Object-Level Representation Learning for Natural and Medical Images

Akash Gokul

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-254
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-254.html

December 1, 2022

Object-Level Representation Learning for Natural and Medical Images

by

Akash Gokul

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science, Plan II

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Trevor Darrell, Advisor
Professor Joseph Gonzalez, Second Reader

Spring 2022

The thesis of Akash Gokul, titled Object-Level Representation Learning for Natural and Medical Images, is approved:

Advisor  Date 05 / 15 / 2022

Joseph E. Gonzalez Date 05 / 16 / 2022

University of California, Berkeley

Object-Level Representation Learning for Natural and Medical Images

Abstract

Object-Level Representation Learning for Natural and Medical Images

by

Akash Gokul

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Trevor Darrell, Advisor

Instance discrimination pretraining has become an effective means of learning transferable visual representations. To date, this paradigm has focused on learning image-level representations. This is not only suboptimal for downstream tasks such as object detection, but can also lead to representations which do not capture the object(s) in the scene [61]. In this thesis, we present two extensions of the instance discrimination paradigm to the object-level. First, we present a method which finds objects in a scene and enforces representational invariance at the object-level (Chapter 2). Next, we apply object-level knowledge to medical images by incorporating anatomical priors into the pretraining pipeline (Chapter 3). These methods provide improvements in downstream performance, efficiency, and interpretability when compared to state-of-the-art instance discrimination pretraining. We conclude with a broader analysis of object-level representation learning and instance discrimination pretraining (Chapter 4).

To my family, Aparna, Gokul, and Akshay.
Thank you.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Dr. Sayna Ebhrahimi, Colorado Reed, and Professor Trevor Darrell for their guidance during the past two years. I would also like to thank my past mentors and advisors: Dr. Oladapo Afolabi, Dr. Allen Yang, Lucas Spangher, Professor Costas Spanos, Dr. Benjamin Caulfield, and Professor Sanjit Seshia. To all my mentors and advisors, thank you for introducing me to the wonderful world of research.

Chapter 1

Introduction

Object perception and recognition is a fundamental part of visual scene understanding. Humans, even as young as infants, possess a remarkable ability to perceive and represent objects solely from visual inputs. Research has shown that infants learn to perceive objects by matching features across occlusions [51, 39, 40, 41]. This framework, of learning invariant representations across views, is similar to the instance discrimination paradigm which has led to state-of-the-art results in self-supervised learning. However, the instance discrimination paradigm treats self-supervised learning similar to supervised image classification, learning a single global representation for an image. As a result, instance discrimination based methods fail to learn object-level representations and instead learn features by matching backgrounds between views [61]. In this thesis, we extend the instance discrimination paradigm to learn object-level representations.

Chapter 2 presents a pretraining method which discovers objects and enforces representational invariance at the object-level. This algorithm is a simple extension of siamese representation learning architectures and does not introduce any new architectural components. Moreover, this method leads to state-of-the-results in downstream performance and efficiency.

In chapter 3, we apply object-level knowledge to medical images. Specifically, we incorporate spatial and/or anatomical knowledge into the self-supervised pretraining pipeline. Models trained with our method are able to discover semantically meaningful regions, e.g. the heart in a chest X-ray, without supervision. Our method also leads to improvements in chest X-ray classification and interpretability.

Finally, chapter 4 provides an analysis of object-level representation learning and the instance discrimination paradigm. We conclude with proposals for future research.

Chapter 2

Refine and Represent: Object-Level Representation Learning via Mask Refinement

The rise of self-supervised learning in computer vision has centered around image-level pretraining on ImageNet [59]. Current methods in the field [32, 31, 27, 8, 9] have demonstrated remarkable downstream transfer performance with state-of-the-art methods [69] outperforming supervised ImageNet baselines. These pretraining methods primarily use the instance discrimination paradigm, training a network to have image-level representations which are invariant to data augmentations.

Image-level representation learning is only the beginning. Tasks such as object detection and instance segmentation require learning object-level features in order to make pixel-level predictions. To achieve the goals of self-supervised learning, namely to learn representations which can effectively transfer to downstream tasks, pretraining methods must also learn object-level feature representations. While image-level pretraining on an object-centric dataset such as ImageNet should lead to object-level features, recent work [61] has shown that state-of-the-art models seldom focus on the object(s) in the scene. This has led to the area of dense self-supervised learning [74]. Methods in this subfield [74, 76, 34, 79, 82, 55] focus on jointly learning local feature representation by extending the image-level representation objective to the per-pixel or region level.

Learning object-level representations requires more than just enforcing invariance to local feature encodings. Objects are more than just pixels or image regions which overlap between two views. Their semantic value allows us to discriminate a contiguous figure even amidst visual clutter. Thus, we propose an object-level pretraining method which jointly discovers objects and learns to represent them given a simple mask prior. Our method does not rely on performant unsupervised segmentation or detection algorithms [33, 72, 89, 76] Instead, we generate a mask by performing pixel-level clustering on the original image (see Figure 2.2). From there, we refine these masks into object-level masks and enforce representation invariance over object-level representations during pretraining (Section 2.2).

Figure 2.1: An overview of the objects discovered by our method. Using a simple prior mask (middle), we are able to refine this mask and perceive objects (right) in an image (see Section 2.2).

Our method extends the siamese representation learning paradigm, allowing us to bypass the need for negative samples during training. Moreover, our method does not introduce any new architectural components. Our method leads to state-of-the-art results in terms of downstream performance (+0.47 on PASCAL semantic segmentation and +0.3 on COCO instance segmentation) and efficiency, surpassing existing methods which pretrain which longer schedules.

2.1 Related Works

Self-supervised learning algorithms aim to learn representations, from unlabelled data, which can effectively transfer to a downstream task [6]. Early works in discriminative representation learning for visual tasks focused on training encoders to predict the position of an image patch [17, 54], predict the angle of rotation [25], and predict the colors in an grayscale image [87]. However, current state-of-the-art methods in this area focus less on explicit prediction and, instead, train an encoder to have representations which are invariant to data augmentations [32, 11, 27, 12, 8, 52]. Instance discrimination pretraining is currently achieved via one of two means: (1) contrastive learning [56] and (2) siamese representation learning. The contrastive objective encourages representations from the same image to be invariant to data augmentations and dissimilar to representations from other images (commonly referred to as negatives). Unlike contrastive objectives, siamese representational learning methods do not rely on negatives and are thus less reliant on batch size [12]. These methods enforce

representational invariance, similar to contrastive learning, but use an asymmetric network architecture to prevent representational collapse. In this work, we build upon BYOL [27] and develop a simple object-level pretraining method which does not rely on negatives.

Recent works [74, 76, 34, 79, 82, 55] have extended the instance discrimination paradigm to concurrently learn local features. These methods are especially effective for downstream tasks, such as object detection and semantic segmentation, which involve pixel-level predictions. Concurrent to this line of work are methods [33, 76, 72, 44, 84, 89] which use object-level priors [21, 70] within the instance discrimination paradigm. The closest related work is Detcon [33] which uses unsupservised segmentation algorithms [21] to train object-level representations. In contrast, our method avoids the need for negatives and allows the encoder to refine these prior masks to better locate objects during training. The goal of improving local feature representations has also led to works [77, 88, 24] which define new data augmentation schemes. These augmentation policies copy-and-paste objects from one image to another to improve object localization within self-supervised and semi-supervised settings.

Our work uses K-means clustering [46] to refine weak prior masks into object-level masks during pretraining. While we are the first work to use K-means clustering in this manner, clustering has been a popular tool in the self-supervised learning toolbox. Clustering has been used as a form of pretraining [7, 9] which enforces consistent cluster assignments similar to the representational invariance seen in the contrastive and siamese paradigms. Recently, K-means clustering has also been used to create semantic segmentation masks from pretrained self-supervised encoders [86]. K-means clustering has also been used in recent unsupervised semantic segmentation methods [38, 35].

2.2 Method

Our method consists of two parts: (1) object segmentation and (2) object-level pretraining. This method can be used in any siamese instance discrmination architectures. For simplicity, we define our method using the BYOL architecture. An overview of our method can be seen in Figure 2.2.

Background

Data Given an image x, we create two views $x_1, x_2 \in \mathbb{R}^{CxHxW}$ using BYOL's data augmentation policy. Additionally, we create $x_0 \in R^{CxHxW}$ which is the original image x resized to the resolution of x_1 and x_2. Finally, we generate a prior mask m_{slic} by clustering x_0 using SLIC [1]. We encode all masks as a multi-channel binary array.

Architecture The BYOL architecture defines two networks– an online network (denoted by parameters θ) and the target network (denoted by parameters ξ). The target network shares the same architecture as the online network and uses an exponential moving average

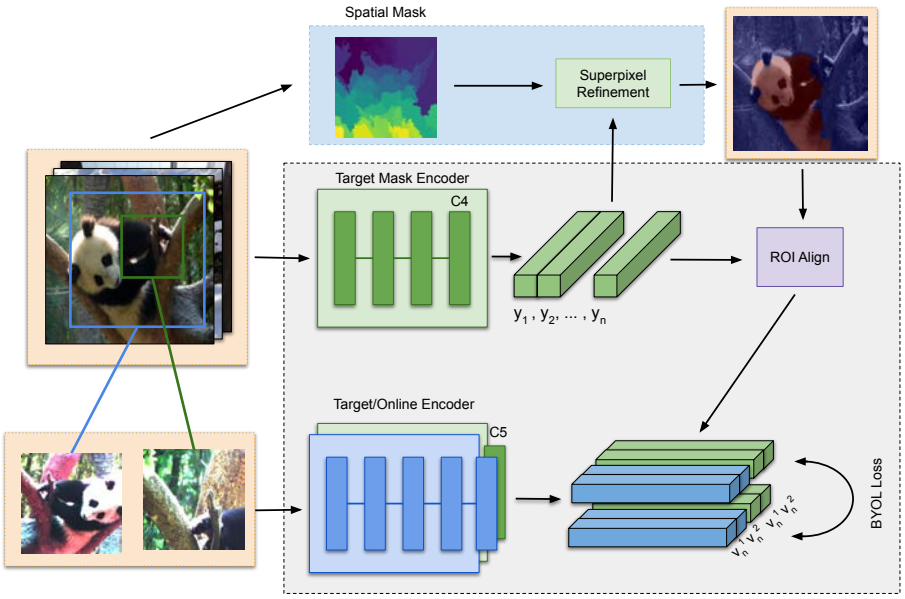

Figure 2.2: An overview of our method. First, we pass the original image (resized to (224, 224)) through our target encoder and cluster the masked-pool encodings. This step generates an object mask. Next, we use this object mask to enforce invariance between object-level representations during pretraining.

of the online network's parameters. The target network is never trained. Going forward, we refer to the encoder of the online network as f_θ and the online network's projector as g_θ. Correspondingly, the target network uses the encoder f_ξ and the projector g_ξ. The online network, unlike the target network, also uses a prediction head denoted q_θ.

Object Segmentation via Mask Refinement

In order to learn object-level representations, we must first find the objects in a scene. Our object localization step uses mask-pooling [33] (equation 2.1) and then performs K-means over the set of mask-pooled encodings. First, we get feature encodings $h = f_\xi(x_0) \in \mathbb{R}^{H'xW'xD}$. Next, we downsample the mask m_{slic} to resolution (H', W') and apply mask-pooling for each of the M channels in the downsampled m_{slic} (equation 2.1). This results in M encodings h_i representing the target encodings for each part of the image. Finally, we perform K-means clustering over the mask-pooled vectors h_m to get a final superpixel mask m_0. The K-means step allows us to refine weak mask priors, such as spatial masks or simple spatial and color based clustering, into object-level masks (see Figure 2.2).

$$\text{maskpool}(h, m) = \frac{1}{\sum_{i,j} m[i][j]} \sum_{i,j} m[i][j] \cdot h[i][j] \tag{2.1}$$

Object-Level Representation Learning

To learn object-level representations, we enforce representational invariance at the object-level. Thus, we replace the commonly used global-pooling operation with the mask-pooling operation. Given mask m_0, we use ROIAlign [30] to create m_1, m_2 corresponding to views x_1, x_2. Next, we compute view encodings $h_\theta = f_\theta(x_1)$ and $h_\xi = f_\xi(x_2)$. We use mask-pooling to create the vectors $h_{i,\theta}$ and $h_{i,\xi}$. We compute the loss as the average BYOL loss (equation 2.4) over the M' mask channels which are present in both views (equation 2.5). We symmetrize this loss, following BYOL, by also computing the loss for $h_\theta = f_\theta(x_2)$ and $h_\xi = f_\xi(x_1)$.

$$h_{i,\theta} = \text{maskpool}(h_\theta, m) \qquad h_{i,\xi} = \text{maskpool}(h_\xi, m) \tag{2.2}$$

$$z_{i,\theta} = g_\theta(h_{i,\theta}) \qquad z_{i,\xi} = g_\xi(h_{i,\xi}) \tag{2.3}$$

$$L_{\text{BYOL}}(z_\theta, z_\xi) = 2 - 2 \cdot \frac{q_\theta(z_\theta) \cdot z_\xi}{\|q_\theta(z_\theta\|_2 \cdot \|z_\xi\|_2} \tag{2.4}$$

$$L(z_\theta, z_\xi; m) = \frac{1}{M'} \sum_{i=1}^{M'} L_{\text{BYOL}}(z_{i,\theta}, z_{i,\xi}) \tag{2.5}$$

2.3 Results

Pretraining Settings

Data Following the pretraining protocol of related works [32, 27, 7, 33, 74], we pretrain on ImageNet. We follow the data augmentation policy of BYOL to generate both views. For our ablation studies, we pretrain our networks on ImageNet100 [68]. We use [71] to generate SLIC masks.

Architecture We use a ResNet-50 [29] architecture for all encoders, similar to BYOL. Our object segmentation step uses the C4 output of the ResNet-50. In our experiments, we cluster the encodings of the target network to generate object masks. We set $K = 64$ when performing K-means. We perform K-means over the entire mini-batch on each GPU. Our object-level representation learning step applies mask-pooling to the C5 output of the ResNet-50, similar to the existing application of global pooling.

Method	Epochs	AP^{bb}	AP^{mk}
Supervised [76]	90	39.6	35.6
MoCo [32]	200	38.5	35.1
MoCo v2 [13]	200	40.4	36.4
BYOL (reproduced)	300	40.6	37.5
VADeR [55]	600	39.2	35.6
DenseCL [74]	200	40.3	36.4
PixPro [82]	400	**41.4**	-
InfoMin [68]	200	40.6	36.7
DetCon_B (reproduced)	300	40.5	37.5
Ours	100	40.4	37.3
Ours	300	40.9	**37.8**

Table 2.1: Performance on COCO object detection and instance segmentation using Mask R-CNN (R50-FPN) following the 1x schedule.

Optimization We follow the optimization details of BYOL. All of our experiments use a batch size of 4096 distributed over 128 NVIDIA V100 GPUs.

Downstream Tasks

We evaluate the efficacy of our pretraining method by evaluating the transfer performance on object detection and instance segmentation on MS COCO [45], and semantic segmentation on PASCAL VOC [20] and Cityscapes [15].

Object Detection and Instance Segmentation We finetune our encoder as the backbone of a Mask-RCNN (R50-FPN) [30] using [78]. We follow the 1x training schedule (12 epochs), training on MS COCO's train2017 dataset and evaluating on the val2017 dataset. We report average precision for bounding box predictions (AP_{bb}) and mask predictions (AP_{mk}).

Table 2.1 details our performance on COCO object detection and instance segmentation. Our method outperforms existing baselines in instance segmentation (+0.3) and is competitive with state-of-the-art in object detection (−0.5). Moreover, our performance after only 100 epochs of pretraining is competitive with existing methods which use double (or even 6x in the case of [55]) epochs of pretraining. Lastly, our method uses the original BYOL hyperparameters and improves BYOL on COCO by +0.3 on object detection and +0.3 on instance segmentation.

Semantic Segmentation We evaluate our semantic segmentation performance by finetuning the encoder as the backbone of a FCN [47] using [14]. For PASCAL VOC, we finetune using the train_aug2012 dataset for 45 epochs. For Cityscapes, we finetune using the train_fine

Method	Epochs	PASCAL VOC	Cityscapes
Supervised [74]	200	67.7	73.7
MoCo v2 [13]	200	67.5	74.5
BYOL (reproduced)	300	73.1	75.2
DenseCL [74]	200	69.4	75.7
PixPro [82]	400	-	**77.2**
InfoMin [68]	200	-	75.6
DetCon_B (reproduced)	100	70.7	73.1
DetCon_B (reproduced)	300	73.8	75.7
Ours	100	72.2	73.9
Ours	300	**74.3**	75.6

Table 2.2: Performance on PASCAL VOC and Cityscapes semantic segmentation (mIoU).

dataset for 160 epochs. Our evaluation pipeline follows BYOL and Detcon. Performance is measured by mean intersection over union (mIoU) on val2012 and val_fine respectively.

As seen in Table 2.2, when pretraining for less than 200 epochs we are able to outperform similar methods (+1.47 on PASCAL and +0.83 on Cityscapes compared to Detcon). After 300 epochs of pretraining, we outperform state-of-the-art on PASCAL VOC (+0.47) while being competitive with existing baselines on Cityscapes. Our method improves BYOL by +1.19 on PASCAL VOC and +0.33 on Cityscapes while using the same hyperparameters.

Ablations

Through our experiments, we used a fixed $K = 64$ when generating object masks. Table 2.3 details the effect of varying K during pretraining. We have found that a larger value of K helps downstream PASCAL VOC segmentation performance. This is surprising as pretraining occurs on ImageNet, an object-centric dataset. Thus, simple masks, i.e. lower values of K, should be able to effectively segment the scene. This leads us to believe that *oversegmentation helps dense or object-level pretraining*. Oversegmented object masks enforces representational invariance at a finer granularity, e.g. each pixel in the $(7, 7, 2048)$ output of the ResNet50. This leads to stronger local features. Thus, allowing the model to demonstrated improved transfer on downstream tasks such as semantic segmentation.

2.4 Conclusion

Here, we have presented a simple means of extending the siamese instance discrimination paradigm to the object-level. Our method bypasses the need for negative samples and can be used in existing architectures with minimal changes. Moreover, our method can actively discover objects by refining a weak prior mask. Overall, this extension leads to the

K	PASCAL VOC (mIoU)
4	65.49
16	65.27
64 (default)	**65.76**

Table 2.3: The impact of varying K, the number of segments generated during our clustering step, on PASCAL VOC semantic segmentation performance (mIoU). Encoders were pretrained on ImageNet100.

state-of-the-art results in terms of downstream performance (+0.47 PASCAL VOC semantic segmentation and +0.3 COCO instance segmentation) and efficiency.

One shortcoming of our method is the use of prior masks. Thus, removing the use of a prior mask in our object segmentation step is a promising area of future work. Moreover, we used K-means across the mini-batch to work to minimize computational overhead. However, this means that object masks are semantically consistent across samples in the mini-batch. This allows for extensions of [19] at the object-level.

2.5 Acknowledgements

This work was done alongside Konstantinos Kallidromitis (Panasonic AI Lab), Shufan Li, Yusuke Kato (Panasonic AI Lab), and Colorado Reed.

Chapter 3

Knowledge-Guided Self-Supervised Vision Transformers for Medical Imaging

The previous chapter and prior work in object-level unsupervised representation learning (2.1) focused on learning objects while pretraining on ImageNet. However, learning semantically meaningful entities is not restricted to natural images. In this chapter, we extend the object-level goals of chapter 2 and present a self-supervised pretraining method which guides the attention module of Vision Transformers (ViT) [18] to learn semantically and spatially meaningful regions within medical images.

Medical data are difficult to acquire and, more importantly, expensive to annotate because of the need for expert knowledge [42]. Thus, making self-supervised pretraining an important part of the medical imaging toolbox. Existing work [66, 2, 23, 3] in the area of self-supervised learning from medical images applies the pretraining methods benchmarked on ImageNet to medical images. However, treating the problem of learning useful representations for medical images as merely another application for Imagenet-based pretraining methods is flawed. Unlike ImageNet, the structures captured in medical images are consistent across images and well-understood after years of advances in fields such as anatomy. Furthermore, understanding such structures is crucial to medical imaging-based tasks and this knowledge may not be correctly learned when following ImageNet based pretraining protocols.

Here, we present a framework called Medical DINO (*MeDINO*) that is built upon the DINO self-supervised vision transformer framework [8] and leads to more interpretable attention heads that perform better on downstream tasks – see Figure 3.1. *MeDINO* incorporates prior knowledge into self-supervised training of vision transformers by regularizing a subset of attention heads in the multi-headed self-attention module so that the attention weights are constrained to be within boundaries corresponding to objects of interest. In other words, a subset of the attention heads are regularized to the objects in the image throughout training. Instead of relying on annotations of these images to determine the boundaries, *MeDINO* uses anatomical relationships to define a template of object boundaries (organs), aligns and

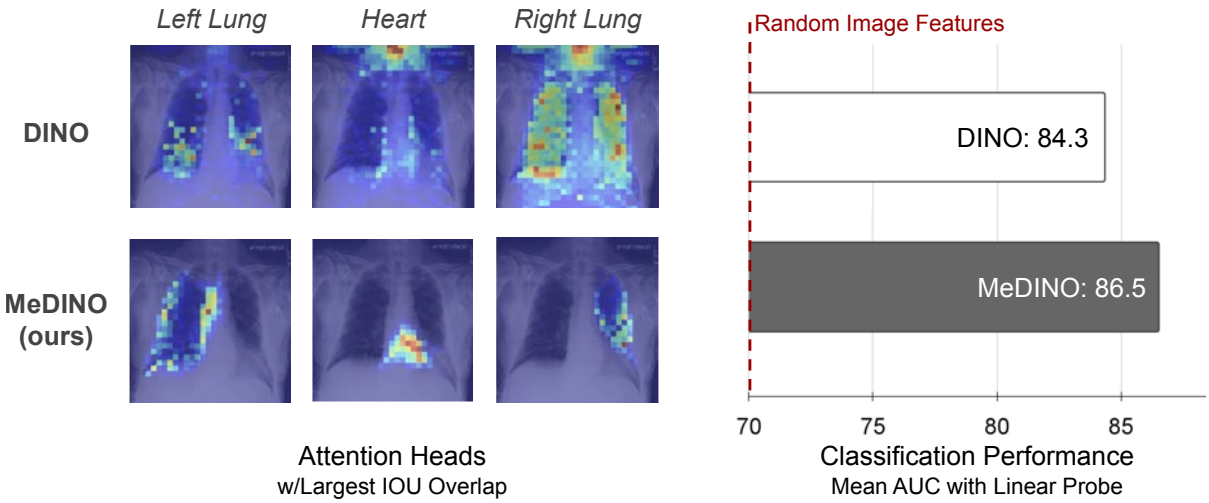

Figure 3.1: **Self-attention from a Vision Transformer on chest X-rays, where the attentions heads with the largest IOU overlap with the lungs/heart are shown.** Existing self-supervised training methods for Vision Transformers, such as DINO, learn scattered attention maps that do not necessarily attend to the constituent objects within the image. *MeDINO*, on the other hand, uses prior knowledge to guide the attention to such regions, as shown by the attention weights constrained to the left lung, heart, and right lung. As indicated in the bar plot on the right, constraining the attention to these semantic components leads to better performing representations as determined by a linear probe, multi-label classification experiment on the CheXpert dataset [36] – see Section 3.3 for details.

registers each image to this template, transforms the object boundaries accordingly, and then uses these object boundaries to regularize the attention heads throughout training – Figure 3.2 provides an overview of this process which is detailed in Section 3.2.

3.1 Related Works

Self-Supervised Learning The performance of machine learning models is heavily contingent on the choice of features and representations from which they learn. Representation learning aims to reveal these intrinsic qualities of data such that they are informative and effective for a desired task [6], such as image classification or object detection. Contemporary methods involve contrastive learning based approaches [32, 11], clustering-based techniques [9], and self-distillation [27, 8]. DINO [8] is an example of a self-supervised learning framework that uses self-distillation, yielding state-of-the-art downstream performance using the Vision Transformer (ViT) architecture [18]. We focus on this framework for two reasons: (i)

the attention modules in ViT allow for greater interpretability than CNN-based approaches that require external tools such as Grad-CAM [73] to extract pixel-level saliency relationships, (ii) in self-supervised training, the DINO attention maps have a demonstrated ability to segment salient foreground objects [8], which, as we show, provide a strong mechanism to regularize salient objects in *MeDINO*.

Most existing self-supervised algorithms are benchmarked based on their performance following pretraining on generic, object-centric image datasets, such as ImageNet, which potentially leads to poor results in domains where the data are dissimilar to these datasets [80, 58]. Furthermore, medical applications of machine learning can benefit from self-supervised pretraining due to the cost and expertise needed to accurately annotate data [66, 23, 3, 2]. Medical data may also be difficult to obtain due to privacy and regulatory issues. For instance, MoCo-CXR [66] adapts MoCo [32] pretraining to chest X-ray data by designing new data augmentations suitable for recognizing subtle differences between X-ray images. We use MoCo-CXR's data augmentations as it uses similar X-ray images as our work. IDEAL [49] focuses on self-supervision and interpretability. However, they use saliency reconstruction to find informative samples for active learning and do not focus purely on learning discriminative and interpretable representations. Finally, many medical applications have adopted a ViT-based self-supervised learning approach for their performance and attention modules with modifications to the attention modules or encoders [60, 75, 67, 26]. *MeDINO* does not require any architectural changes to the backbone and offer a noninvasive method, as we leverage the attention heads that are inherent to Vision Transformers.

Attention-Guided Learning Attention based approaches have been used to improve the explainability of computer vision models through visualizations of attention maps to indicate important regions [83]. Convolutional neural networks use tools such as CAM [90] and GradCAM [73] to create attention maps by looking at the hidden layer activations. Another approach is the Attention Branch Network [22] that generates an attention map based on the extracted features and then uses it to mask out irrelevant features. These attention maps are evaluated through visual checks or against segmentation datasets which are limited in the medical domain. As discussed in Section 3.2, our paper instead uses image registration to align the attention maps with an inductive bias corresponding to a salient region so only a single representative sample is required.

Image Registration is the task of projecting one image onto the coordinate system of another image with similar content [28]. This classical challenge is particularly relevant in medicine as it facilitates the development of atlases, and allows transfer of information across patients. Here, we are particularly interested in using a specific type of image registration, deformable image registration. This is useful as differences in morphological structure of organs in different humans can be modeled using deformable transformations [65]. Traditionally popular techniques to solve this alignment problem involved congealing, optical flow or b-Spline registration. Least squares congealing focuses on obtaining an alignment by iteratively minimizing a misalignment loss function using least squares [16]. Optical flow completes this registration task by looking for possible displacements and solving a minimum energy functional [43]. b-Spline registration operates by modeling the deformation field as

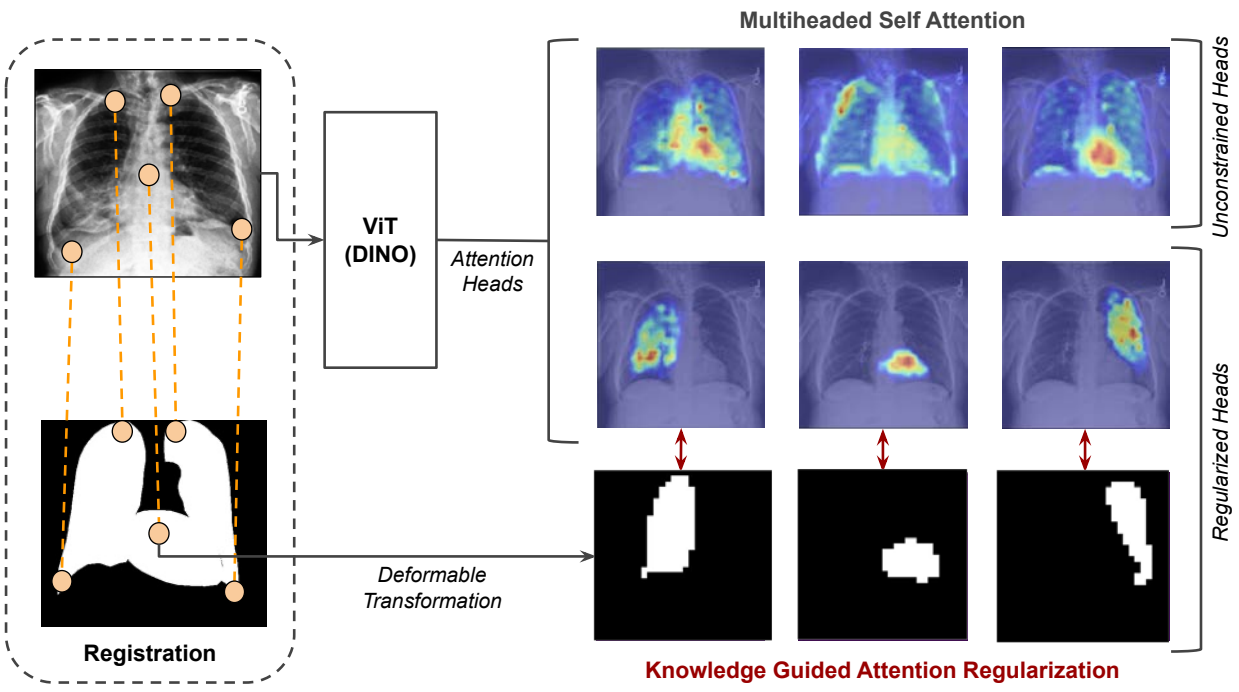

Figure 3.2: **The *MeDINO* framework**. *MeDINO* first registers each image to an exemplar template with known segmentations, the registration outputs a deformable transformation that is applied to the template. During self-supervised pretraining with a ViT model, each component of the template then regularizes an indiviudal attention head in the multiheaded self-attention modules (Regularized Heads). A subset of the attention heads are also unconstrained (Unconstrained Heads).

B-spline curves where each pixel that maps each voxel in the source image to the target image. [62]. More recently, neural networks have become increasingly popular in performing image registration in lieu of the traditional methods [4, 53, 64]. While powerful and increasingly accurate, we opted to use the traditional b-Spline registration over neural-based methods due to simplicity and efficiency. b-Spline registration does not require more than one data example nor does it necessitate any training or GPU compute resources. Mansilla et. al. [50] embeds prior knowledge in the form of anatomical constraints to improve image registration tasks in the form of global constraints. Our work differs as we aim to improve self-supervised pretraining methods using deformable transformation as a means to create anatomically plausible representations rather than an end.

3.2 Method

The goal of *MeDINO* is to incorporate domain knowledge to improve the performance and interpretability of self-supervised pretraining for medical images. To do so, *MeDINO* regularizes transformer attention heads to follow inductive biases on a semantic structure that is common to most images in the dataset, as seen in Figure 3.2. For example, we can incorporate the inductive bias that chest radiographs have expected anatomical relationships between the relative positions of the lungs and heart. In the following, we detail a means of effectively incorporating semantic knowledge, in the form of simple spatial heuristics or even a single instance of ground truth knowledge, into the DINO pretraining of Vision Transformers.

Self-Supervised Vision Transformers with Knowledge Distillation

Caron et al. [8] present a transformer-based knowledge distillation technique, DINO, that we build upon for *MeDINO*. In DINO, a student model g_{θ_s} is trained to match the output of a teacher model g_{θ_t} (parameterized by θ_s and θ_t respectively). This distillation objective is reframed as a representation learning objective where representations are learned for each of n different views of original image X, $\{X_1, ..., X_n\}$, obtained via a set of data augmentations V. The DINO objective encourages the student model to learn "local-to-global" correspondences. This happens by passing in local and global crops of an image to the student and tasking the student model to predict the teacher's representation. The teacher is only given global crops denoted X_1^g and X_2^g. To train the student network, the authors begin by defining probability distributions P_m for the student and teacher model

$$P_m(X) = softmax\left(\frac{g_{\theta_m}(X)}{\tau_m}\right)$$

where τ_m is the model-specific temperature. The overall DINO objective, given below, is the cross-entropy loss $H(p, q) = -p \log q$ over the probability distributions $P_s(X)$ and $P_t(X)$.

$$L(X_1, X_2) = H(P_t(X_1), P_s(X_2))$$

$$_{\theta_s} \sum_{x \in \{X_1^g, X_2^g\}} \sum_{X' \in V} L(X, X')$$

While the original DINO augmentations can be powerful for learning representations from a dataset such as ImageNet, they can fail in domains where local structure is critical to scene understanding. [81] In particular, our preliminary empirical findings (see 4.3 and 4.4) suggest that local crops harm the performance of self-supervised learning on chest radiographs. Following this, we instead use a set of domain-specific augmentations which replace DINO's local crops with other task-relevant data augmentations [66].

Attention

Vision Transformers are perform well on a wide variety of vision tasks and allow for pixel-level relationship introspection due to their built-in attention modules [18]. As input, Vision Transformers take in a sequence of P image patches with fixed size ($p = 16$) which is prepended by a $[CLS]$-token. The $[CLS]$ token enables a corresponding output that allows for downstream tasks such as classification.

Self-attention modules are the key component to Transformer networks. Given embeddings q, k, v calculated from a sequence of inputs, the attention matrix A measures the pairwise similarity between q_i, query value of patch i, in relationship with k_j, key value of patch j. Formally,

$$A = softmax\left(\frac{qk^\intercal}{\sqrt{D_h}}\right)$$

where D_h is defined as the dimensionality of the heads and $A \in \mathbb{R}^{P \times P}$. When probing self-attention, we extract the attention values of each patch with respect to the $[CLS]$ token of the last layer of each of n_h heads and exclude the attention value for the $[CLS]$ token with itself. This tensor is then upsampled via nearest-neighbor interpolation into the shape of the original image resulting in an attention map $A_s \in \mathbb{R}^{w \times h \times n_h}$ where w and h are the dimensions of X.

Knowledge-Guided Regularization

The Vision Transformer's attention module allows us to guide a model given any arbitrary knowledge map K by back-propagating through the model. If $K \in \{0, 1\}^{w \times h}$, where K_{ij} is 1 if the patch at location i, j is considered a useful bias and 0 otherwise, the central idea is to add a penalty when a model's self-attention map $A_s^{(\theta_t)}$ attends outside salient regions and a negative penalty for attending at the salient regions. This yields the following regularization terms that are combined with the DINO objective:

- **Inclusion Loss**: $L_{inclusion}(A, K) = \lambda_1 \sum_i \sum_j a_{ij}^{(\theta_t)} k_{ij}$
- **Exclusion Loss**: $L_{exclusion}(A, K) = -\lambda_2 \sum_i \sum_j a_{ij}^{(\theta_t)} (1 - k_{ij})$

where hyperparameters λ_1, λ_2 denote the respective regularization strengths. Experimentally, we find that both regularizers are needed in order to prevent mode collapse in the attention maps.

Disentanglement

As vision transformers have multiple independent attention heads, we can disentangle them to attend to different discrete entities in X using different knowledge maps $K \in \{0, 1\}^{w \times h \times n}$. This disentanglement allows for task-specificity and functions as a scaffold for interpretability through which failure cases can be deconstructed into explainable task specific entities.

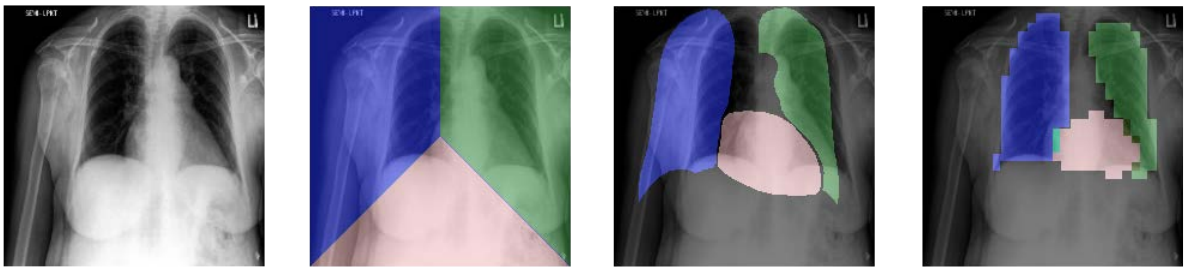

Figure 3.3: **Example templates for encoding spatial and semantic information.** 1st image: a randomly sampled image from the CheXpert dataset. 2nd image: a template based on spatial heuristics. 3rd image: a global prediction-based template. These masks are computed by averaging the predictions made from an external segmentation model. 4th image: deformable registration template. Given an exemplar image with ground-truth segmentation mask, the template is obtained by warping the segmentation using deformable image registration.

Knowledge maps representing a specific task can be assigned arbitrarily to any attention head. Unassigned heads become general attention heads and remain unregularized.

Encoding Prior Knowledge

Embedding knowledge within the knowledge-guided regularization module above comes in many varieties. Our regularization procedure allows for any type of inductive bias that can be translated into knowledge map K. We identified two useful types of inductive biases that are useful in guiding medical vision models: (1) spatial and (2) semantic. These categories are then used to embed prior knowledge, such as anatomical constraints or other assumptions, into a knowledge map K. Intuitively, the goal is to not only assist a model to look at task-relevant features but also to specialize the individual heads. To this end, we assign specific heads $n \in N$ to discretely identified entities of interest. The remainder of the attention heads remain unassigned and hence, are able to attend the whole image X. We use the following three knowledge encoding procedures for *MeDINO* which are depicted in Figure 3.3:

Spatial Heuristic

As a baseline, we explore a simple spatial heuristic that approximately segments the constant relative positioning of organs in the thorax into a knowledge map K. We encode our

knowledge as a tripartite mask with triangular parts corresponding to the left lung, right lung, and the heart.

Predicted Global Template

Instead of relying on generic spatial regions for attention supervision, we calculate a global average of the predicted locations of relevant organs from a pretrained model. We trained a segmentation model (DeepLabv3-ResNet101 [10]) on a held-out subset of the JSRT segmentation data (separate from the interpretability validation data) ($n = 200$). We average the model's inference segmentations over all images in the pretraining image dataset to obtain a single predicted global template. These knowledge maps provide a more robust spatial bias signal than the spatial heuristic.

Semantic Deformable Image Registration

To test the impact of increasingly accurate knowledge templates, we use a single ground-truth segmentation from a different dataset which is adapted to our dataset via deformable image registration. Given a single annotated exemplar pair of image X^e and its ground-truth segmentation S^e, canonical deformable image registration [65] is performed to learn a parameterization ϕ that deforms exemplar image X^e to training image X^i. This learned ϕ is then used to create a deformable knowledge map K, as an estimate to true S_i, by applying ϕ on S_e. In our paper, we use SimpleElastix wrappers [48, 85, 5] that are based on b-Spline deformation models. As seen in Figure 3.3, this procedure results in the most accurate results due to combining both spatial and semantic information.

3.3 Results

In the following experiments, we compare the qualitative and quantitative performance of *MeDINO* with self-supervised vision transformers. The different experiments focus on interpretability and downstream classification performance. The quantitative and qualitative interpretability analyses reveal that *MeDINO* leads to more interpretable representations, and the second set of experiments show the increased downstream classification performance.

Setup

Dataset

We pretrain our models using CheXpert, a medical X-ray dataset with 220k images and 14 disease classes collected from 65,240 unique patients. [36] We exclude the lateral images, as no high-quality lateral image priors are available. This reduces the dataset to 190k images. To validate the learned representations, we evaluate them against two ground-truth segmentation datasets, JSRT [63] and Montgomery [37, 57]. These two datasets are smaller

in size and contain 247 and 138 images respectively. JSRT provides segmentation masks for both lungs and the heart. Montgomery only has annotations for the lungs.

Data Augmentations

In our experiments, we differentiate between domain-agnostic and domain-specific data augmentations. Domain-agnostic data augmentations are based on the default DINO and BYOL augmentations. They contain global crops, local crops, color jittering, Gaussian blur and solarization. These augmentations are solely implemented in baseline runs. *MeDINO* incorporates domain-specific data augmentations, in particular chest X-ray specific data augmentations, are inspired from ChX-MoCo, a framework for Momentum Contrasting in X-rays. These augmentations only perform global crops in addition to translations, rotations, brightness, contrast and sharpness.

Training and Finetuning

Vision transformer architecture configurations are based on the PyTorch Image Models Library. ViTs have different pre-set configurations with respect to their hidden size; there exist 'Large', 'Base' and 'Small' vision transformers. In our experiments, we fixed the backbone of our models to be the small Vision Transformer (ViT-S, 21M parameters) with patch size 16.

Self-supervised pretraining is performed on 8 GPUs (NVIDIA Tesla V100). We train ImageNet pretrained (800 epochs) ViTs using an Adam optimizer, batch size 28, base learning rate of 10^{-3} for 30 epochs. Other hyperparameters are directly implemented from DINO. The best attention regularization hyperparameters λ_1 and λ_2 are chosen using a sweep for values between $[10^{-2}, 10^{-6}]$. For downstream classification tasks, we train a linear layer on top of the frozen learned representations without any sort of data augmentations for 100 epochs.

Attention Head Interpretability

Performance

In Table 1, we compare the different models' attention maps against the ground truth segmentations for the lungs and heart. The Montgomery dataset does not contain ground truth segmentation maps for the heart and hence these results have been omitted. The interpretability results are evaluated using pixel-wise mAP scores that calculate the average precision at different thresholds. For the *MeDINO* trained models, we use the attention maps at the assigned head for evaluation. In DINO tasks where no head was assigned to a specific part, the score represents the maximum across the different heads.

The interpretability results show higher mAP scores in *MeDINO* models compared to DINO pretrained models for all thoracic parts. *MeDINO* with triangular spatial heuristics sees a 30 mAP increase in performance over the DINO baseline pretrained using chest specific

Table 3.1: **Interpretability scores of attention heads.** The evaluation metrics included pixel-wise mAP on external validation sets where groundtruth segmentation masks were available. Due to the lack of heart segmentations in the Montgomery dataset, results of heart interpretability have not been reported. The results indicate that *MeDINO* improves the interpretability over DINO baselines.

Metric	Part	Regimen	JSRT	Montgomery
AP	Heart	DINO	19.1	
		DINO (Chexpert)	5.7	
		DINO (Chexpert Augmentations)	26.4	
		MeDINO (Triangular)	54.5	
		MeDINO (Global Average)	71.6	
		MeDINO (Deformable)	**89.9**	
	Left Lung	DINO	30.1	43.2
		DINO (Chexpert)	22.0	16.5
		DINO (Chexpert Augmentations)	25.1	40.7
		MeDINO (Triangular)	59.2	40.0
		MeDINO (Global Average)	71.3	84.1
		MeDINO (Deformable)	**88.3**	**90.0**
	Right Lung	DINO	46.3	35.0
		DINO (Chexpert)	27.0	15.8
		DINO (Chexpert Augmentations)	36.5	27.6
		MeDINO (Triangular)	45.6	54.8
		MeDINO (Global Average)	82.5	50.3
		MeDINO (Deformable)	**87.3**	**88.4**

augmentations. This further improves with the templates acquired from global average masks, specifically in the right lung. As the templates become more specific, the deformable semantic masks acquired further performance gains yielding 88.5 mAP on average in JSRT. This is a 58.7 mAP increase over the baseline (DINO with chest augmentations). **The key trend we observe is that the more specific information that is encoded in masks, the more higher the interpretability scores**. The results also corroborate that semantic and spatial information ultimately attain the highest performance outcomes, as the deformable parts based mask model gained the highest performance. In general, any attention based model seems to outperform a non-guided model.

The baseline DINO pretrained with domain-agnostic augmentations on X-ray scans has the lowest scores across all body parts. Interestingly, a pretrained model that was not pretrained on chest images outperforms this setup. This suggests that DINO domain-agnostic augmentations (such as the global and local crops) have a large negative impact on pretrain-

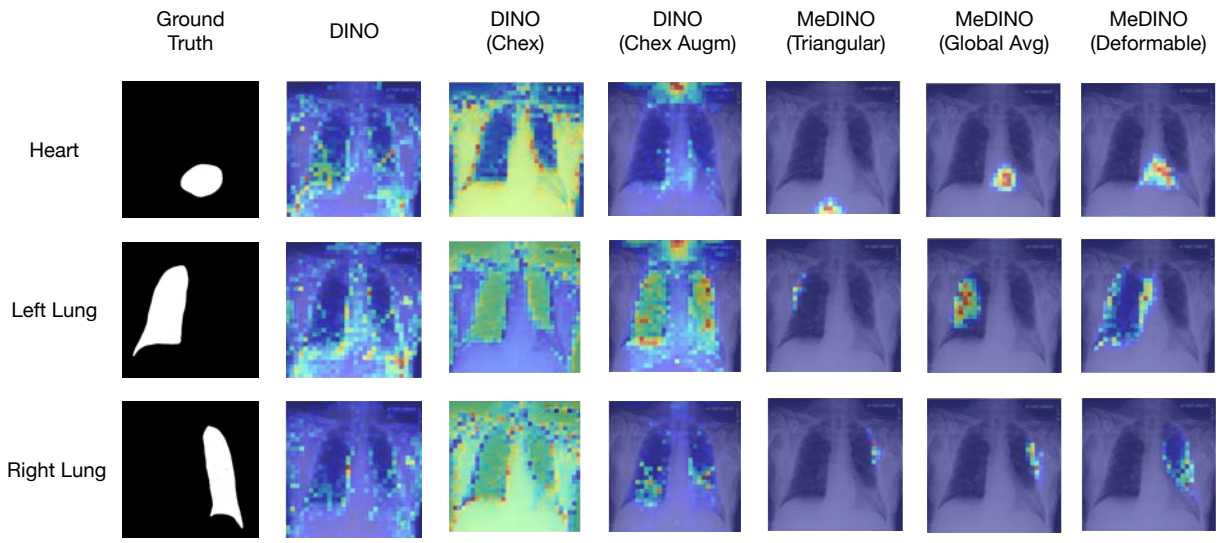

Figure 3.4: **Visualized attention maps from differently pretrained models.** We analyze the visualized attention maps by probing the heads of the respective models and choosing the map with the highest IOU overlap with the ground truth for each model. These maps show that as the prior for attention becomes more specific, the mAP and specialization of attention heads increases. Additionally, they show the inability of DINO to learn interpretable representations without chest-specific augmentations.

ing on non-object-centric tasks where semantics and spatial relationships reveal essential information. This negative performance is restored through the removal of local crops and the addition of domain-specific augmentations. The same patterns hold for both datasets.

Qualitative Assessment

In Figure 3.4, we visualize the attention maps resulting from the different models. Figure 3.4 shows that DINO pretrained Vision Transformers are unable to learn salient representations from X-ray images. DINO even leads to collapse with worse representations than a DINO model not pretrained on chest X-rays at all, though removal of global-local crops and inclusion of medical imaging specific augmentations mitigates this performance drop. Figure 3.4 also shows that *MeDINO* improves the alignment with the segmented regions and also disentangles the constituent attention maps across the heads.

Downstream Disease Classification

Classification performances are assessed using the mean receiver operating area under the curve (mAUC) score averaged over the 6 disease classes in the CheXpert classification chal-

Table 3.2: **Linear disease classification trained on frozen pretrained features.** The pretrained models are used as feature extractors in the CheXpert classification task whereby a linear layer is fine-tuned to predict the presence of six diseases: Atelectasis, Pleural Effusion, Consolidation, Cardiomegaly, No Finding and Edema. The mAUC over all diseases are reported. *MeDINO* outperforms DINO pretraining methods for all different attention priors. DINO pretraining decreases the accuracy performance, which is then restored with the addition of chest-specific augmentations.

	Regimen	mAUC
	Random	69.9
	DINO	83.8
	DINO (Chexpert)	60.4
	DINO (Chexpert Augmentations)	84.3
Ours	MeDINO (Triangular)	84.8
	MeDINO (Global Average)	86.2
	MeDINO (Deformable)	**86.5**

lenge using a hold-out test set of 200 images: Atelectasis, Edema, Pleural Effusion, Cardiomegaly, Consolidation and No Finding. The linear classifier is trained on top of the frozen pretrained representations. The results in Table 2 show that ***MeDINO* has stronger multi-label classification performance compared to all baseline DINO variants**. Specifically, *MeDINO* with the deformable image registration templates attains the highest mAUC score followed by the predicted global templates and the triangular spatial heuristics. This indicates that the more interpretable representations from *MeDINO* also lead to higher downstream performance as well. In comparison, DINO pretrained on ImageNet and CheXpert with domain-agnostic representations attains the lowest classification score. DINO only pretrained on ImageNet performs equally as DINO pretrained on ImageNet and CheXpert with domain-specific augmentations.

3.4 Conclusion and Future Work

We presented *MeDINO*: a framework for knowledge-based self-supervised Vision Transformers, which incorporates useful inductive biases into the training processes that learn more interpretable representations and lead to better performance on downstream classification tasks Medical DINO (*MeDINO*), a method that takes advantage of consistent spatial and semantic structure in unlabeled medical imaging datasets to guide vision transformer attention. Using chest X-ray radiographs as a primary case study, we show that the resulting attention masks are more interpretable than those resulting from domain-agnostic pretraining, producing a 58.7 mAP improvement for lung and heart segmentation following the self-supervised

pretraining. Additionally, *MeDINO* yields a 2.2 mAUC improvement compared to domain-agnostic pretraining when transferring the pretrained model to a downstream chest disease classification task. Our results indicate that the attention heads in self-supervised Vision Transformer can be specialized to attend to different objects and learn more semantically and meaning representations underlying the data by embedding prior knowledge using our attention regularization framework. Follow-up work could focus on generalizing this framework to expand beyond thoracic X-rays or even the medical domain, and exploring incorporating different forms of prior knowledge.

3.5 Acknowledgements

This work was done alongside Kevin Miao, Colorado Reed, Suzie Petryk, and Raghav Singh.

Chapter 4

Conclusion

Object-level representations are a fundamental part of how humans perceive visual scenes. Although, models pretrained using the instance discrimination paradigm fail to learn object-level knowledge. This thesis detailed how the simple addition of object-level knowledge, whether it is through object discovery in natural images or anatomical priors in medical images, leads to improvements in downstream performance, efficiency, and interpretability. Object-level representation learning is the next step for instance discrimination pretraining. However, learning objects is not the only point of improvement. Instance discrimination relies heavily on data augmentation, an implicit form of supervision for unsupervised learning algorithms. To truly uphold the title of unsupervised representation learning, the instance discrimination paradigm must avoid the use of data augmentations. We believe that the combination of data-augmentation free instance discrimination [19] and object-level pretraining provides is an exciting area for future research.

Bibliography

[1] Radhakrishna Achanta et al. "SLIC superpixels compared to state-of-the-art superpixel methods". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2274–2282.

[2] Shekoofeh Azizi et al. "Big self-supervised models advance medical image classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3478–3488.

[3] Wenjia Bai et al. "Self-supervised learning for cardiac mr image segmentation by anatomical position prediction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 541–549.

[4] Guha Balakrishnan et al. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38.8 (Aug. 2019), pp. 1788–1800. ISSN: 1558-254X. DOI: 10.1109/tmi.2019.2897538. URL: http://dx.doi.org/10.1109/TMI.2019.2897538.

[5] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. "Image segmentation, registration and characterization in R with SimpleITK". In: *Journal of statistical software* 86 (2018).

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[7] Mathilde Caron et al. "Deep clustering for unsupervised learning of visual features". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.

[8] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9650–9660.

[9] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9912–9924.

[10] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[11] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[12] Xinlei Chen and Kaiming He. "Exploring simple siamese representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15750–15758.

[13] Xinlei Chen et al. "Improved baselines with momentum contrastive learning". In: *arXiv preprint arXiv:2003.04297* (2020).

[14] MMSegmentation Contributors. *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. `https://github.com/open-mmlab/mmsegmentation`. 2020.

[15] Marius Cordts et al. "The cityscapes dataset for semantic urban scene understanding". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3213–3223.

[16] Mark Cox et al. "Least squares congealing for unsupervised alignment of images". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[17] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised visual representation learning by context prediction". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1422–1430.

[18] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[19] Debidatta Dwibedi et al. "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9588–9597.

[20] Mark Everingham et al. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[21] Pedro F Felzenszwalb and Daniel P Huttenlocher. "Efficient graph-based image segmentation". In: *International journal of computer vision* 59.2 (2004), pp. 167–181.

[22] Hiroshi Fukui et al. "Attention Branch Network: Learning of Attention Mechanism for Visual Explanation". In: *CoRR* (2018).

[23] Matej Gazda et al. "Self-supervised deep convolutional neural network for chest X-ray classification". In: *IEEE Access* 9 (2021), pp. 151972–151982.

[24] Golnaz Ghiasi et al. "Simple copy-paste is a strong data augmentation method for instance segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2918–2928.

[25] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations". In: *arXiv preprint arXiv:1803.07728* (2018).

[26] Ronglin Gong et al. "Self-Supervised Bi-channel Transformer Networks for Computer-Aided Diagnosis". In: *IEEE Journal of Biomedical and Health Informatics* (2022), pp. 1–1. DOI: 10.1109/JBHI.2022.3153902.

[27] Jean-Bastien Grill et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21271–21284.

[28] Grant Haskins, Uwe Kruger, and Pingkun Yan. "Deep learning in medical image registration: a survey". In: *Machine Vision and Applications* 31.1–2 (Jan. 2020). ISSN: 1432-1769. DOI: 10.1007/s00138-020-01060-x. URL: http://dx.doi.org/10.1007/s00138-020-01060-x.

[29] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[30] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[31] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *arXiv preprint arXiv:2111.06377* (2021).

[32] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[33] Olivier J Hénaff et al. "Efficient visual pretraining with contrastive detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10086–10096.

[34] Lang Huang et al. "Learning Where to Learn in Cross-View Self-Supervised Learning". In: *arXiv preprint arXiv:2203.14898* (2022).

[35] Jyh-Jing Hwang et al. "Segsort: Segmentation by discriminative sorting of segments". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 7334–7344.

[36] Jeremy Irvin et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[37] Stefan Jaeger et al. "Pu-Xuan Lu". In: *Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg* 4.6 (2014), pp. 475–477.

[38] Xu Ji, Joao F Henriques, and Andrea Vedaldi. "Invariant information clustering for unsupervised image classification and segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9865–9874.

[39] Scott P Johnson. "Young infants' perception of object unity: Implications for development of attentional and cognitive skills". In: *Current Directions in Psychological Science* 6.1 (1997), pp. 5–11.

[40] Scott P Johnson, Dima Amso, and Jonathan A Slemmer. "Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm". In: *Proceedings of the National Academy of Sciences* 100.18 (2003), pp. 10568–10573.

[41] Scott P Johnson et al. "Infants' perception of object trajectories". In: *Child development* 74.1 (2003), pp. 94–108.

[42] Christopher J Kelly et al. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC medicine* 17.1 (2019), pp. 1–9.

[43] Martin Lefébure and Laurent D. Cohen. In: *Journal of Mathematical Imaging and Vision* 14.2 (2001), pp. 131–147. DOI: 10.1023/a:1011259231755. URL: https://doi.org/10.1023/a:1011259231755.

[44] Guanbin Li and Yizhou Yu. "Deep contrast learning for salient object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 478–487.

[45] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[46] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[47] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[48] BC Lowekamp et al. *The Design of SimpleITK. Front Neuroinform. 2013; 7: 45*. 2013.

[49] Dwarikanath Mahapatra et al. "Interpretability-driven sample selection using self supervised learning for disease classification and segmentation". In: *IEEE transactions on medical imaging* 40.10 (2021), pp. 2548–2562.

[50] Lucas Mansilla, Diego H. Milone, and Enzo Ferrante. "Learning deformable registration of medical images with anatomical constraints". In: *Neural Networks* 124 (Apr. 2020), pp. 269–279. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2020.01.023. URL: http://dx.doi.org/10.1016/j.neunet.2020.01.023.

[51] Denis Mareschal and Scott P Johnson. "Learning to perceive object unity: A connectionist account". In: *Developmental Science* 5.2 (2002), pp. 151–172.

[52] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.

[53] Tony CW Mok and Albert Chung. "Fast symmetric diffeomorphic image registration with convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 4644–4653.

[54] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. Springer. 2016, pp. 69–84.

[55] Pedro O O Pinheiro et al. "Unsupervised learning of dense visual representations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4489–4500.

[56] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[57] Sivaramakrishnan Rajaraman et al. "Improved semantic segmentation of tuberculosis—Consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations". In: *Diagnostics* 11.4 (2021), p. 616.

[58] Colorado J Reed et al. "Self-supervised pretraining improves self-supervised pretraining". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2584–2594.

[59] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[60] Nurbek Saidnassim et al. "Self-supervised Visual Transformers for Breast Cancer Diagnosis". In: *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2021, pp. 423–427.

[61] Ramprasaath R Selvaraju et al. "Casting your model: Learning to localize improves self-supervised representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11058–11067.

[62] James Shackleford, Nagarajan Kandasamy, and Gregory Sharp. "Unimodal B-Spline Registration". In: *High Performance Deformable Image Registration Algorithms for Manycore Processors*. Elsevier, 2013, pp. 13–43. DOI: 10.1016/b978-0-12-407741-6.00002-5. URL: https://doi.org/10.1016/b978-0-12-407741-6.00002-5.

[63] Junji Shiraishi et al. "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules". In: *American Journal of Roentgenology* 174.1 (2000), pp. 71–74.

[64] James M Sloan, Keith A Goatman, and J Paul Siebert. "Learning rigid image registration-utilizing convolutional neural networks for medical image registration". In: (2018).

[65] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. "Deformable medical image registration: A survey". In: *IEEE transactions on medical imaging* 32.7 (2013), pp. 1153–1190.

[66] Hari Sowrirajan et al. "Moco pretraining improves representation and transferability of chest x-ray models". In: *Medical Imaging with Deep Learning*. PMLR. 2021, pp. 728–744.

[67] Yucheng Tang et al. "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis". In: *CoRR* abs/2111.14791 (2021). arXiv: 2111.14791. URL: https://arxiv.org/abs/2111.14791.

[68] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive multiview coding". In: *European conference on computer vision*. Springer. 2020, pp. 776–794.

[69] Nenad Tomasev et al. "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?" In: *arXiv preprint arXiv:2201.05119* (2022).

[70] Jasper RR Uijlings et al. "Selective search for object recognition". In: *International journal of computer vision* 104.2 (2013), pp. 154–171.

[71] Stefan Van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (2014), e453.

[72] Wouter Van Gansbeke et al. "Unsupervised semantic segmentation by contrasting object mask proposals". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10052–10062.

[73] Ella Y. Wang et al. "Interpretable COVID-19 Chest X-Ray Classification via Orthogonality Constraint". In: *CoRR* (2021).

[74] Xinlong Wang et al. "Dense contrastive learning for self-supervised visual pre-training". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3024–3033.

[75] Xiyue Wang et al. "Transpath: Transformer-based self-supervised learning for histopathological image classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 186–195.

[76] Fangyun Wei et al. "Aligning pretraining for detection via object-level contrastive learning". In: *Advances in Neural Information Processing Systems* 34 (2021).

[77] Di Wu et al. "Align Yourself: Self-supervised Pre-training for Fine-grained Recognition via Saliency Alignment". In: *arXiv preprint arXiv:2106.15788* (2021).

[78] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[79] Tete Xiao et al. "Region similarity representation learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10539–10548.

[80] Tete Xiao et al. "What should not be contrastive in contrastive learning". In: *arXiv preprint arXiv:2008.05659* (2020).

[81] Jiahao Xie et al. "Unsupervised object-level representation learning from scene images". In: *Advances in Neural Information Processing Systems* 34 (2021).

[82] Zhenda Xie et al. "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16684–16693.

[83] Huijuan Xu and Kate Saenko. "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering". In: *CoRR* (2015).

[84] Chenhongyi Yang, Lichao Huang, and Elliot J Crowley. "Contrastive Object-level Pretraining with Spatial Noise Curriculum Learning". In: *arXiv preprint arXiv:2111.13651* (2021).

[85] Ziv Yaniv et al. "SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research". In: *Journal of digital imaging* 31.3 (2018), pp. 290–303.

[86] Feihu Zhang et al. "Looking Beyond Single Images for Contrastive Semantic Segmentation Learning". In: *Advances in Neural Information Processing Systems* 34 (2021).

[87] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

[88] Nanxuan Zhao et al. "Distilling localization for self-supervised representation learning". In: *arXiv preprint arXiv:2004.06638* (2020).

[89] Yucheng Zhao et al. "Self-supervised visual representations learning by contrastive mask prediction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10160–10169.

[90] B. Zhou et al. "Learning Deep Features for Discriminative Localization". In: *CVPR* (2016).