopy and fundus screening datasets. Our code is available at https://github.com/tianyu0207/CCD.

Keywords: Anomaly detection · Unsupervised learning · Lesion detection and segmentation · Self-supervised pre-training · Colonoscopy

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12905, pp. 128–140, 2021.
https://doi.org/10.1007/978-3-030-87240-3_13

1 Introduction

Classifying and localising malignant tissues have been vastly investigated in medical imaging [1,11,22–24,26,29,42,43]. Such systems are useful in health screening programs that require radiologists to analyse large quantities of images [35,41], where the majority contain normal (or healthy) cases, and a small minority have abnormal (or unhealthy) cases that can be regarded as anomalies. Hence, to avoid the difficulty of learning from such class-imbalanced training sets and the prohibitive cost of collecting large sets of manually labelled abnormal cases, several papers investigate anomaly detection (AD) with a few or no labels as an alternative to traditional fully supervised imbalanced learning [1,26,28,32,33,37,38,43–45]. UAD methods typically train a one-class classifier using data from the normal class only, and anomalies (or abnormal cases) are detected based on the extent the images deviate from the normal class.

Current anomaly detection approaches [7,8,14,27,37,43,46] train deep generative models (e.g., auto-encoder [19], GAN [15]) to reconstruct normal images, and anomalies are detected from the reconstruction error [33]. These approaches rely on a low-dimensional image representation that must be effective at reconstructing normal images, where the main challenge is to detect anomalies that show subtle deviations from normal images, such as with small lesions [43]. Recently, self-supervised methods that learn auxiliary pretext tasks [2,6,13,17,18,25] have been shown to learn effective representations for UAD in general computer vision tasks [2,13,18], so it is important to investigate if self-supervision can also improve UAD for medical images.

The main challenge for the design of UAD methods for medical imaging resides in how to devise effective pretext tasks. Self-supervised pretext tasks consist of predicting geometric or brightness transformations [2,13,18], or contrastive learning [6,17]. These pretext tasks have been designed to work for downstream classification problems that are not related to anomaly detection, so they may degrade the detection performance of UAD methods [47]. Sohn et al. [40] tackle this issue by using smaller batch sizes than in [6,17] and a new data augmentation method. However, the use of self-supervised learning in UAD for medical images has not been investigated, to the best of our knowledge. Further, although transformation prediction and contrastive learning show great success in self-supervised feature learning, there are no studies on how to properly combine these two approaches to learn more effective features for UAD.

In this paper, we propose Constrained Contrastive Distribution learning (CCD), a new self-supervised representation learning designed specifically to learn normality information from exclusively normal training images. The contributions of CCD are: a) contrastive distribution learning, and b)two pretext learning constraints, both of which are customised for anomaly detection (AD). Unlike modern self-supervised learning (SSL) [6,17] that focuses on learning generic semantic representations for enabling diverse downstream tasks, CCD instead contrasts the distributions of strongly augmented images (e.g., random permutations). The strongly augmented images resemble some types of abnormal images, so CCD is enforced to learn discriminative normality representations

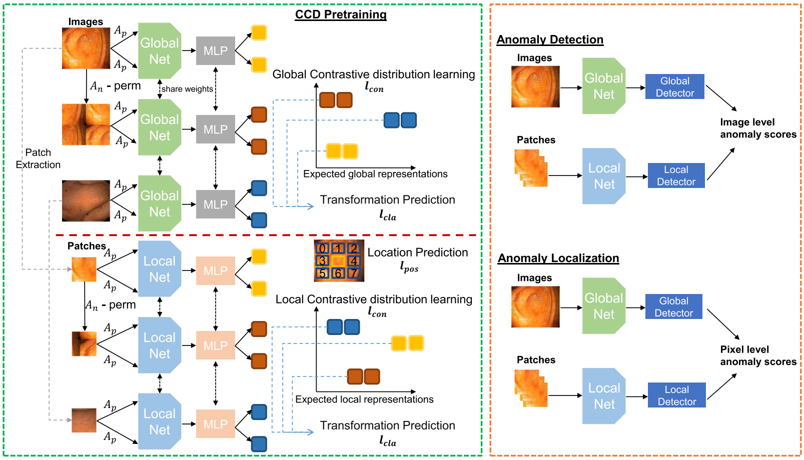

Fig. 1. Our proposed CCD framework. **Left** shows the proposed pre-training method that unifies a contrastive distribution learning and pretext learning on both global and local perspectives (Sect. 2.1), **Right** shows the inference for detection and localisation (Sect. 2.2).

by its contrastive distribution learning. The two pretext learning constraints on augmentation and location prediction are added to learn fine-grained normality representations for the detection of subtle abnormalities. These two unique components result in significantly improved self-supervised AD-oriented representation learning, substantially outperforming previous general-purpose SOTA SSL approaches [2,6,13,18]. Another important contribution of CCD is that it is agnostic to downstream anomaly classifiers. We empirically show that our CCD improves the performance of three diverse anomaly detectors (f-anogan [37], IGD [8], MS-SSIM) [48]). Inspired by IGD [8], we adapt our proposed CCD pretraining on global images and local patches, respectively. Extensive experimental results on three different health screening medical imaging benchmarks, namely, colonoscopy images from two datasets [4,27], and fundus images for glaucoma detection [21], show that our proposed self-supervised approach enables the production of SOTA anomaly detection and localisation in medical images.

2 Method

In this section, we introduce the proposed approach, depicted in the diagram of Fig. 1. Specifically, given a training medical image dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$, with all images assumed to be from the normal class and $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$, our approach aims to learn anomaly detection and localisation using three modules: 1) a self-supervised constrained contrastive feature learner that pre-trains an encoding network $f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ (with $\mathcal{Z} \subset \mathbb{R}^{d_z}$) tailored for anomaly detection, 2) an anomaly classification model $h_\psi : \mathcal{Z} \rightarrow [0, 1]$ that is built upon the pre-trained network, and 3) an anomaly localiser that leverages the classifier $h_\psi(f_\theta(\mathbf{x}_\omega))$ to localise an abnormal image region $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H}\times\hat{W}\times C}$, centred at $\omega \in \Omega$ (Ω is the image lattice) with height $\hat{H} << H$ and width $\hat{W} << W$. The approach is evaluated on a testing set $\mathcal{T} = \{(\mathbf{x}, y, \mathbf{m})_i\}_{i=1}^{|\mathcal{T}|}$, where $y \in \mathcal{Y} = \{\text{normal}, \text{abnormal}\}$, and $\mathbf{m} \in \mathcal{M} \subset \{0,1\}^{H\times W\times C}$ denotes the segmentation mask of the lesion in the image $\mathbf{x}$. For adapting our CCD pretraining on patch representations, we simply crop the training images into patches before applying our method.

2.1 Constrained Contrastive Distribution Learning

Contrastive learning has been used by self-supervised learning methods to pre-train encoders with data augmentation [6,17,47] and contrastive learning loss [39]. The idea is to sample functions from a data augmentation distribution (e.g., geometric and brightness transformations), and assume that the same image, under separate augmentations, form one class to be distinguished against all other images in the batch [2,13]. Another form of pre-training is based on a pretext task, such as solving jigsaw puzzle and predicting geometric and brightness transformations [6,17]. These self-supervised learning approaches are useful to pre-train classification [6,17] and segmentation models [31,49]. Only recently, self-supervised learning using contrastive learning [40] and pretext learning [2,13] have been shown to be effective in anomaly detection. However, these two approaches are explored separately. In this paper, we aim at harnessing the power of both approaches to learn more expressive pre-trained features specifically for UAD. To this end, we propose the novel Constrained Contrastive Distribution learning method (CCD).

Contrastive distribution learning is designed to enforce a non-uniform distribution of the representations in the space $\mathcal{Z}$, which has been associated with more effective anomaly detection performance [40]. Our CCD method constrains the constrastive distribution learning with two pretext learning tasks, with the goal of enforcing further the non-uniform distribution of the representations. The CCD loss is defined as

$$\ell_{CCD}(\mathcal{D};\theta,\beta,\gamma) = \ell_{con}(\mathcal{D};\theta) + \ell_{cla}(\mathcal{D};\beta) + \ell_{pos}(\mathcal{D};\gamma), \tag{1}$$

where $\ell_{con}(\cdot)$ is the contrastive distribution loss, ℓ_{cla} and ℓ_{pos} are two pretext learning tasks added to constrain the optimisation; and θ, β and γ are trainable parameters. The contrastive distribution learning uses a dataset of **weak data augmentations** $\mathcal{A}_p = \{a_l : \mathcal{X} \to \mathcal{X}\}_{l=1}^{|\mathcal{A}_p|}$ and **strong data augmentations** $\mathcal{A}_n = \{a_l : \mathcal{X} \to \mathcal{X}\}_{l=1}^{|\mathcal{A}_n|}$, where $a_l(\mathbf{x})$ denotes a particular data augmentation applied to $\mathbf{x}$, and the loss is defined as

$$\ell_{con}(\mathcal{D};\theta) = \\ -\mathbb{E}\left[\log \frac{\exp\left[\frac{1}{\tau} f_\theta(a(\tilde{\mathbf{x}}^j))^\top f_\theta(a'(\tilde{\mathbf{x}}^j))\right]}{\exp\left[\frac{1}{\tau} f_\theta(a(\tilde{\mathbf{x}}^j))^\top f_\theta(a'(\tilde{\mathbf{x}}^j))\right] + \sum_{i=1}^{M} \exp\left[\frac{1}{\tau} f_\theta(a(\tilde{\mathbf{x}}^j))^\top f_\theta(a'(\tilde{\mathbf{x}}_i^j))\right]}\right], \tag{2}$$

where the expectation is over $\mathbf{x} \in \mathcal{D}$, $\{\mathbf{x}_i\}_{i=1}^{M} \subset \mathcal{D} \setminus \{\mathbf{x}\}$, $a(.), a'(.) \in \mathcal{A}_p$, $\tilde{\mathbf{x}}^j = a_j(\mathbf{x})$, $\tilde{\mathbf{x}}_i^j = a_j(\mathbf{x}_i)$, and $a_j(.) \in \mathcal{A}_n$. The images augmented with the functions from the strong set $\mathcal{A}_n$ carry some 'abnormality' compared to the original images, which is helpful to learn a non-uniform distribution in the representation space $\mathcal{Z}$.

We can then constrain further the training to learn more non-uniform representations with a self-supervised classification constraint $\ell_{cla}(\cdot)$ that enforces the model to achieve accurate classification of the strong augmentation function:

$$\ell_{cla}(\mathcal{D}; \beta) = -\mathbb{E}_{\mathbf{x} \in \mathcal{D}, a(.) \in \mathcal{A}_n} \left[\log \mathbf{a}^\top f_\beta(f_\theta(a(\mathbf{x}))) \right], \tag{3}$$

where $f_\beta : \mathcal{Z} \to [0,1]^{|\mathcal{A}_n|}$ is a fully-connected (FC) layer, and $\mathbf{a} \in \{0,1\}^{|\mathcal{A}_n|}$ is a one-hot vector representing the strong augmentation $a(.) \in \mathcal{A}_n$.

The second constraint is based on the relative patch location from the centre of the training image – this positional information is important for segmentation tasks [20,31]. This constraint is added to learn fine-grained features and achieve more accurate anomaly localisation. Inspired by [10], the positional constraint predicts the relative position of the paired image patches, with its loss defined as

$$\ell_{pos}(\mathcal{D}; \gamma) = -\mathbb{E}_{\{\mathbf{x}_{\omega_1}, \mathbf{x}_{\omega_2}\} \sim \mathbf{x} \in \mathcal{D}} \left[\log \mathbf{p}^\top f_\gamma(f_\theta(\mathbf{x}_{\omega_1}), f_\theta(\mathbf{x}_{\omega_2})) \right], \tag{4}$$

where $\mathbf{x}_{\omega_1}$ is a randomly selected fixed-size image patch from $\mathbf{x}$, $\mathbf{x}_{\omega_2}$ is another image patch from one of its eight neighbouring patches (as shown in 'patch location prediction' in Fig. 1), $f_\gamma : \mathcal{Z} \times \mathcal{Z} \to [0,1]^8$, and $\mathbf{p} = \{0,1\}^8$ is a one-hot encoding of the synthetic class label.

Overall, the constraints in (3) and (4) to the contrastive distribution loss in (2) are designed to increase the non-uniform representation distribution and to improve the representation discriminability between normal and abnormal samples, compared with [40].

2.2 Anomaly Detection and Localisation

Building upon the pre-trained encoder $f_\theta(\cdot)$ using the loss in (1), we fine-tune two state-of-the-art UAD methods, IGD [8] and F-anoGAN [37], and a baseline method, multi-scale structural similarity index measure (MS-SSIM)-based auto-encoder [48]. All UAD methods use the same training set $\mathcal{D}$ that contains only normal image samples.

IGD [8] combines three loss functions: 1) two reconstruction losses based on local and global multi-scale structural similarity index measure (MS-SSIM) [48] and mean absolute error (MAE) to train the encoder $f_\theta(\cdot)$ and decoder $g_\phi(\cdot)$, 2) a regularisation loss to train adversarial interpolations from the encoder [3], and 3) an anomaly classification loss to train $h_\psi(\cdot)$. The anomaly detection score of image $\mathbf{x}$ is

$$s_{IGD}(\mathbf{x}) = \xi \ell_{rec}(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \xi)(1 - h_\psi(f_\theta(\mathbf{x}))), \tag{5}$$

where $\tilde{\mathbf{x}} = g_\phi(f_\theta(\mathbf{x}))$, $h_\psi(f_\theta(\mathbf{x})) \in [0,1]$ returns the likelihood that $\mathbf{x}$ belongs to the normal class, $\xi \in [0,1]$ is a hyper-parameter, and

$$\ell_{rec}(\mathbf{x}, \tilde{\mathbf{x}}) = \rho \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 + (1 - \rho)\left(1 - (\nu m_G(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \nu) m_L(\mathbf{x}, \tilde{\mathbf{x}}))\right), \tag{6}$$

with $\rho, \nu \in [0, 1]$, $m_G(\cdot)$ and $m_L(\cdot)$ denoting the global and local MS-SSIM scores [8]. Anomaly localisation uses (5) to compute $s_{IGD}(\mathbf{x}_\omega)$, $\forall \omega \in \Omega$, where $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$ is an image region–this forms a heatmap, where large values denote anomalous regions.

F-anoGAN [37] combines generative adversarial networks (GAN) and auto-encoder models to detect anomalies. Training involves the minimisation of reconstruction losses in both the original image and representation spaces to model $f_\theta(\cdot)$ and $g_\phi(\cdot)$. It also uses a GAN loss [15] to model $g_\phi(\cdot)$ and $h_\psi(\cdot)$. Anomaly detection for image $\mathbf{x}$ is

$$s_{FAN}(\mathbf{x}) = \|\mathbf{x} - g_\phi(f_\theta(\mathbf{x}))\| + \kappa \|f_\theta(\mathbf{x}) - f_\theta(g_\phi(f_\theta(\mathbf{x})))\|. \tag{7}$$

Anomaly localisation at $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$ is achieved by $\|\mathbf{x}_\omega - g_\phi(f_\theta(\mathbf{x}_\omega))\|$, $\forall \omega \in \Omega$.

For the MS-SSIM auto-encoder [48], we train it with the MS-SSIM loss for reconstructing the training images. Anomaly detection for $\mathbf{x}$ is based on $s_{MSI}(\mathbf{x}) = 1 - (\nu m_G(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \nu) m_L(\mathbf{x}, \tilde{\mathbf{x}}))$, with $\tilde{\mathbf{x}}$ as defined in (5). Anomaly localisation is performed with $s_{MSI}(\mathbf{x}_\omega)$ at image regions $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$, $\forall \omega \in \Omega$. Inspired by IGD [8], we also pretrain a local model using our CCD pretraining approach based on the local patches for F-anogan [37] and MS-SSIM autoencoder [48], respectively.

3 Experiments

3.1 Dataset

We test our framework on three health screening datasets. We test both anomaly detection and localisation on the colonoscopy images of Hyper-Kvasir dataset [4]. On the glaucoma datasets using fundus images [21] and colonoscopy dataset [27] that do not have lesion masks, we test anomaly detection only. Detection is assessed with area under the ROC curve (AUC). Localisation is measured with intersection over union (ioU).

Hyper-Kvasir is a large multi-class public gastrointestinal dataset. The data was collected from the gastroscopy and colonoscopy procedures from Baerum Hospital in Norway. All labels were produced by experienced radiologists. The dataset contains 110,079 images from abnormal (i.e., unhealthy) and normal (i.e., healthy) patients, with 10,662 labelled. We use part of the clean images from the dataset to train our UAD methods. Specifically, 2,100 images from ‘cecum’, ‘ileum’ and ‘bbps-2–3’ are selected as normal, from which we use 1,600 for training and 500 for testing. We also take 1,000 abnormal images and their segmentation masks and stored them in the testing set.

LAG is a large scale fundus image dataset for glaucoma detection [21], containing 4,854 fundus images with 1,711 positive glaucoma scans and 3,143 negative glaucoma scans. We reorganised this dataset for training the UAD methods, with 2,343 normal (negative glaucoma) images for training, and 800 normal images and 1,711 abnormal images with positive glaucoma for testing.

Liu et al.'s colonoscopy dataset is a colonoscopy image dataset for UAD using 18 colonocopy videos from 15 patients [27]. The training set contains 13,250 normal (healthy) images without any polyps, and the testing set contains 967 images, having 290 abnormal images with polyps and 677 normal (healthy) images without polyps.

3.2 Implementation Details

For pre-training, we use Resnet18 [16] as the backbone architecture for the encoder $f_\theta(\mathbf{x})$, and similarly to previous works [6,40], we add an MLP to this backbone as the projection head for the contrastive learning. All images from the Hyper-Kvasir [4] and LAG [21] datasets are resized to 256 × 256 pixels. For the Liu et al.'s colonoscopy dataset, images are resized to 64 × 64 pixels. The batch size is set to 32 and learning rate to 0.01 for the self-supervised pre-training. We investigate the impact of different strong augmentations in $\mathcal{A}_n$ such as rotation, permutation, cutout and Gaussian noise. All weak augmentations in $\mathcal{A}_p$ are the same as SimCLR [6] (i.e., colour jittering, random grey scale, crop, resize, and Gaussian blur). The model is trained using SGD optimiser with temperature 0.2. The encoder $f_\theta(\cdot)$ outputs a 128 dimensional feature in $\mathcal{Z}$. All datasets are pre-trained for 2,000 epochs.

For the training of IGD [8], F-anoGAN [37] and MS-SSIM auto-encoder [8], we use the hyper-parameters suggested by the respective papers. For localisation, we compute the heatmap based on the localised anomaly scores from IGD, where the final map is obtained by summing the global and local maps. In our experiments, the local map is obtained by considering each 32 × 32 image patch as a instance and apply our proposed self-supervised learning to it. The global map is computed based on the whole image sized as 256 × 256. For F-anoGAN and MS-SSIM auto-encoder, we use the same setup as the IGD, where models based the 256 × 256 whole image and the 32 × 32 patches are trained, respectively. Code will be made publicly available upon paper acceptance.

3.3 Ablation Study

In Fig. 2 (right), we explore the influence of strong augmentation strategies, represented by rotation, permutation, cutout and Gaussian noise, on the AUC results on Hyper-Kvasir dataset, based on our self-supervised pre-training with IGD as anomaly detector. The experiment indicates that the use of random permutations as strong augmentations yields the best AUC results. We also explore the relation between batch size and AUC results in Fig. 2 (left). The results suggest that small batch size (equal to 16) leads to a relatively low AUC, which increases for batch size 32, and then decreases for larger batch sizes. Given these results, we use permutation as the strong augmentation for colonoscopy images and training batch size is set to 32. For the LAG dataset, we omit the results, but we use rotation as the strong augmentation because it produced the largest AUC. We also used batch size of 32 for the LAG dataset.

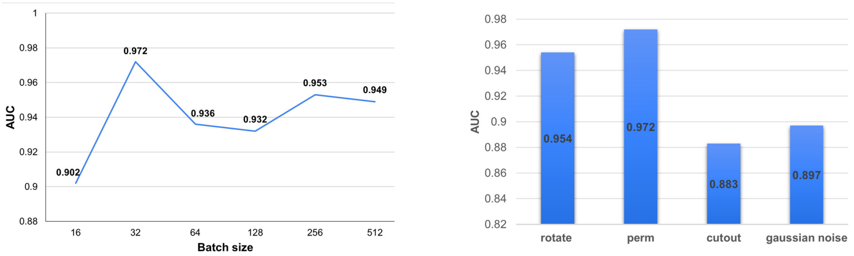

Fig. 2. Left: Anomaly detection performance results based on different batch sizes of self-supervised pre-training. **Right**: Anomaly detection performance in terms of different types of strong augmentations. Both results are on Hyper-Kvasir test set using IGD as anomaly detector.

Table 1. Ablation study of the loss terms in (1) on Hyper-Kvasir, using IGD as anomaly detector.

ℓ_{con} [6,17]	ℓ_{con}	ℓ_{pre}	ℓ_{pat}	AUC - Hyper-Kvasir
✓				0.913
	✓			0.937
	✓	✓		0.964
	✓	✓	✓	**0.972**

Table 2. Anomaly localisation: Mean IoU results on Hyper-Kvasir on 5 different groups of 100 images with ground truth masks. * indicates that we pretrained the geometric transformation-based anomaly detection [13] using IGD [8] as the UAD method.

Supervision	Methods	Localisation - IoU
Supervised	U-Net [36]	0.746
	U-Net++ [50]	0.743
	ResUNet [9]	**0.793**
	SFA [12]	0.611
Unsupervised	RotNet [13]+IGD [8]*	0.276
	CAVGA-R_u [46]	0.349
	Ours - IGD	**0.372**

We also present an ablation study that shows the influence of each loss term in (1) in Table 1, again on Hyper-Kvasir dataset, based on our self-supervised pre-training with IGD. The vanilla contrastive learning in [6,17] only achieves 91.3% of AUC. After replacing it with our distribution contrastive loss from (2), the performance increases by 2.4% AUC. Adding distribution classification and patch position prediction losses boosts the performance by another 2.7% and 0.8% AUC, respectively.

3.4 Comparison to SOTA Models

In Table 3, we show the results of anomaly detection on Hyper-Kvasir, Liu et al.'s colonoscopy dataset and LAG datasets. The IGD, F-anoGAN and MS-SSIM methods improve their baselines (without our self-supervision method) from 3.3% to 5.1% of AUC on Hyper-Kvasir, from −0.3% to 12.2% on Liu et al.'s dataset, and from 0.9% to 7.8% on LAG. The IGD with our pre-trained

Table 3. Anomaly detection: AUC results on Hyper-Kvasir, Liu et al.'s colonocopy and LAG, respectively. * indicates that the model does not use imagenet pre-training.

Methods	Hyper - AUC	Liu et al. - AUC	LAG - AUC
DAE [30]	0.705	0.629 *	–
OCGAN [34]	0.813	0.592 *	–
F-anoGAN [37]	0.907	0.691 *	0.778
ADGAN [26]	0.913	0.730 *	–
CAVGA-R_u [46]	0.928	–	–
MS-SSIM [8]	0.917	0.799	0.823
IGD [8]	0.939	0.787	0.796
RotNet [13]+IGD [8]	0.905	–	–
Ours - MS-SSIM	0.945	0.796	0.839
Ours - F-anoGAN	0.958	0.813	0.787
Ours - IGD	**0.972**	**0.837**	**0.874**

features achieves SOTA anomaly detection AUC on all three datasets. Such results suggest that our self-supervised pre-training can effectively produce good representations for various types of anomaly detectors and datasets. OCGAN [34] constrained the latent space based on two discriminators to force the latent representations of normal data to fall at a bounded area. CAVGA-R_u [46] is a recently proposed approach for anomaly detection and localisation that uses an attention expansion loss to encourage the model to focus on normal object regions in the images. These two methods achieve 81.3% and 92.8% AUC on Hyper-Kvasir, respectively, which are well behind our self-supervised pre-training with IGD of 97.2% AUC.

We also investigate the anomaly localisation performance on Hyper-Kvasir in Table 2. Compared to the SOTA UAD localisation method, CAVGA-R_u [46], our approach with IGD is more than 3% better in terms of IoU. We also compare our results to **fully supervised methods** [9,12,36,50] to assess how much performance is lost by suppressing supervision from abnormal data. The fully supervised baselines [9,12,36,50] use 80% of the annotated 1,000 colonoscopy images containing polyps during training, and 10% for validation and 10% for testing. We validate our approach using the same number of testing samples, but without using abnormal samples for training. The localisation results are post processed by the Connected Component Analysis (CCA) [5]. Notice on Table 2 that we lose between 0.3 and 0.4 IoU for not using abnormal samples for training.

We present visual anomaly localisation results of our IGD with self-supervised pre-training on the abnormal images from Hyper Kvasir [4] test set in Fig. 3. Notice how our model can accurately localise polyps with various size and textures.

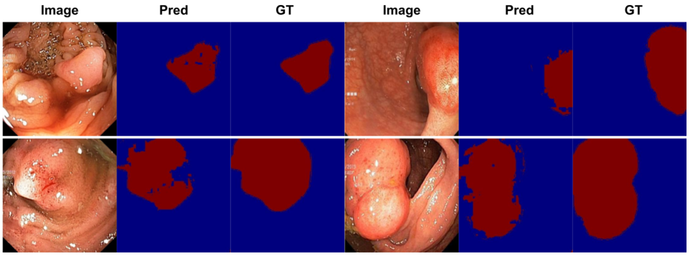

Fig. 3. Qualitative results of our localisation network based on IGD with self-supervised pre-training on the abnormal images from Hyper Kvasir [4] test set.

4 Conclusion

To conclude, we proposed a self-supervised pre-training for UAD named as constrained contrastive distribution learning for anomaly detection. Our approach enforces non-uniform representation distribution by constraining contrastive distribution learning with two pretext tasks. We validate our approach on three medical imaging benchmarks and achieve SOTA anomaly detection and localisation results using three UAD methods. In future work, we will investigate more choices of pretext tasks for UAD.

References

1. Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Scale-space autoencoders for unsupervised anomaly segmentation in brain MRI. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 552–561. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_54
2. Bergman, L., Hoshen, Y.: Classification-based anomaly detection for general data. arXiv preprint arXiv:2005.02359 (2020)
3. Berthelot, D., Raffel, C., Roy, A., Goodfellow, I.: Understanding and improving interpolation in autoencoders via an adversarial regularizer. arXiv preprint arXiv:1807.07543 (2018)
4. Borgli, H., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1), 1–14 (2020)
5. Chai, B.B., Vass, J., Zhuang, X.: Significance-linked connected component analysis for wavelet image coding. IEEE Trans. Image Process. **8**(6), 774–784 (1999)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML, pp. 1597–1607. PMLR (2020)
7. Chen, X., You, S., Tezcan, K.C., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. **64**, 101713 (2020)
8. Chen, Y., Tian, Y., Pang, G., Carneiro, G.: Unsupervised anomaly detection and localisation with multi-scale interpolated gaussian descriptors. arXiv preprint arXiv:2101.10043 (2021)
9. Diakogiannis, F.I., et al.: Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogrammetry Remote. Sens. **162**, 94–114 (2020)

10. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV, pp. 1422–1430 (2015)
11. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26
12. Fang, Y., Chen, C., Yuan, Y., Tong, K.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 302–310. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_34
13. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. arXiv preprint arXiv:1805.10917 (2018)
14. Gong, D., et al.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV, pp. 1705–1714 (2019)
15. Goodfellow, I.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
16. He, K., et al.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
17. He, K., et al.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9729–9738 (2020)
18. Hendrycks, D., et al.: Using self-supervised learning can improve model robustness and uncertainty. arXiv preprint arXiv:1906.12340 (2019)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. CoRR abs/1901.09005 (2019)
21. Li, L., et al.: Attention based glaucoma detection: a large-scale database and CNN model. In: CVPR, pp. 10571–10580 (2019)
22. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
23. Liu, F., Tian, Y., Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: Noisy label learning for large-scale medical image classification. arXiv preprint arXiv:2103.04053 (2021)
24. Liu, F., Tian, Y., et al.: Self-supervised mean teacher for semi-supervised chest x-ray classification. arXiv preprint arXiv:2103.03629 (2021)
25. Liu, F., Jonmohamadi, Y., Maicas, G., Pandey, A.K., Carneiro, G.: Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 594–603. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_58
26. Liu, Y., et al.: Photoshopping colonoscopy video frames. In: ISBI, pp. 1–5 (2020). https://doi.org/10.1109/ISBI45749.2020.9098406
27. Liu, Y., et al.: Photoshopping colonoscopy video frames. In: ISBI, pp. 1–5 (2020)
28. Luo, W., Gu, Z., Liu, J., Gao, S.: Encoding structure-texture relation with p-net for anomaly detection in retinal images
29. LZ, C.T.P., et al.: Computer-aided diagnosis for characterisation of colorectal lesions: a comprehensive software including serrated lesions. Gastrointest. Endosc. **92**(4), 891–899 (2020)
30. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional autoencoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_7

31. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
32. Ouardini, K., et al.: Towards practical unsupervised anomaly detection on retinal images. In: Wang, Q., et al. (eds.) DART/MIL3ID -2019. LNCS, vol. 11795, pp. 225–234. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33391-1_26
33. Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: a review. ACM Comput. Surv. (CSUR) **54**(2), 1–38 (2021)
34. Perera, P., Nallapati, R., Xiang, B.: Ocgan: one-class novelty detection using gans with constrained latent representations. In: CVPR, pp. 2898–2906 (2019)
35. Pu, L., Tao, Z.C., et al.: Prospective study assessing a comprehensive computer-aided diagnosis for characterization of colorectal lesions: results from different centers and imaging technologies. In: Journal of Gastroenterology and Hepatology, vol. 34, pp. 25–26. WILEY 111 RIVER ST, HOBOKEN 07030–5774, NJ USA (2019)
36. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
37. Schlegl, T., et al.: f-anogan: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019)
38. Seeböck, P., et al.: Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. IEEE Trans. Med. Imaging **39**(1), 87–98 (2019)
39. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 1857–1865 (2016)
40. Sohn, K., Li, C.L., Yoon, J., Jin, M., Pfister, T.: Learning and evaluating representations for deep one-class classification. arXiv preprint arXiv:2011.02578 (2020)
41. Tian, Y., otherss: Detecting, localising and classifying polyps from colonoscopy videos using deep learning. arXiv preprint arXiv:2101.03285 (2021)
42. Tian, Y., et al.: One-stage five-class polyp detection and classification. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 70–73. IEEE (2019)
43. Tian, Yu., Maicas, G., Pu, L.Z.C.T., Singh, R., Verjans, J.W., Carneiro, G.: Few-shot anomaly detection for polyp frames from colonoscopy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 274–284. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_27
44. Tian, Y., et al.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. arXiv preprint arXiv:2101.10030 (2021)
45. Uzunova, H., Schultz, S., Handels, H., Ehrhardt, J.: Unsupervised pathology detection in medical images using conditional variational autoencoders. Int. J. Comput. Assist. Radiol. Surg. **14**(3), 451–461 (2018). https://doi.org/10.1007/s11548-018-1898-0
46. Venkataramanan, S., Peng, K.-C., Singh, R.V., Mahalanobis, A.: Attention guided anomaly localization in images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12362, pp. 485–503. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_29
47. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML, pp. 9929–9939. PMLR (2020)

48. Wang, Z., et al.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2, pp. 1398–1402. IEEE (2003)
49. Yi, J., Yoon, S.: Patch svdd: patch-level svdd for anomaly detection and segmentation. In: ACCV (2020)
50. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested u-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1