Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Towards Visually Interpreting Variational Autoencoders

Runze Li

rli047@ucr.edu

UC Riverside

Wenqian Liu

Inria CHROMA

Meng Zheng

United Imaging Intelligence

Max Torop

Northeastern University

Milind Rajadhyaksha

Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-6323-4547

Jennifer Dy

Northeastern University

Kivanc Kose

Memorial Sloan Kettering Cancer Cente https://orcid.org/0000-0003-3185-2639

Srikrishna Karanam

Adobe Research

Ziyan Wu

United Imaging Intelligence

Bir Bhanu

UC Riverside https://orcid.org/0000-0001-8971-6416

Richard Radke

Rensselaer Polytechnic Institute

Octavia Camps

Northeastern University

Article

Keywords: Attention maps, anomaly localization, disentanglement learning, explainable, latent space learned by VAE

Posted Date: May 10th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4331222/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Towards Visually Interpreting Variational Autoencoders

Runze Li[1*†], Wenqian Liu[2*†], Meng Zheng[3], Max Torop[4], Milind Rajadhyaksha[5], Jennifer Dy[4], Kivanc Kose[5], Srikrishna Karanam[6], Ziyan Wu[7], Bir Bhanu[2], Richard J. Radke[3], Octavia Camps[4]

[1]University of California Riverside, Riverside, CA, USA.
[2]Inria CHROMA , Grenoble, France.
[3]Rensselaer Polytechnic Institute, Troy, NY, USA.
[4]Northeastern University, New York, NY, USA.
[5]Memorial Sloan Kettering Cancer Center, Boston, MA, USA.
[6]Adobe Research, Bangalore, India.
[7]United Imaging Intelligence, Cambridge, MA, USA.

*Corresponding author(s). E-mail(s): rli047@ucr.edu; wenqian.liu@inria.fr;
†These authors contributed equally to this work.

Abstract

Recent advances in convolutional neural network (CNN) interpretability have led to a wide-variety of gradient-based visual attention techniques for generating visual attention maps. However, most of these methods require a classification-type design architecture, and consequently concentrate on classification/categorization-type tasks. Extending these methods to generate visual attention maps for other kinds of computer vision models, e.g., variational autoencoders (VAE) is not trivial. In this paper, we present a method that helps bridge this crucial gap, proposing to compute *VAE attention* as a means for interpreting the latent space learned by a VAE. We first present methods to generate visual attention maps from the learned latent space, and then show how they can be used in a variety of applications: localizing anomalies in images, including medical imagery, and improved latent space disentanglement. We conduct extensive experiments on a wide-variety of benchmark datasets to demonstrate the efficacy of the proposed VAE attention.

Keywords: Attention maps, anomaly localization, disentanglement learning, explainable, latent space learned by VAE.

1 Introduction

As computer vision models, driven by deep learning [1–3], begin to be successfully transitioned to real-world tasks like healthcare and robotics [4–6], applications in these areas demand a clear understanding of the algorithm's reasoning apart from performance robustness. This has led to much recent interest in methods that help understand and explain the underlying *what* is driving the output and *why*.

Following the visualization method of Zeiler and Fergus [7], there has been substantial recent progress in designing techniques to visualize CNN feature activations. One such line of work uses the concept of CNN network attention [8, 9], typically described by means of attention maps that show which features are considered by a trained model to be "important" for satisfying the criteria used to train the model. With these methods, one can generate attention maps that visualize object regions, e.g., a *cat*, that help in interpreting why the model classified the image to the *cat* category. Other extensions [10–12] propose methods that use these attention maps as explicit trainable constraints and show improved downstream performance and visual interpretability. However, a key limitation of both class-activation-mapping (CAM)-type methods [8, 9] as well as recent extensions [10–12] is the need for an explicit classification module, typically achieved by means of cross-entropy-type losses, to generate these attention maps, limiting most of their use to problems involving classification/categorization.

Given the aforementioned progress in interpreting classification models, one would naturally like to explain a wider variety of computer vision models. For instance, while progress in algorithms for generative modeling has been rapid [13–21], there is still much room for research in visually interpreting these models. Particularly, variational autoencoders (VAEs) are a widely used class of generative models to solve a variety of computer vision tasks. However, it is apparent from these works that: (i) the latent space learned from VAEs lacks visual interpretability, (ii) while VAEs are foundational generative models, the capabilities of the latent representations learned from VAEs have not been fully utilized, (iii) VAEs are often integrated as a tool for designing specific downstream models, rather than being used in their own right. While there is ongoing work in *using* the concept of visual attention in generative models [22–24], the focus here is on using network attention as an auxiliary information source (e.g., for training), and not to interpret/explain the underlying model. This paper seeks to address the gaps in interpreting the latent space of VAEs and to fully explore capabilities of using representations of learned latent space in VAEs for various applications: localizing anomalies in images such as surveillance videos and medical imagery, as well as achieving improved latent space disentanglement. We distinguish our work from the previous research as outlined below:

- We focus on visually interpreting VAEs by generating gradient-based attention maps from the learned latent space called VAE attention, which has not been done in existing research.
- We introduce two forms of VAE attention, *i.e.*, Sample Attention and Distribution Attention, which have not been explored in previous work.
- We propose a novel approach, to generate gradient-based attention maps from the learned latent space with statistical features and our approach outputs rich interpretations for VAEs, which is distinct from existing methods that focus on deterministic models.
- We propose to use our novel attention maps to improve the disentanglement representations learning on VAEs, which has not been done before.
- We have validated our proposed method in a variety of applications. Our proposed method has achieved significant performance for anomaly localizing on a variety of datasets and given meaningful attention maps to capture anomalous for qualitative analysis. Our proposed method has also benefited disentangled representations learning by using interpretable attention maps.

To the best of our knowledge, this is the first work that systematically and visually interprets the VAEs using gradient-based attention maps, and explores a wide range of applications with generated attention maps.

The key intuition underlying our VAE attention is that the latent space encapsulates all the properties of a trained VAE and that computing attention maps conditioned on these latent variables will help interpret any downstream VAE predictions. Given that a trained VAE that represents a distribution over all the latent variables, we seek to model VAE attention by considering not just a single sample but also the entire learned distribution. To this end, we propose two ways to calculate and generate VAE attention: *Sample Attention* and *Distribution Attention*. Given an input image, our methods first infer a latent space distribution from the encoder of the trained VAE. Given this distribution, *Sample attention* samples one latent code and computes derivatives (with respect to this code) of the encoder's convolutional feature maps at the layer immediately preceding fully connected units, which are then aggregated into the visual attention map for this input image. While this method can help interpret the VAE from the perspective of a single sample of the latent space, there is typically much more information in the entire distribution that what a single sample can capture. As a result, our proposed *Distribution Attention* collects a number of latent codes (by repeated sampling) from the distribution, generates per-code sample attention maps, and fits a pixel-wise Gaussian distribution. The resulting "mean" and "standard deviation" attention maps help visually interpret not only the prediction confidence of the VAE but also the associated uncertainty.

While these visual attention maps help interpret the latent space (which we show with qualitative results), we show how they can be used in many more creative ways. For instance, a classical VAE application involves localizing anomalies, where the idea is, in the learned latent space, any input not from the Gaussian distribution used to train the VAE is anomalous. Our VAE attention maps can naturally be used to "interpret" this behavior by conditioning our attention computation process on the anomalous latent codes. These attention maps can also be used as cues to go a

step further and precisely localize the anomaly in the image. We conduct extensive experiments on multiple natural, industrial, and medical image datasets and present state-of-the-art anomaly localization results with standard VAE models (i.e., no bells and whistles) while also generating attention maps that help interpret predictions.

Another key research topic in the study of VAEs is in disentangling the learned latent space, and there has been much recent progress here as well [25–27]. Our VAE attention maps can naturally be used to improve the disentanglement performance. The key idea is once we compute VAE attention maps conditioned on the learned latent space, they can serve as explicit trainable constraints during the model learning process. To this end, we present a new learning objective called attention disentanglement loss and show how it can be used in conjunction with existing/standard VAE models to improved latent space disentanglement (as measured by standard evaluation metrics). We demonstrate this by means of extensive experiments on the Dsprites [28] and Fashion-MNIST [29] datasets.

To summarize, our key contributions are:

- We take a step towards addressing the relatively unexplored problem of visually interpreting variational autoencoders, presenting techniques to generate visual attention maps, called VAE attention, conditioned on latent codes sampled from the latent space of a trained VAE.
- We demonstrate two different ways to generate VAE attention: Sample Attention (SA) which are generated simply from a sample of latent vector, and Distribution Attention (DA) which are computed from distributions of sampled latent vectors in a statistical manner.
- We show how VAE attention can be put to multipurpose use apart from interpreting the latent space.
  - First, we demonstrate the use of VAE attention in localizing anomalies in images. We show both qualitative results with VAE attention maps and state-of-the-art quantitative performance on multiple natural, industrial, and medical image datasets.
  - Next, we show how the proposed VAE attention maps lead to a new learning objective, called the attention disentanglement loss, that can be used in conjunction with standard VAE models. We demonstrate it leads to improved disentanglement performance by means of extensive experiments on standard benchmark datasets.

2 Results

To prove the validity of our attention map generation methods, we first conduct a qualitative sanity test on the MNIST dataset [30]. All training and testing images are at resolution of 28×28 pixels. We train a one-class VAE model using the training images of one digit only and test on images of other digits from the testing set. For instance, we pick digit "1" as our training set to train a VAE. Then we could feed a digit "7" during testing. We expect that our approach will generate attention maps showing the anomalous regions of a "7" different from a "1".

Qualitative results are shown in Fig. A1 in Appendix A. For instance, with the top-left example, we trained a VAE model with images of digit "1" as the normal class and test on images of all other digits as abnormal classes. For each test image, we infer the latent vector from the encoder of the trained VAE model and generate the attention map by sampling latent codes from the normal difference distribution. As can be observed from the results, the attention maps computed with the proposed method is intuitively satisfying. For example, with test data from class "7" as abnormal samples, our intuition is that a key difference between the "1" and the "7" is the top-horizontal bar in "7", and our generated attention map indeed highlights this region. One can make similar observations from the results of other digits as well (e.g.,"3").

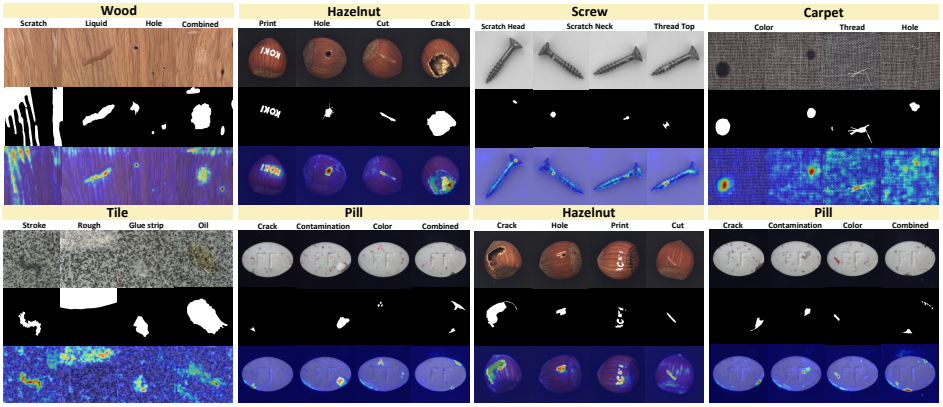
Unsupervised Anomaly Localization

We now present experimental results for unsupervised anomaly localization using proposed Sample Attention and Distribution Attention techniques on multiple datasets. We adopt the commonly used, area under the receiver operating characteristic curve (ROC AUC), for evaluation of all quantitative scores. We define the true positive rate (TPR) as the percentage of samples that are correctly classified as anomalous across the whole testing class, and the false positive rate (FPR) as the percentage of samples that are wrongly classified as anomalous. For image-level performance, the AUROC value is calculated based on the anomaly score generated for each testing sample. For pixel-level performance, the AUROC value is generated based on attention maps and ground truth masks. For each given method, higher AUROC values indicate better performance at detecting anomalies. We will show pixel-level anomaly classification quantitative scores here. But we also conduct image-level anomaly classification experiments shown in Appendix, along side with additional qualitative results. In addition, we also use our Sample Attention technique to compute the best intersection-over-union (IOU) score by searching for the best threshold based on the ROC curve as an extra comparison against other existing methods on the MVTec-AD dataset.

MVTec-AD

We first conduct extensive qualitative and quantitative experiments on MVTec-AD[31] dataset. MVTec-AD is a real-world industrial image anomaly detection dataset with 5354 high resolution images in 15 categories. The training set has 3629 normal images, and the test set contains 1725 normal or abnormal images. The ground-truth in the test set includes both labels and anomaly masks. We follow the original dataset split of MVTec-AD, i.e., use only anomaly-free images in training, and test on both normal and abnormal images.

We use the ResNet18 [2] architecture as the backbone for the feature encoder of the one-class VAE (with a 32-dimensional latent space). As in the original work [31], we also perform random mirroring and random rotation for data augmentation on the training set. After obtaining the SA and DA maps, binary localization maps are computed by using a variety of thresholds on the pixel response values, which is encapsulated in the ROC curve. We then calculate the area under the ROC curve and

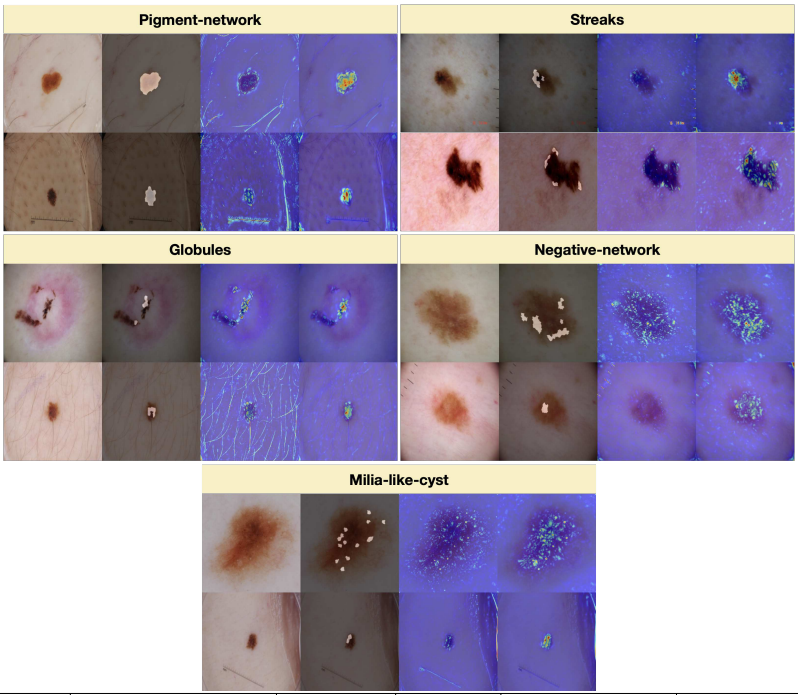


	Category	SSIM-AE[32]	l_2-AE[33]	AnoGAN[34]	CNN-FD[35]	SMAI[36]	GDR[37]	P-Net[38]	PatchNet[39]	**Ours (SA)**	**Ours (DA)**
Texture	Carpet	0.87	0.59	0.54	0.72	0.88	0.74	0.57	**0.96**	0.78	0.8
	Grid	0.94	0.90	0.58	0.59	0.97	0.96	**0.98**	0.78	0.73	0.75
	Leather	0.78	0.75	0.64	0.87	0.86	0.93	0.89	0.90	0.95	**0.96**
	Tile	0.59	0.51	0.50	0.93	0.62	0.65	**0.97**	0.80	0.80	0.80
	Wood	0.73	0.73	0.62	0.91	0.80	0.84	**0.98**	0.81	0.77	0.81
Object	Bottle	0.93	0.86	0.86	0.78	0.86	0.92	**0.99**	0.93	0.87	0.89
	Cable	0.82	0.86	0.78	0.79	0.92	0.91	0.70	**0.94**	0.90	0.93
	Capsule	**0.94**	0.88	0.84	0.84	0.93	0.92	0.84	0.90	0.74	0.74
	Hazelnut	0.97	0.95	0.87	0.72	0.97	**0.98**	0.97	0.84	**0.98**	**0.98**
	Metal Nut	0.89	0.86	0.76	0.82	0.92	0.91	0.79	0.84	**0.94**	**0.94**
	Pill	0.91	0.85	0.87	0.68	0.92	0.93	0.91	**0.93**	0.83	0.83
	Screw	0.96	0.96	0.80	0.87	0.96	0.95	**1.00**	0.96	0.97	0.98
	Toothbrush	0.92	0.93	0.90	0.77	0.96	**0.99**	**0.99**	0.96	0.94	0.95
	Transistor	0.90	0.86	0.80	0.66	0.85	0.92	0.82	**1.00**	0.93	0.93
	Zipper	0.88	0.77	0.78	0.76	0.90	0.87	0.90	**0.99**	0.78	0.82

Fig. 1: **Top Figure: MVTec-AD SA Maps** Qualitative results of anomaly localization with SA from MVTec-AD. For each category, we show four types of defects(anomalies) provided by the dataset. Our anomaly attention maps are able to accurately localize anomalies. **Bottom Table: MVTec-AD AUROC** Quantitative results for pixel-level anomaly classification on 15 categories of MVTec-AD dataset. We adopt comparison scores from [39]. We compare to two sets of baselines. The first set includes *domain-general* methods: SSIM-AE[32], l_2-AE[33], AnoGAN[34], and CNN-FD[35]. The second set includes *domain-specific* methods: SMAI[36], GDR[37], P-Net[38], and PatchNet[39]. SA: Sample Attention. DA: Distribution Attention.

the best IOU number is computed based on the FPR and TPR values from the ROC curve.

The quantitative results of AUROC scores for pixel-level anomaly classification on MVTec-AD are presented in the bottom table of Figure 1. And Table A1 in Appendix A shows the best IOU scores by using Sample Attention only. The performance of our proposed method is evaluated using the same evaluation techniques as in the benchmark paper of Wang *et al.* [39] (note that the baselines here are the same methods as in [31, 39]). We compare to two sets of baselines. The first set includes domain-general methods: SSIM-AE[32], l_2-AE[33], AnoGAN[34], and CNN-FD[35]. The second set includes domain-specific methods: SMAI[36], GDR[37], P-Net[38], and PatchNet[39]. From the bottom table of Figure 1, one can note that the results of our



	Pigment-network	Globules	Streaks	Milia-like-cyst	Negative-network
diffmap	0.74	0.71	0.85	**0.72**	0.67
Ours (SA)	0.75	0.79	0.95	0.67	**0.77**
Ours (DA)	**0.77**	**0.81**	**0.97**	0.71	**0.77**

Fig. 2: **Top Figure: ISIC SA Maps** We show our qualitative attention maps localize on all five attributes. We plot input image, ground-truth mask over lesion, diffmap over lesion and our generated attention map over lesion from left to right. **Bottom Table: ISIC AUROC** AUROC scores for pixel-level anomaly localization accuracy. We compare our gradient based attention maps to the difference maps between input and reconstructed images

method are competitive with respect to these methods, including better performance on some categories. **It is worth noting here that most of these methods are specifically designed for the anomaly localization task with task-specific network designs, whereas we train a standard VAE and generate our VAE attention maps for anomaly localization. Despite this simplicity, our method achieves competitive performance, demonstrating the potential of such an attention generation technique to be useful for tasks other than just model interpretation**. It is also observed that our Distribution Attention can consistently improve anomaly localization performance across most of the categories, which proves that rather than using only *one sample* from the latent space (NDD), considering the collections of attention maps generated with both mean and standard

deviation from the learned distribution (NDD) can further help interpret the VAE model and in turn improve anomaly detection results.

Qualitative results of generated attention maps using SA are shown in the top figure of Fig. 1 (top to bottom: original image, ground truth segmentation mask, SA anomaly attention maps) on a wide range of categories. For each category, we show four types of defects(anomalies) provided by the dataset. One can note that our attention maps are able to accurately localize anomalous regions across these various defect categories. More qualitative results are shown in Fig. A2 in Appendix A that we compare the attention maps generated with SA and DA techniques. We can see that DA can generally localize anomaly regions more accurate than SA.

ISIC 2019 Challenge and Dataset

Besides anomaly detection or diseases classification, deep neural networks (DNNs) have also been widely used in the task of skin lesion segmentation. In this paper, we pursued lesion attribution segmentation for melanoma detection on the ISIC challenge and dataset. The ISIC 2019 dataset includes dermatology images from the HAM10000 [40], BCN20000[41] and the MSK [42] datasets. While the ISIC 2018 dataset only consists of HAM10000 data. The BCN20000 dataset is particularly challenging as it includes uncurated images from day-to-day clinical practice [41]. The training set consists of 25,331 skin lesion images from eight diagnostic categories: Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AK), Benign keratosis (BKL), Dermatofibroma (DF), Vascular lesion (VASC), and Squamous cell carcinoma (SCC). The testing set consists of 8238 images, including 2000 images that are not contained in the diagnostic categories provided in the training split (unknown category, UNK). The ISIC 2018 also comprises the skin lesion attribute detection challenge, which provides response binary mask that identifies 5 dermoscopic attributes for each corresponding lesion image. These 5 attributes are: pigment-network, negative-network, streaks, milia-like-cyst and globules. There are in total 2,594 images with 12,970 ground-truth segmentation masks for the 5 attributes that were provided[43]. Also as mentioned in the official site of the dataset, not all the attributes have ground-truth masks presented for each skin image.

We train our VAE on training samples from the ISIC 2018 dataset, excluding all the images within this task. Our model is trained unsupervised without using any label or meta data. Then we test our model to generate attention maps for these 2,594 samples using the proposed SA and DA techniques. Without imposing any information from the labels, our gradient-based attention maps are highly overlapped with the attribute ground truth masks, highlighting the diagnostically significant lesion structures within the image. We also calculate the difference map between an input image and the its reconstructed image, referred as *diffmap*, to compare with our attention maps. From the bottom table in Fig. 2 we also see using the proposed DA maps can consistently improve pixel-level anomaly localization performance.

For the sake of simplicity, we show qualitative results of our SA maps in the top figure of Fig. 2. We plot input image, ground-truth mask over lesion, diffmap over lesion and our generated attention map over lesion from left to right. From the figure, we show that SA maps accurately heat up regions that coincide with the ground-truth

masks even though trained unsupervised. While the diffmaps are unable to localize the masked regions. We also notice that negative-network and milia-like-cyst tend to be more challenging than the other 3 types of attributes. One of the possible reason is that these attributes have subtle features that are not easy to capture, which localize at very small areas as compared to the whole lesion.

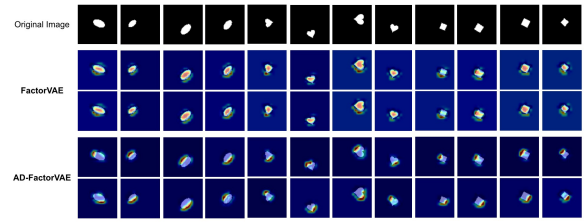
Disentangled Representation Learning

Next, we move to experimental results for disentangled representation learning with our proposed attention disentanglement loss (L_{AD}) on the widely adopted dSprites dataset [28]. There are images of 2D shapes generated from 6 ground truth independent latent factors provided by this dataset. These factors are color, shape, scale, rotation, x and y positions of a sprite. In total, the dataset contains 737,280 binary images with shape of 64×64. And we adopt the disentanglement metric proposed in [26] to evaluate the disentanglement performance of learned VAE latent space.

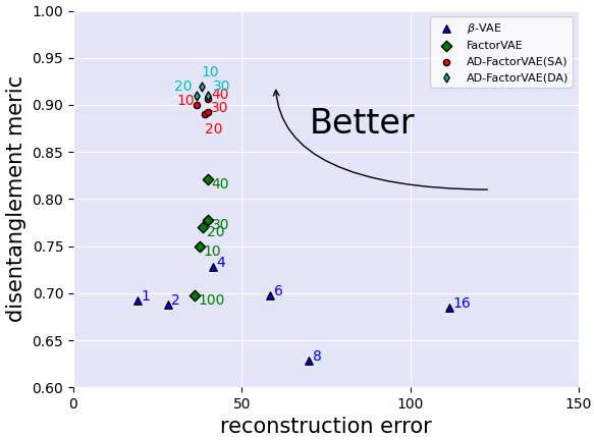
Fig. 3b presents the quantitative comparisons of the best disentanglement performance which is plotted against the reconstruction error between our proposed method (called AD-FactorVAE) with other competing methods: baseline FactorVAE [26] which is trained without attention disentanglement loss and β-VAE[25]. From Fig. 3b, it is observed that training the model by using overall learning objectives with our proposed disentanglement loss L_{AD} results in higher disentanglement scores under the same experimental settings. In more details, our method can give a best disentanglement score of around 0.90, whereas baseline FactorVAE ($\gamma = 40$) gives around 0.82, both with a reconstruction error around 40. Our proposed method also gives a higher disentanglement score compared to the best score of β-VAE which is 0.73 with $\beta = 4$. It is demonstrated that the potential of both our proposed VAE attention and L_{AD} in improving the performance of existing methods in the disentanglement literature.

Fig. 3b also presents experiments results for disentanglement learning with *distribution attention* (called AD-FactorVAE (DA)) on Dsprites dataset. During the training VAE, different from using *one sample* from the learned latent space to generate VAE attention, we generate the *distribution attention* maps $\mu^{\mathbf{M}}$ and $\sigma^{\mathbf{M}}$ after sampling a collection of latent codes and use these two attention maps to calculate the *attention disentanglement* loss. From the Fig. 3b, it is observed that training the model with our proposed disentanglement loss L_{AD} using the DA maps can consistently improve results in higher disentanglement scores under the same experimental settings.

Fig. 3a shows some attention maps generated using the baseline FactorVAE and our proposed AD-FactorVAE. The first row shows 5 input images in different shapes, and the next 4 rows show attention results with the baseline FactorVAE and our proposed method. Row 2 shows attention maps generated with FactorVAE by backpropagating from the latent dimension with the highest response, whereas row 3 shows attention maps generated by backpropagating from the latent dimension with the next highest response. Rows 4 and 5 show the corresponding attention maps generated from the model trained with the proposed AD-FactorVAE. Our intuition and expectation with AD-FactorVAE is that each dimension's attention map will have high responses in different spatial regions of the input. From Fig. 3a, we see that this is indeed the case



(a)

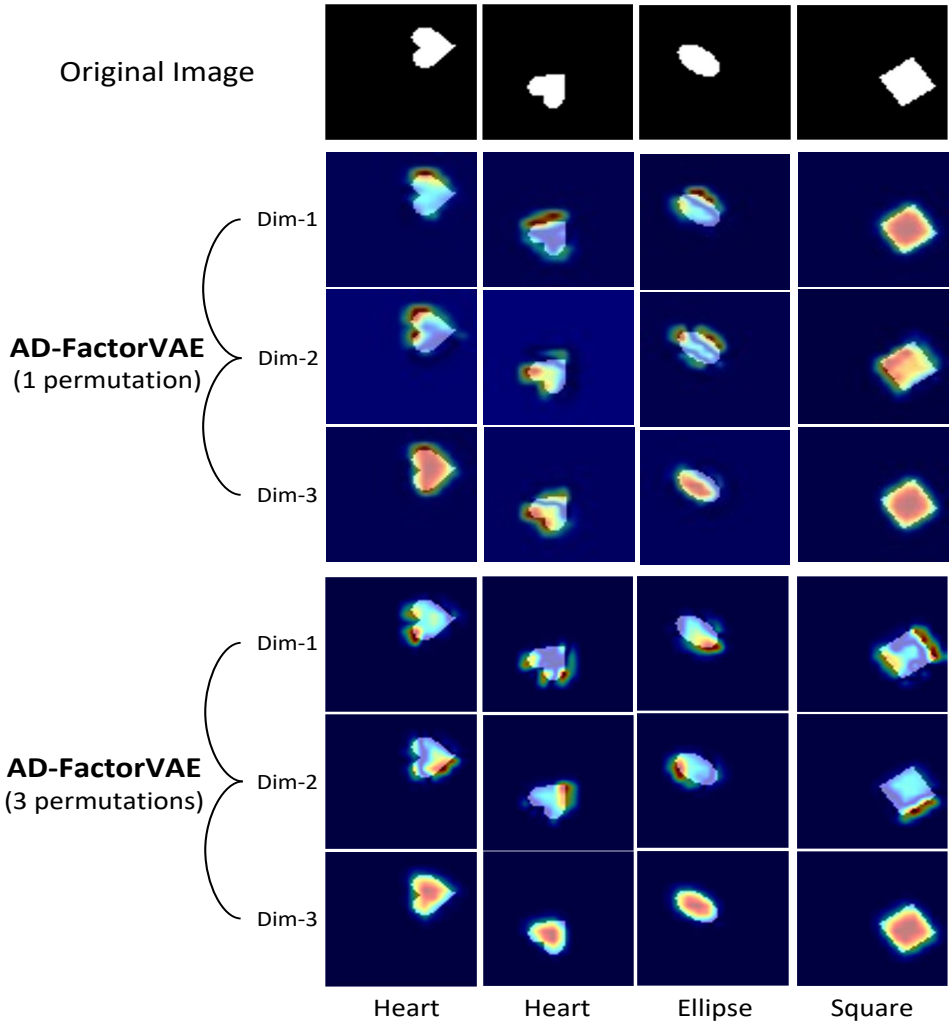


(b)

Fig. 3: **(a) dSprites Qualitative Result with 2-dim Disentanglement.** Attention separation on the Dsprites dataset. Top row: the original shape images. Middle two rows: attention maps from FactorVAE. Bottom two rows: attention maps from AD-FactorVAE. **(b) dSprites Quantitative Result.** Reconstruction error plotted against disentanglement metric [26]. The numbers at each point show β and γ values. We want a low reconstruction error and a high disentanglement metric. AD-FactorVAE (SA): our proposed method of disentanglement learning with *sample attention*. AD-FactorVAE (DA): our proposed method of disentanglement learning with the *distribution attention*

where we observe that, high-response regions in different areas (by backpropogating different dimensions) of latent vectors in attention maps in row 4 and 5 from the proposed method AD-FactorVAE, whereas attentions are overlapped in row 2 and 3 where attention maps are generated from the baseline FactorVAE.

We further perform extensive attention disentanglement experiment on Dsprites dataset by considering more dimensions in implementing learning objectives with L_{AD}. We train the network using our proposed method-attention disentanglement, rather than using a pair of attentions corresponding to two dimensions for training loss, we take additional 2 pairs of attentions corresponding 3 dimension in total for computing the learning objectives during the training. We use the same experimental settings



	γ=10	γ=40
AD-FactorVAE (1 pair of attentions)	0.90	0.91
AD-FactorVAE (3 pairs of attentions)	**0.93**	**0.93**

Fig. 4: **Top Figure: dSprites Qualitative Result with 3-dim Disentanglement** Attention separation on the Dsprites dataset. First row: the original shape images. Middle three rows: attention maps of three dimensions from AD-FactorVAE trained with 1 pair of attention map for attention disentanglement loss. Last three rows: attention maps generated from three different dimensions of latent vector from AD-FactorVAE trained with 3 pair of attention maps for attention disentanglement loss. **Bottom Table: dSprites Quantitative Result with 2-dim or 3-dim Disentanglement** Average disentanglement score on Dsprites Dataset. The higher disentanglement score means the better disentanglement performance.

described above, with different γ values as 10, 40. Both qualitative and quantitative results are presented and discussed below.

The top figure of Fig. 4 shows some attention maps generated using the model trained by our proposed AD-FactorVAE. The first row shows input images in different shapes, and the next 3 rows show attention maps generated with our proposed method by considering 1 pair of attention maps in computing attention disentanglement loss. The last 3 rows show attention maps generated by backpropogating 3 dimensions of latent vector from the model trained with our proposed method by considering 3 pairs of attention maps in calculating attention disentanglement loss. From the top figure of Fig. 4, better separated attention maps are observed in the last three rows, which meet our expectation that different spatial regions would be highlighted in attention maps generated from different dimensions of latent vector.

In the bottom table of Fig. 4, we present the best disentanglement performance of our proposed AD-FactorVAE by considering additional pairs of attention maps generated from different dimensions of latent vector when computing our attention disentanglement loss in training objectives. From the bottom table of Fig. 4, it is observed that training our AD-FactorVAE with additional dimensions in objectives can consistently improve disentanglement scores.

3 Discussion

CNN Visual Explanations. The work of Zeiler and Fergus [7] and Mahendran and Vedaldi [44] have been widely adopted to visualize intermediate CNN feature layers for understanding the activity within the layers of convolutional nets. Some more recent visual-attention-based approaches along this line of work can be categorized into either gradient-based methods or response-based methods, including [8, 9, 45–48]. GradCAM [9] is a widely used gradient-based method which computes and visualizes gradients backpropagated from the decision output to a convolutional feature layer. On the other hand, response-based approaches and its applications [8, 45, 49] typically compute the attention maps by means of extending the original CNN architecture with additional trainable units. The goal of both lines of work is to localize attentive and informative image regions that contribute the most to the model prediction. However, these methods and their extensions [10, 11, 45], while being able to explain classification/categorization models well, cannot be trivially extended to explain deep generative models such as VAEs. In this work, we propose techniques to compute and visualize attention maps directly from the learned latent embedding of the VAE. Furthermore, we make the visual attention maps end-to-end trainable and show how such a change can result in improved latent space disentanglement.

Anomaly Detection. Unsupervised anomaly detection [50] is still a challenging computer vision task. Classification-based [51, 52] or reconstruction-based approaches are two main streams used for anomaly detection in much recent work. Classification-based approaches aim to progressively learn representative one-class decision boundaries like hyperplanes [52] or hyperspheres [51] around the normal-class input distribution to tell outliers/anomalies apart. However, these methods have difficulty dealing with high-dimensional data [53]. Reconstruction-based models, on the

other hand, are proposed based on the hypothesis that input data that are anomalous cannot be reconstructed well by a model that is trained only with normal input data. This principle has been used by several methods based on the traditional PCA [54], sparse representation [55], and more recently deep autoencoders [56, 57] and generative models [58, 59]. In this work, we take a different approach to tackle this problem. We use the attention maps generated by our proposed VAE attention generation method as cues to localize anomalies. Our intuition is that representations of anomalous data should be reflected in latent embedding as being anomalous, and that generating input visual interpretations from such an embedding will give us the information we need to localize the particular anomaly.

Dermoscopy Image Analysis Deep learning models have also been shown to outperform dermatologists in studies on skin cancer detection [60–66]. One of the major challenges is the collection of labeled data, this motivated the creation of the International Skin Imaging Collaboration (ISIC) datasets [40–42, 67, 68] for skin lesions. Class imbalance is another major issue. It has been shown[61, 69] that deep neural networks are prone to high confidence predictions on out-of-distribution samples. Hendrycks and Gimpel [69] used the softmax confidence on the posterior as a baseline to identify out-of-distribution samples. While Lian et al.[67] proposed temperature-scaled softmax for confidence values, and observed that adding small noise to input enhances confidence values. Lee et al.[70] used layer-wise class-conditional Mahalanobis distances to measure how far a given example is from the training distribution. Torop et al.[71] proposed SimCLR-LOF which learns semantically meaningful features and shown competitive results on ISIC 2019 dataset against recent supervised approaches applied on the same data.

VAE Disentanglement Learning. Kingma *et al.* [13] proposed a VAE with a stochastic variational inference and learning algorithm in generative modelings for large dataset. Recently, a number of following work has been expended in understanding latent space disentanglement for generative models. Schmidhuber *et al.* [72] proposed a principal algorithm to disentangle latent representations by minimizing the predictability of one latent dimension given other dimensions. Desjardins *et al.* [73] generalized an approach to factor the latent variables. Chen *et al.* [74] designed the InfoGAN to maximise the mutual information between a subset of latent variables and the observation. Esmaeili *et al.* [75] introduced hierarchically factorized VAEs. β-VAE [25] attempted to explore independent latent factors of variation in observed data. While still a popular unsupervised framework, β-VAE sacrificed reconstruction quality for obtaining better disentanglement. Chen *et al.* [76] extended β-VAE to β-TCVAE by introducing a total correlation-based objective, whereas Mathieu *et al.* [77] explored decomposition of the latent representation into two factors for disentanglement, and Kim *et al.* [26] proposed FactorVAE that encouraged the distribution of representations to be factorial and independent across the dimensions. Xiang *et al.* [78] proposed a multi-factor disentanglement learning framework for multi-conditional generation in a weakly-supervised manner. Yang *et al.* [79] proposed CausalVAE to transform independent exogenous factors for disentanglement learning. While these methods concentrate on factorizing the latent representations provided by each individual latent

neuron, we take a different approach. We enforce learning a disentangled space by formulating disentanglement constraints based on our proposed visual interpretations, *i.e.*, visual attention maps. To this end, we propose a new attention disentanglement learning objective that we quantitatively show, provides superior performance when compared to existing work.

VAE Attention: Insights and Discussion The principle of the VAEs is to learn a latent space for representing the input data. This latent space is defined by Gaussian probability distributions that map input data to latent vectors. The encoded data can be decoded to reconstruct the original input data. Unlike traditional autoencoders (AEs), VAEs allow for the sampling of latent vectors from the latent space. These samples can be used to generate new data points, interpolate between existing ones, or explore the characteristics of the learned representation. Thus, the real power of VAEs lies in the capability of the learned latent space to represent input data with high fidelity. The architecture of the encoder and decoder networks, while important, is secondary to the importance of the learned latent space. In summary, the principle of VAEs is to learn a rich and expressive latent space that enables effective representation and manipulation of input data.

Given that a trained VAE can represent the distribution of the input data as a latent space, we propose two novel ways to calculate and generate VAE attention: *sample attention* and *distribution attention*. Given an input image, our methods first infer a latent space distribution from the encoder of the trained VAE. Given this distribution, *sample attention* samples a latent code and computes derivatives (with respect to this code) of the encoder's convolutional feature maps at the layer immediately preceding fully connected units, which are then aggregated into the visual attention map for this input image. The attention map interprets the distribution of latent space learned on VAEs. For instance, attention maps generated from the normal difference distribution (see Equation (5) in the revised paper) visually highlight the different regions between the abnormal and normal images. While this method can help interpret the VAE from the perspective of a single sample of the latent space, there is typically much more information in the entire distribution than what a single sample can capture. To help take a step towards interpreting this distribution, our proposed *distribution attention* collects a number of latent codes (by repeated sampling) from the distribution, generates per-code sample attention maps, and fits a pixel-wise Gaussian distribution. The resulting "mean" and "standard deviation" attention maps help visually interpret not only the prediction confidence of the VAE but also the associated uncertainty. For instance, in the first four columns of the first row in Fig. 6 of the revised paper, the "mean" attention map ($\mu^{\mathbf{M}}$) represents that the `crack` pixel regions are typically highlighted to be anomalous. The "standard deviation" attention map represent that the model confidently predicts pixels inside the `crack` region to be anomalous while it is relatively less confident about pixels at the edges because it is generally challenging to tell if these boundary pixels are normal or abnormal, hence leading to higher uncertainty.

It is important to note that our approaches are based on the key latent space principles of VAEs noted above. Unlike traditional approaches that focus on the architecture of the encoder or decoder networks, our approaches emphasize the importance

of the latent space in generating attention. This space serves as a bridge between the encoder and decoder networks during both training and testing. Consequently, we contend our approach is indeed VAE attention because both encoders and decoders are needed to learn a meaningful latent space using which we compute attention maps.

A VAE model is trained with both reconstruction loss and K-L divergence loss, which is significantly different from other models, such as autoencoders. A better VAE model implies better capabilities to map the input data into a learned latent space which is a bridge between the encoder for mapping input data and the decoder for reconstructions.

In this paper, we propose to visually interpret the learned latent space of VAEs by generating attention maps to visually interpret the learned latent space of VAEs. Our insight is that a meaningful latent space learned from the input data can be used to generate meaningful attention maps, which will be extensively discussed in the following section. The key factor to obtain meaningful attention maps is the quality of the latent space and for this, we need to optimize both the K-L divergence loss and the reconstruction loss. We leverage this learned latent space to generate anomaly attention maps for anomaly localization and to derive a new training objective for disentanglement representations learning. These results are are dependent on a meaningful latent space which, as noted above, needs both reconstruction and K-L divergence losses. By only optimizing reconstruction loss, we end up with an autoencoder (as opposed to a VAE) which will not lead to the kind of latent space we need to derive our attention maps.

4 Methods

In this work, we present two gradient-based attention map generating approaches towards visually interpreting a variational autoencoder (VAE): the Sample Attention (SA) and the Distribution Attention (DA). As followed, we will apply our generated attention maps to two downstream tasks: Unsupervised Anomaly Localization and Disentangled Representation Learning.

Variational Autoencoder

A vanilla variational autoencoder (VAE) is typically trained with the standard reconstruction loss between the input and the decoded/reconstructed data, along with a variational objective term which attempts to learn a standard normal latent space distribution. The Kullback-Leibler distribution metric, which is typically computed between the latent space distribution and the standard Gaussian distribution, is chosen as the variational objective in VAEs. To be more specific, given the input data $\mathbf{x}$, the conditional distribution $q(\mathbf{z}|\mathbf{x})$ of the encoder, the standard Gaussian distribution $p(\mathbf{z})$, and the reconstructed data $\hat{\mathbf{x}}$, a vanilla VAE aims to optimize:

$$L = L_r(\mathbf{x}, \hat{\mathbf{x}}) + L_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}), p(\mathbf{z})) \tag{1}$$

where L_{KL} is the Kullback-Leibler divergence term and L_r is the reconstruction term, which is generally a mean-squared error between $\mathbf{x}$ and $\hat{\mathbf{x}}$.

Interpreting VAEs with Sample Attention

Our first method, called *sample attention* (SA), for interpreting a VAE is based on calculating gradients of prediction vectors with respect to convolutional feature maps. Our ideas are different from the existing work [8, 9, 12] that compute attention maps by backpropagating the score from a classification model. Instead, in our work, the attention maps generated with SA are computed directly from the learned latent space. As illustrated in Fig 5a and discussed below, we compute a score from the latent space, which is then used to calculate gradients and obtain the attention map.

Given the posterior distribution $q(\mathbf{z}|\mathbf{x})$ inferred by the trained VAE for a data sample $\mathbf{x}$, we sample a latent sample $\mathbf{z}$ from the posterior distribution during the testing time. For each element z_i in $\mathbf{z}$, we backpropagate gradients to the last chosen convolutional feature maps $\mathbf{A} \in \mathbb{R}^{n \times h \times w}$ to generate the attention map $\mathbf{M}_i$ corresponding to z_i. Specifically, $\mathbf{M}_i$ is computed as the linear combination:

$$\mathbf{M}_i = \text{ReLU}(\sum_{k=1}^{n} \alpha_k \mathbf{A}_k) \tag{2}$$

where the scalar $\alpha_k = \text{GAP}(\frac{\partial z_i}{\partial \mathbf{A}_k})$ and $\mathbf{A}_k$ is the k^{th} feature channel ($k = 1, \ldots, n$) of the feature maps $\mathbf{A}$. Note $\frac{\partial z_i}{\partial \mathbf{A}_k}$ is a matrix and so we use the global average pooling (GAP) operation to get the scalar α_k:

$$\alpha_k = \frac{1}{T} \sum_{p=1}^{h} \sum_{q=1}^{w} (\frac{\partial z_i}{\partial A_k^{pq}}) \tag{3}$$

where $T = h \times w$, and A_k^{pq} is the pixel value at location (p, q) of the $h \times w$ matrix $\mathbf{A}_k$. We now repeat this procedure for all the elements $z_1, z_2, \ldots, z_D$ in the $D-$dimensional latent vector $\mathbf{z}$ to yield $\mathbf{M}_1, \ldots, \mathbf{M}_D$ (see Fig 5a). Such a procedure gives one attention map $\mathbf{M}_i$ per latent element. Yet, one can obtain a single overall attention map by adopting any matrix aggregation scheme. For instance, we take the mean of all the element-wise attention maps so the overall attention map is $\mathbf{M} = \frac{1}{D} \sum_i^D \mathbf{M}_i$ which is referred as the VAE attention. We denote such aggregated VAE attention map $\mathbf{M}$ as a *sample attention* (SA) map.

Interpreting VAEs with Distribution Attention

As sample attentions are computed from individual latent codes, they do not help to accurately interpret the entire distribution. To address this issue, we present our second method which is to generate attention maps directly from the distribution, called *distribution attention* (DA).

Specifically, from the mean and standard deviation vectors $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ learned by the VAE , we sample N latent codes $\mathbf{z}^j, j = 1, \ldots, N$ and generate the corresponding SAs $\mathbf{M}^j, j = 1, \ldots, N$. Next, for each pixel location (u, v) in the SA maps, we fit and estimate a Gaussian distribution with these N SA maps values $\mathbf{M}^j_{(u,v)}, j = 1, \ldots, N$:

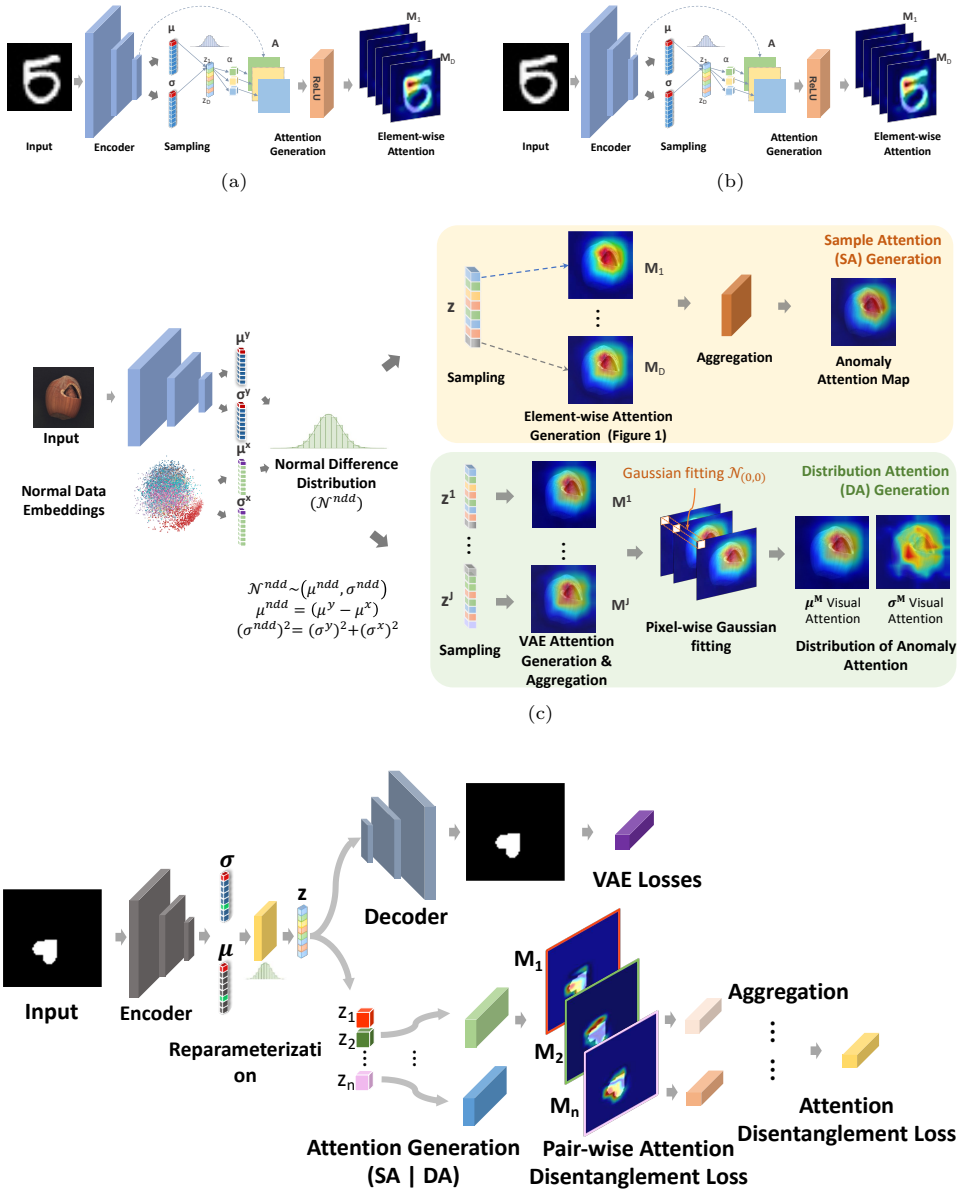


Fig. 5: **(a)** Element-wise attention generation and aggregation with a VAE. **(b)** Generating the distribution of visual attentions with a VAE. **(c)** Attention generation with a one-class VAE. Top branch: generating anomaly attention as the Sample Attention (see Section 4). Bottom branch: visualizing the distribution of anomaly attention as the Distribution Attention (see Section 4) Anomaly Localization Architecture. **(d)** Training a variational autoencoder with the proposed attention disentanglement loss.

$$\mathcal{N}^{\mathbf{M}}_{(u,v)} \sim (\mu^{\mathbf{M}}_{(u,v)}, \sigma^{\mathbf{M}}_{(u,v)}) \tag{4}$$

where $0 \leq u < H$ and $0 \leq v < W$ and $\mathcal{N}^{\mathbf{M}}_{(u,v)}$ is the normal distribution for the pixel location (u, v). By estimating $\mathcal{N}^{\mathbf{M}}_{(u,v)}$ for all the pixel locations, we can obtain a $\mu^{\mathbf{M}}$ map and a $\sigma^{\mathbf{M}}$ map, as *distribution attention* maps. The $\mu^{\mathbf{M}}$ map represents the visualization of the "mean" VAE attention generated from latent vectors in the latent space. The $\sigma^{\mathbf{M}}$ map represents the visualization of the "deviation/variance" of VAE attention map from the "mean" attention map ($\mu^{\mathbf{M}}$) above. For each pixel value ($\sigma^{\mathbf{M}}_{(u,v)}$) in $\sigma^{\mathbf{M}}$, a lower value means a smaller variance and a higher value means a larger variance, with respect to "mean" VAE attention $\mu^{\mathbf{M}}_{(u,v)}$, at the the pixel location (u, v). Both $\mu^{\mathbf{M}}$ and $\sigma^{\mathbf{M}}$ taken together visualize the pixel-wise distributions of attention maps generated from the latent space and *distribution attention* interprets the entire distribution of the latent space. This process is visually summarized in Fig 5b.

The proposed DA maps help provide both a statistical visualization of a VAE's learned latent space while also capturing a more comprehensive and diverse interpretation of the latent distribution when compared to SA maps.

Unsupervised Anomaly Localization

Given a vanilla one-class VAE trained with "normal" data, our hypothesis is that it will infer a distribution that "deviates" from the "normal" distribution when presented with "abnormal" data at test time. Let's take the 10-class digit image dataset MNIST[30] as an example. If we input a test sample from the "normal" class (e.g., digit "1" if the VAE is trained with samples from this class), the inferred latent space will ideally conform to the standard normal distribution. Consequently, given a testing sample from a different class (digit "7" as an example), the latent space inferred by the VAE should deviate considerably from the distribution learned with data from the normal class. This intuition can be formalized mathematically with Gaussian functions by means of both the Sample Attention and Distribution Attention.

Anomaly Localization with Sample Attention Under the aforementioned train/test setting for the anomaly localization task using an one-class VAE, a VAE is trained to approximately model the probability density of the distribution of the training set. During inference time, a latent code that deviates far away from the learned normal data distribution is hence seen as an anomaly. To better represent and study such deviation, we define a normal difference distribution (NDD) function described in the following.

Given the normal data $\mathbf{x} \in \mathbf{X}$ used to train the VAE, one can infer the overall $\mu^{\mathbf{x}}$ and $\sigma^{\mathbf{x}}$ of the normal data distribution. When an abnormal sample $\mathbf{y} \in \mathbf{Y}$ is fed to the trained VAE for testing, a latent code $\mathbf{z_i}$ is inferred from the $\mu^{\mathbf{y}}_{\mathbf{i}}$ and $\sigma^{\mathbf{y}}_{\mathbf{i}}$, where i is the number of samples. Thus, we define the NDD of each latent code $\mathbf{z_i}$ in the latent space as:

$$P_{q(\mathbf{z_i}|\mathbf{x})-q(\mathbf{z_i}|\mathbf{y})}(u) = \frac{e^{-[u-(\mu^{\mathbf{x_i}}-\mu^{\mathbf{y_i}})]^2/[2((\sigma^{\mathbf{x_i}})^2+(\sigma^{\mathbf{y_i}})^2)]}}{\sqrt{2\pi((\sigma^{\mathbf{x_i}})^2+(\sigma^{\mathbf{y_i}})^2)}} \tag{5}$$

For clarity, in the following we will denote the normal difference distribution as $\mathcal{N}^{ndd} \sim (\boldsymbol{\mu}^{ndd}, \boldsymbol{\sigma}^{ndd})$. Following our SA generation procedure introduced in Section 4, we can now compute an anomaly attention map $\mathbf{M}$ corresponding to a random sampled latent vector $\mathbf{z}$ from the normal difference distribution $P_{q(\mathbf{z}|X)-q(\mathbf{z}|Y)}$. This process is visually summarized in the top branch of Fig. 5c.

In Fig. 6a, we show two sets of example results for anomaly SA maps. The first set includes results from the object class *hazelnut* where we train a VAE using all the "normal" (non-defect) *hazelnut* images. An example "normal" image is shown on the top row of the figure for clarity. During testing, inputs include "abnormal" *hazelnut* images with different types of defects, such as holes, cracks, or prints (sample images of different defects are shown from left to right in the middle row of Fig. 6a). We present our anomaly SA maps on the bottom row, where high-response regions are in red and correspond to anomalies in the given images. One can note these attention maps can localize defects accurately. Similar observations can be made in the second set of results for the *leather* object class as well.

To further illustrate how the sampled latent code corresponds to and characterizes the computed SA maps, we generate multiple samples of latent vectors from $\mathcal{N}^{ndd}$ by *traversing* the distribution from from very close to the mean ($\boldsymbol{\mu}^{ndd}$) to very far from it. We then compute the SA map corresponding to each sampled latent code and visualize it. Concretely, we expect that a latent vector $\mathbf{z}$ sampled close to $\boldsymbol{\mu}^{ndd}$ will yield an SA map with higher density of high-response regions (darker red) over anomalous pixels, as opposed to cases when we sample a $\mathbf{z}$ far away from the $\boldsymbol{\mu}^{ndd}$. We show results in Fig. 6c, where one can note that we traverse the distribution by sampling latent vectors from far to close to the mean (from left to right as shown in the figure), the highlighted areas in the SA maps also gradually focus on the appropriate anomalous pixel regions in the image.

Anomaly Localization with Distribution Attention We next discuss how one can use our proposed DA maps for localizing anomalies. As discussed in Section 4, we start by computing a large number of J anomaly SA maps $\mathbf{M}^j, j \in J$ from latent vectors $\mathbf{z}^j, j \in J$ randomly sampled from $\mathcal{N}^{ndd}$. Then at each pixel location (u, v) of these J SAs, we fit and estimate a pixel-wise Gaussian distribution with a mean of $\mu^{\mathbf{M}}_{(u,v)}$ and a standard deviation of $\sigma^{\mathbf{M}}_{(u,v)}$. Finally, we represent these pixel-wise Gaussian representations as two separate attention maps for visualization: one contains only values of $\mu^{\mathbf{M}}_{(u,v)}$ and the other those of $\sigma^{\mathbf{M}}_{(u,v)}$. We refer to them as mean anomaly DA map $\mu^{\mathbf{M}}$ and deviation anomaly DA map $\sigma^{\mathbf{M}}$ respectively. This process is visually summarized in the bottom branch of Fig. 5c.

As a result, each pixel in $\mu^{\mathbf{M}}$ is a statistical summary of all the anomaly SA map responses at the same pixel location. Thus, if high responses at (u, v) are present across SA maps, then $\mu^{\mathbf{M}}_{(u,v)}$ will also have a high response, suggesting the pixel (u, v) is anomalous. On the other hand, each pixel in $\sigma^{\mathbf{M}}$ represents the statistical uncertainty of the resultant SA maps' responses at this location. In other words, if uncertainty levels are high at (u, v) across SA maps, then $\sigma^{\mathbf{M}}_{(u,v)}$ will have a high response as well. This means the uncertainty of the accuracy of the $\mu^{\mathbf{M}}_{(u,v)}$ value is high. Consequently,

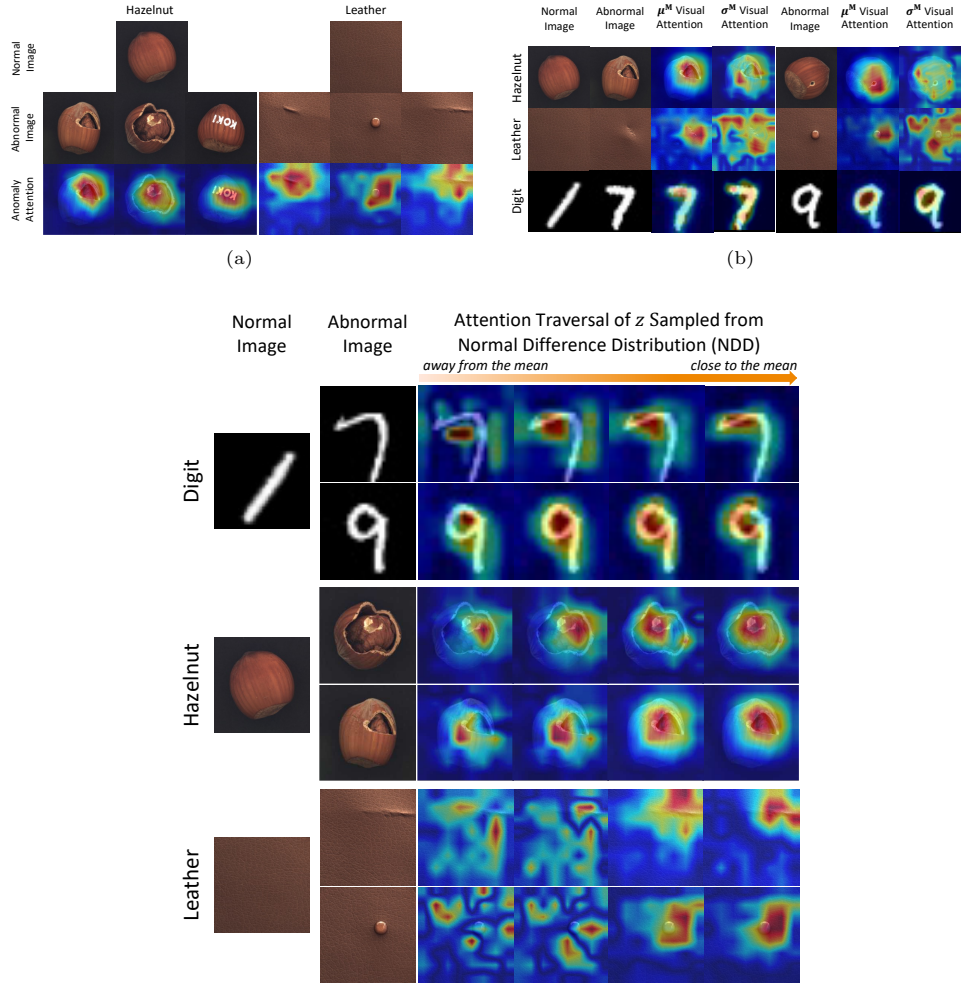


Fig. 6: **(a) SA Maps.** Anomaly sample attention maps generated for two different classes of objects: Hazelnut and Leather. On the top row, we show one sample image of the "normal" data each per class. In the middle row, for each class we show three different types of "abnormal" sample images (images with different types of detects on the objects). While on the bottom row, the corresponding generated SA maps are shown. In these examples, our attention maps correctly localize the anomalous regions (defects) within each testing images. **(b) DA Maps.** $\mu^{\mathbf{M}}$ and $\sigma^{\mathbf{M}}$ visual attention maps of the distribution of anomaly attention over the abnormal images. The $\mu^{\mathbf{M}}$ visual attention maps present the average likelihood of each pixel being detected as anomaly and $\sigma^{\mathbf{M}}$ visual attention maps present the uncertainty of each pixel being detected as anomaly. **(c) Attention Traversal of DA Maps.** Attention traversal by using different $\mathbf{z}$ sampled from $\boldsymbol{\mu}^{ndd}$ given the abnormal data. From the left to the right we show attention maps of different latent codes $\mathbf{z}$ sampled from far to close to the $\boldsymbol{\mu}^{ndd}$. We can see that the abnormal regions are gradually highlighted precisely.

the DA maps $\mu^{\mathbf{M}}$ and $\sigma^{\mathbf{M}}$ taken together provide a more comprehensive interpretation of the anomaly localization predictions of our framework. After obtaining the $\mu^{\mathbf{M}}$ and $\sigma^{\mathbf{M}}$ given the input abnormal images, we use the $\sigma^{\mathbf{M}}$ map to weight the $\mu^{\mathbf{M}}$ map and generate the final anomaly detection result, *i.e.*, $\widetilde{\mu}^{\mathbf{M}} = \mu^{\mathbf{M}} * \frac{1}{1+\exp(-\alpha*(\sigma^{\mathbf{M}}-\beta))}$, α, β are predefined hyperparameters set by cross-validation in all our experiments. Our expectation is that the high response regions (which also indicates high anomaly uncertainties) in $\sigma^{\mathbf{M}}$ visual attention can help remove uncertainties in the $\mu^{\mathbf{M}}$ visual attention so that refined $\widetilde{\mu}^{\mathbf{M}}$ can detect anomalies with better accuracy and less uncertainty as well.

In Fig. 6b, we show some examples of DA maps. For instance, in the first row, we present two examples from *hazlenut* object class. The second through fourth columns show the input (anomalous) *hazelnut* image, the $\mu^{\mathbf{M}}$ map, and the $\sigma^{\mathbf{M}}$ map, respectively. Fifth through seventh columns show another example and its corresponding DA maps. In the first example on the first row, one can note high responses in $\mu^{\mathbf{M}}$ are at the `crack` pixel regions in the image, whereas high responses in $\sigma^{\mathbf{M}}$ are present at the edges of the `crack` pixels. The interpretation from these results is that while the model confidently predicts pixels inside the `crack` region to be anomalous, it is relatively less confident about pixels at the edges. This result is aligned with our intuition that it is generally challenging to tell if these boundary pixels are normal or abnormal, hence leading to higher uncertainty. Similar observations can also be made from other examples as well (including from digits figures). This discussion shows how our proposed $\mu^{\mathbf{M}}$ and $\sigma^{\mathbf{M}}$ pair of DA maps together help both visually localize anomalies as well as interpret the results.

Disentangled Representation Learning

Learning disentangled representations with VAEs has been another intriguing area that has received much progress, as they can increase interpretability and generalization of the models and learn faster on downstream tasks[26, 80–84]. A disentangled representation vector encodes isolated factors of variation in the input data into only a few independent dimensions[85]. We intuitively believe our proposed gradient-based VAE attention generated directly from the latent representation, can help visually explain latent space disentanglement. Furthermore, by leveraging these attention maps as part of trainable constraints, we are able to boost disentangled representation learning.

So far, we have discussed our Sample Attention and Distribution Attention methodologies. We proposed to generate VAE attention maps directly from each dimensions of a given latent vector. We have also shown that the resultant attention maps heat up certain regions (pixels) in the input image corresponding to the latent elements. Hence, we exploited these attention maps to provide visual clues to explain the VAE's latent representations. But here, we want to consider an alternative application to utilize our attention maps in the opposite direction. We seek to design a trainable constraints that attempts to separate the high-response regions in the attention maps generated from different latent vector's dimensions to be as much as possible. And by backpropagating from this attention-separation-driven constraints, the network explicitly forces the latent representation to be disentangled.

To follow this goal, we propose a new training loss called the *attention disentanglement* loss (L_{AD}). As mentioned, L_{AD} is designed to boost disentanglement learning by forcing attention separation. In other words, this loss penalizes on large number of overlapping high-response pixels within different attention maps. Given a pair of attention maps, $\mathbf{M}^p$ and $\mathbf{M}^q$, generated from different latent elements, we mathematically formulate L_{AD} as :

$$L_{AD} = 2 \cdot \sum_{p}^{D} \sum_{q,p \neq q}^{D} \frac{\sum_{ij} min(M_{ij}^p, M_{ij}^q)}{\sum_{ij} M_{ij}^p + M_{ij}^q} \tag{6}$$

D is the dimension of the latent vector, M_{ij}^p and M_{ij}^q are the $(i,j)^{th}$ pixel in the attention maps $\mathbf{M}^p$ and $\mathbf{M}^q$ respectively. One empirical implementation is taken as an example by considering two dimensions from the D where $p = 1$ and $q = 2$. Then two attention maps $\mathbf{M}^1$ and $\mathbf{M}^2$ can be computed from these two dimensions of latent vector, giving *attention disentanglement* loss (L_{AD}) of Equation 6 as:

$$L_{AD} = 2 \cdot \frac{\sum_{ij} min(M_{ij}^1, M_{ij}^2)}{\sum_{ij} M_{ij}^1 + M_{ij}^2} \tag{7}$$

Further, when considering three dimensions from the D of the latent vector, the *attention disentanglement* loss (L_{AD}) is calculated by iterating $p = 1, q = 2$, $p = 2, q = 3$ and $p = 1, q = 3$ so the Equation 6 is implemented as:

$$\begin{aligned} L_{AD} = 2 \cdot (& \frac{\sum_{ij} min(M_{ij}^1, M_{ij}^2)}{\sum_{ij} M_{ij}^1 + M_{ij}^2} + \frac{\sum_{ij} min(M_{ij}^2, M_{ij}^3)}{\sum_{ij} M_{ij}^2 + M_{ij}^3} + \\ & \frac{\sum_{ij} min(M_{ij}^1, M_{ij}^3)}{\sum_{ij} M_{ij}^1 + M_{ij}^3}) \end{aligned} \tag{8}$$

and L_{AD} achieves minimum value when not a single pair of attention maps have overlapping high-response pixels. Our proposed L_{AD} can be easily integrated with other learning objectives on existing VAE-type models. A better illustration of the training framework with our L_{AD} can be viewed in Fig. 5d.

We then generate *sample attention* (SA) or *distribution attention* (DA) as the input attention maps for L_{AD}. In the case of SA, the procedure is straightforward: we choose a pair of attention maps from $\mathbf{M}_1, \ldots, \mathbf{M}_D$ (see Fig. 5a) to compute the loss. While in the case of DA, we first sample N latent codes $\mathbf{z^k}, k = 1, \ldots, N$, each with D dimension $\mathbf{z^k} \in \mathbf{R}^D$. Then, for a chosen pair of dimensions *i.e.*, (p,q), we compute two sets of N element-wise attention maps $\mathbf{M_k}^p, k = 1, \ldots, N$ and $\mathbf{M_k}^q, k = 1, \ldots, N$ corresponding to p and q, respectively. Next, we fit two Gaussian distributions for these two sets and obtain two pairs of element-wise DA maps ($\boldsymbol{\mu}^{\mathbf{M}^p}$, $\boldsymbol{\sigma}^{\mathbf{M}^p}$) and ($\boldsymbol{\mu}^{\mathbf{M}^q}$, $\boldsymbol{\sigma}^{\mathbf{M}^q}$). When computing L_{AD}, we use $\boldsymbol{\sigma}^{\mathbf{M}^p}$ and $\boldsymbol{\sigma}^{\mathbf{M}^q}$ to weight the overlapped high-responsive regions following the weighting function $\widetilde{\mu}^{\mathbf{M}} = \mu^{\mathbf{M}} * \frac{1}{1+\exp(-\alpha*(\sigma^{\mathbf{M}}-\beta))}$ as previously introduced. Our expectations is that by using the *distribution attention*,

attention maps generated from different dimensions should be separated as well as with low uncertainty.

Our proposed L_{AD} can be directly integrated with the existing network FactorVAE, and its standard training objective function L_{FV}. Our final overall learning objective is presented as follows:

$$L = L_{\mathrm{FV}} + \lambda L_{\mathrm{AD}} \tag{9}$$

where λ is the hyperparameter to weight standard FactorVAE training loss and our attention disentanglement loss. Then we train FactorVAE with the overall learning objective Equation (9). With such explicit attention-separation constraints, we expect better disentanglement learning performance than the original FactorVAE pipeline. Note that while we use the FactorVAE [26] for demonstration in this work, the proposed attention disentanglement loss is in no way limited to this model and can be used in conjunction with other models as well (e.g., β-VAE [25]).

5 Conclusion

This work proposed new techniques to visually interpret deep generative models, in particular, variational autoencoders, by means of gradient-based network attention. We presented two ways to calculate visual attention maps. Firstly, unlike most the existing work which perform gradients backpropagation from the classification output, the proposed method computed gradients directly with a latent vector sampled from the latent space on VAEs and generated *sample attention* maps, which bridged the gap between generative models and gradient-base attention visualization mechanism. Secondly, further steps are taken to visually interpret the learned entire distribution, this work proposes to fit and estimate a pixel-wise Gaussian distribution of a collection of *sample attention* maps generated by sampling a group of latent codes and the resulting "mean" and "standard deviation" attention maps are presented as the *distribution attention* maps. Both two types of visual attention maps provided a explainable way to interpret generative models, they can also be used for down-stream computer vision tasks. In this work, two classical VAE applications are taken for experiments to demonstrate the applicability of the resulting VAE attention: anomaly localization and latent space disentanglement. In anomaly localization, the fact that is used is an abnormal input will result in latent variables that do not conform to the standard Gaussian in gradient backpropagation and visual attention generation. Given an abnormal input to a VAE trained on normal data, attention maps generated using the proposed two methods are presented and how these anomaly attention maps are then used as cues to generate pixel-level binary anomaly masks, helping anomalous images classification and anomaly segmentation.

In latent space disentanglement, a new technique is presented which utilizes VAE attention from each latent dimension to enforce new attention disentanglement learning constraints that results in improved attention separability as well as disentanglement performance in training VAE. Experiments are conducted on a wide range of datasets and experimental results validated the effectiveness of the proposed method in visually interpreting variational autoencoders and improving performance of VAEs applications.

This paper has taken the first step to visually interpret VAEs directly from the learned latent space, which freed the gradients-based attention generation from requiring classification/categorization streams in the models. The future work includes studying visual interpretations on a wider range of generative models, for instance, generative adversarial networks (GANs), and these interpretations can potentially boost a variety of down-streaming applications of GANs, e.g., image synthesis, video predictions, etc. In addition, visual interpretations of VAEs can potentially be utilized to design new learning constraints for hybrid deep models which consists of generative modules and discriminative in an end-to-end training for improved performance. Last but not the least, visual attention maps of VAEs bridged the gap of unsupervised learning models' interpretations, and this leveraged broad applications of unsupervised learning with generative models, in particular, in medical procedures, industrial process with limited and scare data.

References

[1] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)

[2] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[3] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)

[4] Jin, D., Guo, D., Ho, T.-Y., Harrison, A.P., Xiao, J., Tseng, C.-K., Lu, L.: Accurate esophageal gross tumor volume segmentation in pet/ct using two-stream chained 3d deep network fusion. In: MICCAI (2019)

[5] Zuo, Y., Qiu, W., Xie, L., Zhong, F., Wang, Y., Yuille, A.L.: Craves: Controlling robotic arm with a vision-based economic system. In: CVPR (2019)

[6] Li, B., Ouyang, W., Sheng, L., Zeng, X., Wang, X.: Gs3d: An efficient 3d object detection framework for autonomous driving. In: CVPR (2019)

[7] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)

[8] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)

[9] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)

[10] Li, K., Wu, Z., Peng, K.-C., Ernst, J., Fu, Y.: Guided attention inference network. T-PAMI **42**(12), 2996–3010 (2019)

[11] Wang, L., Wu, Z., Karanam, S., Peng, K.-C., Singh, R.V., Liu, B., Metaxas, D.N.: Sharpen focus: Learning with attention separability and consistency. In: ICCV (2019)

[12] Zheng, M., Karanam, S., Wu, Z., Radke, R.J.: Re-identification with consistent attentive siamese networks. In: CVPR (2019)

[13] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR **abs/1312.6114** (2013)

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)

[15] Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D.P., Gool, L.V.: Sliced wasserstein generative models. In: CVPR (2019)

[16] Kaneko, T., Ushiku, Y., Harada, T.: Label-noise robust generative adversarial networks. In: CVPR (2019)

[17] Mehrasa, N., Jyothi, A.A., Durand, T., He, J., Sigal, L., Mori, G.: A variational auto-encoder model for stochastic point processes. In: CVPR (2019)

[18] Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)

[19] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: NIPS, (2017)

[20] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: NeurIPS (2018)

[21] Xian, Y., Sharma, S., Schiele, B., Akata, Z.: F-vaegan-d2: A feature generating framework for any-shot learning. In: CVPR (2019)

[22] Tang, Y., Srivastava, N., Salakhutdinov, R.: Learning generative models with visual attention. In: NIPS (2013)

[23] Alami Mejjati, Y., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I.: Unsupervised attention-guided image-to-image translation. In: NeurIPS, (2018)

[24] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: ICML (2019)

[25] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M.M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a

constrained variational framework. In: ICLR (2017)

[26] Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)

[27] Zheng, Z., Sun, L.: Disentangling latent space for vae by label relevant/irrelevant dimensions. In: CVPR (2019)

[28] Matthey, L., Higgins, I., Hassabis, D., Lerchner, A.: dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/ (2017)

[29] Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. https://github.com/zalandoresearch/fashion-mnist

[30] Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)

[31] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: CVPR (2019)

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

[33] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)

[34] Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: IPMI (2017)

[35] Napoletano, P., Piccoli, F., Schettini, R.: Anomaly detection in nanofibrous materials by cnn-based self-similarity. Sensors **18**(1), 209 (2018)

[36] Li, Z., Li, N., Jiang, K., Ma, Z., Wei, X., Hong, X., Gong, Y.: Superpixel masking and inpainting for self-supervised anomaly detection. In: BMVC (2020)

[37] Dehaene, D., Frigo, O., Combrexelle, S., Eline, P.: Iterative energy-based projection on a normal data manifold for anomaly localization. arXiv preprint arXiv:2002.03734 (2020)

[38] Zhou, K., Xiao, Y., Yang, J., Cheng, J., Liu, W., Luo, W., Gu, Z., Liu, J., Gao, S.: Encoding structure-texture relation with p-net for anomaly detection in retinal images. In: ECCV (2020)

[39] Wang, S., Wu, L., Cui, L., Shen, Y.: Glancing at the patch: Anomaly localization with global and local feature comparison. In: CVPR (2021)

[40] Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 1–9 (2018)

[41] Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)

[42] Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., *et al.*: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi). In: ISBI (2018)

[43] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)

[44] Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR (2015)

[45] Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation. In: CVPR (2019)

[46] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV (2018)

[47] desai, s., Ramaswamy, H.G.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: WACV (2020)

[48] Ou, Y., Chen, Z., Wu, F.: Multimodal local-global attention network for affective video content analysis. IEEE Transactions on Circuits and Systems for Video Technology **31**(5), 1901–1914 (2021) https://doi.org/10.1109/TCSVT.2020.3014889

[49] Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. IJCV **126**, 1084–1102 (2016)

[50] Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: ACCV (2018)

[51] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: ICML (2018)

[52] Chalapathy, R., Menon, A.K., Chawla, S.: Anomaly detection using one-class neural networks. arXiv preprint arXiv:1802.06360 (2018)

[53] Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407 (2019)

[54] Kim, J., Grauman, K.: Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: CVPR (2009)

[55] Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.-S.: Spatio-temporal autoencoder for video anomaly detection. In: ACM MM (2017)

[56] Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: ICLR (2018)

[57] Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: KDD (2017)

[58] Zhou, J.T., Zhang, L., Fang, Z., Du, J., Peng, X., Xiao, Y.: Attention-driven loss for anomaly detection in video surveillance. IEEE Transactions on Circuits and Systems for Video Technology **30**(12), 4639–4647 (2020) https://doi.org/10.1109/TCSVT.2019.2962229

[59] Zhang, S., Gong, M., Xie, Y., Qin, A.K., Li, H., Gao, Y., Ong, Y.-S.: Influence-aware attention networks for anomaly detection in surveillance videos. IEEE Transactions on Circuits and Systems for Video Technology **32**(8), 5427–5437 (2022) https://doi.org/10.1109/TCSVT.2022.3148392

[60] Pacheco, A.G., Krohling, R.A.: Recent advances in deep learning applied to skin cancer detection. arXiv preprint arXiv:1912.03280 (2019)

[61] Celebi, M.E., Codella, N., Halpern, A.: Dermoscopy image analysis: overview and future directions. BHI **23**(2), 474–478 (2019)

[62] Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., *et al.*: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology **29**(8), 1836–1842 (2018)

[63] Codella, N.C., Nguyen, Q.-B., Pankanti, S., Gutman, D.A., Helba, B., Halpern, A.C., Smith, J.R.: Deep learning ensembles for melanoma recognition in dermoscopy images. IBM Journal of Research and Development **61**(4/5), 5–1 (2017)

[64] Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.-A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. T-MI **36**(4), 994–1004 (2016)

[65] Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., Lei, B., Wang, T.: Melanoma recognition in dermoscopy images via aggregated deep convolutional features. T-BE **66**(4), 1006–1016 (2018)

[66] Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. MethodsX **7**, 100864 (2020)

[67] Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)

[68] Heaton, J.: Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. Springer (Springer, 2018)

[69] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)

[70] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)

[71] Torop, M., Ghimire, S., Liu, W., Brooks, D.H., Camps, O., Rajadhyaksha, M., Dy, J., Kose, K.: Unsupervised approaches for out-of-distribution dermoscopic lesion detection. arXiv preprint arXiv:2111.04807 (2021)

[72] Schmidhuber, J.: Learning factorial codes by predictability minimization. Neural Comput. **4**(6), 863–879 (1992)

[73] Desjardins, G., Courville, A.C., Bengio, Y.: Disentangling factors of variation via generative entangling. ArXiv **abs/1210.5474** (2012)

[74] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NIPS, (2016)

[75] Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D.H., Dy, J., Meent, J.-W.: Structured disentangled representations. In: AISTATS (2019)

[76] Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: NeurIPS, (2018)

[77] Mathieu, E., Rainforth, T., Siddharth, N., Teh, Y.W.: Disentangling disentanglement in variational autoencoders. In: ICML (2019)

[78] Xiang, S., Gu, Y., Xiang, P., Chai, M., Li, H., Zhao, Y., He, M.: Disunknown: Distilling unknown factors for disentanglement learning. In: ICCV (2021)

[79] Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: Causalvae: Disentangled representation learning via neural structural causal models. In: CVPR (2021)

[80] Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. NeurIPS (2019)

[81] Goodfellow, I., Lee, H., Le, Q., Saxe, A., Ng, A.: Measuring invariances in deep networks. NeurIPS (2009)

[82] Tschannen, M., Bachem, O., Lucic, M.: Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069 (2018)

[83] Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. NeurIPS (2018)

[84] Zheng, Z., Sun, L.: Disentangling latent space for vae by label relevant/irrelevant dimensions. In: CVPR (2019)

[85] Van Steenkiste, S., Locatello, F., Schmidhuber, J., Bachem, O.: Are disentangled representations helpful for abstract visual reasoning? NeurIPS (2019)

[86] Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. T-PAMI **36**(1), 18–32 (2013)

[87] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

[88] Kwon, G., Prabhushankar, M., Temel, D., AlRegib, G.: Backpropagated gradient representations for anomaly detection. In: ECCV (2020)

[89] Clifton, L., Clifton, D.A., Watkinson, P.J., Tarassenko, L.: Identification of patient deterioration in vital-sign data using one-class support vector machines. In: FedCSIS (2011)

[90] Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C., *et al.*: Support vector method for novelty detection. In: NIPS (1999)

[91] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, ??? (2006)

[92] Chen, Z., Yeo, C.K., Lee, B.S., Lau, C.T.: Autoencoder-based network anomaly detection. In: 2018 Wireless Telecommunications Symposium (WTS), pp. 1–5 (2018). IEEE

[93] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[94] Oord, A.v.d., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., Kavukcuoglu, K.: Conditional image generation with pixelcnn decoders. arXiv

preprint arXiv:1606.05328 (2016)

[95] Abati, D., Porrello, A., Calderara, S., Cucchiara, R.: Latent space autoregression for novelty detection. In: CVPR (2019)

[96] Akçay, S., Atapour-Abarghouei, A., Breckon, T.P.: Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In: IJCNN (2019)

[97] Perera, P., Nallapati, R., Xiang, B.: Ocgan: One-class novelty detection using gans with constrained latent representations. In: CVPR (2019)

[98] Yu, Q., Kavitha, M., Kurita, T.: Extensive framework based on novel convolutional and variational autoencoder based on maximization of mutual information for anomaly detection. Neural Computing and Applications, 1–23 (2021)

[99] Combalia, M., Hueto, F., Puig, S., Malvehy, J., Vilaplana, V.: Uncertainty estimation in deep neural networks for dermoscopic image classification. In: CVPR Workshops (2020)

Appendix A Additional Results for Unsupervised Anomaly Localization

A.1 Qualitative Results on MNIST Dataset

We present anomaly localization results from the MNIST data in Fig. A1.

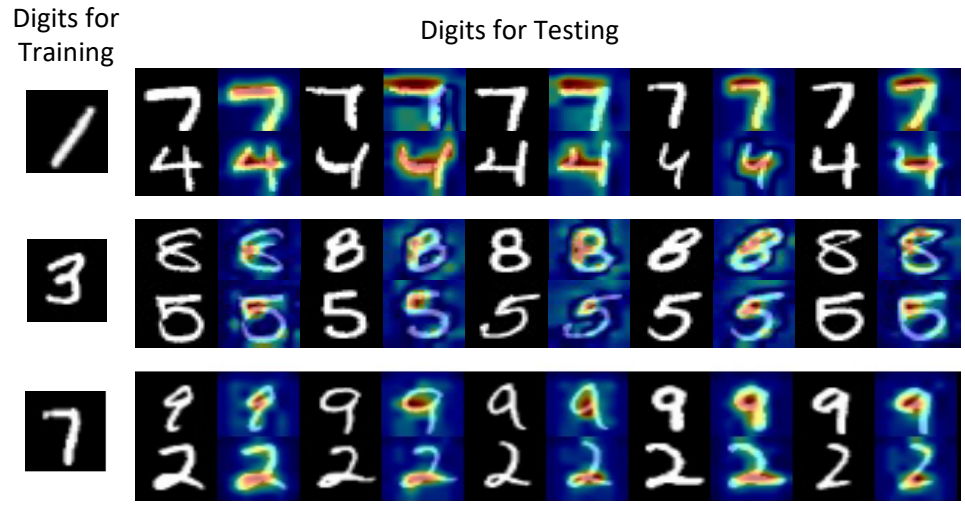


Fig. A1: Anomaly localization results from the MNIST dataset.

A.2 Additional Quantitative Results on MVTec-AD Dataset

In Table A1, we present quantitative results for anomaly localization results in the metric of the best IOU on the MVTec-AD dataset.

Fig. A2 presents qualitative results of anomaly localization with Distribution Attention maps from MVTec-AD data.

A.3 Additional Quantitative Results on USCD Ped 1 Dataset

UCSD Ped 1[86] pedestrian video dataset contains videos captured with a stationary camera to monitor a pedestrian walkway. This dataset includes 34 training sequences and 36 testing sequences, with about 5500 "normal" frames and 3400 "abnormal" frames. We resize the data to 100×100 pixels for training and testing.

We first show qualitative results in Fig A3. In each row (left to right), the first two images are the input and corresponding ground-truth anomaly masks (e.g., *bicycle*, *Car* etc.), the next two images are our results (attention map and mask), and the last two images show the difference maps *diffmap* between the input and its VAE reconstruction output. We note more precise localization of the high-response regions

Table A1: **MVTec-AD bestIOU** Quantitative results of best IOU for pixel level segmentation on 15 categories from MVTec-AD dataset.

	Category	SSIM-AE	l_2-AE	AnoGAN	CNN-FD	**Ours (SA)**
Texture	Carpet	0.69	0.38	0.34	0.2	0.1
	Grid	0.88	0.83	0.04	0.02	0.02
	Leather	0.71	0.67	0.34	0.74	0.24
	Tile	0.04	0.23	0.08	0.14	**0.23**
	Wood	0.36	0.29	0.14	0.47	0.14
Object	Bottle	0.15	0.22	0.05	0.07	**0.27**
	Cable	0.01	0.05	0.01	0.13	**0.18**
	Capsule	0.09	**0.11**	0.04	0.00	**0.11**
	Hazelnut	0.00	0.41	0.02	0.00	**0.44**
	Metal Nut	0.01	0.26	0.00	0.13	**0.49**
	Pill	0.07	0.25	0.17	0.00	0.18
	Screw	0.03	0.34	0.01	0.00	0.17
	Toothbrush	0.08	0.51	0.07	0.00	0.17
	Transistor	0.01	0.22	0.08	0.03	**0.30**
	Zipper	0.10	0.13	0.01	0.00	0.06

in our generated attention maps, and these high-response regions indeed correspond to anomalies in these images.

We also conduct experiments and ablation studies for pixel-level segmentation AUROC score against the baseline method of difference between input data and the reconstruction output. We test our proposed attention generation mechanism with varying levels of spatial resolution by backpropagating from the latent space to each of the encoder's convolutional layers: 50×50, 25×25, and 12×12. The quantitative results in Table A2 show our method gives better performance when compared to the baseline and using the proposed DA maps can consistently improve anomaly localization performance.

Table A2: Results on UCSD Ped1 using pixel-level segmentation AUROC score. We compare results obtained using our anomaly attention generated with different target network layers to reconstruction-based anomaly localization using Vanilla-VAE. SA: Sample Attention. DA: Distribution Attention.

	diffmap	Ours(SA, Conv1)	**Ours(SA, Conv2)**	Ours(SA, Conv3)	**Ours(DA)**
AUROC	0.86	0.89	**0.92**	0.91	**0.95**

Appendix B Image-Level Anomaly Detection

B.1 CIFAR-10 Dataset

Cifar-10 Dataset [87] consists around 60,000 color images in size of 32×32. Images are classified into 10 classes with 6000 samples each. The 10 classes are: Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, Truck. By default splits, 5000

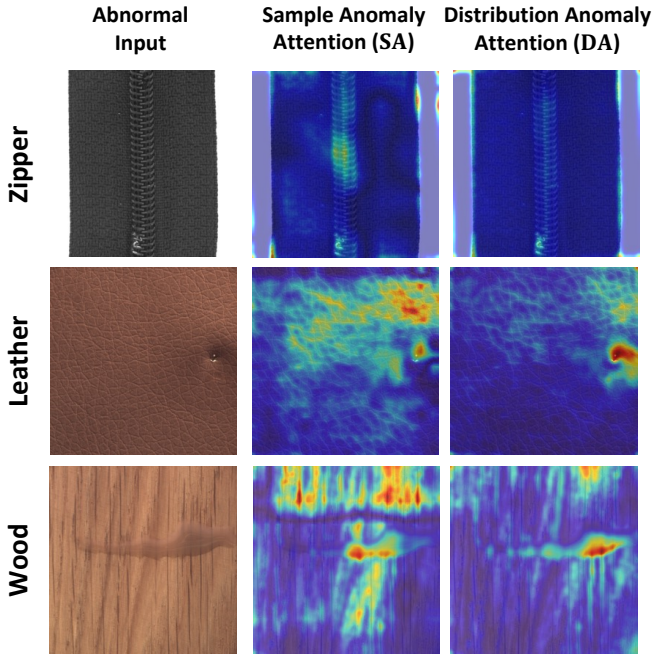


Fig. A2: **MVTec-AD DA Maps** Qualitative results of anomaly localization with DA compared to SA from MVTec-AD. As can be seen here, the DA maps can better localize anomalies than the SA maps.

images are utilized as training samples and 1000 for testing for each class. We adopt the same protocol as described in GradCon[88] to split the dataset for one-class classification task. We train on all the training samples within one class each time. During testing, we take the testing samples for the same class as in-distribution (ID) data, while other testing samples from the other 9 classes as out-of-distribution (OOD) data.

Similarly to our previous experiments, we train a vanilla VAE with a ResNet18 [2] based encoder. We train the model, without data augmentation, then collect the latent representation $\boldsymbol{\mu}$ of training and testing samples to perform anomaly detection as a two-step procedure: we first apply a Local Outlier Factor(LOF) detector onto the training sample features followed by fitting every testing sample features to predict an anomaly score. The lower the LOF score, the more anomalous the testing sample is. Afterwards, we calculate AUROC with these LOF scores.

We report our quantitative image-level anomaly detection performance in Table B3, where we compare to multiple baselines with scores adopted from Wang *et al.* [39]. From the results, we note that our approach performs better than the state-of-art methods on image-level anomaly detection on most of the classes, without using any specific dataset and task oriented design in the network. Despite the simplicity of our VAE model, it has shown competitive performance for one-class anomaly detection.

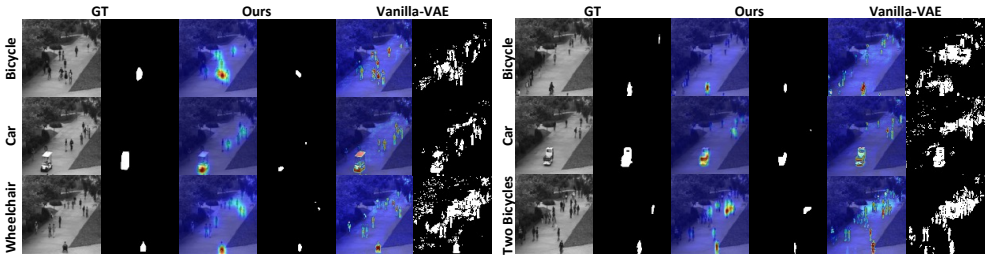


Fig. A3: **UCSD Ped 1 SA Maps** Qualitative results from UCSD Ped1 dataset. L-R: Original test image, ground-truth masks, our anomaly attention localization maps, and difference between input and the VAE's reconstruction. The anomalies in these samples are moving cars, bicycle, and wheelchair.

Table B3: AUROC scores for image-level anomaly detection performance compared to various baselines on CIFAR-10 Dataset. The final column indicates the mean AUROC across all the classes. Bold denotes the best results, and blue the second best. We adopt baseline scores from [39]. As shown in this table, given the simplicity of our vanilla VAE model, we are able to provide the best scores on half of the classes.

	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Average
GMM[89]	0.635	0.433	0.618	0.498	0.733	0.515	0.696	0.540	0.675	0.531	0.587
OC-SVM[90]	0.630	0.440	0.649	0.487	0.735	0.500	0.725	0.533	0.649	0.508	0.586
KDE[91]	0.658	0.520	0.657	0.497	0.727	0.496	0.758	0.564	0.680	0.540	0.610
l_2-AE [33]	0.411	0.478	0.616	0.562	0.728	0.513	0.688	0.497	0.487	0.378	0.536
CAE[92]	0.606	0.271	0.655	0.549	0.701	0.532	0.537	0.408	0.653	0.322	0.523
VAE[93]	0.700	0.386	0.679	0.535	0.748	0.523	0.687	0.493	0.696	0.386	0.583
PixCNN[94]	0.788	0.428	0.617	0.574	0.511	0.571	0.422	0.454	0.715	0.426	0.551
LSA[95]	0.735	0.580	**0.690**	0.542	0.761	0.546	0.751	0.535	0.717	0.548	0.641
GAN[34]	0.708	0.458	0.664	0.510	0.722	0.505	0.707	0.471	0.713	0.458	0.592
SCG[96]	0.717	0.494	0.662	0.527	0.736	0.504	0.726	0.560	0.680	0.566	0.617
AnoGAN[34]	0.671	0.547	0.529	0.545	0.651	0.603	0.585	0.625	0.758	0.665	0.618
DSVDD[51]	0.617	0.659	0.508	0.591	0.609	0.657	0.677	0.673	0.759	0.731	0.648
OCGAN[97]	0.757	0.531	0.640	**0.620**	0.723	0.620	0.723	0.575	**0.820**	0.554	0.657
GradCon[88]	0.760	0.598	0.648	0.586	0.733	0.603	0.684	0.567	0.784	0.678	0.664
CVAE-MMI[98]	0.682	0.614	0.604	**0.620**	0.704	0.562	0.734	0.639	0.756	0.675	0.659
PatchNet[39]	0.791	**0.703**	0.675	0.561	0.739	0.638	0.732	**0.674**	0.814	0.722	0.705
Ours	**0.804**	0.694	**0.690**	0.618	**0.768**	**0.717**	**0.824**	0.611	0.756	**0.746**	**0.723**

B.2 ISIC 2019 Challenge and Dataset

In this section, we show that unsupervised anomaly detection and diagnosis on dermoscopy imaging datasets with our unsupervised VAE model brings concrete and immediate benefits. On one hand, generative representations allow a model to diagnose known classes without overconfident failures on OOD classes. On the other hand, unsupervised approaches significantly reduce the cost of acquiring data by leveraging on large amounts of unlabeled data.

As follows, we start by training and evaluating our anomaly detection pipeline that consists of a one-class vanilla VAE and Local Outlier Factor (LOF), applied on the VAE's latent space, on The International Skin Imaging Collaboration (ISIC)

Table B4: We split the entire training set into 80% for training, 5% for validating and 15% for testing. Each split has all 8 classes, and the class proportions are similar to the overal dataset class proportion. We also remove duplicated images for the same lesion. There can be no lesions in two different splits.

	Images (%)	Lesions (%)	AK %	BCC %	BKL %	DF %	MEL %	NV %	SCC %	VASC %
Entire Set	25,331 (100%)	13,931 (100%)	0.0342	0.1312	0.1036	0.0094	0.1785	0.5083	0.0248	0.0100
Train	20,001 (78.9%)	11,000 (78.9%)	0.0346	0.1333	0.1061	0.0094	0.1749	0.5084	0.0234	0.0098
Val	1,762 (6.9%)	932 (6.7%)	0.0312	0.1271	0.0925	0.0142	0.2015	0.5028	0.0255	0.0051
Test	3,558 (14%)	1,999 (14.3%)	0.0337	0.1214	0.0947	0.0070	0.1875	0.5101	0.0320	0.0135

Challenge 2018/2019 datasets under multiple experimental settings. After this, our anomaly detection pipeline, in conjunction with our gradient-based attention maps generation, are applied to localize different lesion attributes. We conduct multiple experiments to show that our simple unsupervised method can be successfully applied to anomaly detection on skin images, as well as generating meaningful visual maps to assist diagnosis.

B.2.1 Dataset Preparation

We evaluate on the International Skin Imaging Collaboration(ISIC) 2018 and 2019 Challenge datasets. These datasets provide different tasks, including lesion attributes deteciton, lesion diagnosis and classification, OOD detection. The images are coming from 3 sources: Barcelona(BCN)[41], Vienna(HAM)[40] and New York(MSK)[42]. The ISIC 2019 dataset is challenging, as the training set provides 8 diagnostic categories(AK, BCC, BKL, DF, MEL, NV, SCC, VASC) and is built with three different datasets: HAM10000[40], BCN20000[41] and MSK[42], with the images being created by different preprocessing and preparation protocols and end up with different resolutions. HAM10000 dataset, which is also the training set for ISIC Challenge 2018, contains 600×450 images with centered cropping around the lesion, and are color corrected by the curators[40]. BCN 20000, contains images of size 1024×1024, with many images uncropped and lesions present at difficult or uncommon locations[41]. Furthermore, the MSK dataset contains images with various sizes. Lastly, there many legions in the 2019 dataset are represented with multiple images.

Facing all these difficulties, we adopt the image pre-processing algorithm provided by the ISIC Challenge 2019 winner[66] to deal with raw, uncropped samples in the training set. Their procedure including: 1) For uncropped images with large black areas, a rectangular bounding box for cropping is developed based on the center of the mass that covers the relevant field of view. By comparing the substantially difference of the mean intensity between inside and outside the bounding boxes, the algorithm can automatically decide on images for cropping. 2) Referencing to the image resolution in HAM10000 dataset, images are resized when needed, while keeping the aspect ratio, to be at most 600 pixels on the longest side. 3) Finally, resize all images to $512 * 512$

Table B5: AUROC for OOD detection under controlled experimental setups. We achieve AUROC of 0.642 when fitting and testing on ISIC 2019 data, and AUROCs of 0.765 and 0.504 when restricting to HAM10000 and BCN20000 respectively.

	ISIC19	HAM	BCN
AUROC	0.642	0.765	0.504

shapes for future usage. In addition to pre-processing all the images in ISIC 2019 dataset, we also ensure that we take only one image of the same lesion during training to prevent duplicated samples from contaminating the model.

B.2.2 Controlled ISIC 2019 Test

We create a controlled experiment where our OOD detector is trained and tested with diagnostic categories as presented in ISIC 2019 dataset. To start, the training set is split into 3 folds: 80% for training, 5% for validating and 15% for testing. Each fold must have all 8 classes, and the class proportions are similar to the overall dataset class proportions. Detailed split proportions are shown in Table B4. Then, following a similar setup presented in [99], we move a subset of the 8 classes to be the out-of-distribution class, while the rest of classes as in-distribution. To be more specific, we pick the two classes with least proportions, Dermatofibroma(DF) and Vascular lesion(VASC), as OOD cases, and the other 6 categories as ID cases.

We train a vanilla VAE on 6-class ID training samples, using Adam optimization with base learning rate of $1e^{-}5$ and weight decay. The input RGB images are resized to $256*256$ with the following data augmentations: color jittering with 0.4 for brightness, contrast, saturation and 0.1 for hue; random horizontal flip; random resize crop with scale from 0.7 to 0.9. During inference time, we use our model to extract features for each ID samples and fit a LOF detector on ID class. After which, VAE extracts features for all OOD testing samples and LOF generate a corresponding anomaly score. The scores are used to yield AUROC for detecting OOD versus ID. As shown in Table B5, we achieve an AUROC of 0.642 on the whole ISIC 2019 dataset under our split.

When restricting our evaluation to ISIC 2018(HAM) data by fitting and testing LOF with samples from HAM10000, we achieve an AUROC score of 0.765. This value is much higher than our previous score when fitting on all ISIC 2019 data, because ISIC 2018 data come from HAM10000 which is pre-processed well by the creators and hence is relatively simpler than the overall dataset. This result falls within our expectation that LOF detector works better when more representative data are used for fitting.

To further elaborate on this point, we try to restrict on BCN20000 samples alone. As stated before, the BCN20000 data are the most complex and particularly challenging portion among ISIC 2019 dataset. This time, when we fit and test with LOF detector, we receive an AUROC score as 0.504. This very low score on BCN20000 is understandable as fitting LOF to challenging training data points leads to worse performance in anomaly detection.

Table B6: Number of Inlier and Outlier samples in training, validation and test splits on ISIC 2019 official dataset.

	Train	Validation	Test
Inlier	10959	892	6238
Outlier	0	0	2000

Table B7: AUROC for OOD detection on the official ISIC Challenge 2019 dataset using LOF detector for anomaly scores. Comparing to state-of-art **supervised** models Multi-Resolution EfficientNets(EN)[66] and Uncertainty Estimation(UE)[99], our **unsupervised** model can achieve comparable scores.

	EN[66]	UE[99]	Ours(ISIC19)	Ours(HAM)
I AUROC	0.636	0.692	0.606	of 0.648

B.2.3 ISIC 2019 Official Test

Taking the knowledge from our previous experiments under the controlled scenario, we move on to investigate the official ISIC 2019 tests to further evaluate our proposed unsupervised VAE-LOF approach and fulfill the goal to tackle practical real world problems. The challenge provides different tasks, from which we explore in this thesis the task for the identification of the anomalous Unknown(UNK) class samples versus the diagnosed 8-class samples in the testing set. To the best of our knowledge, we are the first to try to solve this problem in an unsupervised setting, as the existing baselines are supervised methods, and many of which even incorporate external data for better training and testing performances.

We tried to adopt as similar split as possible to the state-of-art[99]. An overview of our data split for this task is shown in Table B6. We also apply data augmentation before feeding the training images to our model, including random rotation, random flipping, random color jittering or gray scaling and random resized crop. The latent feature dimension is $256 \times 8 \times 8$ on all the 8 classes training samples. Our scoring protocol remains the same: extracting sample features with trained VAE, fitting LOF detector on training cases, and generating anomaly scores for testing samples. And following the previous section, we evaluate to fit LOF detector on the overall ISIC 2019 data as well as restricting to HAM10000 data.

We report AUROC of our model compared to state-of-art supervised baselines in Tabel B7. The first baseline is provided by Nils Gessert *et al.* [66], in which the authors use multiple resolutions and external dermoscopic images to train different efficient nets and won the 2019 challenge. The second baseline is proposed by Marc Combalia *et al.* [99], in which the authors adopt uncertainty estimation metrics for anomaly detection task. Our final input images are in size of 224 to be consistent to the baselines. As shown in Tabel B7, it is worth noting from our quantitative results that, even though our proposed vanilla VAE is simple and unsupervised, our results are comparable to the SOTA supervised methods[99], and are better than the top algorithms on ISIC 2020 Challenge Leaderboard without using external data.

Appendix C Comparisons of Anomaly Localization with GradCAM on MNIST

We show some qualitative comparisons with GradCAM to demonstrate the benefits of our proposed method. We conduct experiments on the MNIST dataset and compute both GradCAM attention maps and those with our method. Since GradCAM needs classification logits to compute attention maps, we design a VAE-GradCAM baseline by adding a classification branch (with this module trained with the available class label for the input image). For training VAE-GradCAM, we use images and labels of all digits (0-9). The classification branch outputs logits y^c for any class c and a softmax layer is used to calculate cross-entropy loss. During the testing, given an image $\tilde{I}$ with a ground-truth class label $\tilde{c}$, we follow the GradCAM to use the score for class $\hat{c}$, $y^{\hat{c}}$ from the classification branch to calculate the anomaly attention map. The attention map so obtained is expected to highlight regions which show the differences between the class $\hat{c}$ and $\tilde{c}$. For instance, we input an image of digit "7" and the ground-truth label is the class "7". Then we take the classification score of digit "1" to calculate GradCAM attention map. The attention should highlight the regions which distinguish between digit "7" and digit "1", *i.e.*, horizontal bar, in the input image $\tilde{I}$ of digit "7". It should be noted that the aforementioned calculation for attention map by using the GradCAM needs the classification score from the classification branch, which is significantly different from our approach that generates VAE attention from the learned latent space.

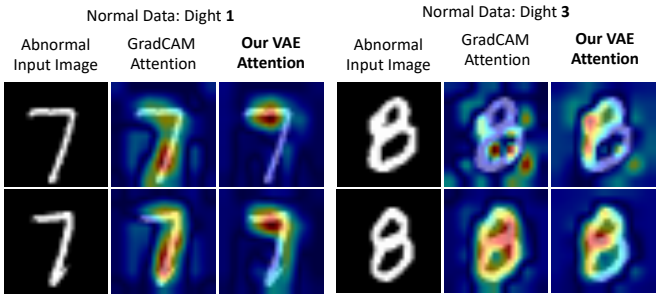


Fig. C4: Comparisons of anomaly attention maps on VAEs generated using our novel approach and Grad-CAM.

Fig. C4 presents comparisons of anomaly attention maps generated by using our proposed method and the aforementioned VAE-GradCAM baseline. We observe that our method gives promising anomaly attention maps that highlight the true anomalies in the input images, while those with VAE-GradCAM lead to failures. We attribute this to our approach to visually interpret the latent space learned on VAEs. In this particular example, our attention maps capture the distribution of anomaly data, leading to significant interpretations with respect to anomaly data and effective anomaly localization results. A trained VAE with images ($\mathbf{x} \in \mathbf{X}$) of digit "1" learns

the distribution as $\mu^{\mathbf{x}}$ and $\sigma^{\mathbf{x}}$. While the VAE outputs a latent space as $\mu^{\mathbf{y}}$ and $\sigma^{\mathbf{y}}$ given testing images of digit "7" ($\mathbf{y} \in \mathbf{Y}$). Specifically, $\mu^{\mathbf{x}}$ and $\sigma^{\mathbf{x}}$ represent the distribution of images of digit "1" for training while the VAE model never "sees" images from different distributions, *e.g.*, images of digit "7" which typically contain a horizontal bar and a vertical line. $\mu^{\mathbf{y}}$ and $\sigma^{\mathbf{y}}$ deviate from the of $\mu^{\mathbf{x}}$ and $\sigma^{\mathbf{x}}$ because images of digit "7" are apparently different from images of digit "1". We define the normal difference distribution as $\mathcal{N}^{ndd} \sim (\boldsymbol{\mu}^{ndd}, \boldsymbol{\sigma}^{ndd})$ to represent the difference between the ($\mu^{\mathbf{y}}$, $\sigma^{\mathbf{y}}$) and the ($\mu^{\mathbf{x}}$, $\sigma^{\mathbf{x}}$) in Equation (5) in the revised paper. Our attention maps are generated by sampling latent vectors $\mathbf{z}$ sampled from the $\mathcal{N}^{ndd} \sim (\boldsymbol{\mu}^{ndd}, \boldsymbol{\sigma}^{ndd})$ which visually capture the difference between images of digit "7" and images of digit "1" and attention maps so obtained are denoted as anomaly attention for anomaly localization. From the Fig. C4, we observe that our VAE attention can accurately interpret the difference of distributions between the normal and abnormal data and attention maps can be used for anomaly localization. For instance, our VAE attention can highlight the different regions between images of digit "7" and digit "1", which the main difference is the horizontal bar. While GradCAM cannot work on latent space in the VAE-GradCAM baseline because it is always requiring discriminative outputs, resulting in that GradCAM has been struggling to provide meaningful attention maps on VAEs. Besides, the requirement of GradCAM for discriminative outputs on generative models, such as VAEs, demands additional training objectives, ground-truth knowledge, data, etc., which makes it inflexible and limits the applications.