Group Anomaly Detection using Deep Generative Models

Raghavendra Chalapathy[1] †, Edward Toth[2] ??, and Sanjay Chawla[3]

[1] The University of Sydney and Capital Markets CRC
[2] School of Information Technologies, The University of Sydney
[3] Qatar Computing Research Institute, HBKU

Abstract. Unlike conventional anomaly detection research that focuses on point anomalies, our goal is to detect anomalous collections of individual data points. In particular, we perform group anomaly detection (GAD) with an emphasis on irregular group distributions (e.g. irregular mixtures of image pixels). GAD is an important task in detecting unusual and anomalous phenomena in real-world applications such as high energy particle physics, social media and medical imaging. In this paper, we take a generative approach by proposing deep generative models: Adversarial autoencoder (AAE) and variational autoencoder (VAE) for group anomaly detection. Both AAE and VAE detect group anomalies using point-wise input data where group memberships are known a priori. We conduct extensive experiments to evaluate our models on real world datasets. The empirical results demonstrate that our approach is effective and robust in detecting group anomalies.

1 Anomaly detection: motivation and challenges

This chapter proposes deep generative models (DGMs) for solving the static group anomaly detection (GAD) problem. Generally, a group consists of a collection of two or more related data instances. Pointwise anomaly detection focuses on individual instances while GAD aims to identify groups or collections of observations that do not conform with the expected group patterns. Many pointwise anomaly detection methods cannot detect a variety of group deviations and thus more specialised techniques for robustly differentiating group behaviours.

A group may be anomalous with respect to a variety of statistical properties. Point-based group anomalies are collections of pointwise anomalies, with deviating statistical properties of location. Distribution-based group anomalies contain points that exhibit seemingly regular behaviour however the statistical properties of these groups deviate in terms of scale, shape, dependence and so on. In image applications, we consider images as a group of pixels or visual features. If it is possible to reduce an image into a representative vector of group properties then pointwise anomaly detection may detect anomalous group behaviours. However in practice, it is difficult to capture statistical properties of rich distributions that are present in image datasets.

†Equal Contribution

GAD is a difficult problem for many real-world applications especially involving more complicated group distributions in image applications. Xiong et al. [**?**] propose a novel method for detecting group anomalies however an improvement in detection performance for image applications is possible. When analysing images, it is not obvious how to accurately characterise key properties of images. For example, it is difficult to distinguish cat images (regular groups) from tiger images (anomalous groups) that possess cat whiskers but also irregular features of tiger stripes. The problem of GAD in image datasets is useful and applicable to similar challenging real-world applications where group distributions are more complex and difficult to characterise.

Figure **??** depicts examples of point-based and distribution-based group anomalies where the innermost circle contains images exhibiting regular behaviours whereas the outer circle conveys group anomalies. Figure **??** (A) displays tiger images as point-based group anomalies as well as rotated cat images as distribution-based group anomalies (180° rotation). In plot (B), distribution-based group anomalies are irregular mixtures of cats and dogs in a single image while Figure **??** (C) portrays anomalous images stitched from different scene categories of cities, mountains and coastlines. Our image data experiments will mainly focus on detecting group anomalies in these scenarios.

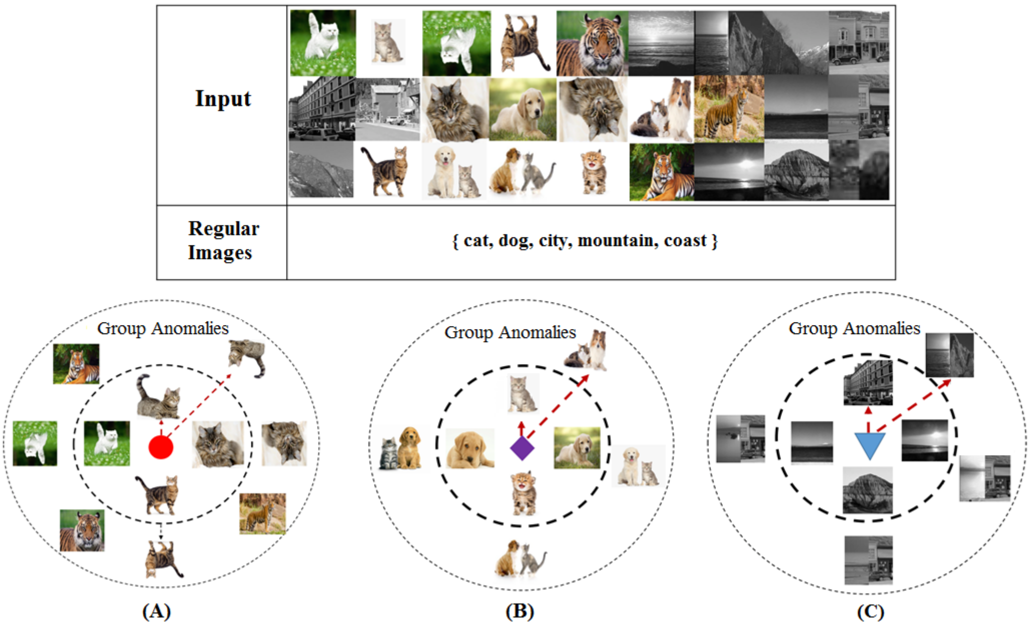


Fig. 1. Examples of point-based and distribution-based group anomalies in our image data experiments. The regular group behaviours represents images in the inner concentric circle while the outer circle contains images that are group anomalies.

Even though the GAD for image datasets may seem like a straightforward comparison of images, many complications and challenges arise. As there is a dependency between the location of pixels in a high-dimensional space, appropriate features in an image are difficult to extract. For effective detection of anomalous images, an adequate characterisation of images is required for model training. Complications in images that potentially arise include low resolution, poor illumination intensity, different viewing angles, different scaling, rotations of images and so on. Like other anomaly detection applications, ground truth labels are usually unavailable for training or evaluation purposes so anomaly injection is recommended. Pre-processing and feature extraction techniques are applicable to resolve some of these challenges.

In order to detect group anomalies in various image applications, we propose using deep generative models (DGMs). The main contributions of this chapter are:

- We formulate DGMs for the problem of detecting group anomalies using a group reference function.
- Although DGMs have been applied in various image applications, they have not been applied to the GAD problem.
- A comparison study is conducted on both synthetic and real-world datasets to demonstrate the effectiveness of DGMs as compared to other GAD techniques.

The rest of the chapter is organised as follows. An overview of related work is provided (Section **??**) and preliminaries for understanding approaches for detecting group anomalies are also described (Section **??**). We formulate our problem and then proceed to elaborate on our proposed solution that involves deep generative models (Section **??**). Our experimental setup and key results are presented in Section **??** and Section **??** respectively. Finally, Section **??** provides a summary of our findings as well as recommends future directions for GAD research.

2 Background and related work on group anomaly detection

3 Related Work

GAD applications are emerging areas of research where most state-of-the-art techniques have been more recently developed. While group anomalies are briefly discussed in anomaly detection surveys such as Chandola et al. [**?**] and Austin [**?**], Xiong [**?**] elaborates on more recent state-of-the-art GAD methods. Yu et al. [**?**] further reviews GAD techniques where group structures are not previously known and clusters are inferred based on pairwise relationships between data instances. Recently Toth and Chawla [**?**] provided a comprehensive overview of GAD methods as well as a detailed description of detecting temporal changes in groups over time. We explore group anomalies where group memberships are known a priori such as in image applications.

Previous studies on image anomaly detection can be understood in terms of group anomalies. Quellec et al. [?] examine mammographic images where point-based group anomalies represent potentially cancerous regions. Perera and Patel [?] learn features from a collection of images containing regular chair objects and detect point-based group anomalies where chairs have abnormal shapes, colors and other irregular characteristics. On the other hand, regular categories in Xiong et al. [?] represent scene images such as inside city, mountain or coast and distribution-based group anomalies are stitched images with a mixture of different scene categories. At a pixel level, Xiong et al. [?] apply GAD methods to detect anomalous galaxy clusters with irregular proportions of RGB pixels. We emphasise detecting distribution-based group anomalies rather than point-based anomalies in our subsequent image applications.

The discovery of group anomalies is of interest to a number of diverse domains. Muandet et al. [?] investigate GAD for physical phenomena in high energy particle physics where Higgs bosons are observed as slight excesses in a collection of collision events rather than individual events. Xiong et al. [?] analyse a fluid dynamics application where a group anomaly represents unusual vorticity and turbulence in fluid motion. In topic modeling, Soleimani and Miller [?] characterise documents by topics and anomalous clusters of documents are discovered by their irregular topic mixtures. By incorporating additional information from pairwise connection data, Yu et al. [?] find potentially irregular communities of co-authors in various research communities. Thus there are many GAD application other than image anomaly detection.

A related discipline to image anomaly detection is video anomaly detection where many deep learning architectures have been applied. Sultani et al. [?] detect real-world anomalies such as burglary, fighting, vandalism and so on from CCTV footage using deep learning methods. In a review, Kiran et al. [?] compare DGMs with different convolution architectures for video anomaly detection applications. Recent work [?,?,?] illustrate the effectiveness of generative models for high-dimensional anomaly detection. Although, there are existing works that have applied deep generative models in image related applications, they have not been formulated as a GAD problem. We leverage autoencoders for DGMs to detect group anomalies in a variety of data experiments.

4 Preliminaries

In this section, a summary of state-of-the-art techniques for detecting group anomalies is provided. We also assess strengths and weaknesses of existing models, compared with the proposed deep generative models.

4.1 Mixture of Gaussian Mixture (MGM) Models

A hierarchical generative approach MGM is proposed by Xiong et al. [?] for detecting group anomalies. The data generating process in MGM assumes that each group follow a Gaussian mixture where more than one regular mixture proportion is possible. For example, an image is a distribution over visual features such as paws and whiskers from a cat image and each image is categorised into possible regular behaviours or genres (e.g. dogs or cats). An anomalous group is then characterised by an irregular mixture of visual features such as a cat and dog in a single image. MGM is useful for distinguishing multiple types of group behaviours however poor results are obtained when group observations do not appropriately follow the assumed generative process.

4.2 One-Class Support Measure Machines (OCSMM)

Muandet et al. [**?**] propose OCSMM to maximise the margin that separates regular class of group behaviours from anomalous groups. Each group is firstly characterised by a mean embedding function then group representations are separated by a parameterised hyperplane. OCSMM is able to classify groups as regular or anomalous behaviours however careful parameter selection is required in order to effectively detect group anomalies.

4.3 One-Class Support Vector Machines (OCSVM)

If group distributions are reduced and characterised by a single value then OCSVM from Schölkopf et al. [**?**] can be applied to the GAD problem. OCSVM separates data points using a parametrised hyperplane similar to OCSMM. OCSVM requires additional pre-processing to convert groups of visual features into pointwise observations. We follow a bag of features approach in Azhar et al. [**?**], where k-means is applied to visual image features and centroids are clustered into histogram intervals before implementing OCSVM. OCSVM is a popular pointwise anomaly detection method however it may not accurately capture group anomalies if the initial group characterisations are inadequate.

4.4 Deep generative models for anomaly detection

This section describes the mathematical background of deep generative models that will be applied for detecting group anomalies. The following notation considers data involving M groups where the mth group is denoted by G_m.

Autoencoders: An autoencoder is trained to learn reconstructions that are close to its original input. The autoencoder consists of encoder f_ϕ to embed the input to latent or hidden representation and decoder g_ψ which reconstructs the input from hidden representation. The reconstruction loss of an autoencoder is defined as the squared error between the input G_m and output $\hat{G}_m$ given by

$$L_r(G_m, \hat{G}_m) = ||G_m - \hat{G}_m||^2 \tag{1}$$

Autoencoders leverage reconstruction error as an anomaly score where data points with significantly high errors are considered to be anomalies.

Variational Autoencoders (VAE): Variational autoencoder (VAE) [**?**] are generative analogues to the standard deterministic autoencoder. VAE impose constraint while inferring latent variable z. The hidden latent codes produced by encoder f_ϕ is constrained to follow prior data distribution $P(G_m)$. The core idea of VAE is to infer $P(z)$ from $P(z|G_m)$ using Variational Inference (VI) technique given by

$$L(G_m, \hat{G}_m) = L_r(G_m, \hat{G}_m) + KL(f_\phi(z|x) \,||\, g_\psi(z)) \tag{2}$$

In order to optimise the Kullback–Leibler (KL) divergence, a simple reparameterisation trick is applied; instead of the encoder embedding a real-valued vector, it creates a vector of means $\boldsymbol{\mu}$ and a vector of standard deviations $\boldsymbol{\sigma}$. Now a new sample that replicates the data distribution $P(G_m)$ can be generated from learned parameters ($\boldsymbol{\mu}$,$\boldsymbol{\sigma}$) and input this latent representation z through the decoder g_ψ to reconstruct the original group observations. VAE utilises reconstruction probabilities [**?**] or reconstruction error to compute anomaly scores.

Adversarial Autoencoders (AAE): One of the main limitations of VAE is lack of closed form analytical solution for integral of the KL divergence term except for few distributions. Adversarial autoencoders (AAE) [**?**] avoid using the KL divergence by adopting adversarial learning, to learn broader set of distributions as priors for the latent code. The training procedure for this architecture is performed using an adversarial autoencoder consisting of encoder f_ϕ and decoder g_ψ. Firstly a latent representation z is created according to generator network $f_\phi(z|G_m)$ and the decoder reconstructs the input $\hat{G}_m$ from z. The weights of encoder f_ϕ and decoder g_ψ are updated by backpropogating the reconstruction loss between $\hat{G}_m$ and G_m. Secondly the discriminator receives z distributed as $f_\phi(z|G_m)$ and z' sampled from the true prior $P(z)$ to compute the score assigned to each ($D(z)$ and $D(z')$). The loss incurred is minimised by backpropagating through the discriminator to update its weights. The loss function for autoencoder (or generator) L_G is composed of the reconstruction error along with the loss for discriminator L_D where

$$L_G = \frac{1}{M'} \sum_{m=1}^{M'} \log D(z_m) \quad \text{and} \quad L_D = -\frac{1}{M'} \sum_{m=1}^{M'} \big[\log D(z'_m) + \log(1 - D(z_m)) \big] \tag{3}$$

where M' is the minibatch size while z represents the latent code generated by encoder and z' is a sample from the true prior $P(z)$.

5 Problem and Model Formulation

Problem Definition: The following formulation follows the problem definition introduced in Toth and Chawla [**?**]. Suppose groups $\mathcal{G} = \big\{\mathbf{G}_m\big\}_{m=1}^{M}$ are observed where M is the number of groups and the mth group has group size N_m with

V-dimensional observations, that is $\mathbf{G}_m \in \mathbb{R}^{N_m \times V}$. In GAD, the behaviour or properties of the mth group is captured by a characterisation function denoted by $f : \mathbb{R}^{N_m \times V} \rightarrow \mathbb{R}^D$ where D is the dimensionality on a transformed feature space. After a characterisation function is applied to a training dataset, group information is combined using an aggregation function $g : \mathbb{R}^{M \times D} \rightarrow \mathbb{R}^D$. A group reference is composed of characterisation and aggregation functions on input groups with

$$\mathcal{G}^{(ref)} = g\Big[\big\{f(\mathbf{G}_m)\big\}_{m=1}^{M}\Big] \tag{4}$$

Then a distance metric $d(\cdot, \cdot) \geq 0$ is applied to measure the deviation of a particular group from the group reference function. The distance score $d\Big(\mathcal{G}^{(ref)}, \mathbf{G}_m\Big)$ quantifies the deviance of the mth group from the expected group pattern where larger values are associated with more anomalous groups. Group anomalies are effectively detected when characterisation function f and aggregation function g respectively capture properties of group distributions and appropriately combine information into a group reference. For example in an variational autoencoder setting, an encoder function f characterises mean and standard deviation of group distributions whereas decoder function g reconstructs the original sample. Further descriptions of functions f and g for VAE and AAE are provided in Algorithm **??**.

5.1 Training the model

The variational and adversarial autoencoder are trained according to the objective function given in Equation (**??**), (**??**) respectively. The objective functions of DGMs are optimised using the standard backpropogation algorithm. Given known group memberships, AAE is fully trained on input groups to obtain a representative group reference $\mathcal{G}^{(ref)}$ described in Equation **??**. While in case of VAE, $\mathcal{G}^{(ref)}$ is obtained by drawing samples using mean and standard deviation parameters that are inferred using VAE as illustrated in Algorithm **??**.

5.2 Predicting with the model

In order to identify group anomalies, the distance of a group from the group reference $\mathcal{G}^{(ref)}$ is computed. The output scores are sorted according to descending order where groups that are furthest from $\mathcal{G}^{(ref)}$ are considered most anomalous. One convenient property of DGMs is that the anomaly detector will be inductive, i.e. it can generalise to unseen data points. One can interpret the model as learning a robust representation of group distributions. An appropriate characterisation of groups results in more accurate detection where any unseen observations either lie within the reference group manifold or deviate from the expected group pattern.

```
Algorithm 1: Group anomaly detection using deep generative models
Input : Groups $\{\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_M\}$ where $\mathbf{G}_m = (X_{ij}) \in \mathbb{R}^{N_m \times V}$
Output: Group anomaly scores $\mathbf{S}$
Train AAE and VAE to obtain encoder $f_\phi$ and decoder $g_\psi$
begin
    switch $C$ do
        case (VAE) do
            $(\mu_m, \sigma_m) = f_\phi(z|\mathbf{G}_m)$ for $m = 1, 2, \ldots, M$
            $(\mu, \sigma) = \frac{1}{M}\sum_{m=1}^{M}(\mu_m, \sigma_m)$
            draw a sample from $z \sim \mathcal{N}(\mu,\, \sigma)$
        end
        case (AAE) do
            draw a random latent representation $z \sim f_\phi(z|\mathbf{G}_m)$
            for $m = 1, 2, \ldots, M$
        end
    end
    for (m = 1 to M) do
        reconstruct sample using decoder $\mathcal{G}^{(ref)} = g_\psi(\mathbf{G}_m|z)$
        compute the score $s_m = d\Big(\mathcal{G}^{(ref)}, \mathbf{G}_m\Big)$
    end
    sort scores in descending order $\mathbf{S}=\{s_{(M)} > \cdots > s_{(1)}\}$
    groups that are furthest from $\mathcal{G}^{(ref)}$ are more anomalous.
    return S
end
```

6 Experimental setup

In this section we show the empirical effectiveness of deep generative models over the state-of-the-art methods on real-world data. Our primary focus is on non-trivial image datasets, although our method is applicable in any context where autoencoders are useful e.g. speech, text.

6.1 Methods compared

We compare our proposed technique using deep generative models (DGMs) with the following state-of-the art methods for detecting group anomalies:

- **Mixture of Gaussian Mixture (MGM) Model**, as per [?].
- **One-Class Support Measure Machines (OCSMM)**, as per [?].
- **One-Class Support Vector Machines (OCSVM)**, as per [?].
- **Variational Autoencoder (VAE)** [?], as per Equation (**??**).
- **Adversarial Autoencoder (AAE)** [?], as per Equation (**??**).

We used Keras [?], TensorFlow [?] for the implementation of AAE and VAE [§]. MGM [¶], OCSMM [‖] and OCSVM [**] are applied using publicly available code.

6.2 Datasets

We compare all methods on the following datasets:

- `synthetic` data follows Muandet et al. [?] where regular groups are generated by bivariate Gaussian distributions while anomalous groups have rotated covariance matrices.
- `cifar-10` [?] consists of 32×32 color images over 10 classes with 6000 images per class.
- `scene` image data following Xiong et al. [?] where anomalous images are stitched from different scene categories.
- `Pixabay` [?] is used to obtain tiger images as well as images of cats and dogs together. These images are rescaled to match dimensions of cat images in `cifar-10` dataset.

The real-world data experiments are previously illustrated in Figure **??**.

6.3 Parameter Selection

We now briefly discuss the model and parameter selection for applying techniques in GAD applications. A pre-processing stage is required for state-of-the-art GAD methods when dealing with images where feature extraction methods such as SIFT [?] or HOG [?] represent images as a collection of visual features. In MGM, the number of regular group behaviours T and number of Gaussian mixtures L are selected using information criteria. The kernel bandwidth smoothing parameter in OCSMM [?] is chosen as median$\{||\mathbf{G}_{m,i} - \mathbf{G}_{l,j}||^2\}$ for all $i, j \in \{1, 2, \ldots, N_m\}$ and $m, l \in 1, 2, \ldots, M$ where $\mathbf{G}_{m,i}$ represents the ith random vector in the mth group. In addition, the parameter for expected proportions of anomalies in OCSMM and OCSVM is set to the true value in the respective datasets.

When applying VAE and AAE, there are four existing network parameters that require careful selection; (a) number of convolutional filters, (b) filter size, (c) strides of convolution operation and (d) activation function. We tuned via grid search of additional hyper-parameters including the number of hidden-layer nodes $H \in \{3, 64, 128\}$ and regularisation λ within range $[0, 100]$. The learning drop-out rates and regularisation parameter μ were sampled from a uniform distribution in the range $[0.05, 0.1]$. The embedding and initial weight matrices are all sampled from uniform distribution within range $[-1, 1]$.

[§] `https://github.com/raghavchalapathy/gad`

[¶] `https://www.cs.cmu.edu/~lxiong/gad/gad.html`

[‖] `https://github.com/jorjasso/SMDD-group-anomaly-detection`

[**] `https://github.com/cjlin1/libsvm`

7 Experimental results

In this section, we explore a variety of GAD experiments. As anomaly detection is an unsupervised learning problem, model evaluation is highly challenging. We employ anomaly injection where known group anomalies are injected into real-world image datasets. The performances of DGMs are evaluated against state-of-the-art GAD methods using area under precision-recall curve (AUPRC) and area under receiver operating characteristic curve (AUROC). AUPRC is more appropriate than AUROC for binary classification under class imbalanced datasets such as in GAD applications [?]. However in our experiments, a high AUPRC score indicates the effectiveness of accurately identifying regular groups while AUROC accounts for the false positive rate of detection methods.

7.1 Synthetic Data: Rotated Gaussian distribution

Firstly we generate synthetic data where regular behaviour consists of bivariate Gaussian samples while anomalous groups have rotated covariance structures. More specifically, $M = 500$ regular group distributions have correlation $\rho = 0.7$ while 50 anomalous groups are generated with correlation $\rho = -0.7$. The mean vectors are randomly sampled from uniform distributions while covariances of group distributions are given by

$$\boldsymbol{\Sigma}_m = \begin{cases} \begin{pmatrix} 0.2 & 0.14 \\ 0.14 & 0.2 \end{pmatrix}, & m = 1, 2, \ldots, 500 \\ \begin{pmatrix} 0.2 & -0.14 \\ -0.14 & 0.2 \end{pmatrix}, & m = 501, 502, \ldots, 550 \end{cases} \tag{5}$$

with each group having $N_m = 1536$ observations. Since we configured the proposed DGMs with an architecture suitable for 32×32 pixels for 3 dimensions (red, green, blue), our dataset is constructed such that each group has bivariate observations with a total of 3072 values.

Parameter settings: GAD methods are applied on the raw data with various parameter settings. MGM is trained with $T = 1$ regular scene types and $L = 3$ as the number of Gaussian mixtures. The expected proportion of group anomalies as true proportion in OCSMM and OCSVM is set to $\nu = 50/\mathrm{M}$ where $M = 550$ or $M = 5050$. In addition, OCSVM is applied by treating each Gaussian distribution as a single high-dimensional observation.

Results: Table **??** illustrates the results of detecting distribution-based group anomalies for different number of groups. For smaller number of groups $M = 550$, state-of-the-art GAD methods achieve a higher performance than DGMs however for a larger training set with $M = 5050$, deep generative models achieve the highest performance. AAE and VAE attain similar results for both synthetic datasets. This conveys that DGMs require larger number of group observations in order to train an appropriate model.

Methods	M=550		M=5050	
	AUPRC	AUROC	AUPRC	AUROC
AAE	0.9060	0.5000	1.0000	1.0000
VAE	0.9001	0.5003	1.0000	1.0000
MGM	0.9781	0.8180	0.9978	0.8221
OCSMM	0.9426	0.6097	0.9943	0.6295
OCSVM	0.9211	0.5008	0.9898	0.5310

Table 1. Task results for detecting rotated Gaussian distributions in synthetic datasets where AAE and VAE attain poor detection results for smaller datasets while they achieve the highest performances (as highlighted in gray) given a larger number of groups.

7.2 Detecting tigers within cat images

Firstly we explore the detection of point-based group anomalies (or image anomalies) by injecting 50 anomalous images of tigers among 5000 cat images. From Figure **??**, the visual features of cats are considered as regular behaviour while characteristics of tigers are anomalous. The goal is to correctly detect images of tigers (point-based group anomalies) in an unsupervised manner.

Parameter settings: In this experiment, HOG extracts visual features as inputs for GAD methods. MGM is trained with $T = 1$ regular cat type and $L = 3$ as the number of mixtures. Parameters in OCSMM and OCSVM are set to $\nu = 50/5050$ and OCSVM is applied with k-means ($k = 40$). Following the success of the Batch Normalisation architecture [**?**] and Exponential Linear Units (elu) [**?**], we have found that convolutional+batch-normalisation+elu layers for DGMs provide a better representation of convolutional filters. Hence, in this experiment the autoencoder of both AAE and VAE adopts four layers of (conv-batch-normalisation-elu) in the encoder part and as well as in the decoder portion of the network. AAE network parameters such as (number of filter, filter size, strides) are chosen to be (16,3,1) for first and second layers while (32,3,1) for third and fourth layers of both encoder and decoder layers. The middle hidden layer size is set to be same as rank $K = 64$ and the model is trained using Adam [**?**]. The decoding layer uses sigmoid function in order to capture the nonlinearity characteristics from latent representations produced by the hidden layer. Similar parameter settings are selected for DGMs in subsequent experiments.

Results: From Table **??**, AAE attains the highest AUROC value of 0.9906 while OCSMM achieves a AUPRC of 0.9941. MGM, OCSMM, OCSVM are associated with high AUPRC as regular groups are correctly identified but their low AUROC scores indicate poor detection of group anomalies. Figure **??**(a) further investigates the top 10 anomalous images detected by these methods and finds that AAE correctly detects all images of tigers while OCSMM erroneously captures regular cat images.

7.3 Detecting cats and dogs

We further investigate GAD detection where images of a single cat and dog are considered as regular groups while images with both cats and dogs are distribution-based group anomalies. The constructed dataset consists of 5050 images; 2500 single cats, 2500 single dogs and 50 images of cats and dogs together. As previously illustrated in Figure **??**(B), our goal is to detect all images with irregular mixtures of cats and dogs in an unsupervised manner.

Parameter settings: In this experiment, HOG extracts visual features as inputs for GAD methods. MGM is trained with $T = 2$ regular cat type and $L = 3$ as the number of mixtures while OCSVM is applied with k-means ($k = 30$).

Results: Table **??** highlights (in gray) that AEE achieves the highest AUPRC and AUROC values. Other state-of-the-art GAD methods attain high AUPRC however AUROC values are relatively low. From Figure **??**(a), the top 10 anomalous images with both cats and dogs are correctly detected by AAE while OCSMM erroneously captures regular cat images. In fact, OCSMM incorrectly but consistently detects regular cats with similar results to subsection **??**.

7.4 Discovering rotated entities

We now explore the detection of distribution-based group anomalies with 5000 regular cat images and 50 images of rotated cats. As illustrated in Figure **??**(A), images of rotated cats are anomalous compared to regular images of cats. Our goal is to detect all rotated cats in an unsupervised manner.

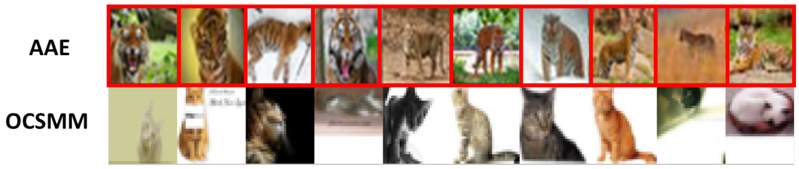
Parameter settings: In this experiment involving rotated entities, HOG extracts visual features because SIFT is rotation invariant. MGM is trained with $T = 1$ regular cat type and $L = 3$ mixtures while OCSVM is applied with k-means ($k = 40$).

Results: In Table **??**, AAE and VAE achieve the highest AUROC with AAE having slightly better detection results. MGM, OCSMM and OCSVM achieve a high AUPRC but low AUROC. Figure **??** illustrates the top 10 most anomalous groups where AAE correctly detects images containing rotated cats while MGM incorrectly identifies regular cats as anomalous.

7.5 Detecting stitched scene images

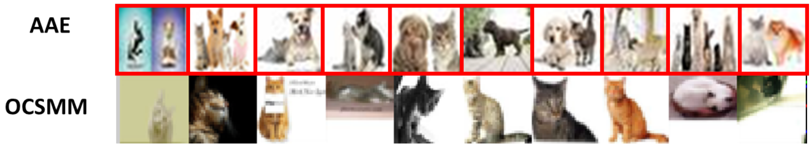
A scene image dataset is also explored where 100 images originated from each category “inside city”, “mountain” and “coast”. 66 group anomalies are injected where images are stitched from two scene categories. Illustrations are provided in Figure **??**(C) where a stitched image may contain half coast and half city street view. These anomalies are challenging to detect since they have the same local features as regular images however as a collection, they are anomalous. Our objective is detect stitched scene images in an unsupervised manner.

Parameter settings: State-of-the-art GAD methods utilise SIFT feature extraction in this experiment. MGM is trained with $T = 3$ regular scene types and $L = 4$ Gaussian mixtures while OCSVM is applied with k-means ($k = 10$).



1
(a)
b

Fig. 2. Tigers within cat images from `cifar-10` dataset.



1
(a)
b

Fig. 3. Images of cats and dogs within single cat and dog images using `cifar-10` dataset.

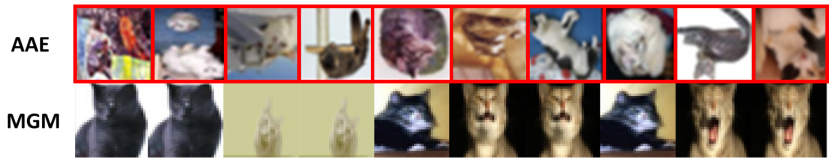
Fig. 4. Top 10 anomalous groups are presented for AAE and the best GAD method respectively where red boxes outlining images represent true group anomalies. AAE has an accurate detection of anomalous tigers injected into the `cifar-10` dataset as well as for anomalous images of both cats and dogs. On the other hand, OCSMM consistently but erroneously identifies similar cat images as the most anomalous images.

The scene image dimensions are rescaled to enable the application of an identical architecture for DGMs as implemented in previous experiments. The parameter settings for both AAE and VAE follows setup as described in Section **??**.

Results: In Table **??**, OCSMM achieves the highest AUROC score while DGMs are less effective in detecting distribution-based group anomalies in this experiment. We suppose that this is because only $M = 366$ groups are available for training in the `scene` dataset as compared to $M = 5050$ groups in previous experiments. Figure **??**(b) displays the top 10 most anomalous images where OCSMM achieves a better detection results than AAE.

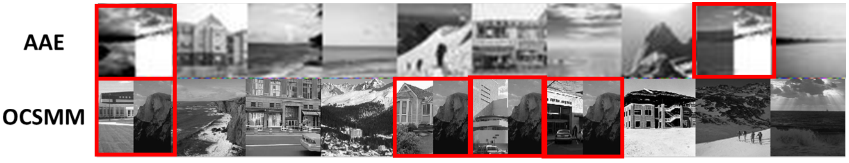
7.6 Results Summary and Discussion

Table **??** summarises the performance of detection methods in our experiments. AAE usually achieves better results than VAE as AAE has the advantage of the embedding coverage in the latent space [**?**]. AAE enforces a better mapping of input variables to embedding space and hence captures more robust input features. Thus AAE achieves the highest detection performance in most experiments however poor results are obtained for `scene` image data due to the limited number of groups. As demonstrated in our `synthetic` data and `scene` images, DGMs have a significantly worse performance on a dataset with a smaller number of groups. Thus given sufficient number of group observations for training,



1
(a)
b

Fig. 5. Rotated cats amongst regular cats in the `cifar-10` dataset.



1
(a)
b

Fig. 6. Stitched Images amongst the `scene` dataset.

Fig. 7. Top 10 anomalous groups are presented where red boxes outlining images represent true group anomalies in the given datasets. AAE performs well in (a) with number of groups $M = 5050$ however does not effectively detect group anomalies in (b) where number of groups is $M = 366$. MGM is unable to correctly detect any rotated cats while OSCMM is able to group anomalies in the `scene` dataset.

DGMs are effective in detecting group anomalies however poor detection occurs for a small number of groups.

Comparison of training times: We add a final remark about applying the proposed DGMs on GAD problems in terms of computational time and training efficiency. For example, including the time taken to calculate SIFT features on the small-scale `scene` dataset, MGM takes 42.8 seconds for training, 3.74 minutes to train OCSMM and 27.9 seconds for OCSVM. In comparison, the computational times for our AAE and VAE are 6.5 minutes and 8.5 minutes respectively. All the experiments involving DGMs were conducted on a MacBook Pro equipped with an Intel Core i7 at 2.2 GHz, 16 GB of RAM (DDR3 1600 MHz). The ability to leverage recent advances in deep learning as part of our optimisation (e.g. training models on a GPU) is a salient feature of our approach. We also note that while MGM and OCSMM are faster to train on small-scale datasets, they suffer from at least $O(N^2)$ complexity for the total number of observations N. It is plausible that one could leverage recent advances in fast approximations of kernel methods [**?**] for OCSMM and studying these would be of interest in future work.

Methods	**Tigers**		**Cats and Dogs**		**Rotated Cats**		**Scene**	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
AAE	0.9449	0.9906	1.0000	1.0000	1.0000	1.0000	0.9449	0.5906
VAE	0.9786	0.9092	0.9998	0.9999	0.9999	0.9999	0.8786	0.3092
MGM	0.9881	0.5740	0.9906	0.5377	0.9919	0.6240	0.8835	0.6639
OCSMM	0.9941	0.6461	0.9930	0.5876	0.9917	0.6128	0.9140	0.7162
OCSVM	0.9909	0.5474	0.9916	0.5549	0.9894	0.5568	0.8650	0.5733

Table 2. Summary of results for various data experiments where first two rows contains deep generative models and the later techniques are state-of-the-art GAD methods. The highest values of performance metrics are shaded in gray.

8 Conclusion

Group anomaly detection is a challenging area of research especially when dealing with complex group distributions such as image data. In order to detect group anomalies in various image applications, we clearly formulate deep generative models (DGMs) for detecting distribution-based group anomalies. DGMs outperform state-of-the-art GAD techniques in many experiments involving both synthetic and real-world image datasets however DGMs require a large number of group observations for model training. To the best of our knowledge, we are the first to formulate and apply DGMs to the problem of detecting group anomalies. A future direction for research involves using recurrent neural networks to detect temporal changes in a group of time series.