When Text and Images Don't Mix: Bias-Correcting Language-Image Similarity Scores for Anomaly Detection

Adam Goodge[1]
goodge_adam_david@i2r.a-star.edu.sg
Bryan Hooi[2]
dcsbhk@nus.edu.sg
Wee Siong Ng[1]
wsng@i2r.a-star.edu.sg

[1] Institute for Infocomm Research (I2R), A*STAR, Singapore
[2] School of Computing, National University of Singapore, Singapore

Abstract

Contrastive Language-Image Pre-training (CLIP) achieves remarkable performance in various downstream tasks through the alignment of image and text input embeddings, and holds great promise for anomaly detection. However, our empirical experiments show that the embeddings of text inputs unexpectedly tightly cluster together, far away from image embeddings, contrary to the model's contrastive training objective to align image-text input pairs. We show that this phenomenon induces a 'similarity bias' - in which false negative and false positive errors occur due to bias in the similarities between images and the normal class label text embeddings. To address this bias, we propose a novel methodology called BLISS which directly accounts for this similarity bias through the use of an auxiliary, external set of text inputs. BLISS is simple, it does not require strong inductive biases about anomalous behaviour nor an expensive training process, and it significantly outperforms baseline methods on benchmark image datasets, even when access to normal data is extremely limited.

1 Introduction

Anomaly detection (AD) is an important task in many vision-related applications, such as medical diagnosis and industrial defect detection. Neural networks are trained to embed input images into a latent space where anomalies are more easily detected. Vision-language models (VLM) such as CLIP [27], which are contrastively trained to align the latent simlarities of image-text input pairs, have surged in popularity for various downstream tasks including anomaly detection recently due to their strong performance. A query image should have high similarity to the text embedding of its class label and low similarity to unrelated text inputs, and this property is exploited for predicting class membership.

In this work, we conduct empirical experiments to examine the latent space learnt by CLIP. We find that, contrary to its contrastive training objective, all text inputs are highly clustered together regardless of the differences in their semantic content, which we call the 'text clustering effect'. The result is that text embeddings of different class labels exhibit significantly higher similarity to other unrelated text inputs than they do to their associated image inputs. This phenomenon violates the behaviour expected of the model after contrastive training, and we find that it also impacts anomaly detection performance by inducing a 'similarity bias' in anomaly scores which are based on the similarity of images to the normal class text embeddings, as some images are more likely to exhibit higher or lower similarity to text inputs regardless of their label. We show in our experiments that this bias is revealed when we measure the similarity of different images to a large, diverse set of general text embeddings, not restricted to only the labels of normal classes. As such, we propose to address this issue through a novel methodology called '**B**ias-corrected **L**anguage **I**mage **S**imilarity **S**coring' (BLISS).

BLISS aims to account for the identified similarity bias by directly incorporating the similarity of test images to general text embeddings into its anomaly scoring mechanism. BLISS is simple, it avoids strong inductive biases about anomalous behaviour, and it is highly efficient; it is entirely inference-based and requires no model training. We show in comprehensive experiments that BLISS outperforms comparable baselines on key benchmark datasets and it exhibits strong robustness to different modifications and problem settings.

In summary, the main contributions of this paper are:

1. We identify a 'text clustering effect' in the latent space of CLIP and analyse how this effect induces a 'similarity bias' in anomaly scoring.
2. We propose a novel methodology called BLISS which accounts for this bias in its anomaly scoring and outperforms competing methods.
3. We examine the performance of different components of our methodology and demonstrate its effectiveness in several experimental settings.

2 Related Work

Existing methods in anomaly detection exploit properties such as the distance to normal samples [3, 12] or a normal hyper-sphere [30] in the latent space to detect anomalies. Some methods train models to perform auxiliary tasks, such as reconstruction [2, 5, 7], generation [1, 31, 32, 36] or classification [11, 13, 33] instead, expecting the model to generalise to other normal samples in the test set but not to anomalies.

We distinguish between 'semantic' anomaly detection, where normal data belongs to a set of known normal class(es) and anomalies are any sample outside of these classes, and another task in which anomalies are seen as small but significant aberrations from normality. The latter task is often seen in the context of industrial defect detection [6, 17, 21, 29], for which some CLIP-based methods have emerged recently [14, 19]. However, these methods typically rely on a high degree of uniformity between normal samples, and also often require some level of prior knowledge about anomalous behaviour, which limits their applicability when anomalies are unknown and unpredictable. Indeed, even the strongest methods in industrial defect detection tend to perform poorly in semantic AD [22].

In this work, we choose to focus on the semantic AD task and make no assumptions about the nature of potential anomalies in the test set. Existing CLIP-based methods are zero-shot [10, 23, 25], meaning they do not exploit information from labelled examples of normal data. In practical settings, it is widely acknowledged that prior examples of normal data are often accessible. and that these examples are highly valuable in capturing the learned notion of normality by the model. Our methodology is designed to exploit this information, not to fine-tune the model parameters but to enhance the anomaly scoring process at test-time.

3 Motivation

In this section, we examine the latent representations of both image and text inputs as learnt by CLIP and identify a phenomenon we call the 'text clustering effect'. We show that this phenomenon induces a 'similarity bias' on the anomaly scores assigned to images based on their similarity to normal class labels, resulting in both false positive and false negative errors and degraded performance. We begin by defining our anomaly detection task.

3.1 Problem Statement

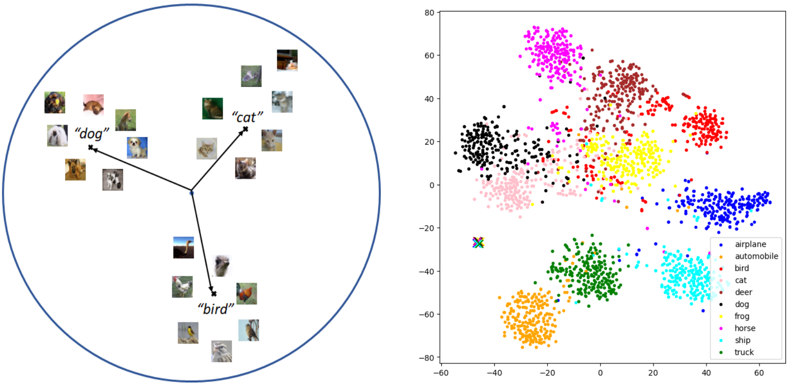
We have labelled normal samples $\mathbf{X}^{(train)} = \{(\mathbf{x}_1^{(train)}, y_1^{(train)}), \ldots, (\mathbf{x}_n^{(train)}, y_n^{(train)})\}$, where $y_i^{(train)}$ is one of the normal classes with text labels $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$. We also have a set of unlabelled test images $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, each of which may be normal (belonging to a class in $\mathcal{C}$) or anomalous. Our goal is to define a function $s : \mathbf{X} \to \mathbb{R}$ which maps an input image $\mathbf{x}$ to an anomaly score $s(\mathbf{x})$ which is low for normal images and high for anomalies.

3.2 Text Clustering Effect

Based on its contrastive training objective, a reasonable expectation of the latent space learnt by CLIP is illustrated in Figure 1(a). Using CIFAR-10 [15] for demonstration, the text embeddings for class labels {'*dog*', '*cat*' and '*bird*'} are distant from each other (exaggerated for clarity) as they are semantically different, while the images tightly cluster around their associated labels. In Figure 1(b), we show that this expectation is not met in reality. We plot the t-SNE projections [34] of the real embeddings of CIFAR-10 class labels (crosses) and images (dots). Following previous work, we convert class labels into text prompts with the template: "*This is a photo of a {class label}*" before being embedded via CLIP. All embeddings are normalized to the unit hyper-sphere.

We immediately see that the text and image embeddings occupy highly separated regions of the latent space. The images embeddings are broadly clustered by class, and these clusters are dispersed throughout the latent space. On the other hand, the text embeddings are tightly clustered together (they all visually overlap at approximately (-45,-25) in the plot). This is despite the fact that most labels are not at all related, e.g. "*airplane*" vs. "*cat*" labels. We call this phenomenon the 'text clustering effect'.

Figure 2 supports this observation. The orange line shows the distribution of cosine similarities of each CIFAR-10 image to their correct class label, which is generally around 0.25 on average. On the other hand, the blue line shows the average similarity between the CIFAR-10 class labels and a large set of general text inputs (we use ImageNet [8] class labels), which we refer to as the 'dictionary'. As the dictionary contains a very large and broad range of concepts, the vast majority of entries should be unrelated to any individual label. We therefore may expect the average similarity of each class label to the dictionary to be relatively low. However, we see that it is actually vastly higher than it is for the image inputs, around 0.75 on average. This means that the class label text inputs are significantly more similar to other text inputs regardless of their content than they are to the images they are supposed to describe. This is despite the fact that CLIP was directly trained to maximise



(a) (b)

Figure 1: (a) The naïve expectation of CLIP's latent space based on its contrastive objective. Text labels which are not semantically similar are separated from each other and the corresponding images cluster around them. (b) t-SNE projections of the true embeddings learnt by CLIP. Text labels (crosses) are highly clustered together, far away from images (dots).

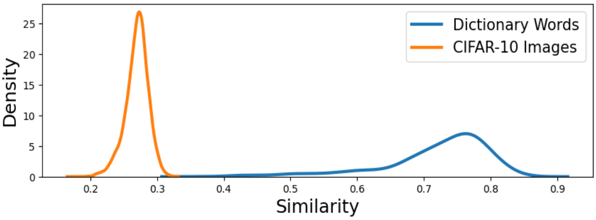


Figure 2: Average cosine similarities of the normalized CLIP embeddings of CIFAR-10 text class labels to the dictionary (blue), and to their associated CIFAR-10 image embeddings (orange).

the similarity between images and their associated text captions. Similar plots in the supplementary material with embeddings from another VLM (BLIP [18]) shows that this text clustering effect is not unique to CLIP but a common feature of vision-language modelling.

3.3 Similarity Bias in Anomaly Scoring

A simple approach to anomaly scoring would be to measure the similarity of a query image to the normal class label(s). A higher similarity suggests the sample is normal, while a lower similarity indicates an anomaly. However, the text clustering effect identified above means that normal class labels are tightly clustered with unrelated text labels, potentially including labels of anomalies. We now investigate the impact of this finding on AD performance.

To this end, we measure the average similarity of CIFAR-10 images to the dictionary mentioned earlier. As the dictionary is large and conceptually diverse, this average similarity can be interpreted as a measure of the similarity of an image to general text embeddings, rather than to the normal class labels specifically. In theory, any given image, whether normal or anomalous, should be meaningfully related to a small number of dictionary entries and unrelated to the vast majority, meaning the average dictionary similarity should between different samples should not contain meaningful patterns. However, in Figure 3, we see that the

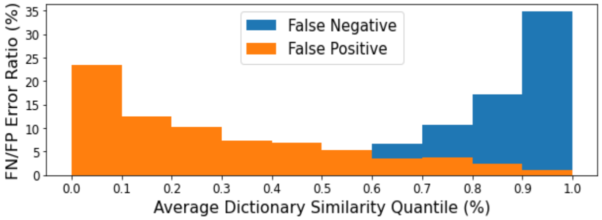


Figure 3: Proportion of false negative (blue) and false positive (orange) errors in each quantile of samples sorted by average similarity to the dictionary.

average similarity of the images to the dictionary words is highly correlated with the proportion of prediction errors. In particular, false negative errors (anomalies classified as normal) are much more frequent amongst images with high average dictionary similarity (right side of the plot). Conversely, false positive errors (normal samples classified as anomalies) are much more frequent in images with low average dictionary similarity (left side of the plot). In other words, anomalies that are relatively more similar to general text embeddings are also more similar to the normal class labels, which makes them appear normal, and vice versa for normal samples that are relatively less similar. We call this phenomenon the 'similarity bias', as the discrepancy in similarities of different images to general text embeddings causes bias in their anomaly scores and degrades AD performance. We aim to address this similarity bias in our proposed methodology by directly accounting for similarity to general text embeddings.

4 Methodology

BLISS consists of two components; an **internal class score** and **external text score**. We rely on the fixed, pre-trained CLIP backbone model with image encoder $\mathcal{I}$ and text encoder $\mathcal{T}$. We firstly obtain all of the embeddings of labelled normal images using the image encoder and store them in a memory bank:

$$\mathbf{Z}^{(train)} = \{\mathbf{z}_1^{(train)}, ..., \mathbf{z}_n^{(train)}\}, \text{ where } \mathbf{z}_i^{(train)} = \mathcal{I}(\mathbf{x}_i^{(train)}), \tag{1}$$

Note that the encoders are fixed and we do not fine-tune any model parameters. For a given unlabelled test image, $\mathbf{x}$, we similarly find its own embedding:

$$\mathbf{z} = \mathcal{I}(\mathbf{x}). \tag{2}$$

We now define the two components, starting with the internal class score.

4.1 Internal Class Score

The internal class score measures the normality of the test image based on its similarity to the normal class label(s). We obtain all of the text embeddings from the N normal class labels with the fixed CLIP text encoder:

$$\mathbf{C} = \{\mathbf{C}_1, .., \mathbf{C}_N\}, \text{ where } \mathbf{C}_i = \mathcal{T}(\mathcal{C}_i). \tag{3}$$

For each normal class, we collect all of the embeddings of samples belonging to that class in $\mathbf{Z}^{(train)}$ and compute their mean (mean) and standard deviation (std) similarity (sim) to $\mathbf{C}_i$:

$$\mu_i = \underset{y_i^{(train)} = \mathcal{C}_i}{\mathsf{mean}} \; sim(\mathbf{z}_i^{(train)}, \mathbf{C}_i), \qquad \sigma_i = \underset{y_i^{(train)} = \mathcal{C}_i}{\mathsf{std}} \; sim(\mathbf{z}_i^{(train)}, \mathbf{C}_i) \tag{4}$$

The internal class score of $\mathbf{x}$ is then the normalized similarity $sim(\mathbf{z}, \mathbf{C}_i)$ with respect to these statistics:

$$IC(\mathbf{x}, \mathcal{C}_i) = -\frac{sim(\mathbf{z}, \mathbf{C}_i) - \mu_i}{\sigma_i + \varepsilon}, \tag{5}$$

where ε is a small constant to avoid division by zero. If $\mathbf{x}$ is normal and belongs to $\mathcal{C}_i$, then $sim(\mathbf{z}, \mathbf{C}_i)$ should be high and therefore $IC(\mathbf{x}, \mathcal{C}_i)$ should be low, and vice versa if $\mathbf{x}$ does not belong to $\mathcal{C}_i$.

4.2 External Text Score

The internal class score measures the similarity of images to the normal class labels. However, we saw that this is vulnerable to similarity bias, as some images are more similar to text embeddings in general, regardless of their class membership. The external text score is designed to address similarity bias by accounting for the similarity of images to general text inputs from a dictionary. We embed each entry in the dictionary $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_t\}$ with the CLIP text encoder:

$$\mathbf{D} := \{\mathbf{D}_1, .., \mathbf{D}_t\}, \text{ where } \mathbf{D}_i = \mathcal{T}(\mathcal{D}_i). \tag{6}$$

For the given test image $\mathbf{x}$, with embedding $\mathbf{z}$, we denote its top K closest matches with the highest similarity in $\mathbf{D}$ as $\mathbf{D}^*$:

$$\mathbf{D}^* = \mathsf{topK}\{sim(\mathbf{z}, \mathbf{D}_i) : \mathbf{D}_i \in \mathbf{D}\} \tag{7}$$

Note that the chosen elements of $\mathbf{D}^*$ are specific to $\mathbf{x}$. As before, we find the mean and standard deviation statistics from the labelled normal images to each $\mathbf{d}^* \in \mathbf{D}^*$:

$$\mu_i(\mathbf{d}^*) = \underset{y_i^{(train)} = \mathcal{C}_i}{\mathsf{mean}} \; sim(\mathbf{z}_i^{(train)}, \mathbf{d}^*) \qquad \sigma_i(\mathbf{d}^*) = \underset{y_i^{(train)} = \mathcal{C}_i}{\mathsf{std}} \; sim(\mathbf{z}_i^{(train)}, \mathbf{d}^*). \tag{8}$$

The external text score is then similarity computed as the normalized mean of $sim(\mathbf{z}, \mathbf{d}^*)$ over all $\mathbf{d}^* \in \mathbf{D}^*$:

$$ET(\mathbf{x}, \mathcal{C}_i) = \underset{\mathbf{d}^* \in \mathbf{D}^*}{\mathsf{mean}} \frac{sim(\mathbf{z}, \mathbf{d}^*) - \mu_i(\mathbf{d}^*)}{\sigma_i(\mathbf{d}^*) + \varepsilon}, \tag{9}$$

Intuitively, $\mu_i(\mathbf{d}^*)$ and $\sigma_i(\mathbf{d}^*)$ capture the learned notion of normality using statistics from each normal class. After correcting for these statistics, test samples with high external text scores are those with high similarity to the dictionary words. From our observation of similarity bias, such samples tend to appear normal in terms of internal class score, and thus have a higher chance of being false negatives. The external text score corrects this by 'subtracting away' this bias, leaving behind a more meaningful view of the similarity between an image and the normal class labels. As our method only requires forward passes through the fixed backbone model, it is light-weight and efficient to compute even with a very large dictionary size.

4.3 BLISS Score

We take a linear combination of the internal class and external text scores, regulated by the hyper-parameter λ:

$$s(\mathbf{x}, \mathcal{C}_i) = IC(\mathbf{x}, \mathcal{C}_i) + \lambda ET(\mathbf{x}, \mathcal{C}_i). \tag{10}$$

In the case of multiple normal classes, i.e. $|\mathcal{C}| > 1$, we repeat scoring over every $\mathcal{C}_i \in \mathcal{C}$ and take the minimum as the final anomaly score:

$$s(\mathbf{x}) = \min_{\mathcal{C}_i \in \mathcal{C}} s(\mathbf{x}, \mathcal{C}_i). \tag{11}$$

In summary, we *jointly* consider similarity not only to the normal class labels (internal) but also to general text embeddings (external), both standardized to account for the learned distribution of normal data.

5 Experiments

We now perform experiments to answer the following questions about our methodology:
RQ1 (Performance): Does BLISS outperform baseline methods in anomaly detection on benchmark datasets?
RQ2: (Ablation Study) How do the components of BLISS perform in different experimental settings?

Datasets We use the most popular datasets in semantic AD: CIFAR-10, CIFAR-100 [15] and TinyImageNet [16]. For CIFAR-10, we partition the 10 classes into 1/6/9 normal classes and 9/4/1 anomaly classes respectively. For 1 and 9 normal class experiments, we run 10 trials - one for each permutation of normal and anomaly class splits. For 6 normal classes, and for CIFAR-100 and TinyImageNet experiments, we follow the same class splits as in [10] over 5 trials.

Model Setup We use the publicly available pre-trained ViT-B/16 [9] CLIP model and do not fine-tune any model parameters. Images are resized, center-cropped and normalized. We use ImageNet class labels as the external dictionary. Classes containing multiple labels, e.g. "*goldfish, Carassius auratus*" are treated as separate entries "*goldfish*" and "*Carassius auratus*", resulting in 1850 dictionary entries. The text prompt "*This is a photo of a* $\{$*class label*$\}$" is formulated for each class label and dictionary entry. We try different prompt formulations in the supplementary material and find little effect on performance. For both modalities, we normalized their 512-dimensional embeddings to the unit hyper-sphere. We set $K = 10$ for the external text score and $\lambda = 0.5$ to approximately match the range of the two scores and we do not conduct hyper-parameter search for optimal performance.

Baselines We reprint the results from [10] of ZOC, CSI [33] and CAC [24], for which some experiments are missing (marked by -). We also implement our own baseline methods as follows. To measure the role of multi-modal learning, we use features from the image-only ViT-B/16 model pre-trained on ImageNet-21k from HuggingFace [35] with AD methods: DN2 (K = 5) [4], LUNAR [12] and Gaussian mixture model (GMM). DINO-FT uses KNN

Dataset $\vert\mathcal{C}\vert$	CIFAR-10 1 class	CIFAR-10 6 classes	CIFAR-10 9 classes	CIFAR-100 20 classes	TinyImageNet 20 classes
CAC [24]	-	80.1 ± 3.0	75.4 ± 6.0	76.1 ± 0.7	76.0 ± 1.5
CSI [33]	-	87.0 ± 4.0	-	80.4 ± 1.0	76.9 ± 1.2
DN2 [4]	97.2 ± 1.2	89.3 ± 3.9	86.0 ± 8.8	83.5 ± 0.8	84.7 ± 1.6
LUNAR[12]	96.7 ± 1.6	88.5 ± 4.2	85.8 ± 8.9	82.9 ± 1.0	83.4 ± 1.7
GMM	97.7 ± 0.9	91.7 ± 3.8	88.0 ± 11.2	86.2 ± 0.6	89.0 ± 1.4
DINO-FT [28]	98.7 ± 0.9	94.0 ± 3.0	89.5 ± 8.3	88.1 ± 0.7	90.2 ± 1.3
CLIP-based Methods					
CAC [24]	-	89.3 ± 2.0	-	83.5 ± 1.2	84.6 ± 1.7
DN2 [4]	96.2 ± 1.9	87.6 ± 5.2	83.7 ± 11.5	78.2 ± 2.8	79.4 ± 2.4
LUNAR [12]	96.3 ± 2.1	87.7 ± 3.9	85.0 ± 10.6	79.0 ± 2.9	80.4 ± 2.2
GMM	97.8 ± 0.9	91.7 ± 2.6	89.5 ± 4.9	77.3 ± 2.8	79.8 ± 2.4
ZOC [10]	-	93.0 ± 1.7	-	82.1 ± 2.1	84.6 ± 1.0
MCM [25]	-	89.8 ± 6.8	87.4 ± 13.6	82.7 ± 1.2	82.6 ± 1.2
BCE-CL [23]	98.6 ± 0.8	93.8 ± 3.5	91.7 ± 6.1	85.6 ± 0.5	87.4 ± 1.4
Biased-CLIP	97.8 ± 0.9	93.2 ± 2.0	91.2 ± 3.3	82.6 ± 2.6	81.3 ± 6.2
BLISS	**99.1 ± 0.4**	**97.3 ± 0.8***	**96.0 ± 1.9***	**89.4 ± 0.6***	**91.1 ± 1.8**

Table 1: AUROC performance of BLISS compared against baselines on several datasets. We show the mean and standard deviation of scores across trials detailed in section 5. The highest score for each dataset is highlighted in bold. * indicates statistical significance at $p = 0.05$ against the next best performing baseline with the one-sided T-test.

with the same ViT architecture but trained via DINO self-distillation [28]. Out of CLIP-based methods, CAC, DN2, LUNAR and GMM are the same methods as above but using CLIP embeddings. MCM [25] and BCE-CL [23] are two contemporary methods which also use text embeddings for anomaly scores, though note that these are zero-shot and do not utilise labelled normal samples. There is no result for MCM in CIFAR-10 with one normal class as the method requires more than one normal class. We present their performance primarily for reference. Biased-CLIP is the internal class score alone, which is vulnerable to similarity bias. For fairness, we do not consider methods which (i) require prior knowledge about anomalies or (ii) fine-tune the parameters of a pre-trained model. We also do not consider older baselines due to poor performance (<90 AUROC on CIFAR-10 one-class experiments). We use PyTorch in Windows with an Nvidia GeForce RTX 2080 Ti GPU.

5.1 RQ1: (Performance)

Table 1 shows the mean and standard deviation in AUROC scores. We use the AUROC metric to avoid choosing an anomaly score threshold. We see that BLISS outperforms all baselines in all datasets and often by a significant margin, a result of accounting for the similarity bias as well as exploiting normality statistics from labelled normal samples.

As TinyImageNet is a subset of ImageNet, all of its class labels are also present in the dictionary. This could be seen as a violation of the assumption of no prior knowledge, therefore we also measured performance after removing all TinyImageNet labels from the dictionary. We find only a small drop in performance from 91.1 to 90.4 AUROC, and this still

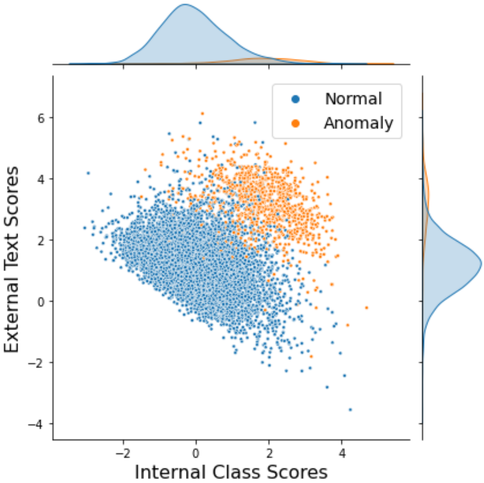


Figure 4: Scatter plot of internal class scores (x-axis) vs. external text scores (y-axis) for normal (blue) and anomaly (orange) samples.

outperforms all baselines. This issue does not affect CIFAR-10 or CIFAR-100 experiments.

Figure 4 shows the internal class and external text scores of both normal (blue) and anomalous (orange) samples independently. Using only the internal class information for scoring is akin to discriminating anomalies based only on the horizontal axis. We see that the optimal decision boundary is closer to a diagonal line, which is achieved through combining the information from both internal class and external text scores.

5.2 RQ2: (Ablation Study)

Weight hyper-parameter Table 2 shows the performance of BLISS as λ in (10) is varied. A higher value gives more weighting to the external text score. As the internal class score is generally smaller in magnitude than the external text score, we heuristically set $\lambda = 0.5$ to approximately equalise their weight. However, we see that $\lambda = 0.75$ is better for performance across all datasets. Overall, performance is robust within a reasonable range of λ.

	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-100	TinyImageNet
λ	1 class	6 classes	9 classes	20 classes	20 classes
0.1	98.34	94.66	92.90	84.94	84.85
0.25	98.82	96.12	94.69	87.4	88.42
0.5	99.14	97.27	96.01	89.43	91.13
0.75	**99.18**	**97.62**	**96.45**	**89.73**	**91.92**
1	99.10	97.59	96.38	89.09	91.88
2	98.32	96.20	94.34	84.13	89.52

Table 2: Mean AUROC scores of BLISS for different values of λ in weighting the two scores. We see that the optimal performance is achieved around $\lambda = 0.75$ across all datasets, but performance is highly robust for a large range.

	CLIP			BLIP			SLIP		
	w/o (5)	w (5)	BLISS	w/o (5)	w (5)	BLISS	w/o (5)	w (5)	BLISS
1 class	94.69	97.80	**99.14**	96.47	96.47	**97.93**	96.08	96.08	**97.96**
6 classes	86.17	93.19	**97.27**	86.50	89.86	**94.20**	84.96	88.46	**92.33**
9 classes	82.16	91.20	**96.02**	85.41	88.13	**92.60**	84.25	86.28	**90.56**

Table 3: Mean AUROC performance of the internal class score without calibration (Eq. 5), (ii) the internal class score with calibration, and (iii) the full BLISS score including external text score, using different backbone VLMs.

Performance with other backbone VLMs [20] shows the embedding gap between different modalities in several multi-modal models. We now examine whether BLISS improves AD performance with other backbone VLMs. In Table 3, we show the performance of the internal class score without normalization using labelled normal samples ("w/o (5)"), with normalization ("w (5)"), and our full BLISS score. We see that BLISS achieves similar performance gains over the internal class score alone with BLIP [18] and SLIP [26] backbone VLMs too, proving the generalisation of our methodology.

Our method diverges from recent zero-shot CLIP-based methods in its use of labelled normal samples. This could be seen as a limitation in settings where labelled normal samples are difficult to obtain. In the supplementary material, we measure the impact of restricting the number of labelled samples. We find that BLISS performance remains highly robust and out-performs the other methods even with as few as one sample per normal class, proving its suitability for few-shot anomaly detection. This shows the importance of correcting for similarity bias. We also find it is robust to changing K in (7). We also test BLISS with other dictionary sources and find that performance is aided by larger and broader dictionaries. Finally, we show that BLISS is highly effective in challenging cases by specially choosing semantically close normal vs. anomaly classes (e.g. horse vs. deer classes). We also measure the false positive rate at 95% recall (FPR95) performance, which is an important metric for AD in practice, and find BLISS also improves performance in this metric.

6 Conclusion

In this paper, we examined the embeddings of image and text inputs in the latent space learnt by CLIP, and discovered that the text embeddings tightly cluster together despite their semantic differences, and this cluster is highly separated from their related images. We investigated how this phenomenon impacts anomaly detection performance using CLIP, identifying the 'similarity bias' in anomaly scores, and this may also have important implications for other downstream tasks. We propose a novel anomaly scoring approach called BLISS to address this bias by additionally measuring the similarity to a large, external source of general text inputs from outside the normal class labels. BLISS is fast, anomaly-agnostic and achieves state-of-the-art performance on popular benchmark datasets. More broadly, our findings highlight the need for greater understanding of multi-modal learning and the mitigation of their unintended and unexpected characteristics, which is a goal of fundamental importance.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018.

[2] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

[3] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *Principles of Data Mining and Knowledge Discovery: 6th European Conference, PKDD 2002 Helsinki, Finland, August 19–23, 2002 Proceedings 6*, pages 15–27. Springer, 2002.

[4] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020.

[5] Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. Autoencoder-based network anomaly detection. In *2018 Wireless telecommunications symposium (WTS)*, pages 1–5. IEEE, 2018.

[6] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021.

[7] Ailin Deng, Adam Goodge, Lang Yi Ang, and Bryan Hooi. Cadet: Calibrated anomaly detection for mitigating hardness bias. IJCAI, 2022.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, 2022.

[11] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018.

[12] Adam Goodge, Bryan Hooi, See-Kiong Ng, and Wee Siong Ng. Lunar: Unifying local outlier detection methods via graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6737–6745, 2022.

[13] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[14] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[16] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[17] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[19] Yiting Li, Adam Goodge, Fayao Liu, and Chuan-Sheng Foo. Promptad: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1093–1102, 2024.

[20] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[21] Jingyi Liao, Xun Xu, Manh Cuong Nguyen, Adam Goodge, and Chuan Sheng Foo. Coft-ad: Contrastive fine-tuning for few-shot anomaly detection. *IEEE Transactions on Image Processing*, 2024.

[22] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023.

[23] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *arXiv preprint arXiv:2205.11474*, 2022.

[24] Dimity Miller, Niko Suenderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3570–3578, 2021.

[25] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

[26] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pages 529–544. Springer, 2022.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[28] Tal Reiss, Niv Cohen, Eliahu Horwitz, Ron Abutbul, and Yedid Hoshen. Anomaly detection requires better representations. *arXiv preprint arXiv:2210.10773*, 2022.

[29] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

[30] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

[31] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[32] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

[33] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[35] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

[36] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.