DiRA: Discriminative, Restorative, and Adversarial Learning for Self-supervised Medical Image Analysis

Fatemeh Haghighi[1,*], **Mohammad Reza Hosseinzadeh Taher**[1,*], **Michael B. Gotway**[2], **Jianming Liang**[1]

[1]Arizona State University

[2]Mayo Clinic

Abstract

Discriminative learning, restorative learning, and adversarial learning have proven beneficial for self-supervised learning schemes in computer vision and medical imaging. Existing efforts, however, omit their synergistic effects on each other in a ternary setup, which, we envision, can significantly benefit deep semantic representation learning. To realize this vision, we have developed **DiRA**, the first framework that unites discriminative, restorative, and adversarial learning in a unified manner to collaboratively glean complementary visual information from unlabeled medical images for fine-grained semantic representation learning. Our extensive experiments demonstrate that DiRA (1) encourages collaborative learning among three learning ingredients, resulting in more generalizable representation across organs, diseases, and modalities; (2) outperforms fully supervised ImageNet models and increases robustness in small data regimes, reducing annotation cost across multiple medical imaging applications; (3) learns fine-grained semantic representation, facilitating accurate lesion localization with only image-level annotation; and (4) enhances state-of-the-art restorative approaches, revealing that DiRA is a general mechanism for united representation learning. All code and pretrained models are available at https://github.com/JLiangLab/DiRA.

1. Introduction

Self-supervised learning (SSL) aims to learn generalizable representations without using any expert annotation. The representation learning approaches in the SSL paradigm can be categorized into three main groups: (1) *discriminative learning*, which utilizes encoders to cluster instances of the same (pseudo) class and distinguish instances from different (pseudo) classes; (2) *restorative learning*, which utilizes generative models to reconstruct original images from their distorted versions; and (3) *adversarial learning*, which utilizes adversary models to enhance restorative learning. In computer vision, discriminative SSL approaches, especially contrastive learning [8, 12, 13, 15, 21, 24, 27, 27, 34, 44, 53], currently offer state-of-the-art (SOTA) performance, surpassing standard supervised

---

fhaghigh@asu.edu .

*Equal contributors ordered alphabetically.

ImageNet models in some tasks. In medical imaging, however, restorative SSL methods [10, 25, 26, 43, 55, 57] compared to discriminative approaches [3, 56] presently reach a new height in performance. Naturally, we contemplate: *What contributes to the popularity differences between discriminative and restorative methods in computer vision and in medical imaging?* Furthermore, from our extensive literature review, we have discovered that *no* SSL method exploits all three learning components simultaneously; therefore, we ponder: *Can discriminative, restorative, and adversarial learning be seamlessly integrated into a single framework to foster collaborative learning for deep semantic representation, yielding more powerful models for a broad range of applications?* In seeking answers to the two questions, we have gained the following insights.

Computer vision and medical imaging tasks embrace the spirit of evil in opposite ways, originating from the *marked* differences between photographic and medical images. Photographic images, particularly those in ImageNet, have large foreground objects with apparent discriminative parts, residing in varying backgrounds (e.g. zebra and daisy images in Fig. 2). Thus, object recognition tasks in photographic images are primarily based on high-level features captured from discriminative regions. In contrast, medical images generated from a particular imaging protocol exhibit consistent anatomical structures (e.g, chest anatomy in Fig. 2), with clinically relevant information dispersed over the entire image [26]. In particular, high-level structural information, i.e., anatomical structures and their relative spatial orientations, are essential for the identification of normal anatomy and various disorders. Importantly, medical tasks require much stronger attention to fine-grained details within images as identifying diseases, delineating organs, and isolating lesions rely on subtle, local variations in texture [29]. Therefore, recognition tasks in medical images desire complementary high-level and fine-grained discriminative features captured throughout images.

According to our systematical analysis, we have gained the following understandings: (1) discriminative learning excels in capturing high-level (global) discriminative features, (2) restorative learning is good at conserving fine-grained details embedded in local image regions, and (3) adversarial learning consolidates restoration by conserving more fine-grained details. Putting these understandings and fundamental differences between photographic and medical images together would explain why restorative learning is preferred in medical imaging while discriminative learning is preferred in computer vision. More importantly, we have acquired a new and intriguing insight into *trio* of discriminative, restorative, and adversarial learning to excavate effective features required for medical recognition tasks—not only high-level anatomical representations but also fine-grained discriminative cues embedded in the local parts of medical images.

Based on the insights above, we have designed a novel self-supervised learning framework, called DiRA, by uniting discriminative learning, restorative learning, and adversarial learning in a unified manner to glean complementary visual information from unlabeled medical images. Our extensive experiments demonstrate that (1) DiRA encourages collaborative learning among three learning components, resulting in more generalizable representation across organs, diseases, and modalities (see Fig. 4); (2) DiRA outperforms

fully supervised ImageNet models and increases robustness in small data regimes, thereby reducing annotation cost in medical imaging (Tab. 1 and Tab. 2); (3) DiRA learns fine-grained representations, facilitating more accurate lesion localization with only image-level annotations (Fig. 5); and (4) DiRA enhances SOTA restorative approaches, showing that DiRA is a general framework for united representation learning (Tab. 3).

In summary, we make the following contributions:

- The insights that we have gained into the *synergy* of discriminative, restorative, and adversarial learning in a <u>ternary</u> setup, realizing a new paradigm of collaborative learning for SSL.

- The first self-supervised learning framework that seamlessly unites discriminative, restorative, and adversarial learning in a unified manner, setting a new SOTA for SSL in medical imaging.

- A thorough and insightful set of experiments that demonstrate not only DiRA's generalizability but also its potential to take a fundamental step towards developing *universal* representations for medical imaging.

2.   Related works

Discriminative self-supervised learning.

Discriminative methods can be divided into *class-level* and *instance-level* discrimination. *Class-level* discrimination methods [7, 8, 17, 22, 35, 54] group images based on certain criteria, assign a pseudo label to each group, and train a model to discriminate the images based on their pseudo labels, such as rotation degrees [22] and cluster assignments [7, 8, 54]. On the other hand, *instance-level* discrimination methods [8, 12, 13, 15, 21, 24, 27, 34, 44, 48, 52, 53] treat each image as a distinct class, and maximize the similarity of representations derived from different views of the same image, seeking to learn transformation invariant representations. *Instance-level* discriminative learning has been investigated in various forms, including contrastive learning [12, 14, 27, 49], asymmetric networks [15, 24], and redundancy reduction [21, 53]. However, both *class-level* and *instance-level* approaches in discriminative learning have shown failures in tasks that require finer-grained features [47,50,51]. Our DiRA addresses this limitation by incorporating restorative and adversarial learning, which not only improves discriminative learning but also yields fine-grained representations required for medical imaging tasks.

Restorative and adversarial self-supervised learning.

The key objective for a restorative method is to faithfully reconstruct the distribution of data [36,48]. In the SSL context, multiple pretext tasks are formulated to reconstruct the perturbed images using generative models [33, 37, 45]. The advance of GANs [23] has led to a new line of research in unsupervised learning, using adversarial learning to generate transferable representations [18, 19]. While recent works [11,18] have demonstrated impressive results by employing large-scale generative models, it remains unclear to what extent generative models can encapsulate high-level structures. Our DiRA alleviates this

limitation by bringing the advantages of discriminative learning into generative models. Through discriminating image samples, generative models are encouraged to capture global discriminative representations rather than superficial representations, leading to a more pronounced embedding space.

Self-supervised learning in medical imaging.

Due to the lack of large annotated datasets, SSL created substantial interest in medical imaging. Motivated by the success in computer vision, recent discriminative methods concentrate on instance-level discrimination. A comprehensive benchmarking study in [29] evaluated the efficacy of existing instance discrimination methods pre-trained on ImageNet for diverse medical tasks. Several other works adjusted contrastive-based methods on medical images [3, 9, 56]. A large body of work, on the other hand, focuses on restorative approaches, which can be categorized into restorative only [10,57], restorative and adversarial [43], and discriminative and restorative [26,30,55]. Among these groups, the most recent study on TransVW [25, 26] demonstrated superiority by combining discriminative and restorative components into a single SSL framework. DiRA distinguishes itself from all previous works by demonstrating two key advances: (1) employing discriminative, restorative, and adversarial learning simultaneously in a unified framework; and (2) providing a general representation learning framework that is compatible with existing discriminative and restorative methods, regardless of their objective functions.

3. DiRA framework

As shown in Fig. 3, DiRA is a SSL framework comprised of three key components: (1) Discrimination (Di) that aims to learn high-level discriminative representations, (2) Restoration (R) that aims to enforce the model to conserve fine-grained information about the image by focusing on more localized visual patterns, and (3) Adversary (A) that aims to further improve feature learning through the restoration component. By integrating these components into a *unified* framework, DiRA captures comprehensive information from images, providing more powerful representations for various downstream tasks. In the following, we first introduce each component by abstracting a common paradigm and then describe the joint training loss.

3.1. Discriminative learning

Discriminative learning can be thought of as training an encoder to maximize agreement between instances of the same (pseudo) class in the latent space via a discriminative loss. As illustrated in Fig. 3, the discriminator branch is comprised of two twin backbone networks f_θ and f_ξ, and projection heads h_θ and h_ξ. f_θ is a regular encoder, while f_ξ can be a momentum encoder [24, 27] or share weights with f_θ [15, 26, 53]. Given two patches x_1 and x_2, which are cropped from the same image or different images, we first apply an augmentation function $\mathcal{T}(.)$ on them. The two augmented patches are then processed by f_θ and f_ξ to generate latent features $y_1 = f_\theta(\mathcal{T}(x_1))$ and $y_2 = f_\xi(\mathcal{T}(x_2))$. The projection heads h_θ and h_ξ projects the latent features to a unit sphere and output projections $z_1 = h_\theta(y_1)$ and $z_2 =$

$h_\xi(y_2)$. The discriminator's objective is to maximize the similarity between the embedding vectors obtained from two samples of the same (pseudo) class:

$$\mathcal{L}_{dis} = \ell\ (z_1, z_2) \tag{1}$$

where $\ell(z_1, z_2)$ is the similarity/distance function that measures compatibility between z_1 and z_2. DiRA is a general framework that allows various choices of discrimination tasks without any constraint. As such, the declaration of class might range from considering every single image as a class (instance discrimination) to clustering images based on a similarity metric (cluster discrimination). Accordingly, x_1 and x_2 can be two views of the same image or two samples from the same cluster. Based on the nature of the discrimination task, the instantiation of $\mathcal{L}_{dis}$ can be crossentropy [22,26,35,58], contrastive [3,8,12,27], redundancy reduction [21,53], etc.

3.2. Restorative learning

Our restorative learning branch aims to enhance discrimination learning by leveraging fine-grained visual information. As shown in Fig. 3, the restoration branch is comprise of an encoder f_θ and decoder g_θ, where encoder f_θ is shared with the discrimination branch. Given the input sample x_1 distorted by $\mathcal{T}$, the f_θ and g_θ aims to map the distorted sample back to the original one, i.e., $f_\theta, g_\theta : (x, \mathcal{T}) \mapsto x$. f_θ and g_θ are trained by minimizing the distance between the original sample and the restored one at pixel-level:

$$\mathcal{L}_{res} = \mathbb{E}_x\, dist(x_1, x_1') \tag{2}$$

where $x_1' = g_\theta(f_\theta(\mathcal{T}(x_1)))$ denotes the restored image. dist (x_1, x'_1) presents the distance function that measures similarity between x_1 and x'_1, such as L_1 or L_2.

3.3. Adversarial learning

Adversarial learning aims to reinforce f_θ by measuring how realistic the restored images are. As such, the adversarial discriminator D_ϕ is formulated to distinguish (discriminate) the set of training images from the set of synthesized images, guiding encoder f_θ to capture more informative features from images so that g_θ can reproduce the original images effectively. Therefore, the encoder f_θ and decoder g_θ play a minimax game with adversarial discriminator D_ϕ, and are optimized jointly with an adversarial loss [6, 36]:

$$\mathcal{L}_{adv} = \mathbb{E}_x\big[log\, D_\phi(x_1)\big] + \mathbb{E}_x\big[log\big(1 - D_\phi(x_1')\big)\big] \tag{3}$$

3.4. Joint training

Finally, the combined objective for the proposed DiRA framework becomes:

$$\mathcal{L} = \lambda_{dis} * \mathcal{L}_{dis} + \lambda_{res} * \mathcal{L}_{res} + \lambda_{adv} * \mathcal{L}_{adv} \tag{4}$$

where λ_{dis}, λ_{res}, and λ_{adv} are multiplication factors that determine the relative importance of different losses. Through our unified training scheme, DiRA learns a representation

that preserves fine-grained details within the samples while being discriminative among the image classes. In particular, the formulation of $\mathscr{L}_{dis}$ encourages the model to capture high-level discriminative features. Moreover, $\mathscr{L}_{res}$ forces the model to encode fine-grained information from the images by focusing on pixel-level visual patterns. This results in more descriptive feature embeddings that elevate the discrimination task. Finally, $\mathscr{L}_{adv}$ elevates restoration based learning through capturing more informative features.

4. Implementations details

4.1. Pre-training protocol

DiRA is a general framework that is compatible with existing self-supervised discriminative and restorative methods, regardless of their objective functions. To assess the effectiveness of our framework, we adapt recent SOTA 2D and 3D self-supervised methods into DiRA, as described in the following. The pretrained models with DiRA are identified as DiRA subscripted by the original method name.

2D image pretraining settings.—We apply DiRA to MoCo-v2 [14], Barlow Twins [53], and SimSiam [15] for 2D image self-supervised learning. All DiRA models are pretrained from scratch on the training set of ChestX-ray14 [46] dataset. For each of these three discrimination tasks [14,15,53], we follow the original methods in the formulation of $\mathscr{L}_{dis}$, projection head architecture, and hyperparameters settings. Furthermore, we optimize the encoder and decoder networks f_θ and g_θ following the optimization setups in [14, 15, 53]. For all methods, we employ a 2D U-Net [38] with a standard ResNet-50 [28] backbone as the f_θ and g_θ. We adopt mean square error (MSE) as the $\mathscr{L}_{res}$. The adversarial discriminator network D_ϕ consists of four convolutional layers with the kernel size of 3×3 [37], trained using the Adam optimizer with a learning rate of 2e-4 and $(\beta_1, \beta_2) = (0.5, 0.999)$. We use batch size 256 distributed across 4 Nvidia V100 GPUs. λ_{res}, λ_{adv}, λ_{dis} are empirically set to 10, 0.001, and 1, respectively. Input images are first randomly cropped and resized to 224×224; the image augmentation function $\mathscr{T}(.)$ includes random horizontal flipping, color jittering, and Gaussian blurring. Additionally, we apply cutout [16, 37] and shuffling [10] to make the restoration task more challenging. More details are provided in the Appendix.

3D volume pretraining settings.—We apply DiRA to TransVW [26], the SOTA method for 3D self-supervised learning in medical imaging. We adapt TransVW in DiRA by adding an adversarial discriminator D_ϕ into its training scheme. For fair comparisons, we follow the publicly available TransVW code for setting instance discrimination and restoration tasks. Moreover, similar to publicly released TransVW, DiRA models are pre-trained from scratch using 623 chest CT scans in the LUNA [40] dataset. We use 3D U-Net [20] as the encoder-decoder network and a classification head including fully-connected layers. The adversarial discriminator D_ϕ includes four convolutional blocks with the kernel size 3×3 ×3. λ_{res}, λ_{adv}, λ_{dis} are empirically set to 100, 1, and 1, respectively. f_θ, g_θ, and D_ϕ are optimized for 200 epochs using Adam with a learning rate of 1e-3 and batch size of 8. More details are provided in the Appendix.

4.2. Transfer learning protocol

Target tasks and datasets.—We evaluate the effectiveness of DiRA's representations in transfer learning to a diverse suite of 9 common but challenging 2D and 3D medical imaging tasks, including: ChestX-ray14, CheXPert [31], SIIM-ACR [1], and NIH Montgomery [32] for 2D models, and LUNA, PE-CAD [41], LIDC-IDRI [2], LiTS [5], and BraTS [4] for 3D models (see Appendix for dataset details). These tasks encompass various label structures (multi-label classification and pixel-level segmentation), diseases (brain tumors and thoracic diseases, such as lung nodules, pulmonary emboli, and pneumothorax), organs (lung, liver, brain), and modalities (X-ray, CT, MRI). Moreover, these tasks contain many hallmark challenges encountered when working with medical images, such as imbalanced classes, limited data, and small-scanning areas for the pathology of interest [3, 29]. We use the official data split of these datasets when available; otherwise, we randomly divide the data into 80%/20% for training/testing.

Fine-tuning settings.—We transfer the pre-trained (1) encoder (f_θ) to the classification tasks, and (2) encoder and decoder (f_θ and g_θ) to segmentation tasks. We evaluated the generalization of DiRA representations by fine-tuning all the parameters of downstream models. We use the AUC (area under the ROC curve), and the IoU (Intersection over Union) and Dice coefficient for evaluating classification and segmentation performances, respectively. Following [29], we strive to optimize each downstream task with the best performing hyperparameters (details in Appendix). We employ the early-stop mechanism using 10% of the training data as the validation set to avoid over-fitting. We run each method ten times on each downstream task and report the average, standard deviation, and statistical analysis based on an independent two-sample *t*-test.

5. Results

We conduct extensive experiments to better understand not only the properties of our framework but also its generalizability across 9 downstream tasks. Through the following groups of experiments, we establish that DiRA (1) enriches existing discriminative approaches, capturing a more diverse visual representation that generalizes better to different tasks; (2) addresses the annotation scarcity challenge in medical imaging, providing an annotation-efficient solution for medical imaging; (3) learns fine-grained features, facilitating more accurate lesion localization with only image-level annotation; and (4) improves SOTA restorative approaches, demonstrating that DiRA is a general framework for united representation learning.

5.1. DiRA enriches discriminative learning

Experimental setup: To study the flexibility and efficacy of our proposed self-supervised framework, we apply DiRA to three recent SOTA self-supervised methods with diverse discrimination objectives: MoCo-v2, Barlow Twins, and SimSiam. To evaluate the quality of our learned representations and ascertain the generality of our findings, we follow [29] and consider a broader range of four target tasks, covering classification (ChestX-Ray14 and CheXpert) and segmentation (SIIM-ACR and Montgomery).

Results: As seen in Fig. 4, utilizing our DiRA framework consistently enhances its underlying discriminative method across all tasks (1) ChestX-ray14, (2) CheXpert, (3) SIIM-ACR, and (4) NIH Montgomery. Compared with the original methods, $DiRA_{MoCo-v2}$ showed increased performance by 0.76%, 1.17%, 1.35%, and 0.21%, respectively; Similarly, $DiRA_{Barlow\ Twins}$ showed increased performance by 0.43%, 0.60%, 0.16%, and 0.03%. Finally, $DiRA_{SimSiam}$ showed increased performance by 0.82%, 2.22%, 1.18%, and 0.45%. These results imply that DiRA is a comprehensive representation learning framework that encourages existing self-supervised instance discriminative approaches to retain more fine-grained information from images, enriching their visual representation and allowing them to generalize to different medical tasks more effectively.

5.2. DiRA improves robustness to small data regimes

Experimental setup: We investigate the robustness of representations learned with DiRA in small data regimes to determine if the learned representation can serve as a proper foundation for fine-tuning. We randomly select 1%, 25%, and 50% of training data from ChestX-ray14, CheXpert, and Montgomery, and fine-tune the self-supervised pre-trained models on these training-data subsets.

Results: As shown in Tab. 1, our DiRA pre-trained models outperform their counterparts' original methods in all subsets, 1%, 25%, and 50%, across ChestX-ray14, CheXpert, and Montgomery. In particular, the average of improvement for MoCo-v2 and SimSiam across all three downstream tasks in each underlying subset garnering: (1) 5.6 % and 7% when using 1%, (2) 2.9 % and 1.3% when using 25%, and (3) 2.2 % and 1% when using 50%. As seen in 1%, DiRA outperforms its counterparts MoCo-v2 and SimSiam by a large margin, demonstrating our framework's potential for combating overfitting in extreme low data regimes. Although the Barlow Twins is more resistant to low data regimes than the previous two approaches, DiRA still improves its performance by 0.5%, 0.5%, and 0.6% on average across all three datasets when using 1%, 25%, and 50% of labeled data, respectively. In summary, our results in the low-data regimes demonstrate our framework's superiority for providing more robust and transferable representations that can be harnessed for downstream tasks with limited amounts of data, thereby reducing annotation costs.

5.3. DiRA improves weakly-supervised localization

Experimental setup: We investigate our DiRA framework in a weakly supervised setting, comparing its applicability for localizing chest pathology to underlying discriminative methods. Given this goal, we follow [46] and use the ChestX-ray14 dataset, which contains bounding box annotations for approximately 1,000 images. For training, we initialize models with our DiRA pre-trained models, and train downstream models using only image-level disease labels. Following [39, 46], bounding boxes are only used as ground truth to evaluate disease localization accuracy in the testing phase. To generate heatmaps, we leverage Grad-CAM [39]. Heatmaps indicate the spatial location of a particular thoracic disease.

Results: As seen in Fig. 5, our framework learns more fine-grained representations, enabling it to localize diseases more accurately. In particular, heatmaps generated by MoCo-v2, Barlow Twins, and SimSiam models are highly variable, whereas DiRA models consistently achieve more robust and accurate localization results over each corresponding original method. Through the production of more interpretable activation maps, our DiRA framework demonstrates possible clinical potential for post-hoc interpretation by radiologists. Quantitative disease localization results are provided in the Appendix.

5.4. DiRA outperforms fully-supervised baselines

Experimental setup: Following the recent transfer learning benchmark in medical imaging [29], we compare the transferability of DiRA models, pre-trained solely on unlabeled images from ChestX-ray14, with two fully-supervised representation learning approaches: (1) supervised ImageNet model, the most common transfer learning pipeline in medical imaging and (2) supervised model pretrained on ChestX-ray14, the upper-bound in-domain transfer learning baseline. The supervised baselines benefit from the same encoder as DiRA, namely ResNet-50. We fine-tune all pre-trained models for 4 distinct medical applications ranging from target tasks on the source dataset to the tasks with comparatively significant domain-shifts in terms of data distribution and disease/object of interest.

Results: As shown in Tab. 2, DiRA models achieves significantly better or on-par performance compared with both supervised ImageNet and ChestX-ray14 models across four downstream tasks. In particular, DiRA$_{MoCo-v2}$ and DiRA$_{Barlow\ Twins}$, outperforms both supervised baselines in CheXpert, SIIM-ACR, and Montgomery, respectively. Moreover, DiRA$_{SimSiam}$ outperforms the supervised ImageNet and the ChestX-ray14 pre-trained models in SIIM-ACR and Montgomery, respectively. These results indicate that our framework, with zero annotated data, is capable of providing more generic features for different medical tasks.

5.5. DiRA sets a new state-of-the-art for self-supervised learning in 3D medical imaging

Experimental setup: We further investigate the effectiveness of our framework for enhancing restorative representation learning by applying DiRA to TransVW [26], the state-of-the-art SSL approach for 3D medical imaging. We select TransVW as representative of restorative self-supervised methods because it shows superior performance over discriminative [42, 58], restorative only [10, 57], and restorative and adversarial [43] methods. Following the common evaluation pipeline [26], we evaluate our learned representations by transfer learning to five common and challenging 3D downstream tasks, including classification (LUNA and PE-CAD) and segmentation (LIDC, LiTS, and BraTS).

Results: As shown in Tab. 3, DiRA framework consistently enhances TransVW across all downstream tasks. In particular, DiRA improves TransVW in LUNA, LIDC-IDRI, LiTS, and BraTS, and offers equivalent performance in PE-CAD. These results imply that by utilizing three learning components in tandem, image-based self-supervision approaches capture a more diverse visual representation that generalizes better to different downstream tasks.

6. Ablation study

Experimental setup:

We conduct a thorough ablation study to show how each component contributes to DiRA. To do so, we only vary the loss function of DiRA. For each underlying self-supervised method, i.e. MoCo-v2, Barlow Twins, and SimSiam (referred to as the base), we start with the discrimination component and incrementally add restorative and the adversarial learning. When all three components are unified, they represent the completed DiRA models. All models are pretrained on the ChestX-ray14 dataset and fine-tuned for four downstream tasks, including ChestX-ray14, CheXpert, SIIM-ACR, and Montgomery.

Results:

We draw the following observations from the results in Tab. 4: (1) Expanding discriminative self-supervised methods by adding a restoration task consistently enhances the original methods. In particular, incorporating $\mathscr{L}_{res}$ into training objectives of MoCo-v2, Barlow Twins, and SimSiam outperforms the corresponding original methods, with the exception of SimSiam in ChestX-ray14, which shows slight performance degradation. Note that this gap later compensates after adding $\mathscr{L}_{adv}$, which signifies collaborative learning among restorative and adversary components in our framework. (2) The overall trend showcases the advantage of the adversarial discriminator when added to the restoration component, improving the performance of all methods in four downstream tasks. Our findings indicate that unifying the three components in DiRA models significantly enhances the original self-supervised methods by retaining more fine-grained information from images.

7. Conclusion and discussion

We propose DiRA, the *first* SSL framework that unites discriminative, restorative, and adversarial learning in a unified manner. The key contribution of our DiRA arises from the insights that we have gained into the *synergy* of these three SSL approaches for collaborative learning. Given DiRA's generalizability, we envisage it will take a fundamental step towards developing universal representations for medical imaging. Our DiRA achieves remarkable performance gains, though we fixed the restorative learning tasks in all experiments when examining various formulations of discriminative learning. In the future, examining various choices of restoration tasks and searching for optimal collaborative learning strategies may lead to even stronger representations for medical imaging. In this paper, we have focused on medical imaging, but we envision that DiRA can also offer outstanding performance for vision tasks that demand fine-grained details.

Appendix

A. Weakly-supervised localization

In this section, we provide quantitative and additional qualitative results for weakly-supervised localization, discussed in the Sec. 5.3 of the main paper. Our quantitative results in Tab. 5, together with the qualitative results in Fig. 5 and Fig. 6, demonstrate the capability of our framework in learning fine-grained representations that can be used for more accurate pathology localization when just image-level annotations are available.

A.1. Quantitative results

Experimental setup:

Following the common protocol [11–13], we quantitatively evaluate the applicability of our DiRA framework in a weakly supervised setting using ChestX-ray14 dataset. First, we use min-max normalization to normalize each heatmap; then, following [11], we binarize the heatmaps by thresholding at $\{60, 180\}$, and generate bounding boxes around the isolated regions. To evaluate localization accuracy, we compute the intersection over union (IoU) between the generated and ground truth bounding boxes. According to [11, 12], a localization is correct when the bounding box prediction overlaps with the ground truth box with IoU δ. Following [11], we investigate the accuracy of localization under various δ values, from 10% to 60%. We run each method ten times and report the average accuracy across all runs.

Result:

Tab. 5 shows the pathology localization accuracy of our DiRA and underlying discriminative models. As seen, in each of the six IoU thresholds, DiRA models significantly outperform the corresponding discriminative models. In particular, the average of improvement for MoCo-v2, Barlow Twins, and SimSiam across all IoU thresholds is 2.38%, 5.4%, and 9.4%, respectively.

A.2. Qualitative results

Experimental setup:

During training, we initialize models with our DiRA pre-trained models, and fine-tune downstream models using only image-level disease labels. We use heatmaps to approximate the spatial location of a particular thorax disease. We generate heatmaps using Grad-CAM [13], a technique for highlighting the important regions in the image for predicting the pathology class.

Results:

Fig. 6 presents the visualizations of heatmaps generated by DiRA and the corresponding discriminative models for 8 thorax pathologies in ChestX-ray14 dataset. As seen, DiRA models provide more accurate pathology localizations compared to the underlying discriminative methods. These results demonstrate the impact of restorative learning in providing fine-grained features that are useful for disease localization.

B. Datasets and tasks

We have examined our framework in a diverse suite of 9 downstream tasks, including classification and segmentation in X-ray, CT, and MRI modalities. In this section, we provide the details of each dataset and the underlying task, as well as the evaluation metric for each task.

ChestX-ray14:

ChestX-ray14 is a large open source dataset of de-identifie chest X-ray images. The dataset includes 112K chest images taken from 30K unique patients. The ground truth consists of a label space of 14 thorax diseases. We use the official patient-wise split released with the dataset, including 86K training images and 25K testing images. The models are trained to predict 14 pathologies in a multi-label classification setting. The mean AUC score over 14 diseases is used to evaluate the classification performance. In addition to image-level labels, ChestX-ray14 provides bounding box annotations for approximately 1,000 test images. Of this set of images, bounding box annotations are available for 8 out of 14 thorax diseases. During testing, we use bounding box annotations to assess the accuracy of pathology localization in a weakly-supervised setting. The mean accuracy over 8 diseases is used to evaluate the localization performance.

CheXpert:

CheXpert is a hospital-scale publicly available dataset with 224K chest X-ray images taken from 65K unique patients. We use the official data split released with the dataset, including 224K training and 234 test images. The ground truth for the training set includes 14 thoracic pathologies that were retrieved automatically from radiology reports. The testing set is labeled manually by board-certified radiologists for 5 selected thoracic pathologies— Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion. The models are trained to predict five pathologies in a multi-label classification setting. The mean AUC score over 5 diseases is used to evaluate the classification performance.

SIIM-ACR:

This open dataset is provided by the Society for Imaging Informatics in Medicine (SIIM) and American College of Radiology, including 10K chest X-ray images and pixel-wise segmentation mask for Pneumothorax disease. We randomly divided the dataset into training (80%) and testing (20%). The models are trained to segment pneumothorax from chest

radiographic images (if present). The segmentation performance was measured by the mean Dice coefficient score.

NIH Montgomery:

This publicly available dataset is provided by the Montgomery County's Tuberculosis screening program, including 138 chest X-ray images. There are 80 normal cases and 58 cases with Tuberculosis (TB) indications in this dataset. Moreover, ground truth segmentation masks for left and right lungs are provided. We randomly divided the dataset into a training set (80%) and a test set (20%). The models are trained to segment left and right lungs in chest scans. The segmentation performance is evaluated by the mean Dice score.

Table 5.

Weakly-supervised pathology localization accuracy under different IoU thresholds (δ): DiRA models provide stronger representations for pathology localization with only image-level annotations. For each method, we report the average performance over ten runs. The green arrows show the improvement of DiRA models compared with the underlying discriminative method in each IoU threshold.

Method	$\delta = 10\%$	$\delta = 20\%$	$\delta = 30\%$	$\delta = 40\%$	$\delta = 50\%$	$\delta = 60\%$
MoCo-v2 [3]	54.89	39.43	24.81	14.59	7.58	2.68
DiRA$_{MoCo-v2}$	**58.13** (↑ 3.2)	**42.74** (↑ 3.3)	**27.52** (↑ 2.7)	**16.25** (↑ 1.7)	**9.30** (↑ 1.7)	**4.35** (↑ 1.7)
Barlow Twins [4]	50.54	38.01	26.36	16.93	9.31	4.69
DiRA$_{BarlowTwins}$	**58.98** (↑ 8.4)	**45.26** (↑ 7.2)	**32.71** (↑ 6.3)	**21.71** (↑ 4.8)	**13.62** (↑ 4.3)	**6.26** (↑ 1.6)
SimSiam [5]	30.24	19.80	11.46	5.62	2.30	0.79
DiRA$_{SimSiam}$	**51.07** (↑ 20.8)	**34.24** (↑ 14.4)	**20.64** (↑ 9.2)	**11.32** (↑ 5.7)	**6.46** (↑ 4.2)	**2.90** (↑ 2.1)

LUNA:

This publicly-available dataset consists of 888 lung CT scans with a slice thickness of less than 2.5mm. The dataset were divided into training (445 cases), validation (178 cases), and test (265 cases) sets. The dataset provides a set of 5M candidate locations for lung nodule. Each location is labeled as true positive (1) or false positive (0). The models are trained to classify lung nodule candidates into true positives and false positives in a binary classification setting. We evaluate the classification accuracy by Area Under the Curve (AUC) score.

PE-CAD:

This dataset includes 121 computed tomography pulmonary angiography (CTPA) scans with a total of 326 pulmonary embolism (PE). The dataset provides a set of candidate locations for PE and is divided at the patient-level into training and test sets. Training set contains 434 true positive PE candidates and 3,406 false positive PE candidates. Test set contains 253 true positive PE candidates and 2,162 false positive PE candidates. We pre-processed the

3D scans as suggested in [6]. The 3D models are trained to classify PE candidates into true positives and false positives in a binary classification setting. We evaluate the classification accuracy by Area Under the Curve (AUC) score at candidate-level.

LIDC-IDRI:

The Lung Image Database Consortium image collection (LIDC-IDRI) dataset is created by seven academic centers and eight medical imaging companies. The dataset includes 1,018 chest CT scans and marked-up annotated lung nodules. The dataset is divided into training (510), validation (100), and test (408) sets. We pre-processed the data by re-sampling the 3D volumes to 1-1-1 spacing and then extracting a $64 \times 64 \times 32$ crop around each nodule. The models are trained to segment long nodules in these 3D crops. The segmentation accuracy is measured by the Intersection over Union (IoU) metric.

LiTS:

The dataset is provided by MICCAI 2017 LiTS Challenge, including 130 CT scans with expert ground-truth segmentation masks for liver and tumor lesions. We divide dataset into training (100 patients), validation (15 patients), and test (15 patients) sets. The models are trained to segment liver in 3D scans. The segmentation accuracy is measured by the Intersection over Union (IoU) metric.

BraTS:

The dataset includes brain MRI scans of 285 patients (210 HGG and 75 LGG) and segmentation ground truth for necrotic and non-enhancing tumor core, peritumoral edema, GD-enhancing tumor, and background. For each patient, four different MR volumes are available: native T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 fluid attenuated inversion recovery (FLAIR). We divide dataset at patient-level into training (190 patients) and testing (95 patients) sets. The models are trained to segment brain tumors (background as negatives class and tumor sub-regions as positive class). The segmentation accuracy is measured by the Intersection over Union (IoU) metric.

C. Implementation

C.1. Pre-training settings

We apply DiRA to four existing self-supervised methods [1, 3–5]. To be self-contained, we'll explain each method briefly here. Also, we provide additional pretraining details that supplements Sec. 4.1.

MoCo-v2 [3]:

We adopt MoCo-v2— a popular representative of *contrastive learning* methods, into our framework. MoCo leverages a momentum encoder to ensure the consistency of negative samples as they evolve during training. Moreover, a queue $K = \{k_1, k_2, \dots k_N\}$ is utilized to store the representations of negative samples. The discrimination task is to contrast

representations of positive and negative samples. As MoCo-v2 is adopted in DiRA, the encoder f_θ and projection head h_θ are updated by back-propagation, while f_ξ and h_ξ are updated by using an exponential moving average (EMA) of the parameters in f_θ and h_θ, respectively. The discrimination branch is trained using InfoNCE loss [7], which for a pair of positive samples x_1 and x_2 defined as follows:

$$\mathcal{L}_{dis} = -log\frac{exp(z_1 \cdot z_2/\tau)}{\sum_{n=0}^{N} exp(z_1 \cdot k_n/\tau)} \tag{5}$$

where $z_1 = h_\theta(f_\theta(x_1))$ and $z_2 = h_\xi(f_\xi(x_2))$, τ is a temperature hyperparameter, and N is the queue size. Following [3], f_θ is a standard ResNet-50 and h_θ is a two-layer MLP head (hidden layer 2048-d, with ReLU). Moreover, when adopting MoCo-v2 in DiRA, f_θ, h_θ, and g_θ are optimized using SGD with an initial learning rate of 0.03, weight decay 0.0001, and the SGD momentum 0.9.

(a) MoCo-v2 vs. DiRA_MoCo-v2

(b) Barlow Twins vs. DiRA_Barlow Twins
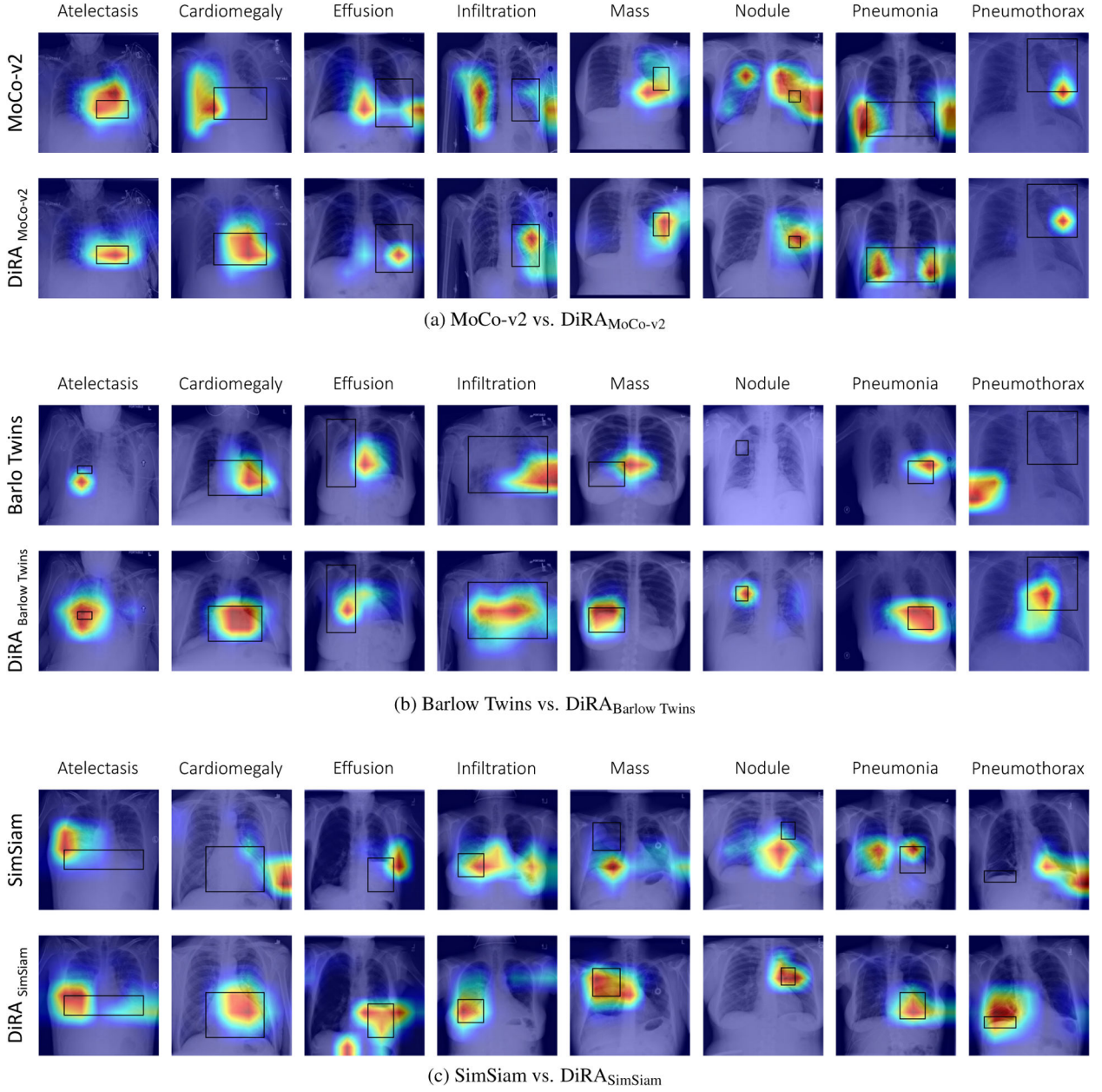

(c) SimSiam vs. DiRA_SimSiam

Figure 6. Visualization of Grad-CAM heatmaps:
We provide the heatmap examples for 8 thorax diseases in each column. The first row in each sub-figure represents the results for the original self-supervised method, while the second row represents the original method when adopted in DiRA framework. The black boxes represents the localization ground truths.

SimSiam [5]:

We adopt SimSiam— a popular representative of *asymmetric instance discrimination* methods, into our framework. SimSiam trains the model without negative pairs and directly maximizes the similarity of two views from an image using a simple siamese network

followed by a predictor head. To prevent collapsing solutions, a stopgradient operation is utilized. As such, the model parameters are only updated using one distorted version of the input, while the representations from another distorted version are used as a fixed target. As SimSiam is adopted in DiRA, the encoder f_θ and projection head h_θ share weights with f_ξ and h_ξ, respectively. The model is trained to maximize the agreement between the representations of positive samples using negative cosine similarity, defined as follows:

$$\mathscr{D}(z_1, y_2) = -\frac{z_1}{\|z_1\|_2} \cdot \frac{y_2}{\|y_2\|_2}$$

(6)

where $z_1 = h_\theta(f_\theta(x_1))$ and $y_2 = f_\xi(x_2)$. The discrimination branch is trained using a symmetrized loss as follows:

$$\mathscr{L}_{dis} = \frac{1}{2}\mathscr{D}(z_1, stopgrad(y_2)) + \frac{1}{2}\mathscr{D}(z_2, stopgrad(y_1))$$

(7)

where stopgrad means that y_2 is treated as a constant in this term. Following [5], f_θ is a standard ResNet-50 and h_θ is a three-layer projection MLP head (hidden layer 2048-d), followed by a two-layer predictor MLP head. Moreover, when adopting SimSiam in DiRA, f_θ, h_θ, and g_θ are optimized using SGD with a linear scaling learning rate (lr×BatchSize/ 256). The initial learning rate is 0.05, weight decay is 0.0001, and the SGD momentum is 0.9.

Barlow Twins [4]:

We adopt Barlow Twins— a popular representative of *redundancy reduction instance discrimination learning* methods, into our framework. Barlow Twins makes the cross-correlation matrix computed from two siamese branches close to the identity matrix. By equating the diagonal elements of the cross-correlation matrix to 1, the representation will be invariant to the distortions applied to the samples. By equating the off-diagonal elements of the cross-correlation matrix to 0, the different vector components of the representation will be decorrelated, so that the output units contain non-redundant information about the sample. The discrimination loss is defined as follows:

$$\mathscr{L}_{dis} = \sum_i (1 - \mathscr{C}_{ii})^2 + \lambda \sum_i \sum_{i \neq j} \mathscr{C}_{ij}^2$$

(8)

where $\mathscr{C}$ is the cross-correlation matrix computed between the outputs of the h_θ and h_ξ networks along the batch dimension. λ is a coefficient that determines the importance of the invariance term and redundancy reduction term in the loss. Following [4], f_θ is a standard ResNet-50 and h_θ is a three-layer MLP head. Moreover, when adopting Barlow Twins in DiRA, f_θ, h_θ, and g_θ are optimized using LARS optimizer with the learning rate schedule similar to [4].

TransVW [1]:

TransVW defines the similar anatomical patterns within medical images as anatomical visual words, and combines the discrimination and restoration of visual words in a single loss objective. As TransVW is adopted in DiRA, the encoder f_θ and projection head h_θ are identical to f_ξ and h_ξ, respectively. In particular, the discrimination branch is trained to classify instances of visual words according to their pseudo class labels using the standard crossentropy loss:

$$\mathscr{L}_{dis} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C}\mathscr{Y}_{bc}\log\mathscr{P}_{bc} \tag{9}$$

where B denotes the batch size; C denotes the number of visual words classes; $\mathscr{Y}$ and $\mathscr{P}$ represent the ground truth (onehot pseudo label vector obtained from visual word classes) and the prediction of h_θ, respectively. Following [1], we use 3D U-Net as the f_θ and g_θ. h_θ includes a set of fully-connected layers followed by a classification head. f_θ and g_θ are trained with the same setting as [1].

Joint training process:

Following [8, 9], we perform the overall pre-training with the discrimination, restoration, and adversarial losses in a gradual evolutionary manner. First, the encoder f_θ along with projector h_θ are optimized using the discrimination loss $\mathscr{L}_{dis}$ according to the learning schedule of the original discriminative methods [1, 3–5], empowering the model with an initial discrimination ability. Then, the restoration and adversarial losses are further fused into the training process incrementally. To stabilize the adversarial training process and reduce the noise from imperfect restoration at initial epochs [9], we first warm up the f_θ and g_θ using the $\mathscr{L}_{dis} + \mathscr{L}_{res}$, and then add the adversarial loss $\mathscr{L}_{adv}$ to jointly train the whole framework; the optimization of the framework by incorporation of $\mathscr{L}_{res}$ and $\mathscr{L}_{adv}$ takes up to 800 epochs. Following [2], we use the early-stop technique on the validation set, and the checkpoints with the lowest validation loss are used for fine-tuning.

C.2. Fine-tuning settings

Preprocessing and data augmentation:

Following [10], for 2D target tasks on X-ray datasets (ChestX-ray14, CheXpert, SIIM-ACR, and Montgomery), we resize the images to 224×224. For thorax diseases classification tasks on ChestX-ray14 and CheXpert, we apply standard data augmentation techniques, including random cropping and resizing, horizontal flipping, and rotating. For segmentation tasks on SIIM-ACR and Montgomery, we apply random brightness contrast, random gamma, optical distortion, elastic transformation, and grid distortion. For 3D target tasks, we use regular data augmentations including random flipping, transposing, rotating, and adding Gaussian noise.

Training parameters:

We endeavour to optimize each downstream task with the best performing hyperparameters. In all 2D and 3D downstream tasks, we use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We use early-stop mechanism using the 10% of the training data as the validation set to avoid over-fitting. For 2D classification tasks on ChestX-ray14 and CheXpert datasets, we use a learning rate $2e-4$ and *ReduceLROnPlateau* as the learning rate decay scheduler. For 2D segmentation tasks on SIIM-ACR and Montgomery, we use a learning rate $1e-3$ and *cosine* learning rate decay scheduler. For all 3D downstream tasks, we use *ReduceLROnPlateau* as the learning rate decay scheduler. For downstream tasks on LUNA, PECAD, LIDC, and LiTS, we use a learning rate $1e-2$. For BraTS dataset, we use a learning rate of $1e-3$.

References

[1]. Siim-acr pneumothorax segmentation, 2019.

[2]. Armato Samuel G III, McLennan Geoffrey, Bidaut Luc, McNitt-Gray Michael F, Meyer Charles R, Reeves Anthony P, Zhao Binsheng, Aberle Denise R, Henschke Claudia I, Hoffman Eric A, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics, 38(2):915–931, 2011. [PubMed: 21452728]

[3]. Azizi Shekoofeh, Mustafa Basil, Ryan Fiona, Beaver Zachary, Freyberg Jan, Deaton Jonathan, Loh Aaron, Karthikesalingam Alan, Kornblith Simon, Chen Ting, Natarajan Vivek, and Norouzi Mohammad. Big self-supervised models advance medical image classification. arXiv:2101.05224, 2021.

[4]. Bakas Spyridon, Reyes Mauricio, Jakab Andras, Bauer Stefan, Rempfler Markus, Crimi Alessandro, Shinohara Russell Takeshi, Berger Christoph, Ha Sung Min, Rozycki Martin, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv:1811.02629, 2018.

[5]. Bilic Patrick, Christ Patrick Ferdinand, Vorontsov Eugene, Chlebus Grzegorz, Chen Hao, Dou Qi, Fu Chi-Wing, Han Xiao, Heng Pheng-Ann, Hesser Jurgen, et al. The liver tumor segmentation benchmark (lits). arXiv:1901.04056, 2019.

[6]. Cao Bing, Zhang Han, Wang Nannan, Gao Xinbo, and Shen Dinggang. Auto-gan: Self-supervised collaborative learning for medical image synthesis. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):10486–10493, Apr. 2020.

[7]. Caron Mathilde, Bojanowski Piotr, Joulin Armand, and Douze Matthijs. Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision, pages 132–149, 2018.

[8]. Caron Mathilde, Misra Ishan, Mairal Julien, Goyal Priya, Bojanowski Piotr, and Joulin Armand. Unsupervised learning of visual features by contrasting cluster assignments. arXiv:2006.09882, 2021.

[9]. Chaitanya Krishna, Erdil Ertunc, Karani Neerav, and Konukoglu Ender. Contrastive learning of global and local features for medical image segmentation with limited annotations. In Advances in Neural Information Processing Systems, volume 33, pages 12546–12558. Curran Associates, Inc., 2020.

[10]. Chen Liang, Bentley Paul, Mori Kensaku, Misawa Kazunari, Fujiwara Michitaka, and Rueckert Daniel. Self-supervised learning for medical image analysis using image context restoration. Medical image analysis, 58:101539, 2019. [PubMed: 31374449]

[11]. Chen Mark, Radford Alec, Child Rewon, Wu Jeffrey, Jun Heewoo, Luan David, and Sutskever Ilya. Generative pretraining from pixels. In Daumé Hal III and Singh Aarti, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 1691–1703. PMLR, 13–18 Jul 2020.

[12]. Chen Ting, Kornblith Simon, Norouzi Mohammad, and Hinton Geoffrey. A simple framework for contrastive learning of visual representations. In Daumé Hal III and Singh Aarti, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 1597–1607. PMLR, 13–18 Jul 2020.

[13]. Chen Ting, Kornblith Simon, Swersky Kevin, Norouzi Mohammad, and Hinton Geoffrey. Big self-supervised models are strong semi-supervised learners, 2020.

[14]. Chen Xinlei, Fan Haoqi, Girshick Ross, and He Kaiming. Improved baselines with momentum contrastive learning, 2020.

[15]. Chen Xinlei and He Kaiming. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15750–15758, June 2021.

[16]. DeVries Terrance and Taylor Graham W.. Improved regularization of convolutional neural networks with cutout, 2017.

[17]. Doersch Carl, Gupta Abhinav, and Efros Alexei A. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, pages 1422–1430, 2015.

[18]. Donahue Jeff and Simonyan Karen. Large scale adversarial representation learning. In Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, and Garnett R, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[19]. Dumoulin Vincent, Belghazi Ishmael, Poole Ben, Lamb Alex, Arjovsky Martín, Mastropietro Olivier, and Courville Aaron C.. Adversarially learned inference. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

[20]. Ellis David G. and Aizenberg Michele R.. Trialing u-net training modifications for segmenting gliomas using open source deep learning framework. In Crimi Alessandro and Bakas Spyridon, editors, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pages 40–49, Cham, 2021. Springer International Publishing.

[21]. Ermolov Aleksandr, Siarohin Aliaksandr, Sangineto Enver, and Sebe Nicu. Whitening for self-supervised representation learning. In Meila Marina and Zhang Tong, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 3015–3024. PMLR, 18–24 Jul 2021.

[22]. Gidaris Spyros, Singh Praveer, and Komodakis Nikos. Unsupervised representation learning by predicting image rotations. arXiv:1803.07728, 2018.

[23]. Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, David Warde-Farley, Ozair Sherjil, Courville Aaron, and Bengio Yoshua. Generative adversarial nets. In Ghahramani Z, Welling M, Cortes C, Lawrence N, and Weinberger KQ, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.

[24]. Grill Jean-Bastien, Strub Florian, Altché Florent, Tallec Corentin, Richemond Pierre, Buchatskaya Elena, Doersch Carl, Pires Bernardo Avila, Guo Zhaohan, Azar Mohammad Gheshlaghi, Piot Bilal, kavukcuoglu koray, Munos Remi, and Valko Michal. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle H, Ranzato M, Hadsell R, Balcan MF, and Lin H, editors, Advances in Neural Information Processing Systems, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[25]. Haghighi Fatemeh, Taher Mohammad Reza Hosseinzadeh, Zhou Zongwei, Gotway Michael B., and Liang Jianming. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pages 137–147, Cham, 2020. Springer International Publishing.

[26]. Haghighi Fatemeh, Taher Mohammad Reza Hosseinzadeh, Zhou Zongwei, Gotway Michael B., and Liang Jianming. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. IEEE Transactions on Medical Imaging, 40(10):2857–2868, 2021. [PubMed: 33617450]

[27]. He Kaiming, Fan Haoqi, Wu Yuxin, Xie Saining, and Girshick Ross. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[28]. He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[29]. Taher Mohammad Reza Hosseinzadeh, Haghighi Fatemeh, Feng Ruibin, Gotway Michael B., and Liang Jianming. A systematic benchmarking analysis of transfer learning for medical image analysis. In Albarqouni Shadi, Cardoso M. Jorge, Dou Qi, Kamnitsas Konstantinos, Khanal Bishesh, Rekik Islem, Rieke Nicola, Sheet Debdoot, Tsaftaris Sotirios, Xu Daguang, and Xu Ziyue, editors, Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health, pages 3–13, Cham, 2021. Springer International Publishing.

[30]. Taher Mohammad Reza Hosseinzadeh, Haghighi Fatemeh, Gotway Michael B., and Liang Jianming. CAiD: a self-supervised learning framework for empowering instance discrimination in medical imaging. Proceedings of Machine Learning Research, 2022.

[31]. Irvin Jeremy, Rajpurkar Pranav, Ko Michael, Yu Yifan, Ciurea-Ilcus Silviana, Chute Chris, Marklund Henrik, Haghgoo Behzad, Ball Robyn, Shpanskaya Katie, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031, 2019.

[32]. Jaeger Stefan, Candemir Sema, Antani Sameer, Wáng Yí-Xiáng J, Lu Pu-Xuan and Thoma George. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery, 4(6), 2014.

[33]. Larsson Gustav, Maire Michael1, and Shakhnarovich Gregory. Colorization as a proxy task for visual understanding. In CVPR, 2017.

[34]. Li Junnan, Zhou Pan, Xiong Caiming, and Hoi Steven C. H.. Prototypical contrastive learning of unsupervised representations, 2021.

[35]. Noroozi Mehdi and Favaro Paolo. Unsupervised learning of visual representations by solving jigsaw puzzles. In European Conference on Computer Vision, pages 69–84. Springer, 2016.

[36]. Parmar Gaurav, Li Dacheng, Lee Kwonjoon, and Tu Zhuowen. Dual contradistinctive generative autoencoder. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 823–832, June 2021.

[37]. Pathak Deepak, Krahenbuhl Philipp, Donahue Jeff, Darrell Trevor, and Efros Alexei A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2536–2544, 2016.

[38]. Ronneberger Olaf, Fischer Philipp, and Brox Thomas. U-net: Convolutional networks for biomedical image segmentation. In Navab Nassir, Hornegger Joachim, Wells William M., and Frangi Alejandro F., editors, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pages 234–241, Cham, 2015. Springer International Publishing.

[39]. Selvaraju Ramprasaath R., Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, and Batra Dhruv. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[40]. Setio Arnaud Arindra Adiyoso, Traverso Alberto, De Bel Thomas, Berens Moira SN, van den Bogaard Cas, Cerello Piergiorgio, Chen Hao, Dou Qi, Fantacci Maria Evelina, Geurts Bram, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis, 42:1–13, 2017. [PubMed: 28732268]

[41]. Tajbakhsh Nima, Gotway Michael B, and Liang Jianming. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 62–69. Springer, 2015.

[42]. Taleb Aiham, Loetzsch Winfried, Danz Noel, Severin Julius, Gaertner Thomas, Bergner Benjamin, and Lippert Christoph. 3d self-supervised methods for medical imaging. In Larochelle H, Ranzato M, Hadsell R, Balcan MF, and Lin H, editors, Advances in Neural Information Processing Systems, volume 33, pages 18158–18172. Curran Associates, Inc., 2020.

[43]. Tao Xing, Li Yuexiang, Zhou Wenhui, Ma Kai, and Zheng Yefeng. Revisiting rubik's cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation.

In Martel Anne L., Abolmaesumi Purang, Stoyanov Danail, Mateus Diana, Zuluaga Maria A., Zhou S. Kevin, Racoceanu Daniel, and Joskowicz Leo, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pages 238–248, Cham, 2020. Springer International Publishing.

[44]. Tian Yonglong, Sun Chen, Poole Ben, Krishnan Dilip, Schmid Cordelia, and Isola Phillip. What makes for good views for contrastive learning?, 2020.

[45]. Vincent Pascal, Larochelle Hugo, Bengio Yoshua, and Manzagol Pierre-Antoine. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, page 1096–1103. Association for Computing Machinery, 2008.

[46]. Wang Xiaosong, Peng Yifan, Lu Le, Lu Zhiyong, Bagheri Mohammadhadi, and Summers Ronald M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2097–2106, 2017.

[47]. Wang Xinlong, Zhang Rufeng, Shen Chunhua, Kong Tao, and Li Lei. Dense contrastive learning for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3024–3033, June 2021.

[48]. Wu Zhirong, Xiong Yuanjun, Yu Stella X., and Lin Dahua. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[49]. Wu Zhirong, Xiong Yuanjun, Yu Stella X, and Lin Dahua. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3733–3742, 2018.

[50]. Xie Enze, Ding Jian, Wang Wenhai, Zhan Xiaohang, Xu Hang, Sun Peize, Li Zhenguo, and Luo Ping. Detco: Unsupervised contrastive learning for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 8392–8401, October 2021.

[51]. Xie Zhenda, Lin Yutong, Zhang Zheng, Cao Yue, Lin Stephen, and Hu Han. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16684–16693, June 2021.

[52]. Ye Mang, Zhang Xu, Yuen Pong C., and Chang Shih-Fu. Unsupervised embedding learning via invariant and spreading instance feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[53]. Zbontar Jure, Jing Li, Misra Ishan, LeCun Yann, and Deny Stéphane. Barlow twins: Self-supervised learning via redundancy reduction. arXiv:2103.03230, 2021.

[54]. Zhan Xiaohang, Xie Jiahao, Liu Ziwei, Ong Yew-Soon, and Loy Chen Change. Online deep Computer Vision and Pattern Recognition (CVPR), 2020.

[55]. Zhou Hong-Yu, Lu Chixiang, Yang Sibei, Han Xiaoguang, and Yu Yizhou. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3499–3509, October 2021.

[56]. Zhou Hong-Yu, Yu Shuang, Bian Cheng, Hu Yifan, Ma Kai, and Zheng Yefeng. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In Martel Anne L., Abolmaesumi Purang, Stoyanov Danail, Mateus Diana, Zuluaga Maria A., Zhou S. Kevin, Racoceanu Daniel, and Joskowicz Leo, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pages 398–407, Cham, 2020. Springer International Publishing.

[57]. Zhou Zongwei, Sodha Vatsal, Pang Jiaxuan, Gotway Michael B., and Liang Jianming. Models genesis. Medical Image Analysis, 67:101840, 2021. [PubMed: 33188996]

[58]. Zhuang Xinrui, Li Yuexiang, Hu Yifan, Ma Kai, Yang Yujiu, and Zheng Yefeng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In Shen Dinggang, Liu Tianming, Peters Terry M., Staib Lawrence H., Essert Caroline, Zhou Sean, Yap Pew-Thian, and

Khan Ali, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pages 420–428, Cham, 2019. Springer International Publishing.

References

[1]. Haghighi Fatemeh, Taher Mohammad Reza Hosseinzadeh, Zhou Zongwei, Gotway Michael B., and Liang Jianming. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. IEEE Transactions on Medical Imaging, 40(10):2857–2868, 2021. [PubMed: 33617450]

[2]. Zhou Hong-Yu, Lu Chixiang, Yang Sibei, Han Xiaoguang, and Yu Yizhou. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3499–3509, October 2021.

[3]. Chen Xinlei, Fan Haoqi, Girshick Ross, and He Kaiming. Improved baselines with momentum contrastive learning, 2020.

[4]. Zbontar Jure, Jing Li, Misra Ishan, LeCun Yann, and Deny Stéphane. Barlow twins: Self-supervised learning via redundancy reduction. arXiv:2103.03230, 2021.

[5]. Chen Xinlei and He Kaiming. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15750–15758, June 2021.

[6]. Zhou Zongwei, Sodha Vatsal, Pang Jiaxuan, Gotway Michael B., and Liang Jianming. Models genesis. Medical Image Analysis, 67:101840, 2021. [PubMed: 33188996]

[7]. van den Oord Aaron, Li Yazhe, and Vinyals Oriol. Representation learning with contrastive predictive coding, 2019.

[8]. Chaitanya Krishna, Erdil Ertunc, Karani Neerav, and Konukoglu Ender. Contrastive learning of global and local features for medical image segmentation with limited annotations. In Advances in Neural Information Processing Systems, volume 33, pages 12546–12558. Curran Associates, Inc., 2020.

[9]. Chen Hao, Wang Yaohui, Lagadec Benoit, Dantcheva Antitza, and Bremond Francois. Joint generative and contrastive learning for unsupervised person reidentification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2004–2013, June 2021.

[10]. Taher Mohammad Reza Hosseinzadeh, Haghighi Fatemeh, Feng Ruibin, Gotway Michael B., and Liang Jianming. A systematic benchmarking analysis of transfer learning for medical image analysis. In Albarqouni Shadi, Cardoso M. Jorge, Dou Qi, Kamnitsas Konstantinos, Khanal Bishesh, Rekik Islem, Rieke Nicola, Sheet Debdoot, Tsaftaris Sotirios, Xu Daguang, and Xu Ziyue, editors, Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health, pages 3–13, Cham, 2021. Springer International Publishing.

[11]. Wang Xiaosong, Peng Yifan, Lu Le, Lu Zhiyong, Bagheri Mohammadhadi, and Summers Ronald M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2097–2106, 2017.

[12]. Choe Junsuk, Oh Seong Joon, Lee Seungho, Chun Sanghyuk, Akata Zeynep, and Shim Hyunjung. Evaluating weakly supervised object localization methods right. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[13]. Selvaraju Ramprasaath R., Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, and Batra Dhruv. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.
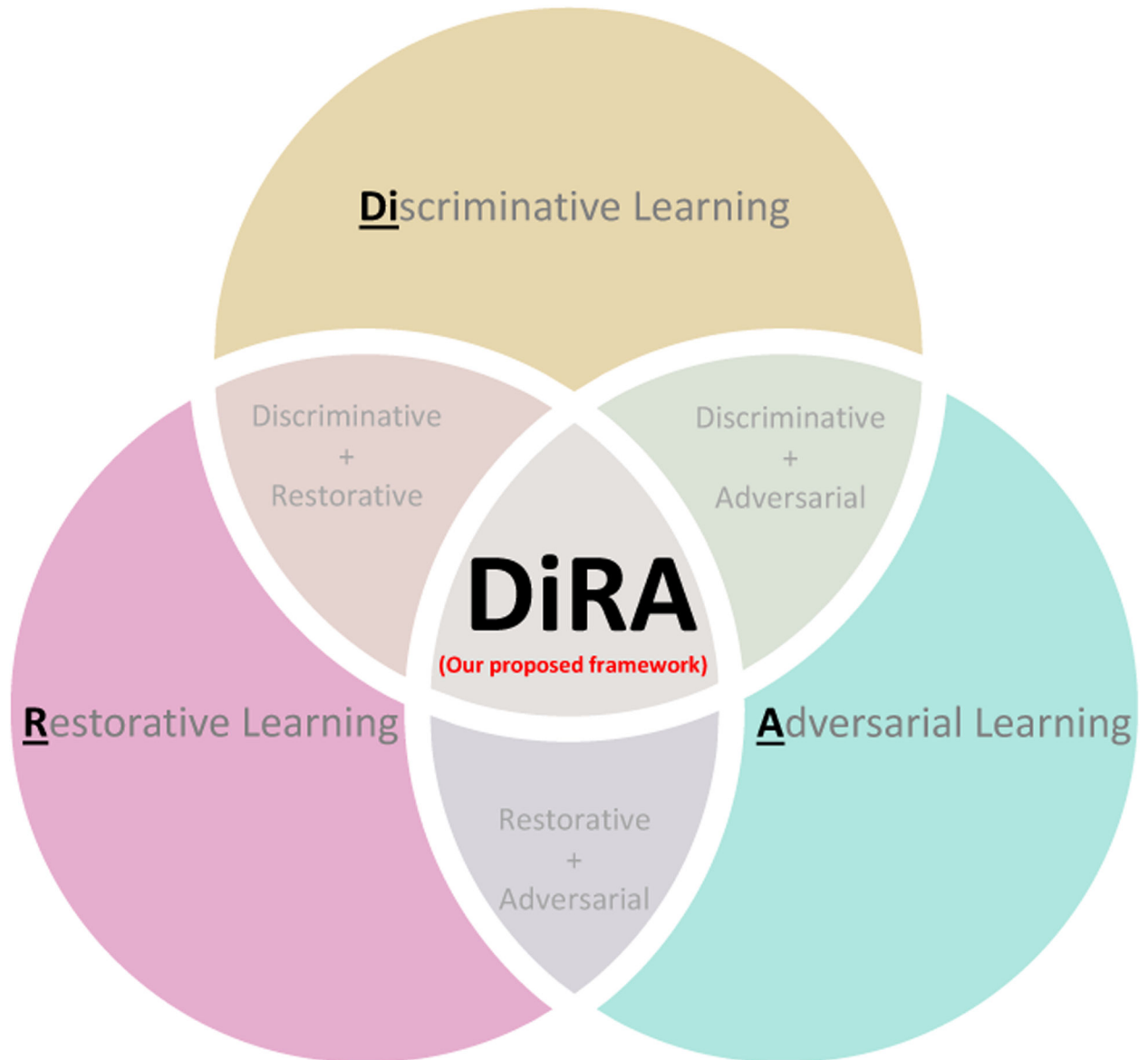
Figure 1.
Despite the critical contributions of discriminative, restorative, and adversarial learning to SSL performance, yet no SSL method simultaneously employs all three learning ingredients. Our proposed DiRA, a novel SSL framework, unites discriminative, restorative, and adversarial learning in a unified manner to collaboratively glean complementary visual information from unlabeled data for fine-grained semantic representation learning.
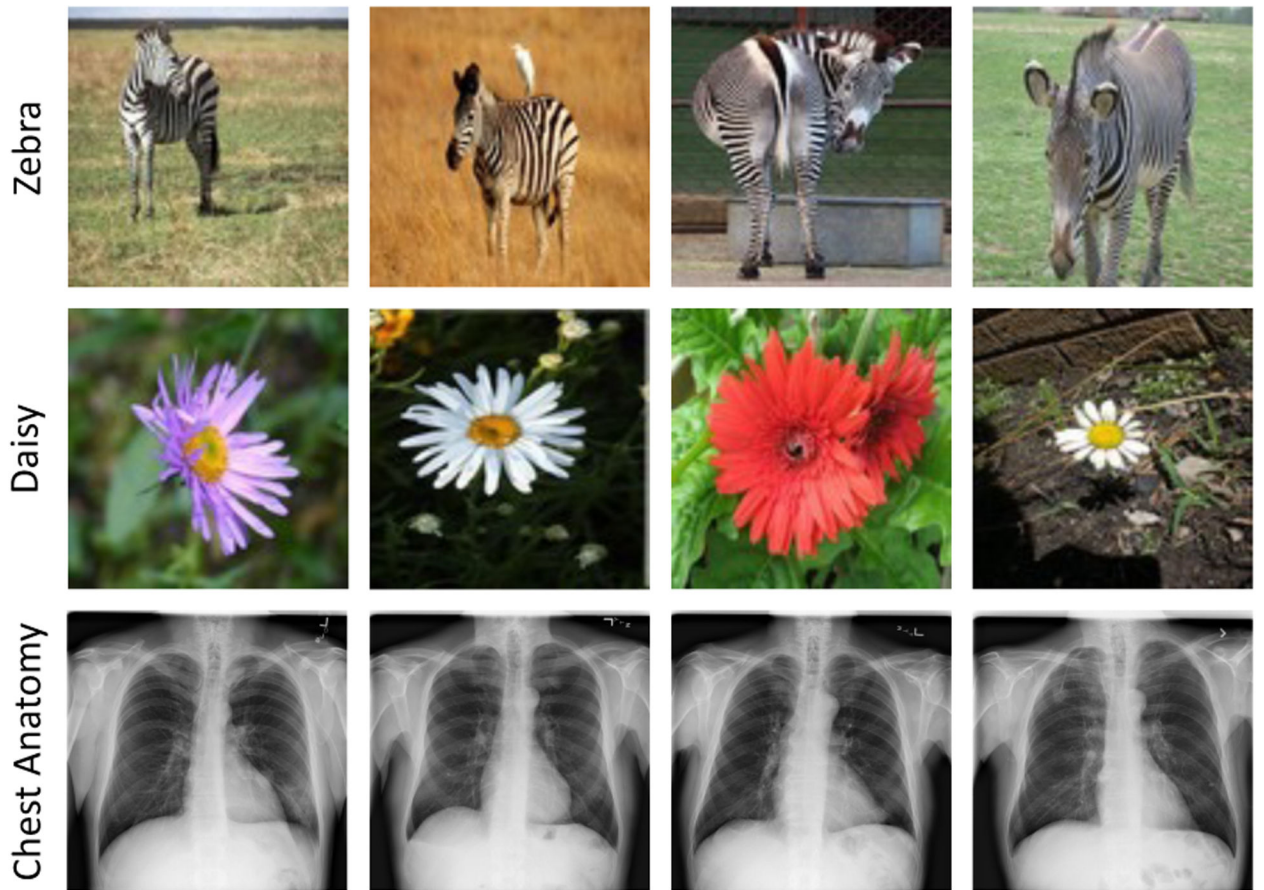
Figure 2.
Photographic images typically have large foreground objects with apparent discriminative parts, whereas medical images contain consistent anatomical structures with semantic information dispersed over the entire images. As a result, recognition tasks in photographic images are mainly based on high-level features, while medical tasks demand holistic fine-grained discriminative features captured throughout images.
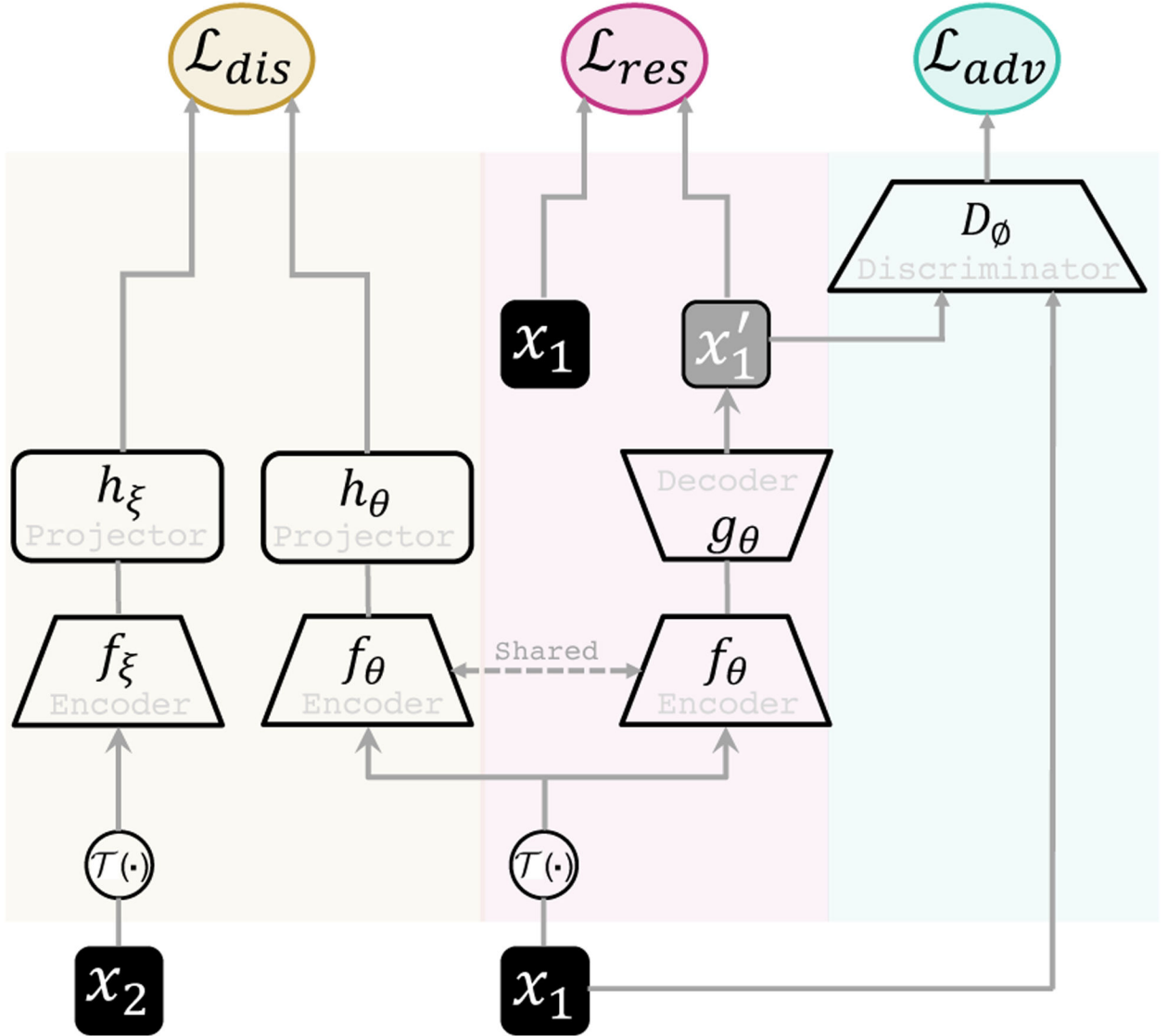
Figure 3. Our proposed framework.

DiRA consists of three learning components: discriminative, restorative, and adversary. Given two input patches x_1 and x_2, we perturb them with $\mathcal{T}(.)$ and provide them as input to discrimination and restoration branches. The discrimination branch consists of encoders f_θ and f_ξ, and projectors h_θ and h_ξ, and maximizes the agreement between (high-level) embedding vectors of samples from the same (pseudo) class. The restoration branch consists of encoder f_θ and decoder g_θ, and maximizes the (pixel-level) agreement between original sample x_1 and restored x'_1. Adversarial discriminator D_ϕ contrasts the original samples with the restored ones, reinforcing the restoration to preserve more fine-grained details.
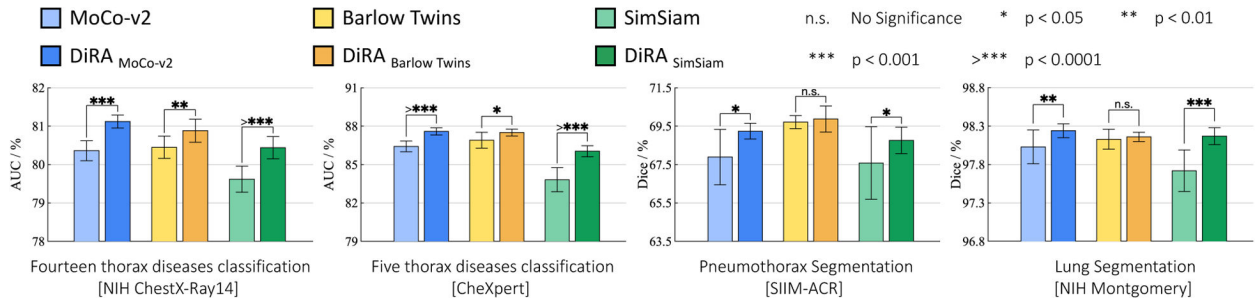
Figure 4. Comparison with discriminative self-supervised methods:
We apply our DiRA to three representative SOTA self-supervised methods with different discrimination objectives: MoCo-v2 [14], Barlow Twins [53], and SimSiam [15]. DiRA empowers discriminative methods to capture more fine-grained representations, yielding significant ($p < 0.05$) performance gains on four downstream tasks.
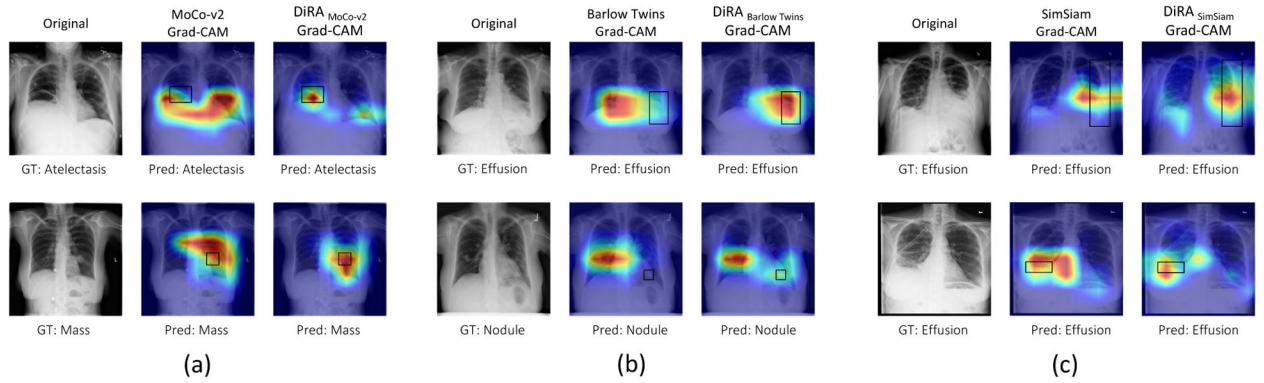
Figure 5. Visualization of Grad-CAM heatmaps

for (a) MoCo-v2 vs. DiRA$_{MoCo-v2}$, (b) Barlow Twins vs. DiRA$_{Barlow Twins}$, and (c) SimSiam vs. DiRA$_{SimSiam}$. Ground truth bounding box annotations are shown in black. Training with DiRA leads to improvements in weakly-supervised disease localization. While both DiRA and underlying models predict the correct disease label on the test images, DiRA models capture the diseased locations more precisely than the baselines which attune to larger regions of the image (e.g. (c), second row) or provide inaccurate localization with no overlap with the ground truth (e.g. (b), second row).

Table 1.

Transfer learning under different downstream label fractions:

DiRA models combat overfitting in low data regimes and provide stronger representations for downstream tasks with limited annotated data. For each downstream task, we report the average performance over multiple runs. (↑) shows the improvement of DiRA models compared with the underlying discriminative method.

| Method | ChestX-ray14 [AUC (%)] | | | CheXpert [AUC (%)] | | | Montgomery [Dice (%)] | | |
| | Label fraction | | | Label fraction | | | Label fraction | | |
	1%	25%	50%	1%	25%	50%	1%	25%	50%
MoCo-v2 [14]	52.99	74.89	76.71	76.87	81.70	83.23	63.69	96.44	97.60
DiRA$_{MoCo-v2}$	59.39 (↑ 6.4)	77.55 (↑ 2.6)	78.74 (↑ 2.0)	78.43 (↑ 1.5)	87.12 (↑ 5.4)	87.31 (↑ 4.0)	72.53 (↑ 8.8)	97.06 (↑ 0.62)	98.14 (↑ 0.5)
Barlow Twins [53]	62.43	76.23	77.59	82.85	83.74	84.66	86.79	97.49	97.68
DiRA$_{Barlow Twins}$	62.51 (↑ 0.08)	77.18 (↑ 0.9)	78.46 (↑ 0.8)	83.12 (↑ 0.2)	84.20 (↑ 0.4)	85.32 (↑ 0.6)	87.25 (↑ 0.4)	97.62 (↑ 0.1)	98.15 (↑ 0.4)
SimSiam [15]	51.07	73.05	75.20	65.39	80.05	81.46	48.20	94.86	97.21
DiRA$_{SimSiam}$	53.42 (↑ 2.3)	74.38 (↑ 1.3)	76.43 (↑ 1.2)	70.46 (↑ 5.0)	81.03 (↑ 1.0)	82.70 (↑ 1.2)	61.86 (↑ 13.6)	96.61 (↑ 1.7)	97.91 (↑ 0.7)

Table 2.

Comparison with fully-supervised transfer learning:

DiRA models outperform fully-supervised pre-trained models on ImageNet and ChestX-ray14 in three downstream tasks. The best methods are bolded while the second best are underlined. ↑ and ↑ present the statistically significant ($p < 0.05$) improvement compared with supervised ImageNet and ChestX-ray14 baselines, respectively, while * and * presents the statistically equivalent performances accordingly. For supervised ChestX-ray14 model, transfer learning to ChestX-ray14 is not applicable since pre-training and downstream tasks are the same, denoted by "—".

Method	Pretraining Dataset	Classification [AUC (%)]		Segmentation [Dice (%)]	
		ChestX-ray14	CheXpert	SIIM-ACR	Montgomery
Random	-	80.31±0.10	86.62±0.15	67.54±0.60	97.55±0.36
Supervised	ImageNet	**81.70±0.15**	87.17±0.22	67.93±1.45	98.19±0.13
Supervised	ChestX-ray14	-	87.40±0.26	68.92±0.98	98.16±0.05
DiRA$_{MoCo-v2}$	ChestX-ray14	81.12±0.17	**87.59±0.28**↑	69.24±0.41 *↑	**98.24±0.09**↑
DiRA$_{Barlow\ Twins}$	ChestX-ray14	80.88±0.30	87.50±0.27↑*	**69.87±0.68**↑↑	98.16±0.06 **
DiRA$_{SimSiam}$	ChestX-ray14	80.44±0.29	86.04±0.43	68.76±0.69 **	98.17±0.11 **

Table 3.
Comparison with restorative self-supervised method:

We apply our DiRA to the TransVW as the SOTA restorative self-supervised method. DiRA enhances TransVW by conserving more fine-grained details, resulting in performance boosts in four 3D downstream tasks.

Dataset	Method		
	Random	TransVW [26]	DiRA$_{\text{TransVW}}$
LUNA	94.25±5.07	98.46±0.30	**98.87±0.61**(↑ 0.41)
LIDC-IDRI	74.05±1.97	77.33±0.52	**77.51±1.36**(↑ 0.18)
LiTS	79.76±5.42	86.53±1.30	**86.85±0.81**(↑ 0.32)
BraTS	59.87±4.04	68.82±0.38	**69.57±1.13**(↑ 0.75)
PE-CAD	80.36±3.58	87.07±2.83	86.91±3.27

Table 4.

Ablation study on different components of DiRA:

We study the impact of each component of DiRA, including discrimination, restoration, and adversary, in four downstream tasks. Adding restorative learning ($\mathcal{L}_{res}$) to discriminative learning leads to consistent performance improvements. Furthermore, equipping models with adversarial learning ($\mathcal{L}_{adv}$) yields performance boosts across all tasks.

Base	Pretraining dataset	$\mathcal{L}_{dis}$	$\mathcal{L}_{res}$	$\mathcal{L}_{adv}$	Classification [AUC (%)]		Segmentation [Dice (%)]	
					ChestX-ray14	CheXpert	SIIM-ACR	Montgomery
MoCo-v2	ChestX-ray14	✓	×	×	80.36±0.26	86.42±0.42	67.89±1.14	98.03±0.22
		✓	✓	×	80.72±0.29 ↑	86.86±0.37 ↑	68.16±1.07 ↑	98.19±0.08 ↑
		✓	✓	✓	81.12±0.17 ↑	87.59±0.28 ↑	69.24±0.41 ↑	98.24±0.09 ↑
Barlow Twins	ChestX-ray14	✓	×	×	80.45±0.29	86.90±0.62	69.71±0.34	98.13±0.13
		✓	✓	×	80.86±0.16 ↑	87.44±0.33 ↑	69.83±0.29 ↑	98.15±0.14 ↑
		✓	✓	✓	80.88±0.30 ↑	87.50±0.27 ↑	69.87±0.68 ↑	98.16±0.06 ↑
SimSiam	ChestX-ray14	✓	×	×	79.62±0.34	83.82±0.94	67.58±1.89	97.72±0.27
		✓	✓	×	79.41±0.42 ↑	84.45±0.46 ↑	68.35±1.16 ↑	98.02±0.21 ↑
		✓	✓	✓	80.44±0.29 ↑	86.04±0.43 ↑	68.76±0.69 ↑	98.17±0.11 ↑