Knowledge-Guided Self-Supervised Vision Transformers for Medical Imaging

Kevin Miao
Colorado Reed
Akash Gokul
Suzanne Petryk
Raghav Singh
Kurt Keutzer
Joseph Gonzalez
Trevor Darrell

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-56
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-56.html

May 10, 2022

Knowledge-Guided Self-Supervised Vision Transformers for Medical Imaging

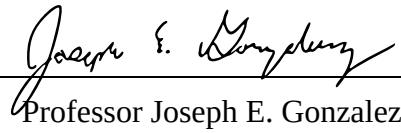
Kevin Miao

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

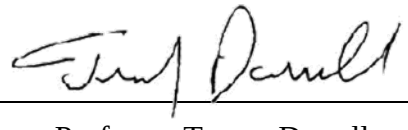


Professor Joseph E. Gonzalez
Research Advisor

5/10/2022

(Date)

* * * * * * *



Professor Trevor Darrell
Second Reader

5/10/22

(Date)

Abstract

Recent trends in self-supervised representation learning have focused on removing inductive biases from the training process. However, inductive biases can be useful in certain settings, such as medical imaging, where domain expertise can help define a prior over semantic structure. We present Medical DINO (MeDINO), a method that takes advantage of consistent spatial and semantic structure in unlabeled medical imaging datasets to guide vision transformer attention. MeDINO operates by regularizing attention masks from separate transformer heads to follow various priors over semantic regions. These priors can be derived from data statistics or are provided via a single labeled sample from a domain expert. 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, our method yields a 2.2 mAUC improvement compared to domain-agnostic pretraining when transferring the pretrained model to a downstream chest disease classification task.

Knowledge-Guided Self-Supervised Vision Transformers for Medical Imaging

Kevin Miao[1*], Colorado Reed[1], Akash Gokul[1], Suzie Petryk[1], Raghav Singh[1], Kurt Keutzer[1], Trevor Darrell[1], and Joseph E. Gonzalez[1]

[1]UC Berkeley, Electrical Engineering and Computer Science, Berkeley CA, USA
{kevinmiao,cjrd,akashgokul,spetryk,raghavsingh,keutzer,
trevordarrell,jegonzal}@berkeley.edu

Abstract. Recent trends in self-supervised representation learning have focused on removing inductive biases from the training process. However, inductive biases can be useful in certain settings, such as medical imaging, where domain expertise can help define a prior over semantic structure. We present Medical DINO (*MeDINO*), a method that takes advantage of consistent spatial and semantic structure in unlabeled medical imaging datasets to guide vision transformer attention. *MeDINO* operates by regularizing attention masks from separate transformer heads to follow various priors over semantic regions. These priors can be derived from data statistics or are provided via a single labeled sample from a domain expert. 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, our method yields a 2.2 mAUC improvement compared to domain-agnostic pretraining when transferring the pretrained model to a downstream chest disease classification task.

Keywords: self-supervised; interpretable; medical imaging

1 Introduction

Recent works in unsupervised learning have largely focused on removing inductive biases from the training process: transformer-based methods have successfully removed the scale-and-shift invariance from CNNs [17] and autoencoders have successfully removed the hardcoded augmentation-based invariances from contrastive learning methods [23]. However, inductive biases can capture knowledge that would otherwise be difficult to infer strictly from observed data and are particularly beneficial when there is not enough data to allow for generalization to unseen scenarios [4, 8, 34] or when domain-specific knowledge can provide information that can be used to model the underlying data distribution [36, 16]. For example, in a medical imaging application, the human anatomy can provide information about the expected positioning of the elements within the image [9].

Medical imaging is one instance where data are expensive to acquire and store, require exhaustive labeling procedures by experts, and acquisition of data

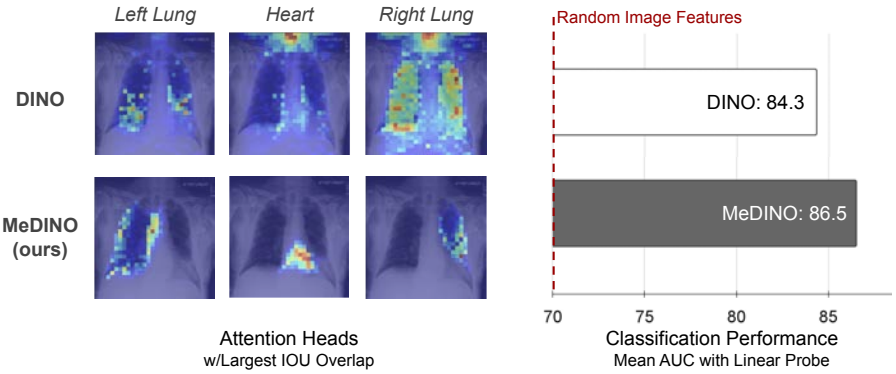


Fig. 1. Self-attention from a Vision Transformer on chest X-rays, where the attentions heads with the largest IOU overlap with the lungs/heart are shown. Existing self-supervised training methods for Vision Transformers, such as DINO, learn scattered attention maps that do not necessarily attend to the 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 [26] – see Section 4 for details.

is challenging due to privacy or regulatory concerns [28]. While medical images are often limited in availability, they share a common underlying anatomical structure that is well understood. This underlying structure can provide a strong inductive bias that a model might not be able to learn from data alone. For instance, the anatomy informs about the presence and relative positioning of organs and the type of image (i.e. X-ray, CT, MRI) reveals information about the characteristics about the pixel intensities. So while one thread of research seeks to remove such inductive biases and learn directly from data, a complementary thread can seek useful inductive biases to guide the training process, specifically in structured domains such as medical imaging. In this work, we investigate an intersection of these two threads for medical imaging applications where we leverage both a self-supervised transformer learning framework as well as anatomical knowledge to guide the training.

We present a framework called Medical DINO (*MeDINO*) that is built upon the DINO self-supervised vision transformer framework [11] and leads to more interpretable attention heads that perform better on downstream tasks – see Figure 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 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 2 provides an overview of this process which is detailed in Section 3.

Compared to DINO, *MeDINO* leads to better transfer performance to other datasets and tasks, and more interpretable attention weights. Specifically *MeDINO* improves the interpretability of learned representations by 58.7 mAP and yields 2.2 mAUC improve compared to domain-agnostic pre-training. We summarize our contributions as follows:

1. We present a novel knowledge guided-attention regularization framework in self-supervised Vision Transformers that leverages the inherent attention heads to learn disentangled and meaningful representations for medical radiographs.
2. We establish a range of procedures that incorporate medical prior knowledge and inductive biases into templates when annotated data are sparse or these priors reveal information that cannot be learned from the data alone.
3. We find that encoding prior knowledge using attention regularization increases the interpretability of representations by 58.7 mAP compared to domain-agnostic pretraining. This leads to a 2.2 mAUC increase in downstream disease classification tasks.

2 Related Works

MeDINO incorporates domain knowledge to improve the performance and interpretability of self-supervised pretraining for medical images. *MeDINO* builds upon works in self-supervised learning, knowledge-guided and interpretable methods, and image registration and alignment, which we detail below.

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 [24, 13], clustering-based techniques [10], and self-distillation [21, 11]. DINO [11] 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 [17]. 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 [47] to extract pixel-level saliency relationships, (ii) in self-supervised training, the DINO attention maps have a demonstrated ability to segment salient foreground objects [11], 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 [49, 38]. Furthermore, medical applications of machine learning can benefit from self-supervised pretraining due to the cost and expertise needed to accurately annotate data [45, 19, 2, 1]. Medical data may also be difficult to obtain due to privacy and regulatory issues. For instance, MoCo-CXR [45] adapts MoCo [24] 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 [31] 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 [40, 48, 46, 20]. *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.

Knowledge-Guided Learning seeks to incorporate prior knowledge in such a way that it leads to better performance, efficiency, or interpretability for the learned model. Several papers have incorporated first order logic rules [25, 39, 53] as well as anatomical constraints for pose estimation [35, 7, 7]. Neural networks have also been combined with physics-based models to capture and enforce the relationship between variables through an additional loss [15], which is similar to our work in that it also captures the alignment between the guided attention mechanism and domain knowledge via an attention-based loss mechanism. Different from these works, *MeDINO* uses attention-based regularization in transformer models to enforce such constraints, rather than external architectural or optimization mechanisms.

Furthermore, attention based approaches have been used to improve the explainability of computer vision models through visualizations of attention maps to indicate important regions [51]. Convolutional neural networks use tools such as CAM [55] and GradCAM [47] to create attention maps by looking at the hidden layer activations. Another approach is the Attention Branch Network [18] 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, 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 [22]. 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 [44]. 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 [14]. Optical flow completes this registration task by looking for possible displacements and solving a minimum energy functional [29]. b-Spline registration operates by modeling the deformation field as B-spline curves where each pixel that maps each voxel in the source image to the target image. [41]. More recently, neural networks have become increasingly popular in performing image registration in lieu of the traditional methods [3, 33, 43]. 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. [32] 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 Incorporating Knowledge into Self-Supervised Vision Transformers

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 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.

3.1 Self-Supervised Vision Transformers with Knowledge Distillation

Caron et al. [11] 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

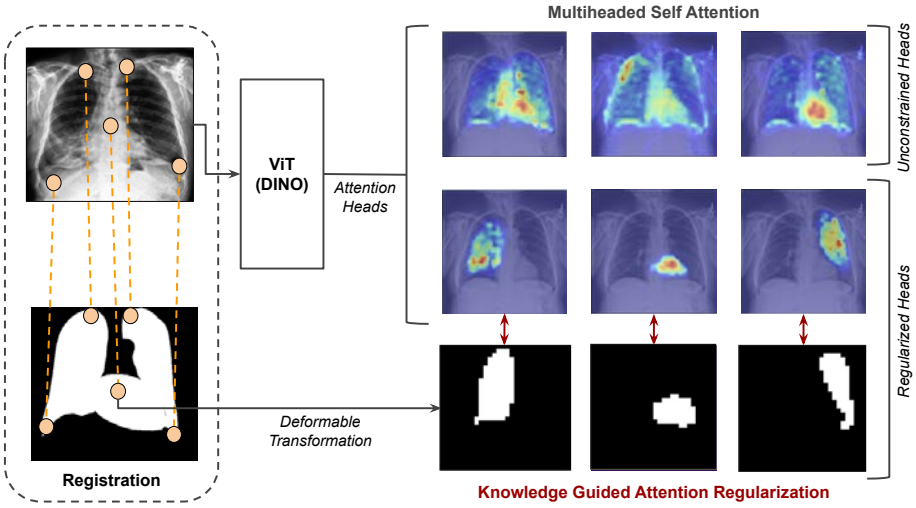


Fig. 2. The *MeDINO* framework. *MeDINO* first registers each image to an exemplar template with known segmentations, the registration outputs a deformable transformation that is applied to the template. During self-supervised pretraining with a ViT model, each component of the template then regularizes an indiviudal attention head in the multiheaded self-attention modules (Regularized Heads). A subset of the attention heads are also unconstrained (Unconstrained Heads).

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 $P_t(X)$.

$$L(X_1, X_2) = H(P_t(X_1), P_s(X_2))$$
$$\arg\min_{\theta_s} \sum_{x\in\{X_1^g, X_2^g\}} \sum_{X'\in V} L(X, X')$$

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. [50] 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 (see the Appendix for details).

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 [17]. 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.

3.2 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

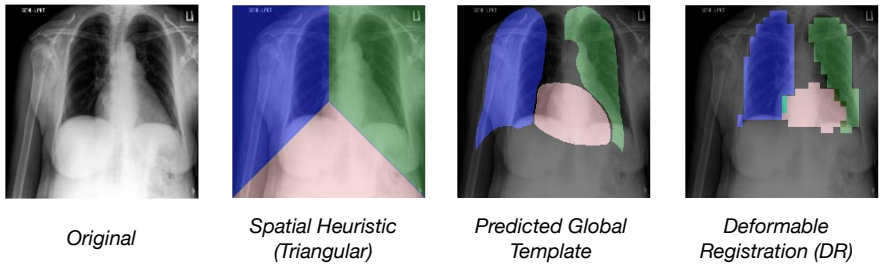


Fig. 3. Example templates for encoding spatial and semantic information. 1st image: a randomly sampled image from the CheXpert dataset. 2nd image: a template based on spatial heuristics. 3rd image: a global prediction-based template. These masks are computed by averaging the predictions made from an 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.

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

3.3 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:

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 [12]) 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 [44] 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 [30, 52, 5] that are based on b-Spline deformation models. As seen in Figure 3, this procedure results in the most accurate results due to combining both spatial and semantic information.

4 Experiments

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.

4.1 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. [26] 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 [42] and Montgomery [27, 37]. 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

Table 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**

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.

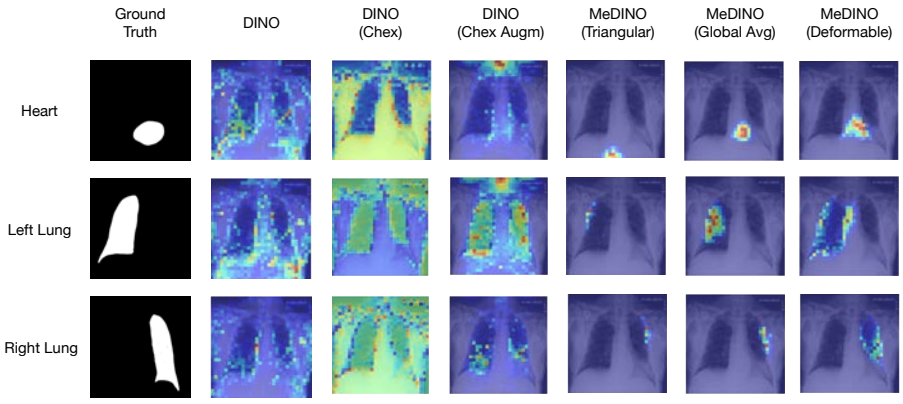


Fig. 4. Visualized attention maps from differently pretrained models. We analyze the visualized attention maps by probing the heads of the respective models and choosing the map with the highest IOU overlap with the ground truth for each model. These maps show that as the prior for attention becomes 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.

4.2 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 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 pretraining 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. An additional analysis using an alternate metric, the pointing game, used for interpretability assessment can be found in the appendix.

Qualitative Assessment In Figure 4, we visualize the attention maps resulting from the different models. Figure 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 4 also shows that *MeDINO* improves the alignment with the segmented regions and also disentangles the constituent attention maps across the heads.

Table 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**

4.3 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 challenge 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 line with the results of Section 4.2, 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.

5 Conclusion

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.

References

1. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3478–3488 (2021)
2. Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D.: Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 541–549. Springer (2019)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (Aug 2019). https://doi.org/10.1109/tmi.2019.2897538, http://dx.doi.org/10.1109/TMI.2019.2897538
4. Baxter, J.: A model of inductive bias learning. Journal of artificial intelligence research **12**, 149–198 (2000)
5. Beare, R., Lowekamp, B., Yaniv, Z.: Image segmentation, registration and characterization in r with simpleitk. Journal of statistical software **86** (2018)
6. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
7. Bigalke, A., Hansen, L., Diesel, J., Heinrich, M.P.: Domain adaptive 3d human pose estimation through anatomical constraints (2021)
8. Bouvier, V., Very, P., Chastagnol, C., Tami, M., Hudelot, C.: Robust domain adaptation: Representations, weights and inductive bias. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 353–377. Springer (2020)
9. Boveiri, H.R., Khayami, R., Javidan, R., Mehdizadeh, A.: Medical image registration using deep neural networks: a comprehensive review. Computers & Electrical Engineering **87**, 106767 (2020)
10. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
12. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
13. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
14. Cox, M., Sridharan, S., Lucey, S., Cohn, J.: Least squares congealing for unsupervised alignment of images. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
15. Daw, A., Karpatne, A., Watkins, W., Read, J.S., Kumar, V.: Physics-guided neural networks (PGNN): an application in lake temperature modeling. CoRR (2017)
16. Desai, S., Strachan, A.: Parsimonious neural networks learn interpretable physical laws. Scientific reports **11**(1), 1–9 (2021)

17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
18. Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation. CoRR (2018)
19. Gazda, M., Plavka, J., Gazda, J., Drotar, P.: Self-supervised deep convolutional neural network for chest x-ray classification. IEEE Access **9**, 151972–151982 (2021)
20. Gong, R., Han, X., Wang, J., Ying, S., Shi, J.: Self-supervised bi-channel transformer networks for computer-aided diagnosis. IEEE Journal of Biomedical and Health Informatics pp. 1–1 (2022). https://doi.org/10.1109/JBHI.2022.3153902
21. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
22. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Machine Vision and Applications **31**(1–2) (Jan 2020). https://doi.org/10.1007/s00138-020-01060-x, http://dx.doi.org/10.1007/s00138-020-01060-x
23. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
24. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
25. Hu, Z., Ma, X., Liu, Z., Hovy, E.H., Xing, E.P.: Harnessing deep neural networks with logic rules. CoRR (2016)
26. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., 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, pp. 590–597 (2019)
27. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.: Pu-xuan lu. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg **4**(6), 475–477 (2014)
28. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC medicine **17**(1), 1–9 (2019)
29. Lefébure, M., Cohen, L.D.: Journal of Mathematical Imaging and Vision **14**(2), 131–147 (2001). https://doi.org/10.1023/a:1011259231755, https://doi.org/10.1023/a:1011259231755
30. Lowekamp, B., Chen, D., Ibáñez, L., Blezek, D.: The design of simpleitk. front neuroinform. 2013; 7: 45 (2013)
31. Mahapatra, D., Poellinger, A., Shao, L., Reyes, M.: Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE transactions on medical imaging **40**(10), 2548–2562 (2021)
32. Mansilla, L., Milone, D.H., Ferrante, E.: Learning deformable registration of medical images with anatomical constraints. Neural Networks **124**, 269–279 (Apr 2020). https://doi.org/10.1016/j.neunet.2020.01.023, http://dx.doi.org/10.1016/j.neunet.2020.01.023
33. Mok, T.C., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4644–4653 (2020)

34. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
35. Ning, G., Zhang, Z., He, Z.: Knowledge-guided deep fractal neural networks for human pose estimation. IEEE Transactions on Multimedia **20**(5), 1246–1259 (2017)
36. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational physics **378**, 686–707 (2019)
37. Rajaraman, S., Folio, L.R., Dimperio, J., Alderson, P.O., Antani, S.K.: Improved semantic segmentation of tuberculosis—consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations. Diagnostics **11**(4), 616 (2021)
38. Reed, C.J., Yue, X., Nrusimha, A., Ebrahimi, S., Vijaykumar, V., Mao, R., Li, B., Zhang, S., Guillory, D., Metzger, S., et al.: Self-supervised pretraining improves self-supervised pretraining. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2584–2594 (2022)
39. Roychowdhury, S., Diligenti, M., Gori, M.: Regularizing deep networks with prior knowledge: A constraint-based approach. Knowledge-Based Systems **222**, 106989 (04 2021)
40. Saidnassim, N., Abdikenov, B., Kelesbekov, R., Akhtar, M.T., Jamwal, P.: Self-supervised visual transformers for breast cancer diagnosis. In: 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 423–427 (2021)
41. Shackleford, J., Kandasamy, N., Sharp, G.: Unimodal b-spline registration. In: High Performance Deformable Image Registration Algorithms for Manycore Processors, pp. 13–43. Elsevier (2013). https://doi.org/10.1016/b978-0-12-407741-6.00002-5, https://doi.org/10.1016/b978-0-12-407741-6.00002-5
42. Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K.: 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. American Journal of Roentgenology **174**(1), 71–74 (2000)
43. Sloan, J.M., Goatman, K.A., Siebert, J.P.: Learning rigid image registration-utilizing convolutional neural networks for medical image registration (2018)
44. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE transactions on medical imaging **32**(7), 1153–1190 (2013)
45. Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P.: Moco pretraining improves representation and transferability of chest x-ray models. In: Medical Imaging with Deep Learning. pp. 728–744. PMLR (2021)
46. Tang, Y., Yang, D., Li, W., Roth, H., Landman, B.A., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. CoRR **abs/2111.14791** (2021), https://arxiv.org/abs/2111.14791
47. Wang, E.Y., Som, A., Shukla, A., Choi, H., Turaga, P.K.: Interpretable COVID-19 chest x-ray classification via orthogonality constraint. CoRR (2021)
48. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X.: Transpath: Transformer-based self-supervised learning for histopathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 186–195. Springer (2021)
49. Xiao, T., Wang, X., Efros, A.A., Darrell, T.: What should not be contrastive in contrastive learning. arXiv preprint arXiv:2008.05659 (2020)

50. Xie, J., Zhan, X., Liu, Z., Ong, Y., Loy, C.C.: Unsupervised object-level representation learning from scene images. Advances in Neural Information Processing Systems **34** (2021)
51. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. CoRR (2015)
52. Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R.: Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research. Journal of digital imaging **31**(3), 290–303 (2018)
53. Yin, C., Zhao, R., Qian, B., Lv, X., Zhang, P.: Domain knowledge guided deep learning with electronic health records. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 738–747. IEEE (2019)
54. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126**(10), 1084–1102 (2018)
55. Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. CVPR (2016)

A Appendix

A.1 Notations and Definitions

Notation	Definition
$g_{\theta s}$	Student model
$g_{\theta t}$	Teacher model
X	Input image
w	Image width
h	Image height
n_h	The number of attention heads
n	The number of regularized attention heads
p	Patch size
P	The total number of patches
V	The set of data augmentations to apply to each image
$\{X_1, ..., X_n\}$	The set of augmented images acquired resulting from applying V on X
X^g	Global crop
$P(x)$	Output probability distribution
τ	Temperataure
H	Cross-entropy
A	Attention matrix
A_s	Self-attention map
K	Knowledge map
λ_1, λ_2	Magnitude for inclusion and exclusion regularization
X^e	Exemplar image
X^i	Training image
S^e	Ground-truth segmentation for exemplar
S^i	Ground-truth segmentation for training sample
ϕ	Deformable transformation

A.2 Data Augmentation Details

DINO augmentations pretrained on chest images result in worse representations, as observed in Table 1. Therefore, we follow the radiograph-specific data augmentations proposed by Sowrirajan et al. [45]. These augmentations refrain from using local crops, color jittering and randomized grayscaling. Such augmentations do not apply well to gray-scale chest radiographs and are eliminated from the pretraining procedure. The remaining augmentations are rotations, translations, and random global crops. We expanded them by adding brightness, sharpness and contrast. Each augmentation is applied with a random probability. When applied, the strength of the augmentation is sampled uniformly at random from the range of values specified in Table 3.

For the classification tasks, we do not apply any data augmentations. Pretrained self-supervised Vision Transformers are able to generalize well without

Table 3. Data augmentations and their respective parameters.

Data augmentation	Probability	Range [min, max]
Rotation	0.2	[-30, 30]
Contrast	0.1	[0, 0.2]
Brightness	0.1	[0, 0.2]
Sharpness	0.1	[0, 0.2]
Random Crop	1	224 x 224

augmentations when used as a feature extractor for linear classification evaluation. [11] We only normalize the images prior to training.

A.3 Training and Experiment Details

The hyperparameters for training follow the defaults from DINO [11] where possible. Pretraining is performed on the subset of frontal CheXpert images. Validation CheXpert images are used for saliency map evaluations. Additionally, JSRT and Montgomery validation images are used to score the interpretability of attention heads. For pretraining, all CheXpert images are rescaled to (224, 224). Their original sizes vary between 300-400 pixels. During inference, we rescale the image to (480, 480). All *MeDINO* models are pretrained for 30 epochs. The batch size is kept constant at 28. The default learning rate for DINO is $0.0005 * batchsize/256$. However, we found that training *MeDINO* with a learning rate of 0.0001 created slightly better representations. λ values are found via grid-search on a *log*-scale. The λ values leading to the highest interpretability scores are chosen.

A.4 Additional Results

Interpretability The pointing game is a metric that measures the interpretability of attention maps compared to a groundtruth segmentation or bounding boxes. [54] It evaluates whether the maximum entry of a saliency map falls within the region of interest, also called a hit. The metric is calculated as follows:

$$PointingGame = \frac{\#hits}{\#hits + \#misses}$$

Table 4 shows that all *MeDINO* pretrained models outperform DINO by a large margin for both heart (80+%) and lung (60+%) interpretability. The results from the Montgomery dataset follow the same patterns from Table 1 where the increasingly specific templates in *MeDINO* allowed for better performance. This is generally the case for the JSRT results as well, but the results between the global average templates outperform the deformable model for the left lung and ties for the interpretability of the heart.

Table 4. Additional interpretability scores for the attention heads with highest overlap. In this additional analysis, we included the pointing game evaluation scores, another metric for interpretability of attention maps. The figure shows that *MeDINO* outperforms all variants of DINO pre-training.

Metric	Part	Regimen	JSRT	Montgomery
Pointing	Heart	DINO	9.2	
		DINO (Chexpert)	0	
		DINO (Chexpert Augmentations)	10.7	
		MeDINO (Triangular)	62.0	
		MeDINO (Global Average)	98.0	
		MeDINO (Deformable)	**99.0**	
	Left Lung	DINO	26.9	57.97
		DINO (Chexpert)	0	0
		DINO (Chexpert Augmentations)	38.1	67.4
		MeDINO (Triangular)	96.4	35.0
		MeDINO (Global Average)	**97.9**	100
		MeDINO (Deformable)	97.5	**99.3**
	Right Lung	DINO	65.9	39.3
		DINO (Chexpert)	0	0
		DINO (Chexpert Augmentations)	34.5	43.5
		MeDINO (Triangular)	79.7	69.0
		MeDINO (Global Average)	**100.0**	90.58
		MeDINO (Deformable)	**100.0**	**98.6**

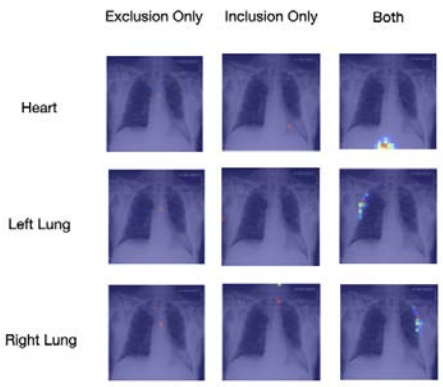


Fig. 5. Collapse in the absence of either regularization terms. This graphic shows the effect of the absence of either regularization terms. As can be seen, all the mass of the attention maps reside in one pixel when one regularization term is included.

Regularization & Collapse In this experiment, we investigate the effect of excluding the inclusion or the exclusion regularization loss terms. The results in Table 5 show that in the absence of either terms the pretrained models perform worse. Closer inspection of these attention heads in Figure 5 reveals empirically that pretraining collapses when either term is absent. This issue is mitigated when both loss terms are present.

Table 5. Decreased interpretability in the absence of either regularization terms. The mAP scores show that the presence of both the inclusion and exclusion loss terms contribute to the interpretability of the representations. When either is absent, the representations are lower. This analysis is performed using attention maps from *MeDINO* (Triangular) with JSRT scans as input.

Regularization Term	Heart	Left Lung	Right Lung
Exclusion Only	34.7	21.1	13.0
Inclusion Only	14.9	52.0	41.1
Both	**54.5**	**59.2**	**45.6**

A.5 Self-Attention Visualizations

We sample two validation images from the CheXpert dataset and provide the attention maps over all 6 attention heads in Figure 6. The Figure shows that the

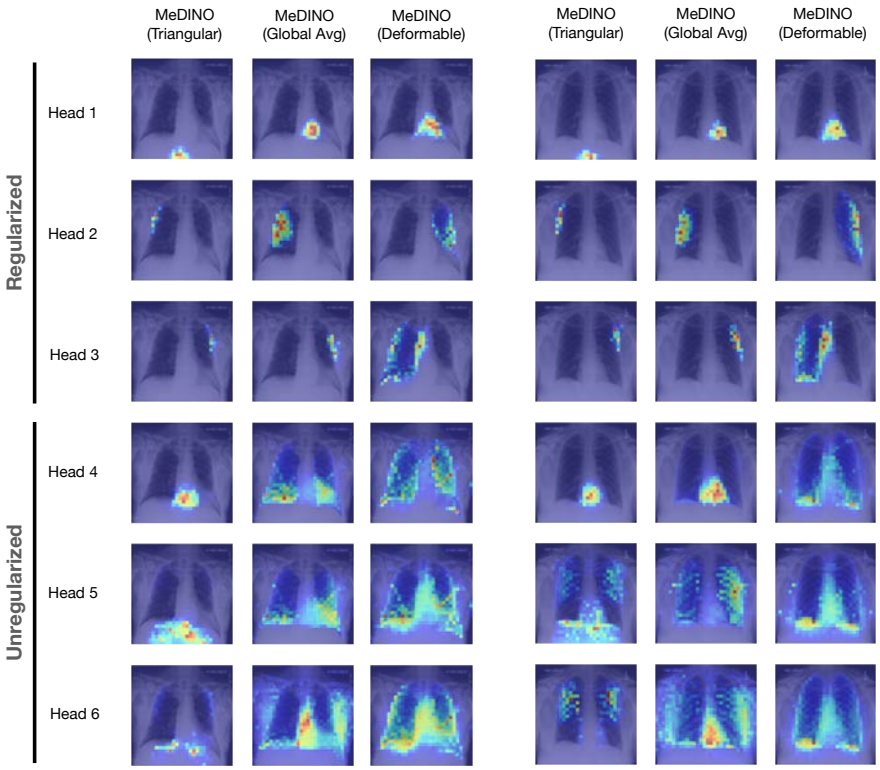


Fig. 6. Visualized self-attention over all heads. This Figure shows the attention maps from *MeDINO* given two samples from the CheXpert validation set. The regularized heads attend to regions and semantics. The unregularized heads also learn emergent, interpretable representations.

unregularized heads become more interpretable as the quality of the representations in the regularized heads increase. We even observe emergent properties in unregularized heads as a result of this.