DAug: Diffusion-based Channel Augmentation for Radiology Image Retrieval and Classification

Ying Jin[1,2]   Zhuoran Zhou[1]   Haoquan Fang[1]   Jenq-Neng Hwang[1]

[1]University of Washington   [2]Microsoft

Abstract

Medical image understanding requires meticulous examination of fine visual details, with particular regions requiring additional attention. While radiologists build such expertise over years of experience, it is challenging for AI models to learn where to look with limited amounts of training data. This limitation results in unsatisfying robustness in medical image understanding. To address this issue, we propose Diffusion-based Feature Augmentation (DAug), a portable method that improves a perception model's performance with a generative model's output. Specifically, we extend a radiology image to multiple channels, with the additional channels being the heatmaps of regions where diseases tend to develop. A diffusion-based image-to-image translation model was used to generate such heatmaps conditioned on selected disease classes. Our method is motivated by the fact that generative models learn the distribution of normal and abnormal images, and such knowledge is complementary to image understanding tasks. In addition, we propose the Image-Text-Class Hybrid Contrastive learning to utilize both text and class labels. With two novel approaches combined, our method surpasses baseline models without changing the model architecture, and achieves state-of-the-art performance on both medical image retrieval and classification tasks.

1 Introduction

Shortages and burnout of radiologists are significant problems worldwide and leave risks to patient care [1, 2]. Training a radiologist takes thirteen to fifteen years [3], making AI models assisting in diagnostics a scalable solution. Notably, Chest X-ray (CXR) classification and retrieval are fundamental problems that have solid clinical values, as classification can cross-check with doctors, and retrieval allows comparison with historical cases for more accurate diagnoses.

In recent years, the performance of various image understanding tasks are meaningfully improved by leveraging pre-trained vision models. Contrastive models like CLIP [4] provide a strong baseline for understanding the common, scenic images. However, when transferred to medical images, their pre-trained capabilities are under-explored for two reasons. First, medical images require meticulous examination of fine details, which differs from scenic images where salient objects are large and have clear boundaries. Second, radiology images like Chest X-rays are monochrome, preventing full utilization of the pre-trained model's capability to utilize all three color channels. Consequently, vision models trained on limited radiology images may not attend to the correct regions for accurate understanding. Previous work [5] has shown that drawing visual marks on the input image improves model performance on tasks related to visual grounding. Therefore, we are motivated to explore highlighting potential disease areas in medical images to improve the performance of medical imaging tasks. In this study, we propose a new method to achieve this goal by adding visual grounding hints as additional image channels. Specifically, we augment the image feature with abnormality heatmaps as additional channels alongside the original monochrome medical image. The augmented feature
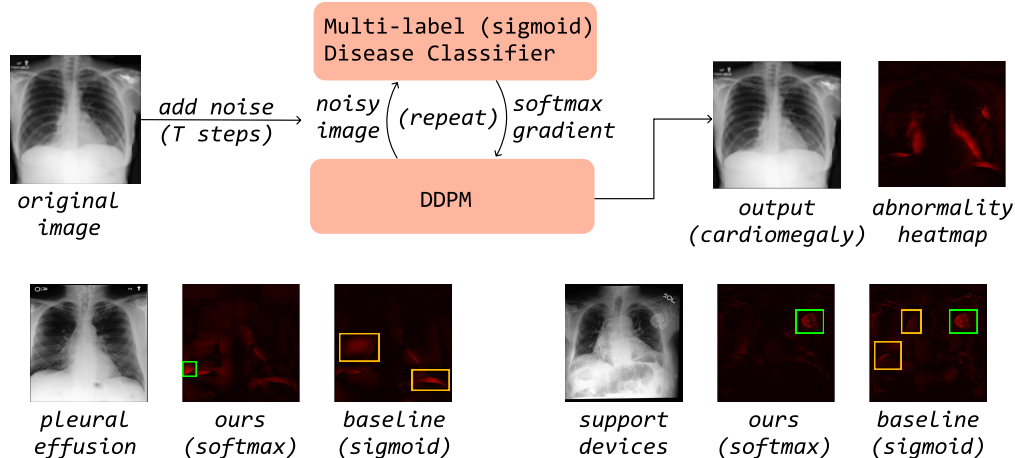
Figure 1: Diffusion-based Feature Augmentation (DAug) pipeline. The original image is translated into a diseased or healthy version with a classifier-guided diffusion model. The upper row shows an example where an image with a healthy heart is turned into a cardiomegaly (enlarged heart) version. The difference between the input and output images produces a heatmap highlighting the potential area of the corresponding disease. In the case of cardiomegaly, the heatmap correctly highlights the boundary of the heart. The heatmaps are added to the original monochrome radiology image as additional image channels, resulting in an augmented input feature that can improve the performance of downstream tasks. The DAug features support multiple disease categories (two examples in the second row), and our softmax-based approach generates more accurate heatmaps than the existing baseline. Green and orange bounding boxes indicate correctly and wrongly highlighted regions, respectively.

provides additional information to the model on areas requiring extra attention, mimicking the routine of experienced radiologists. This method, dubbed Diffusion-based Feature Augmentation (DAug), is portable onto a wide range of model architectures and leverages their native compatibility with multi-channel (RGB) images. An example of such heatmaps (e.g. for the disease of cardiomegaly) can be found in Figure 2.

Concretely, we train an image-to-image translation model to convert an input image to heatmaps of selected disease classes. Each heatmap highlights the region where the disease could potentially occur. The heatmap, as an additional feature channel, explicitly directs the model's attention to clinically significant areas, a skill difficult to acquire through conventional model training methods. The image-to-image translation model is implemented by repurposing a classifier-guided diffusion model [6]. First, we add Gaussian noise into a Chest X-ray image, making a noised version of the original image as the model input. Second, the diffusion model, guided by a disease classifier, removes the noise toward a direction where a disease is mitigated or exacerbated. In Figure 2, we show an example where cardiomegaly is worsened in the diffusion output. The difference between the original image and the output yields an attention heatmap that highlights the potential disease area (e.g. the boundaries of the heart for cardiomegaly). Through this generative learning process, the model acquires knowledge beyond what is typically learned by directly training a classification or retrieval model, and enhances the performance of these tasks. Utilizing generative learning as the mechanism, our method produces heatmaps in a self-supervised manner, eliminating the need for human-annotated heatmaps. Our diffusion-based heatmap generation is inspired by [7], and we extensively improved the method to handle the co-occurrence of multiple diseases and to generate cleaner and more accurate outputs.

In addition, different from existing work which trains retrieval and classification models separately, we are motivated by the synergy of learning two tasks together. Particularly, the text labels for retrieval provide richer textual context whereas the class labels have a well-defined discrete format. Note that the text labels often omit negative medical findings (i.e. diseases that are not found won't be mentioned) but the class labels always state these explicitly. Therefore, we design an Image-Text-Class Hybrid Contrastive learning criterion which contrasts both image-text and image-class pairs during training. This method not only improves the performance on each task as it leverages both

text and class labels, but also results in a single model that supports both classification and retrieval tasks, making real-world deployment at ease.

To validate the proposed two methods, we combine both DAug and the Image-Text-Class Hybrid Contrastive loss and evaluate on the largest Chest X-ray dataset, MIMIC-CXR [8]. Our model outperforms existing state of the arts on both retrieval and classification tasks.

To sum up, the contribution of this paper is three-fold:

- We propose DAug, a portable feature augmentation method which improves medical image understanding performance by adding abnormality heatmaps as additional image channels.
- We introduce Image-Text-Class Hybrid Contrastive learning which leverages both image-text and image-class labels to improve performance on both retrieval and classification tasks.
- We deliver a single model which is capable of both retrieval and classification tasks with state-of-the-art performance. The proposed methods can be applied to standard pre-trained models (such as CLIP ViT [4]), making real-world deployment at ease.

2 Related Work

2.1 Generation for Perception

Perception and generation models have long been regarded as two distinct paradigms in machine learning. In computer vision, perception tasks such as classification, retrieval, and segmentation require training data of <image, ground truth> pairs labeled in a defined format, which often leads to a limitation in the amount of training data. Generative models, on the other hand, reconstruct the original image during training without requiring annotations, enabling the use of larger amounts of image-only training data. In order to generate new samples, image generation models learn the distribution of images $q(\mathbf{x}_0)$, and this knowledge could aid in the understanding of images as well. This insight has led us to explore the use of a diffusion-based generation model in aiding medical image understanding.

Existing works in image understanding have been benefiting from generative models. One stream of work explores using generative models to augment or synthesize training data [9, 10, 11]. Another stream of work uses generative modeling as a pre-training task. For example, Masked AutoEncoders (MAE) [12] reconstructs the original image from a partial input during pretraining, and finetunes on downstream classification tasks. In terms of medical images, [7] proposed medical anomaly detection with classifier-guided diffusion. When applied on Chest X-rays, their work produces reasonable anomaly heatmaps for easy cases like pleural effusion, but cannot handle the co-occurrence of multiple diseases reliably. Our work is based on [7], and we make significant algorithmic improvements to turn the output into a useful feature that aids downstream tasks. Our improvements are detailed in section 3.

2.2 Image Retrieval and Classification

In the era of transformer models, pretraining on web-scale data boosts the performance on downstream tasks like retrieval and classification. CLIP [4] proposes image-text contrastive learning, which aligns the feature spaces of an image encoder and a text encoder. Retrieval methods base on the cosine similarity of CLIP features have become dominant since then. Due to the significant domain gap between medical images and the common scenic images, CLIP shows limited zero-shot performance in the medical domain. A stream of work [13, 14] aims to fine-tune CLIP on medical image-text data.

In this study, we address the challenges of medical image retrieval and classification from two perspectives, including DAug from the feature augmentation perspective and the Image-Text-Class Hybrid Contrastive learning from the loss function perspective. Related to feature augmentation, previous studies [15, 16] show that maintaining feature banks for prototypes built on the whole dataset is often effective for tasks with limited training data, and this applies to medical images. Recently, X-TRA [17] proposes improving radiology multi-modal retrieval and classification by a retrieval-based feature augmentation. They use a CLIP model to select the top-K similar samples from the dataset and construct an augmented feature for each input sample. Differently, DAug, our method, leverages generative models and outperforms existing methods.

Related to our Image-Text-Class Hybrid Contrastive loss, UniCL [18] also combines image-text and image-class datasets for contrastive pre-training. However, each training sample is assumed to have either a text label or a class label, but not both. This assumption is different from the medical domain, where datasets such as MIMIC-CXR and CheXpert [8, 19, 20] often contain both text (medical report) and class (multiple disease classes) labels at the same time. Our Image-Text-Class Hybrid Contrastive loss function utilizes this observation by combining both image-text contrastive loss and image-class contrastive loss into a single loss term per sample. Intuitively, the method treats image classification as a special image-to-class retrieval task. It not only enables the synergy of learning both retrieval and classification tasks together, but also delivers a single model that can perform both two tasks. Therefore, both model performance and deployment ease are improved.

3   DAug: Diffusion-based Feature Augmentation

We propose a unified model for multi-modal retrieval and image classification in the radiology domain. Two proposed methods improve the input feature and the training criterion, respectively, and are both portable to a suite of pre-trained transformer architectures. This section walks through Diffusion-based Feature Augmentation (DAug) in detail. We first introduce the training of the classifier which guides diffusion. Then, we describe the use of a diffusion-based generative model to create heatmaps and the methods of using these additional features.

3.1   Multi-label Image Classifier

The disease classifier is trained separately with the diffusion model to provide guidance on generating an image targeting a disease class. The inputs include a noisy image $\mathbf{x}_t$ and the time step t. We use sigmoid activation and binary cross-entropy loss to handle multiple labels per image.

For experiments on both retrieval and classification tasks, we use MIMIC-CXR [8], the largest Chest X-ray image-text dataset. As class labels are unavailable on this dataset, we follow existing work to generate them with CheXbert [20], a text classification model that converts text into a pre-defined set of 14 disease classes.

3.1.1   Disease Super-classes

To support our goal of generating heatmaps to enhance the feature for downstream tasks, we extend previous study [7] from a single disease class that is relatively easy (pleural effusion) to multiple disease classes that develop in different regions. We find that simply training a classifier on the 14 CheXbert classes results in unsatisfying result when using the classifier to guide a diffusion model. This is because some of the 14 defined disease classes may share the same or similar visual features. For example, the class "Consolidation" includes another class "Pneumonia", and "Pneumonia" has similar visual features with "Edema", although they are triggered by different causes. In sum, the definition of 14 CheXbert classes includes non-visual clues. Therefore, it hurts the performance of image generation when the classifier is forced to distinguish them. In addressing these challenges, we select representative classes and group them into seven super-classes, and trained the classifier on the super-classes. For example, "Pleural Effusion" and "Pleural Other" are grouped together to generate a heatmap on abnormalities near the pleural areas. The rationale for using super-classes is two-fold: it eliminates ambiguity in distinguishing between a parent class and its child classes, and it acknowledges that some diseases share similar visual features but differ in their causes. This approach enables the classifier to focus on distinguishing visual features rather than the underlying causes.

The grouping of child classes should be decided based on the dataset and the tasks. Too many or too few super-classes result in under-fitting and over-fitting to the grouped classes, respectively. We find seven super-classes is a balanced choice for the scenarios in this paper. Please find the list of super-classes and the classifier performance in the supplementary material.

3.2   Classifier-Guided Diffusion Model

We train a Denoising Diffusion Probabilistic Model (DDPM) model [21] on Chest X-ray images. Given an input image I, we generate a series of noisy images $\{x_0, x_1, \ldots, x_T\}$ by gradually adding Gaussian noise for T steps. x_0 is the original image with no noise and $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A U-Net model
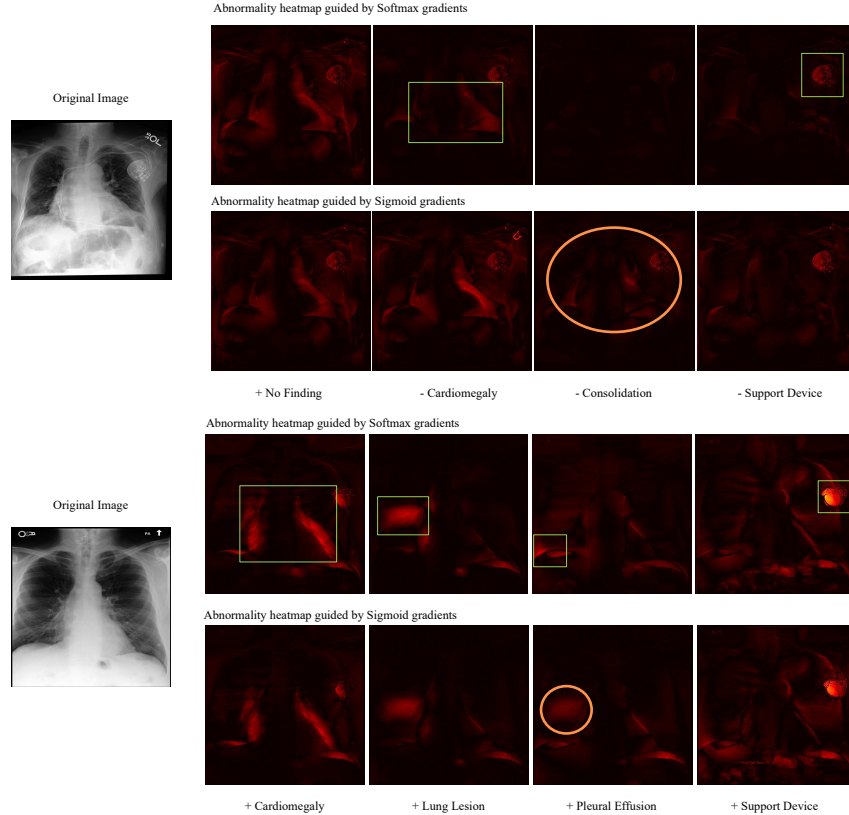
Figure 2: Example output of abnormality heatmaps generated by DAug. We use two chest X-rays as examples, each one covers four types of abnormalities (cardiomegaly, consolidation, etc. in each column). The **plus** $(+)$ and **minus** $(-)$ signs indicate the direction of the classifier gradient, meaning amplifying the disease and reducing the disease, respectively. **"+ No Findings"** reduces the probability of all potential diseases. For each input, the **first row** shows the output heatmap guided by gradients of **softmax** probabilities, and the **second row** shows the results guided by gradients of the **sigmoid** probabilities. The **green bounding boxes** shows that our method correctly highlights the region of the disease, which can help the mode to establish better image-text correspondence. Also, using softmax gradient is better than sigmoid gradient as guidance, as softmax successfully removes false positives (see orange circles). **Orange circle** in the second row highlights false positive areas of consolidation, and the orange circle in the last row highlights a wrong activation of lung lesion when it is supposed to detect pleural effusion. The corresponding softmax version (green box) makes correct detections.

is trained to reverse the process by estimating the noise added in any timesteps. During evaluation, the model can recover the original image x_0 by removing the estimated noise for T steps from Gaussian noise X_T. We set $T = 1,000$.

To translate a radiology image into its healthy or diseased variant, the generated image must retain the same anatomic structure as the input image. For this purpose, following [7], we add noise by only 500 steps to the original input, resulting in x_{500} where the anatomic structures are still visible. The diffusion model is used to recover the original input from x_{500}. To produce a diseased or healthy version of the input, the denoising process is guided by the image classifier in subsection 3.1 that assigns an X-ray to defined disease classes. The gradients for a selected class are used to condition the denoising process, making the output image maximize or minimize the probability of the class in the classifier.

For Chest X-ray images, each image may contain multiple diseases, making the binary classifier from the single-class baseline [7] unsuitable. From our evaluation, particularly challenging cases involve diseases that often co-occur (e.g., pleural effusion and lung opacity). In such instances, the
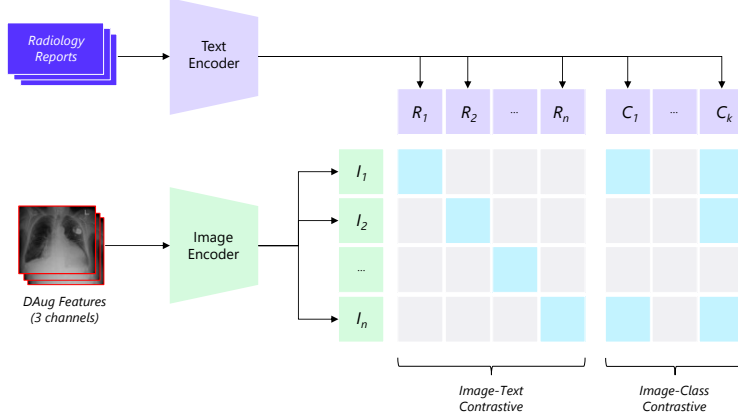
Figure 3: Model architecture and Image-Text-Class Hybrid Contrastive Loss. The inputs are pairs of radiology reports and DAug features (3-channel images including both medical image and abnormality heatmap channels). The image and text encoders are pre-trained by CLIP. R and C are text embeddings for the reports and class prompts, respectively. The hybrid contrastive loss includes both the image-text CLIP loss and image-class contrastive loss. Blue cells are positive pairs, in which the image-class matchings are derived from ground-truth class labels.

generated heatmap highlights areas for both diseases without accurately distinguishing them. This problem arises because the image classifier cannot differentiate diseases that frequently co-exist, where the presence of one disease serves as a strong bias for the presence of the other. To address these challenges, we propose the improvements below.

3.2.1 Sigmoid for Training and Softmax for Testing

When using the classifier as guidance, we employ sigmoid activation for training the classifier but softmax for guiding the diffusion model during testing. As multiple diseases can co-exist on a medical image, we train the classifier with sigmoid activation to use multiple disease class labels. During inference, however, we observed that the generated outputs using sigmoid gradients as denoising guidance tend to highlight false positive regions. This issue can be intuitively explained with an example. Suppose that we are generating a diseased version of the input image where the disease class is cardiomegaly. We want the generated output to contain only cardiomegaly so that the difference heatmap will accurately highlight this particular disease. During the denoising process, we update the noisy image toward a direction that increases the sigmoid probability z_i for cardiomegaly. As the sigmoid equation in Equation 1 considers only one class, the generated output is not guaranteed to be free of other diseases (i.e. minimized $z_j, j \neq i$). In fact, as some diseases tend to co-exist (e.g. cardiomegaly and pleural effusion), the generated image tends to have both diseases, which introduces false positives. Using softmax as the activation during inference solves this issue because mathematically Equation 2 enforces maximized z_i while minimizing probability of other classes ($z_j, j \neq i$). Intuitively, the task transforms from "generating an image where the chance of cardiomegaly is maximized" to "generating an image where the chance of cardiomegaly, compared to other diseases, is maximized". We validate the effectiveness with qualitative comparisions, as shown in Figure 1 (bottom row) and Figure 2.

$$\text{sigmoid}(z_i) = \frac{1}{1 + e^{-z_i}} \tag{1}$$

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} = \frac{1}{\sum_{j=1}^{K} e^{z_j - z_i}} \tag{2}$$

3.2.2 Channel-wise Feature Augmentation

The generated heatmaps can be integrated into the downstream perception model in various ways. A simple method involves processing them through a vision backbone and then concatenating or adding

Table 1 columns and data:

	No Finding	Enl. Cardiomed.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	Avg	wAvg
Zhang et al. [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.485	-
X-TRA *CLIP* ($\mathcal{L}_{CLIP}$)	.62	.52	.93	.88	.50	.60	.29	.44	.75	.54	.85	.50	.36	.71	.723	.606
X-TRA *CNN+BERT* ($r \rightarrow x$)	**.77**	.54	.73	.91	.52	**.83**	.39	**.87**	.77	.63	.74	.23	.61	.73	.735	.662
X-TRA *PubmedCLIP*	.75	.65	.92	**.99**	.23	.79	.21	.51	.59	**.72**	.81	.56	.43	.67	.720	.645
X-TRA *CLIP*	.63	.62	**.96**	.94	.62	.69	.47	.61	**.85**	.69	**.91**	.57	.46	**.82**	**.779**	.703
DAug *CLIP (ours)*	**.77**	**.86**	.83	.89	**.64**	.79	**.80**	.71	.81	.64	.79	**.66**	.76	.79	.767	**.799**
Yu et al. [23]	-	.65	.75	.72	.43	.80	.73	.60	.76	.76	.85	.43	.16	.86	.680	-
X-TRA[17] *CLIP* ($\mathcal{L}_{CLIP}$)	.71	.52	.74	.78	.39	.79	.39	.40	.76	.42	.67	.44	.43	.64	.761	.578
X-TRA *CNN+BERT* ($x \rightarrow x$)	.87	.63	.88	**.90**	.49	.90	.57	.60	.85	**.85**	.83	.29	.47	.82	.769	.678
X-TRA *PubmedCLIP*	**.90**	.63	.82	.83	.39	.86	.45	.63	.87	.53	**.90**	.48	.51	.79	.795	.685
X-TRA *PubmedCLIP*	**.90**	.63	.82	.83	.39	.86	.45	.63	.87	.53	**.90**	.48	.51	.79	.795	.685
X-TRA *CLIP*	.84	.62	**.89**	.89	**.56**	**.91**	.55	.59	**.89**	.60	.86	.49	.57	**.84**	**.840**	.713
DAug *CLIP (ours)*	.72	**.86**	.83	.88	.53	.78	**.79**	**.65**	.81	.47	.77	**.59**	**.69**	.74	.721	**.771**

Table 1: **Retrieval performance** for both report-to-image (r-x) and image-to-image(x-x) scenarios, measured with retrieval mAP@K, where K=5. wAvg and Avg are the weighted average of classes and the average of classes, respectively. **wAvg** is the primary metric as long-tailed challenges is not a focus of this paper. Our model, DAug-CLIP surpasses existing methods by a clear margin in the prevalent report-to-image scenario. In terms of image-to-image retrieval, DAug-CLIP achieves state-of-the-art performance on the most common classes and in terms of wAvg. This could be attributed to the reliability of the DAug feature in these classes, as the classifier guiding the diffusion model is affected by class imbalance.

	Aug	No Finding	Enl. Cardiomed.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	Avg	wAvg
CNN+BERT	-	.81	.63	.73	.67	.62	.83	.69	.59	.68	.75	.83	.70	.58	.84	.79	.71
	X-TRA[17]	.81	.74	.75	.69	.63	.81	.72	.63	.75	.75	.83	.69	.63	.85	.82	.73
PubmedCLIP	-	.78	.65	.72	.66	.61	.82	.70	.61	.73	.76	.81	.62	.54	.84	.78	.70
	X-TRA[17]	.84	.76	.78	.69	.64	.83	.73	.64	.76	.75	.82	.75	.67	.85	.83	.75
CLIP	-	.77	.65	.71	.67	.62	.85	.73	.61	.72	.75	.80	.59	.51	.83	.80	.70
	X-TRA[17]	.82	.78	.74	.70	.71	.82	.75	.63	.79	.78	.86	.74	.72	.91	.85	.77
	DAug, ours	**.93**	**.90**	**.90**	**.91**	**.85**	**.89**	**.89**	**.85**	**.89**	**.90**	**.90**	**.87**	**.84**	**.89**	**.89**	**.89**

Table 2: **Image classification** performance on MIMIC-CXR dataset. Results are in AUC-ROC. wAvg is the weighted average by number of samples per class. Avg is the average. Augmentation method X-TRA improves baseline performance, with CLIP achieving the best result. DAug-CLIP, our method, surpasses the best setting of X-TRA by a clear margin in all classes.

the heatmap features to the image features for downstream tasks. This straightforward approach necessitates modifications to the model architecture. We propose an alternative, channel-wise feature augmentation method, which incorporates the heatmaps as additional image channels alongside the medical image.

This channel-wise feature augmentation offers two main advantages. First, it leverages the capabilities of powerful pre-trained vision models designed for RGB three-channel input. As radiology images are typically monochrome, the channel-wise augmentation makes better use of the computational budget for three channels. Second, channel-wise augmentation does not require any changes to the model architecture, making it easier to utilize a wide range of pre-trained transformers. This approach is complementary to other methods, enhancing performance on downstream tasks without the need for architectural modifications.

4 Image-Text-Class Hybrid Contrastive Learning

Image-to-text retrieval and image classification are deeply interconnected tasks. Essentially, image classification can be seen as a retrieval problem focusing on a smaller, reused set of targets. When the classification head is a linear layer without bias, the class logits become unscaled cosine similarities between the image feature and the weights in the linear layer. In image-to-text retrieval, the weights

Method	Feature Aug	Criterion	No Finding	Enl. Cardiomed.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	Avg	wAvg
Image Retrieval: $r \rightarrow x$																	
X-TRA (Baseline)	X-TRA	CLIP	.62	.52	.93	.88	.50	.60	.29	.44	.75	.54	.85	.50	.36	.71	.723	.606
DAug (Ablation)	DAug	CLIP	**.77**	.85	.80	.86	.60	.76	.78	.67	.79	.61	.76	.65	.75	**.79**	.745	.778
DAug (Full)	DAug	Hybrid Contrastive	**.77**	**.86**	**.83**	**.89**	**.64**	**.79**	**.80**	**.71**	**.81**	**.64**	**.79**	**.66**	**.76**	**.79**	**.767**	**.799**
Image Retrieval: $x \rightarrow x$																	
X-TRA (Baseline)	X-TRA	CLIP	.71	.52	.74	.78	.39	.79	.39	.40	.76	.42	.67	.44	.43	.64	.761	.578
DAug (Ablation)	DAug	CLIP	.71	**.86**	.81	**.88**	.49	.76	.78	.62	**.81**	.44	.76	.58	**.69**	.74	.711	.765
DAug (Full)	DAug	Hybrid Contrastive	**.72**	**.86**	**.83**	**.88**	**.53**	**.78**	**.79**	**.65**	**.81**	**.47**	**.77**	**.59**	**.69**	.74	**.721**	**.771**
Image Classification																	
CLIP (Baseline)	-	CLIP	.77	.65	.71	.67	.62	.85	.73	.61	.72	.75	.80	.59	.51	.83	.80	.70
X-TRA (Baseline)	X-TRA	CLIP	.82	.78	.74	.70	.71	.82	.75	.63	.79	.78	.86	.74	.72	.91	.85	.77
DAug (Ablation)	DAug	CLIP	.77	.86	.83	.89	.64	.79	.80	.71	.81	.64	.79	.66	.76	.79	.77	.80
DAug (Full)	DAug	Hybrid Contrastive	**.93**	**.90**	**.90**	**.91**	**.85**	**.89**	**.89**	**.85**	**.89**	**.90**	**.90**	**.87**	**.84**	**.89**	**.89**	**.89**

Table 3: **Ablation studies** on image retrieval and classification tasks. Compared to baselines, our feature augmentation method DAug gains performance over no augmentation (CLIP Baseline) and the X-TRA augmentation (X-TRA Baseline). Besides, using the proposed Image-Text-Class Hybrid Contrastive loss outperforms the baselines using the original CLIP loss. Baseline results from [17].

in the linear layer are replaced by text embeddings dynamically generated for each target text. Motivated by the potential benefits of jointly training both retrieval and classification tasks, we integrate image-text and image-class labels into our training loss. Unlike existing methods [24, 18] that expand contrastive learning to class labels, our approach uniquely addresses scenarios where a single sample is associated with both text and class labels, aiming to train a unified model for retrieval and classification. To distinguish from existing work, we name our method image-text-class hybrid contrastive learning.

As illustrated in Figure 3, we first transform each class into a fixed set of texts by converting each class into prompts. For instance, the class "Cardiomegaly" is rephrased as "A photo of a Chest X-ray image with cardiomegaly". During training, as depicted in Figure 3, class prompts and radiology reports are transformed into text embeddings C and R, respectively. We then calculate two sets of losses and take the weighted average:

$$\mathcal{L} = w * \mathcal{L}_{CLIP} + (1 - w) * \mathcal{L}_{i2c}, \tag{3}$$

where $\mathcal{L}_{CLIP}$ is the CLIP loss including both image-to-text and text-to-image cross-entropy loss, and $\mathcal{L}_{i2c}$ is the image-to-class binary cross-entropy loss. Specifically,

$$\mathcal{L}_{i2c} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\text{sim}(i,j)) + (1 - y_i) \log(1 - \text{sim}(i,j)) \right], \tag{4}$$

where $\text{sim}(i,j)$ is the cosine similarity between the image embedding I_i and text embedding for class prompts C_i. During training, embedding C_i is regenerated with each update to the text encoder. Several works discussed the connection between contrastive learning and cross-entropy loss [24, 18]. In our scenario, $\mathcal{L}_{i2c}$ is essentially contrasting image I with the class prompts, where multiple positive pairs could exist determined by the ground truth class labels. The inherent cross-entropy nature of both loss terms facilitates training both tasks simultaneously without causing the embedding space to diverge for each task.

4.1 Multi-modal Retrieval and Classification

After the model is trained, we leverage the model in different ways for the classification and retrieval tasks. For classification, we use only the image encoder and connect it with a linear classifier. The linear classifier consists of a single layer where the weight vectors per class are populated with the text embeddings for class prompts generated by the final text encoder. The bias is set to zeros. Essentially, the output logit per class is equivalent to the unscaled cosine similarity between the image and classes $\text{logit}_i = sigmoid(I \cdot C_i)$.

For retrieval tasks, both the image and text encoders are used to convert each sample into embeddings. We use the cosine similarity of the embeddings to rank the association for retrieval.

5 Experiments

5.1 Implementation details

To benchmark our ideas, we need an image-text-class dataset on medical images. We select **MIMIC-CXR**, the largest Chest X-ray (CXR) medical report dataset. It contains 227,835 image-text pairs, where the texts are radiology reports which list the normal and abnormal findings. We follow the official training and testing splits. As class labels are unavailable, we generate pseudo-class labels with the CheXbert labeler [25], which is a standard practice in the field. It is a text classification model that converts a radiology report into binary labels on 14 disease classes. One of them is "No Findings", indicating a healthy case.

For DAug feature augmentation, we construct a three-channel image with the first two channels containing the medical image and the third channel filled with the diffusion-generated heatmap guided by the class "No Findings". In our experiments, we evaluate on all 14 disease classes and compare the results with the original CLIP as a baseline. Therefore, we select the heatmap for "No Finding" which combines all diseases. This requires no change in the model architecture for a fair comparison with the vanilla CLIP. In real-world applications, heatmaps for individual disease groups (e.g., cardiomegaly) can be selected for optimal performance gain according to the scenario.

Following existing work [17], we resize images to 256×256 and use a CLIP pretrained ViT-Base/32 model for fair comparison. We fine-tune the model with a cosine learning rate scheduler with a base learning rate of $2e^{-5}$ for 10 epochs. We set the weight w in Equation 3 to 0.7 based on the validation performance of two tasks. We use a batch size of 256 over eight V-100 GPUs. With all images resized in advance, training and evaluation take around 2 hours. Abnormality heatmaps are pre-generated.

5.2 Results

We compare our method, DAug with existing methods on both multi-modal retrieval and image classification on the MIMIC dataset. Table 1 demonstrates that our method outperforms existing state-of-the-art approaches in retrieving radiology images with medical reports, a critical clinical scenario where radiologists refer to previous cases to confirm diagnoses. Table 2 shows the performance of the same model on image classification, which outperforms existing methods by a clear margin. Specifically, we find that X-TRA benefits more in the tail classes, while DAug improves the overall performance which is reflected in the higher wAvg score.

DAug produces heatmaps without requiring ground truth. Unfortunately, the lack of ground truth also constrains us on measuring the quality of the heatmaps. Considering that our end goal is to improve the performance of downstream tasks, we validate the heatmap quality by measuring the improvement of these tasks when using the heatmaps. Therefore, we conduct ablation studies in Table 3. The results show that performance on both tasks surpasses the baseline when only changing the feature augmentation approach to DAug, which proves that the heatmaps are helpful features. When both the DAug feature augmentation and the proposed Hybrid Contrastive criterion are used, the performance on both tasks is improved. This comparison validates that both methods effectively aid in medical image understanding tasks.

6 Conclusion

In this paper, we present a single model that achieves state-of-the-art performance in both medical image retrieval and classification tasks. We propose DAug, a diffusion-based channel-wise feature augmentation method that empirically directs the model where to look for clinical diagnoses. We enable multi-class support and reduce the false positive issue found in previous work by using super-classes and softmax gradients for diffusion guidance. In addition, we propose the Image-Text-Class Hybrid Contrastive learning criterion, which leverages both image-text and image-class labels and enables a unified model for two tasks with easy deployment. Our method requires no modifications to the model architecture, making it portable to a wide range of pre-trained models.

References

[1] D. J. Cao, C. Hurrell, and M. N. Patlas, "Current status of burnout in canadian radiology," *Canadian Association of Radiologists Journal*, vol. 74, no. 1, pp. 37–43, 2023.

[2] A. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ: British Medical Journal (Online)*, vol. 359, 2017.

[3] C. Clinic, "Radiologist." `https://my.clevelandclinic.org/health/articles/24184-radiologist`, 2022.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[5] J. Yang, H. Zhang, F. Li, X. Zou, C. yue Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *ArXiv*, vol. abs/2310.11441, 2023.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[7] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pp. 35–45, Springer, 2022.

[8] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[9] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *arXiv preprint arXiv:1806.03384*, 2018.

[10] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pp. 1–11, Springer, 2018.

[11] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293, IEEE, 2018.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[13] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," *arXiv preprint arXiv:2210.10163*, 2022.

[14] S. Eslami, G. de Melo, and C. Meinel, "Does clip benefit visual question answering in the medical domain as much as it does in the general domain?," *arXiv preprint arXiv:2112.13906*, 2021.

[15] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pp. 563–579, Springer, 2022.

[16] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," *arXiv preprint arXiv:2204.13258*, 2022.

[17] T. van Sonsbeek and M. Worring, "X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation," in *International Conference on Information Processing in Medical Imaging*, pp. 471–482, Springer, 2023.

[18] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19163–19173, 2022.

[19] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[20] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert," *arXiv preprint arXiv:2004.09167*, 2020.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[22] Y. Zhang, W. Ou, J. Zhang, and J. Deng, "Category supervised cross-modal hashing retrieval for chest x-ray and radiology reports," *Computers & Electrical Engineering*, vol. 98, p. 107673, 2022.

[23] Y. Yu, P. Hu, J. Lin, and P. Krishnaswamy, "Multimodal multitask deep learning for x-ray image retrieval," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 603–613, Springer, 2021.

[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.

[25] S. Jain, A. Smit, S. Q. Truong, C. D. Nguyen, M.-T. Huynh, M. Jain, V. A. Young, A. Y. Ng, M. P. Lungren, and P. Rajpurkar, "Visualchexbert: addressing the discrepancy between radiology report labels and image labels," in *Proceedings of the Conference on Health, Inference, and Learning*, pp. 105–115, 2021.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[28] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

A Details on Image Generation and Image-to-Image Translation

Common image generation methods include Generative Adversarial Networks (GANs) [26], Variational Autoencoders (VAEs) [27] and Diffusion Models [28]. Diffusion models have become mainstream due to the ease of training and the superior image quality. During training, the diffusion model learns to remove noise from a noisy input and therefore can gradually turn a Gaussian noise into an image. Such denoising steps can be guided by an image classifier trained separately, whose gradients are used to determine the direction of the denoising process, encouraging the output to maximize the probability of a certain class based on the classifier. The result will be an image of the chosen class.

To train a diffusion model, we first conduct a forward diffusion process which gradually adds Gaussian noise to the original image. The forward process has T steps (usually, $T = 1000$), producing a sequence of noisy samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$. $\mathbf{x}_0 = \mathbf{I}$ is the original image and $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ becomes a Gaussian noise. Then, a U-Net model is trained to predict the noise added per step, in order to reverse the steps by removing the added noise from the noisy image. The reverse process gradually turns a Gaussian noise into an image. Diffusion models can be guided by an image classifier to generate an image of a particular class. The image classifier is trained to produce class probabilities given a noisy image $f_\phi(y|\mathbf{x}_t, t)$. The gradients of the classifier are used to alter the denoising at each step so that the output maximizes the probability of the target class.

A classifier-guided diffusion model can be used for image-to-image translation. Specifically, we obtain a half-noised image at time step $x = 500$, and then conduct denoising guided by a classifier. As $\mathbf{x}_{500}$ maintains key distinguishable features of the original image, the output still maintains the identity of the original input, but changes it in a way that will be classified to the target class. Related to medical images, this is about converting a healthy CXR to a diseased one, and vice versa.

B Disease Super-classes

In our classifier-guided diffusion model, the classifier is trained on disease super-classes, with each super-class consisting of one or multiple related diseases. The super-classes were defined with radiologists to align with medical knowledge. For the additional rationale behind this definition, CheXpert [20] provides a hierarchical structure of the 14 disease classes, which aligns with our super-class definition. Take super-class #4 as an example, they are grouped together because they are shown as increased density in the X-ray, although for different reasons. The goal is to let the classifier focus only on the appearance instead of attempting to distinguish the cause. For example, in [20], Atelectasis is another type of lung opacity abnormality. We categorize it separately because it looks different (an absence of density).

We show the classifier performance on the super-classes in Table 4. Upon empirical examinations, we found that the quality of the heatmaps is highly correlated with the classifier performance of the selected class. This observation supports the decision to group sub-classes together to improve the classifier's robustness. Please note that the performance in the table is expected to be low, as input to the classifier is noisy images instead of the original image (see Appendix A for details about classifier-guided diffusion).

C Pseudo-label Quality

The class labels are not human annotated but are generated with CheXpert, a text classification model that converts a radiology report into disease classes. According to human evaluation in [20], the label quality is claimed to have a 96.9% F1 score.

Super Class	Disease classes	AP
1	No Finding	0.631
2	Enlarged Cardiomediastinum, Cardiomegaly	0.885
3	Lung Lesion	0.407
4	Consolidation, Edema, Pneumonia	0.857
5	Atelectasis	0.778
6	Pleural Effusion, Pleural Other	0.746
7	Support Device	0.633

Table 4: Multi-label classifier performance in Average Precision (AP). Please note that this is the classifier that takes in the noisy image and is trained to guide the diffusion model. There are totally seven classes, and each one is a super-class consisting of disease classes with similar visual features. For example, super class 4 includes Edema and Pneumonia, which are sub-categories of Consolidation. Training the classifier with merged classes reduces class imbalance and improves performance, and therefore provides better guidance for the diffusion model.

D   Ethical Considerations and Limitations

Our use of the MIMIC-CXR dataset was approved through PhysioNet[1]. All authors who accessed the data have obtained permission.

We identify two limitations of our work. First, to be compatible with the pre-trained models, we configure the input image to be 3 channels. The method may achieve even better results if heatmaps of all supported super-classes are used. Second, the use of a diffusion model introduces significant computational overhead. To address this limitation, it worth exploring to use the heatmaps as an additional supervision to improve a student model. We will take these directions as our future work.

---

[1]https://physionet.org/content/mimic-cxr/2.0.0/