Joint Learning of Localized Representations from Medical Images and Reports

Philip Müller[1(✉)], Georgios Kaissis[1,2,3], Congyu Zou[4], and Daniel Rueckert[1,3]

[1] Institute of Artificial Intelligence in Medicine, Technical University of Munich, 81675 Munich, Germany
philip.j.mueller@tum.de
[2] Institute of Radiology, Technical University of Munich, 81675 Munich, Germany
[3] Department of Computing, Imperial College London, London SW7 2BX, UK
[4] Department for Internal Medicine I, Klinikum Rechts der Isar, Technical University of Munich, 81675 Munich, Germany

Abstract. Contrastive learning has proven effective for pre-training image models on unlabeled data with promising results for tasks such as medical image classification. Using paired text (like radiological reports) during pre-training improves the results even further. Still, most existing methods target image classification downstream tasks and may not be optimal for localized tasks like semantic segmentation or object detection. We therefore propose ***Lo****calized representation learning from* ***V****ision and* ***T****ext (LoVT)*, a text-supervised pre-training method that explicitly targets localized medical imaging tasks. Our method combines instance-level image-report contrastive learning with local contrastive learning on image region and report sentence representations. We evaluate LoVT and commonly used pre-training methods on an evaluation framework of 18 localized tasks on chest X-rays from five public datasets. LoVT performs best on 10 of the 18 studied tasks making it the preferred method of choice for localized tasks.

Keywords: Representation learning · Contrastive learning · Text supervision

1 Introduction and Motivation

In medical applications of computer vision, high-quality annotated data is scarce and expensive to acquire, as it typically requires trained physicians to manually label samples [37]. Therefore, the requirement for large labeled datasets can become quite problematic and may limit the applications of deep learning in this

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19809-0_39.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13686, pp. 685–701, 2022.
https://doi.org/10.1007/978-3-031-19809-0_39

field. One approach to overcome this problem is to utilize radiological reports that are paired with medical images. Such reports are produced routinely in clinical practice and are typically written by medical experts (e.g. radiologists). They thus provide a valuable source of semantic information that is available with little additional cost. Rule-based Natural Language Processing (NLP) models like CheXpert [19] extract labels from these reports allowing the automatic creation of large datasets but they also have some significant limitations. Most importantly, such approaches are typically limited to classification tasks. They generate overall labels for reports (and therefore the paired images) but relating these labels to specific image regions is nontrivial so they cannot be used for localized tasks like semantic segmentation or object detection. Also, rule-based NLP models have to be manually created and cannot generalize to different classification tasks or even different report writing styles [19]. Instead of using these reports to generate classification labels, the reports can be utilized directly in the pre-training method, as was first proposed in the ConVIRT method [51]. Here, the semantic information contained in the reports is used as weak supervision to pre-train image models that are then fine-tuned on labeled downstream tasks, where results can be improved or the number of labeled samples can be reduced. We argue that while this approach is quite promising it is not designed for localized downstream tasks. For example, ConVIRT [51] only works on per-sample image representations and does not explicitly provide more localized representations that might be beneficial for localized tasks like semantic segmentation and object detection. In this work, we therefore propose ***Lo****calized representation learning from* ***V****ision and* ***T****ext (LoVT)*, a pre-training method that utilizes the structure of radiological reports (where each sentence typically describes a single property of the image) to pre-train image models for localized tasks. It extends ConVIRT [51] and outperforms it on most localized downstream tasks.

Our contributions are as follows:

- We split each report into sentences and each image into regions (i.e. patches), jointly encode all sentences of the report to get representations per sentence and jointly encode all patches to get region representations.
- We align sentence and region representations using an attention mechanism and local contrastive learning.
- We show that this can be effectively achieved using our novel local contrastive loss that encourages spatial smoothness and sensitivity.
- We evaluate our method trained using MIMIC-CXR [13,22–24] on a downstream evaluation framework [30] with 18 localized tasks on chest X-rays, including object detection and semantic segmentation on five public datasets. We compare it with several self- and text-supervised methods and with transfer from classification in more than 1400 evaluation runs. Our method LoVT proves as the most successful method outperforming all other methods on 10 out of 18 tasks.

2 Related Work

In recent years, contrastive learning [2–4,6,7,11,14,15,17,18,25,29,31,47,50], has become the state-of-the-art approach for self-supervised representation learning on images. It has been successfully applied as pre-training method in medical imaging including downstream tasks such as image classification on chest X-rays [12,41,42].

Most contrastive learning approaches use, unlike our method, only instance-level contrast, i.e. represent each view of the image by a single vector. While the resulting representations are well-suited for global downstream tasks, they are not designed for localized downstream tasks. Therefore, there is a number of recent approaches that use region-level contrast [5,28,32,46,48,49], i.e. they act on representations of image regions. Unlike our method, these methods do not utilize paired text.

Recently however, there is much focus on self-supervised representation learning methods that pre-train image models for downstream tasks by taking advantage of the companion text [8,21,27,33,38,51]. VirTex [8] and ICMLM [38] use image captioning tasks (generative tasks). ConVIRT [51], CLIP [33] and ALIGN [21] on the other hand use multiview contrastive learning [1]. These approaches have been found to be more effective for discriminative downstream tasks [33]. ConVIRT, CLIP, and ALIGN all follow the same general framework where an image and a text encoder are trained jointly using the NT-Xent loss (which is also used in SimCLR) on image and text views. The text views are based on single sentences from companion text, in the case of ConVIRT it is a sentence sampled from the radiology report. The main difference between these methods is the datasets they are studied on, ConVIRT is trained on chest X-rays while the other methods use natural images. Additionally, CLIP uses attention pooling to compute image representations from feature maps while the other methods use the default pooling method from the image encoder (average pooling in the case of ResNet50 [16]). Our method follows a similar framework but adds local contrastive losses for better performance on localized tasks. Also, it encodes the whole report instead of sampling a single sentence and uses attention pooling in the image and text encoders. LocTex [27] does localized pre-training on natural images with companion text and predicts alignment of text and image regions. Unlike our method, it uses supervision generated by mouse gazes instead of learning the alignment implicitly using a local contrastive loss. Most related to our work is the recently published local Mutual Information approach [26] that performs contrastive learning on report sentences and image regions but targets classification instead of localized tasks and does therefore neither encourage contrast between regions nor spatial smoothness.

3 Method

3.1 Assumptions and Intuition

As shown in Fig. 1, a radiology report is typically split into several sections, including a *Findings* section, describing related radiological images, and an *Assessment* section, interpreting the findings. As these sections describe medical aspects observed (*Findings*) in one or more related images and conclusions (*Assessment*) drawn from it, they provide supervision for identifying relevant patterns in the images and interpretations of these patterns. Both sections can be split into sentences and each of these sentences typically describes one or a few aspects of which we assume that most are related to one or a few very localized regions in a paired image. We randomly sample one of the images related to a given report and split it into 7×7 equally-sized regions. More precisely, we augment and resize the image to a size of 224×224, feed it into a convolutional neural network, and use the output feature map of size 7×7 as region representations. A language model encodes the tokens of the report as contextualized (i.e. considering their meaning in the whole report) vector representations from which we compute sentence representations. A many-to-many alignment model is then used to compute *cross-modal representations* from *uni-modal representations*, i.e. image region representations from sentence representations and vice-versa. We argue that by aligning cross-modal and uni-modal representations, the image region representations are encouraged to contain the high-level semantics present in the report.

EXAMINATION: CHEST (PA and LAT)

INDICATION: ___ year old woman with ?pleural effusion

FINDINGS:
Cardiac size cannot be evaluated.
Large left pleural effusion is new.
Small right effusion is new.
The upper lungs are clear.
Right lower lobe opacities are better seen in prior CT.
There is no pneumothorax.
There are mild degenerative changes in the thoracic spine.

IMPRESSION:
Large left pleural effusion.

Fig. 1. Example radiology report describing chest X-Rays. Taken from the MIMIC-CXR [13,23,24] dataset.

3.2 Model Overview

Figure 2 shows the general architecture of our proposed LoVT model. Each training sample $\boldsymbol{x}_i$ is a pair of an image $\boldsymbol{x}_i^{\mathcal{I}} \in \mathbb{R}^{224 \times 224}$ and the related report $\boldsymbol{x}_i^{\mathcal{R}}$

consisting of M_i sentences. Both, $\boldsymbol{x}_i^{\mathcal{I}}$ and $\boldsymbol{x}_i^{\mathcal{R}}$, are encoded independently into two global representations, for image and report respectively, and multiple local representations per sample, corresponding to image regions and report sentences, respectively. An attention-based alignment model then computes cross-modal representations (i.e. sentence representations from image regions and vice-versa) which are aligned with the local uni-modal representations using local contrastive losses. Additionally, the global representations are aligned using a global contrastive loss. The encoders and the alignment model are trained jointly on batches of image-report pairs $\boldsymbol{x}_i$. The details of the model and the loss function will be described in the following sections.

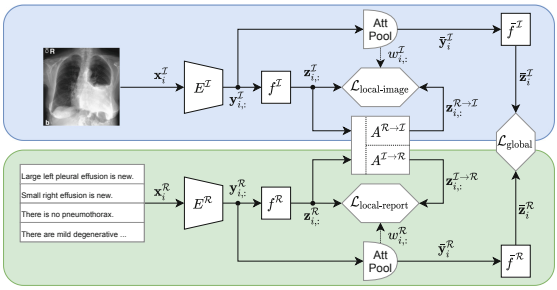

Fig. 2. Architecture of LoVT. Given an image $\boldsymbol{x}_i^{\mathcal{I}}$ and the related report $\boldsymbol{x}_i^{\mathcal{R}}$, the encoders $E^{\mathcal{I}}$ and $E^{\mathcal{R}}$ compute image region and report sentence representations, respectively, which are projected using $f^{\mathcal{I}}$ and $f^{\mathcal{R}}$. The alignment models $A^{\mathcal{R}\to\mathcal{I}}$ and $A^{\mathcal{I}\to\mathcal{R}}$ compute cross-modal report-to-image ($\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$) and image-to-report ($\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$) representations which are aligned with the uni-modal representations ($\boldsymbol{z}_{i,k}^{\mathcal{I}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{R}}$) using the local losses $\mathcal{L}_{\text{local-image}}$ and $\mathcal{L}_{\text{local-report}}$, respectively. Global image ($\bar{\boldsymbol{y}}_i^{\mathcal{I}}$) and report ($\bar{\boldsymbol{y}}_i^{\mathcal{R}}$) representations are computed using attention pooling on the local representations, are then projected using $\bar{f}^{\mathcal{I}}$ and $\bar{f}^{\mathcal{R}}$ and aligned using the global loss $\mathcal{L}_{\text{global}}$.

3.3 Encoding

Each image $\boldsymbol{x}_i^{\mathcal{I}}$ is encoded into $K = H \times W$ (we use $K = 7 \times 7$) region representations $\boldsymbol{y}_{i,k}^{\mathcal{I}} \in \mathbb{R}^{d^{\mathcal{I}}}$ using the image encoder $E^{\mathcal{I}}$, where k is the index of the image region, and $d^{\mathcal{I}}$ is the dimension of the image region representation space. Our approach is encoder agnostic, i.e. any model encoding image regions into vector representations can be used for $E^{\mathcal{I}}$. We use a ResNet50 [16] and take the feature map before global average pooling as region representations. Similarly, each report $\boldsymbol{x}_i^{\mathcal{R}}$ is encoded into M_i sentence representations $\boldsymbol{y}_{i,m}^{\mathcal{R}} \in \mathbb{R}^{d^{\mathcal{R}}}$ using the report encoder $E^{\mathcal{R}}$. Here M_i is the number of sentences of report sample i, m is the index of the sentence, and $d^{\mathcal{R}}$ is the dimension of the report sentence representation space. Note that while K is constant, M_i may be different for each sample. Any model encoding sentences into vector representations can be

used for $E^{\mathcal{R}}$. We use BERT_base [10] to jointly encode the tokens of the concatenated sentences of each report and then perform max pooling over the token representations of each sentence to get sentence representations.

The global (i.e. per-sample) representations $\bar{\boldsymbol{y}}_i^{\mathcal{I}}$ and $\bar{\boldsymbol{y}}_i^{\mathcal{R}}$ are each computed by an attention pooling layer (not shared between modalities) on the region and sentence representations, respectively. It is implemented using multi-head query-key-value attention [44] where the query is computed from the globally averaged region or sentence representations. This pooling approach was first proposed for the image encoder of CLIP [33].

Following previous works [6,14,51], we compute projected local representations $\boldsymbol{z}_{i,k}^{\mathcal{I}} \in \mathbb{R}^{d^{\mathcal{Z}}}$ and $\boldsymbol{z}_{i,m}^{\mathcal{R}} \in \mathbb{R}^{d^{\mathcal{Z}}}$, and projected global representations $\bar{\boldsymbol{z}}_i^{\mathcal{I}} \in \mathbb{R}^{\bar{d}^{\mathcal{Z}}}$ and $\bar{\boldsymbol{z}}_i^{\mathcal{R}} \in \mathbb{R}^{\bar{d}^{\mathcal{Z}}}$ from the representations $\boldsymbol{y}_{i,k}^{\mathcal{I}}$, $\boldsymbol{y}_{i,m}^{\mathcal{R}}$, $\bar{\boldsymbol{y}}_i^{\mathcal{I}}$, and $\bar{\boldsymbol{y}}_i^{\mathcal{R}}$, using the (non-shared) nonlinear transformations $f^{\mathcal{I}}$, $f^{\mathcal{R}}$, $\bar{f}^{\mathcal{I}}$, and $\bar{f}^{\mathcal{R}}$, respectively, where $d^{\mathcal{Z}}$ is the dimension of the shared local and $\bar{d}^{\mathcal{Z}}$ of the shared global representation space (we use 512 for both). Note that for local representations the projections are applied to each region k or sentence m independently.

3.4 Alignment Model

Following our assumptions (see Sect. 3.1), we compute an alignment of image regions and sentences and compute cross-modal representations using the alignment models $A^{\mathcal{I}\to\mathcal{R}}$ and $A^{\mathcal{R}\to\mathcal{I}}$, which are based on single-head query-key-value attention [44].

For each sentence m the cross-modal representation $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$ is computed by letting $\boldsymbol{z}_{i,m}^{\mathcal{R}}$ attend to all image region representations $\boldsymbol{z}_{i,k}^{\mathcal{I}}$ (of the related image). We therefore compute the probability $\alpha_{i,m,k}^{\mathcal{I}\to\mathcal{R}}$ that sentence m is aligned with region k based on the scaled dot product scores of their projected representations, i.e. $\alpha_{i,m,k}^{\mathcal{I}\to\mathcal{R}} = \mathrm{softmax}_k\left(\frac{(\boldsymbol{Q}\boldsymbol{z}_{i,m}^{\mathcal{R}})^T(\boldsymbol{Q}\boldsymbol{z}_{i,k}^{\mathcal{I}})}{\sqrt{d^{\mathcal{Z}}}}\right)$, where the linear query-key projection $\boldsymbol{Q}$ is a learned matrix. Then the alignment model $A^{\mathcal{I}\to\mathcal{R}}$ uses $\alpha_{i,m,k}^{\mathcal{I}\to\mathcal{R}}$ to compute $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$ as projected weighted sum of the image region representations $\boldsymbol{z}_{i,k}^{\mathcal{I}}$:

$$\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}} = \boldsymbol{O}\left(\sum_{k=1}^{K} \alpha_{i,m,k}^{\mathcal{I}\to\mathcal{R}} \left(\boldsymbol{V}\boldsymbol{z}_{i,k}^{\mathcal{I}}\right)\right), \tag{1}$$

where the value projection $\boldsymbol{V}$, and the output projection $\boldsymbol{O}$ are learned matrices.

In a similar fashion the cross-modal representations $\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$ are computed by $A^{\mathcal{R}\to\mathcal{I}}$:

$$\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}} = \boldsymbol{O}\left(\sum_{m=1}^{M_i} \alpha_{i,k,m}^{\mathcal{R}\to\mathcal{I}} \left(\boldsymbol{V}\boldsymbol{z}_{i,m}^{\mathcal{R}}\right)\right), \tag{2}$$

with $\alpha_{i,k,m}^{\mathcal{R}\to\mathcal{I}} = \mathrm{softmax}_m\left(\frac{(\boldsymbol{Q}\boldsymbol{z}_{i,k}^{\mathcal{I}})^T(\boldsymbol{Q}\boldsymbol{z}_{i,m}^{\mathcal{R}})}{\sqrt{d^{\mathcal{Z}}}}\right)$. Note that as $A^{\mathcal{R}\to\mathcal{I}}$ and $A^{\mathcal{I}\to\mathcal{R}}$ share the same matrices $\boldsymbol{Q}$, $\boldsymbol{V}$, and $\boldsymbol{O}$, the only difference between $\alpha_{i,k,m}^{\mathcal{R}\to\mathcal{I}}$ and $\alpha_{i,m,k}^{\mathcal{I}\to\mathcal{R}}$ is transposition and the index over which softmax is applied.

3.5 Loss Function

Global Alignment. For global alignment we follow ConVIRT [51] and maximize the cosine similarity between paired image and report representations while minimizing the similarity between non-paired (i.e. from different samples) representations. The loss consists of a image-report part, where all non-paired report representations from the batch are used as negatives:

$$\ell_{\text{global}}^{\mathcal{I}\|\mathcal{R}} = -\log \frac{e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{I}}, \bar{\boldsymbol{z}}_i^{\mathcal{R}}\right)/\tau}}{\sum_j e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{I}}, \bar{\boldsymbol{z}}_j^{\mathcal{R}}\right)/\tau}}, \tag{3}$$

and a report-image part, defined analogously:

$$\ell_{\text{global}}^{\mathcal{R}\|\mathcal{I}} = -\log \frac{e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{R}}, \bar{\boldsymbol{z}}_i^{\mathcal{I}}\right)/\tau}}{\sum_j e^{\cos\left(\bar{\boldsymbol{z}}_i^{\mathcal{R}}, \bar{\boldsymbol{z}}_j^{\mathcal{I}}\right)/\tau}}, \tag{4}$$

where τ is the similarity temperature (we use 0.1) and all logarithms are natural. Both parts are combined using the hyperparameter $\lambda \in [0, 1]$ (we use 0.75):

$$\mathcal{L}_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} \left[\lambda \cdot \ell_{\text{global}}^{\mathcal{I}\|\mathcal{R}} + (1 - \lambda) \cdot \ell_{\text{global}}^{\mathcal{R}\|\mathcal{I}} \right]. \tag{5}$$

Local Alignment. The global alignment loss does not only align the global representations but it also prevents the global representations from collapsing to a constant vector using negative samples to contrast the positive pairs. Similarly, we propose local alignment losses encouraging spatial (sentence) sensitivity through negatives from the same sample, i.e. preventing the local representations to be similar for all regions (sentences) of an image (report). We use two NT-Xent-based [6] local losses: $\mathcal{L}_{\text{local-image}}$, aligning region representations $\boldsymbol{z}_{i,k}^{\mathcal{I}}$ with $\boldsymbol{z}_{i,k}^{\mathcal{R}\to\mathcal{I}}$, and $\mathcal{L}_{\text{local-report}}$, aligning sentence representations $\boldsymbol{z}_{i,m}^{\mathcal{R}}$ with $\boldsymbol{z}_{i,m}^{\mathcal{I}\to\mathcal{R}}$.

Some regions or sentences may not be relevant for aligning a sample (e.g. background regions or sentences not related to the image). Therefore, we introduce region weights $w_{i,k}^{\mathcal{I}}$ and sentence weights $w_{i,m}^{\mathcal{R}}$, which are computed as the attention probabilities from the respective attention pooling layer (which was used to compute global representations), averaged over all attention heads. These weights are used in the local loss functions such that irrelevant representations do not have to be aligned. Note that we do not backpropagate through the region or sentence weights.

The loss $\mathcal{L}_{\text{local-image}}$ allows for having multiple positive pairs within each sample by giving each pair of regions (k, l) a positiveness probability $p_{k,l}^{\mathcal{I}} \in [0, 1]$. We then treat each positive pair as its own (weighted) example and contrast it with all other pairs (again all logarithms are natural):

$$\ell_{\text{local-image}}^{\mathcal{I}\|\mathcal{R}\to\mathcal{I}} = -\sum_{l=1}^{K} p_{k,l}^{\mathcal{I}} \log \frac{e^{\cos\left(\boldsymbol{z}_{i,k}^{\mathcal{I}}, \boldsymbol{z}_{i,l}^{\mathcal{R}\to\mathcal{I}}\right)/\tau'}}{\sum_{k'} e^{\cos\left(\boldsymbol{z}_{i,k}^{\mathcal{I}}, \boldsymbol{z}_{i,k'}^{\mathcal{R}\to\mathcal{I}}\right)/\tau'}} \tag{6}$$

$$\ell^{\mathcal{R}\to\mathcal{I}\|\mathcal{I}}_{\text{local-image}} = -\sum_{l=1}^{K} p^{\mathcal{I}}_{k,l} \log \frac{e^{\cos\left(\boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{I}}_{i,l}\right)/\tau'}}{\sum_{k'} e^{\cos\left(\boldsymbol{z}^{\mathcal{R}\to\mathcal{I}}_{i,k}, \boldsymbol{z}^{\mathcal{I}}_{i,k'}\right)/\tau'}} \tag{7}$$

$$\mathcal{L}_{\text{local-image}} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{k=1}^{K} w^{\mathcal{I}}_{i,k} \cdot \left[\ell^{\mathcal{I}\|\mathcal{R}\to\mathcal{I}}_{\text{local-image}} + \ell^{\mathcal{R}\to\mathcal{I}\|\mathcal{I}}_{\text{local-image}}\right] . \tag{8}$$

Here τ' is the similarity temperature and is set to 0.3. We assume that nearby image regions are often similar and that therefore nearby regions are more likely to be positives while distant regions are more likely to be negatives. Thus, we define the positiveness probability $p^{\mathcal{I}}_{k,l}$ of two image regions as the complementary cumulative exponential distribution of $d_{\boldsymbol{x}}$ (their spatial ℓ_2-distance in 2D space normalized by the length of the diagonal $\sqrt{H^2 + W^2}$) and set $p^{\mathcal{I}}_{k,l}$ to zero above cutoff threshold $T \in [0, \infty)$:

$$p^{\mathcal{I}}_{k,l} = \frac{\mathbb{1}_{[d_x(k,l) \le T]} \cdot e^{-d_x(k,l)/\beta}}{\sum_{k'} \mathbb{1}_{[d_x(k,k') \le T]} \cdot e^{-d_x(k,k')/\beta}} . \tag{9}$$

Here $\beta \in (0, \infty)$ is a sharpness hyperparameter. We set $\beta = 1$ and $T = 0.5$. Note that the normalization of $d_{\boldsymbol{x}}$ is equal to rescaling T and β, i.e. it allows us to define both hyperparameters independently of the image size.

The definition of $p^{\mathcal{I}}_{k,l}$ is derived by modeling the occurrence of related features at specific distances in the image as a Poisson point process, such that the ℓ_2-distance of related features follows the exponential distribution. We assume a Poisson process due to its property of being memoryless, i.e. knowing that a feature is already related to another feature at some distance does not change how distant additional related features can be found. Also, the probability density function of the exponential distribution is decreasing (with support on the interval $[0, \infty)$), which seems reasonable as it is typically more likely that related features are near than far. Its cumulative distribution function then describes the probability that two related features are within a given radius and its complementary function that of being outside a given radius. The threshold T assures that very distant pairs do not count as positives. The loss $\mathcal{L}_{\text{local-image}}$ thus encourages spatial smoothness of image regions while maintaining spatial sensitivity through negative samples. Note that it is related to the pixel-contrast loss proposed in [49], where the main novelty of our work is the partly smooth definition of $p^{\mathcal{I}}_{k,l}$ based on the exponential distribution.

The local report loss $\mathcal{L}_{\text{local-report}}$ is defined similarly but we do not assume prior knowledge about the similarity of sentences and therefore only have a single positive pair per sentence (again all logarithms are natural):

$$\ell^{\mathcal{R}\|\mathcal{I}\to\mathcal{R}}_{\text{local-report}} = -\log \frac{e^{\cos\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m}\right)/\tau'}}{\sum_{m'} e^{\cos\left(\boldsymbol{z}^{\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m'}\right)/\tau'}} \tag{10}$$

$$\ell^{\mathcal{I}\to\mathcal{R}\|\mathcal{R}}_{\text{local-report}} = -\log \frac{e^{\cos\left(\boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{R}}_{i,m}\right)/\tau'}}{\sum_{m'} e^{\cos\left(\boldsymbol{z}^{\mathcal{I}\to\mathcal{R}}_{i,m}, \boldsymbol{z}^{\mathcal{R}}_{i,m'}\right)/\tau'}} \tag{11}$$

$$\mathcal{L}_{\text{local-report}} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{m=1}^{M_i} w_{i,m}^{\mathcal{R}} \cdot \left[\ell_{\text{local-report}}^{\mathcal{R} \| \mathcal{I} \to \mathcal{R}} + \ell_{\text{local-report}}^{\mathcal{I} \to \mathcal{R} \| \mathcal{R}} \right] \tag{12}$$

Total Loss. The total loss $\mathcal{L}$ is computed as the weighted sum of global and local losses:

$$\mathcal{L} = \gamma \cdot \mathcal{L}_{\text{global}} + \mu \cdot \mathcal{L}_{\text{local-image}} + \nu \cdot \mathcal{L}_{\text{local-report}} \,, \tag{13}$$

where γ, μ, and ν are loss weights to balance the individual losses and are set to 1.0, 0.75, and 0.75, respectively. We determined these loss weights by running small grid searches (see supplementary material for details).

4 Evaluation

4.1 Downstream Tasks and Experimental Setup

We evaluate our method on a downstream evaluation framework [30] with 18 localized tasks on chest X-rays, which we will shortly describe here. For more details, we refer to the supplementary material.

Evaluation Protocols. We only evaluate the pre-trained ResNet50 (from the image encoder). For semantic segmentation tasks we use the following evaluation protocols: (i) **U-Net Finetune**, where the ResNet50 is used as the backbone of a U-Net [35] and is finetuned jointly with all other layers, (ii) **U-Net Frozen**, where the ResNet50 is used as the frozen backbone of a U-Net [35] and only the non-backbone layers are trained, and (iii) **Linear**, where an element-wise linear layer is trained that is applied to the feature map of the frozen ResNet50, and then results are upsampled to the segmentation resolution.

For object detection tasks we use the following protocols: (i) **YOLOv3 Finetune**, where the ResNet50 is used as the backbone of a YOLOv3 [34] model and is finetuned jointly with all other layers, (ii) **YOLOv3 Frozen**, where the ResNet50 is used as the frozen backbone of a YOLOv3 [34] model and only the non-backbone layers are trained, and (iii) **Linear**, where the object detection ground truth is converted to segmentation masks and the *Linear* evaluation protocol is applied.

Downstream Datasets. We evaluate the pre-trained ResNet50 on several medical datasets, namely (i) **RSNA Pneumonia Detection** [39,45], with more than 260000 frontal-view chest X-rays with detection targets for pneumonia opacities. We use the *YOLOv3 Finetune*, *YOLOv3 Frozen*, and *Linear* protocols, each with 1%, 10%, and 100% of the training samples; (ii) **COVID Rural** [9,43], with more than 200 frontal-view chest X-rays with segmentation masks for COVID-19 lung opacity regions. We use the *UNet Finetune*, *UNet Frozen*, and *Linear* protocols; (iii) **SIIM-ACR Pneumothorax Segmentation** [40], with more than 12000 frontal-view chest X-rays with segmentation masks for pneumothorax. We use the *UNet Finetune*, *UNet Frozen* protocols, but due not use *Linear* due to the fine-grained nature of the segmentation masks; (iv) **Object CXR** [20] with

9000 frontal-view chest X-rays with detection targets for foreign objects. We use the *YOLOv3 Finetune*, *YOLOv3 Frozen*, and *Linear* protocols; (v) **NIH CXR** [45], with almost 1000 frontal-view chest X-rays with detection targets for eight pathologies (Atelectasis, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, and Pneumothorax). Due to the limited data per class, we only use the *Linear* protocol. The different evaluation protocols are complementary to each other, where the *U-Net Finetune* and *YOLOv3 Finetune* protocols evaluate how well suited the pre-trained image models are for fine-tuning as used in practical applications and the *Linear* protocols evaluate the quality of learned local representations (i.e. feature maps) while adding only a few parameters and therefore mostly omitting the variance introduced by random initialization during downstream evaluation. The *U-Net Frozen* and *YOLOv3 Frozen* protocols are a trade-off, where representations are frozen but evaluated in a more practical setting (with several randomly initialized layers).

Tuning and Evaluation Procedure. All baselines and our models have been tuned only on a single downstream task, *RSNA YOLOv3 Frozen 10%*, where a single fixed downstream learning rate was used (determined in preliminary experiments) and the results of five runs have been averaged. Other downstream tasks have not been evaluated during tuning to make sure that models are not biased towards the downstream tasks. After tuning, we evaluated each model on all downstream tasks: The learning rates were tuned individually per model and task (using single evaluation runs) before running five evaluations per task (all using the tuned learning rate). We report the average results of these five runs and their 95%-confidence interval (where each evaluation run is considered a sample).

Pre-Training Dataset. We train our method on MIMIC-CXR [13,22–24] (version 2) as, to our best knowledge, it is the largest and most commonly used dataset of this kind. Since all downstream tasks contain only frontal views, we remove all lateral views, such that roughly 21000 training samples remain, each with a report and one or more frontal images.

Baselines. We compare our method against several baseline methods:

- **Random Init.**: The ResNet50 is initialized using its default random initialization
- **ImageNet** [36] **Init.**: The ResNet50 is initialized with weights pre-trained on the ImageNet ILSVRC-2012 task [36];
- **CheXpert** [19]: The ResNet50 is pre-trained using supervised multi-label binary classification with CheXpert [19] labels on frontal chest X-rays of MIMIC-CXR
- **Global image pre-training methods**: The ResNet50 is pre-trained using the self-supervised pre-training methods SimCLR [6] or BYOL [14] on frontal chest X-rays of MIMIC-CXR. We decided to include SimCLR as is uses a similar loss function as LoVT and we include BYOL because of its widespread use.
- **Local image pre-training methods**: The ResNet50 is pre-trained using the self-supervised pre-training method PixelPro [49] on frontal chest X-rays

Table 1. Results on the RSNA pneumonia detection tasks with different training set sizes. All results are averaged over five evaluation runs and the 95%-confidence interval is shown. The best results per task are underlined, the second-best results are dash-underlined and the best results per pre-training category (general initialization, pre-training on 30% and 100%) are highlighted in bold. Note that the *YOLOv3 Frozen 10%* task (task 5) was used for tuning of all methods and may therefore not be representative as methods may overfit on this task.

	RSNA YOLOv3 Finetune mAP (%)			RSNA YOLOv3 Frozen mAP (%)			RSNA Lin. Seg. Dice (%)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
General initialization methods									
Random	2.4 ± 0.5	5.1 ± 1.2	14.9 ± 1.7	1.0 ± 0.2	4.0 ± 0.3	8.9 ± 0.9	21.9 ± 1.2	5.3 ± 0.0	5.3 ± 0.0
ImageNet [36]	**5.0 ± 0.7**	**12.4 ± 0.8**	**19.0 ± 0.2**	**3.6 ± 1.4**	**8.0 ± 0.1**	**15.7 ± 0.3**	**27.5 ± 0.6**	**38.3 ± 0.0**	**43.3 ± 0.0**
Pre-Training on 30 % of frontal MIMIC-CXR									
CheXpert [19]	**8.3 ± 0.8**	12.4 ± 1.6	**21.3 ± 0.3**	7.0 ± 1.0	14.8 ± 0.8	18.8 ± 0.4	38.9 ± 0.2	45.5 ± 0.2	48.1 ± 0.0
BYOL [14]	7.0 ± 1.0	11.9 ± 1.1	18.8 ± 0.2	9.6 ± 0.2	14.0 ± 1.2	**21.0 ± 0.2**	42.9 ± 0.1	47.8 ± 0.2	50.0 ± 0.0
SimCLR [6]	6.7 ± 0.5	**12.9 ± 0.5**	20.4 ± 1.8	7.9 ± 1.0	11.9 ± 0.1	19.9 ± 0.2	43.1 ± 0.0	46.0 ± 0.0	48.2 ± 0.0
PixelPro [49]	4.8 ± 0.6	12.6 ± 1.2	19.8 ± 0.4	3.1 ± 0.2	6.4 ± 0.5	13.4 ± 0.3	25.9 ± 0.2	34.6 ± 0.0	39.8 ± 0.1
ConVIRT [51]	7.4 ± 1.3	12.7 ± 1.5	18.3 ± 0.4	**9.8 ± 0.3**	14.8 ± 1.1	18.4 ± 1.1	42.1 ± 0.1	47.1 ± 0.2	50.2 ± 0.0
CLIP [33]*	7.2 ± 0.8	12.8 ± 1.2	19.7 ± 0.5	9.3 ± 0.4	16.1 ± 1.1	19.6 ± 1.4	44.3 ± 0.1	48.8 ± 0.1	50.7 ± 0.0
LoVT (Ours)	7.7 ± 1.0	11.7 ± 0.5	17.2 ± 1.3	8.6 ± 1.5	**17.9 ± 0.4**	18.0 ± 0.1	**46.0 ± 0.0**	**49.4 ± 0.0**	**51.5 ± 0.0**
Pre-Training on 100 % of frontal MIMIC-CXR									
CheXpert [19]	10.0 ± 1.9	12.4 ± 0.9	**22.2 ± 0.4**	5.8 ± 0.4	11.9 ± 0.7	20.0 ± 0.2	40.0 ± 0.1	44.3 ± 0.0	46.9 ± 0.0
BYOL [14]	5.6 ± 0.8	11.0 ± 0.2	17.3 ± 1.1	6.8 ± 1.6	12.1 ± 1.1	15.9 ± 0.6	41.9 ± 0.0	45.1 ± 0.0	46.8 ± 0.0
SimCLR [6]	7.1 ± 0.7	12.2 ± 0.8	18.8 ± 1.0	5.4 ± 0.2	13.1 ± 0.2	17.3 ± 1.6	43.0 ± 0.0	45.1 ± 0.0	47.0 ± 0.0
PixelPro [49]	4.8 ± 0.3	11.0 ± 1.5	17.4 ± 1.7	4.6 ± 1.6	5.4 ± 1.1	12.6 ± 1.3	23.9 ± 0.4	34.8 ± 0.2	40.2 ± 0.1
ConVIRT [51]	**10.7 ± 1.1**	**13.3 ± 0.8**	18.5 ± 0.4	8.2 ± 0.9	15.6 ± 1.2	17.9 ± 0.3	44.6 ± 0.1	48.5 ± 0.0	50.4 ± 0.3
CLIP [33]*	7.0 ± 1.5	10.7 ± 1.1	19.9 ± 0.8	**11.9 ± 0.7**	15.0 ± 1.1	18.7 ± 0.0	45.2 ± 0.0	49.3 ± 0.1	51.1 ± 0.0
LoVT (Ours)	8.5 ± 0.8	13.2 ± 0.6	18.1 ± 3.2	9.6 ± 1.2	**16.4 ± 1.3**	**20.5 ± 1.0**	**46.3 ± 0.0**	**50.1 ± 0.0**	**51.8 ± 0.0**
Task Nr.	1	2	3	4	5	6	7	8	9

* Modified to use the same image and text encoders as ConVIRT and LoVT.

of MIMIC-CXR. We include PixelPro to study the effect of local contrastive losses when using only images.

– **Global image-text pre-training methods**: The ResNet50 is pre-trained using the image-text methods ConVIRT [51] or CLIP [33] on frontal MIMIC-CXR. Note that for comparability we adapted CLIP to use the same image and text encoders as ConVIRT such that the main difference between CLIP and ConVIRT is that CLIP uses attention pooling to compute the scan representation while ConVIRT uses average pooling. We include both methods as LoVT builds upon a similar general framework, where we include ConVIRT because it targets chest X-rays (like LoVT) and include CLIP because of its widespread use and as it uses (like LoVT) attention pooling in the image encoder. We decided not to include VirTex [8] and ICMLM [38] as they use generative tasks, which have been found to be less effective for discriminative downstream tasks [33].

4.2 Downstream Results

We present the downstream results of our model LoVT and the baselines, with pre-training on 100% and 30% of MIMIC-CXR. Table 1 shows the results on

Table 2. Results on downstream tasks on the COVID Rural, SIIM Pneumothorax, Object CXR, and NIH CXR datasets. All results are averaged over five evaluation runs and the 95%-confidence interval is shown. The best results per task are underlined, the second-best results are dash-underlined and the best results per pre-training category (general initialization, pre-training on 30% and 100%) are highlighted in bold.

	COVID Rural			SIIM-ACR Pneumoth.		Object CXR			NIH CXR
	UNet Finetune	UNet Frozen	Linear	UNet Finetune	UNet Frozen	YOLOv3 Finetune	YOLOv3 Frozen	Linear	Linear
	Dice (%)	Dice (%)	Dice (%)	Dice (%)	Dice (%)	fROC (%)	fROC (%)	Dice (%)	Avg Dice (%)
General initialization methods									
Random	34.0 ± 1.1	32.2 ± 1.8	6.0 ± 0.0	23.2 ± 1.0	23.9 ± 1.6	49.5 ± 1.2	28.4 ± 1.4	6.9 ± 0.0	0.5 ± 0.4
ImageNet [36]	**43.9 ± 2.0**	**41.9 ± 1.7**	**32.6 ± 0.7**	**38.5 ± 0.9**	**36.9 ± 0.7**	**62.5 ± 0.4**	**52.7 ± 1.3**	**37.8 ± 0.0**	**2.6 ± 1.6**
Pre-Training on 30 % of frontal MIMIC-CXR									
CheXpert [19]	43.5 ± 4.9	44.1 ± 3.2	32.1 ± 2.0	38.9 ± 0.9	40.7 ± 0.7	62.2 ± 0.6	46.3 ± 1.9	16.5 ± 7.7	8.7 ± 0.6
BYOL [14]	46.2 ± 1.6	47.5 ± 1.6	36.9 ± 1.7	43.1 ± 0.6	42.9 ± 0.3	59.6 ± 1.0	55.7 ± 1.0	32.3 ± 0.1	6.0 ± 0.1
SimCLR [6]	44.9 ± 2.9	41.4 ± 3.7	33.0 ± 0.0	42.6 ± 0.4	39.2 ± 0.7	61.9 ± 0.8	54.3 ± 1.0	33.2 ± 0.1	13.3 ± 0.5
PixelPro [49]	47.0 ± 3.4	38.5 ± 3.9	26.6 ± 0.4	39.3 ± 0.8	39.1 ± 0.3	**63.1 ± 0.7**	46.3 ± 0.2	29.9 ± 0.2	1.8 ± 0.0
ConVIRT [51]	48.8 ± 2.2	44.2 ± 3.1	45.0 ± 3.0	42.5 ± 1.0	42.5 ± 0.2	62.5 ± 0.1	54.0 ± 0.7	37.7 ± 0.1	11.4 ± 0.8
CLIP [33]*	49.3 ± 2.0	46.5 ± 2.3	46.2 ± 0.3	42.8 ± 1.5	42.5 ± 0.6	62.9 ± 0.8	55.5 ± 2.1	**39.0 ± 0.0**	12.5 ± 1.0
LoVT (Ours)	**49.5 ± 1.3**	**49.2 ± 4.6**	**49.2 ± 0.2**	**43.4 ± 0.7**	**43.1 ± 0.6**	61.0 ± 1.3	**55.8 ± 1.1**	37.6 ± 0.2	**13.4 ± 0.8**
Pre-Training on 100 % of frontal MIMIC-CXR									
CheXpert [19]	46.2 ± 1.7	45.9 ± 3.9	37.7 ± 0.4	34.2 ± 0.8	37.7 ± 0.3	57.5 ± 1.1	39.8 ± 2.4	19.4 ± 0.1	15.2 ± 0.0
BYOL [14]	50.7 ± 2.7	42.0 ± 3.0	32.9 ± 0.0	42.6 ± 0.7	40.7 ± 0.7	60.6 ± 1.1	53.1 ± 0.8	21.8 ± 0.1	5.7 ± 0.0
SimCLR [6]	48.1 ± 2.5	44.1 ± 2.1	35.3 ± 0.0	41.2 ± 0.8	38.7 ± 0.5	61.1 ± 0.7	48.7 ± 0.5	30.0 ± 0.0	11.8 ± 0.0
PixelPro [49]	42.4 ± 4.4	37.7 ± 1.0	18.9 ± 6.4	39.4 ± 1.2	38.7 ± 0.6	**65.0 ± 0.5**	46.2 ± 1.2	29.7 ± 0.1	1.8 ± 0.0
ConVIRT [51]	47.9 ± 0.7	46.0 ± 1.1	42.7 ± 2.0	39.3 ± 0.3	43.1 ± 0.3	60.6 ± 1.2	52.5 ± 1.0	36.0 ± 0.0	**18.6 ± 0.1**
CLIP [33]*	48.6 ± 2.4	45.8 ± 4.1	41.7 ± 0.1	44.0 ± 0.7	**45.0 ± 0.5**	62.8 ± 0.5	56.9 ± 1.4	39.4 ± 0.0	11.4 ± 0.8
LoVT (Ours)	**51.2 ± 2.5**	**46.2 ± 2.4**	**44.0 ± 0.8**	**44.1 ± 0.3**	43.9 ± 0.7	62.1 ± 0.5	**57.4 ± 0.5**	**39.9 ± 0.0**	9.4 ± 0.5
Task Nr.	10	11	12	13	14	15	16	17	18

* Modified to use the same image and text encoders as ConVIRT and LoVT.

different subsets of the RSNA dataset and Table 2 shows the results on the remaining downstream datasets, i.e. on COVID Rural, SIM-ACR Pneumothorax, Object CXR, and NIH CXR.

Comparison of Methods. We found that there is no single pre-training method performing best on all evaluated downstream tasks. On most tasks (15 out of 18) image-text self-supervised methods (i.e. LoVT, CLIP, or ConVIRT) outperform the other methods, such that they should be preferred if paired text is available.

Our model LoVT is the best method (over all pre-training settings) on 10 of 18 tasks, and significantly outperforms all other methods in 6 of these tasks, while the second-best method CLIP significantly outperforms all other methods only on 2 tasks. LoVT outperforms image-only methods (i.e. BYOL, SimCLR, and PixelPro) on 14 tasks, where the localized image-only method PixelPro outperforms LoVT only on one task (task 15). On 11 tasks LoVT outperforms other text-supervised methods (i.e. ConVIRT and CLIP), on 14 tasks it outperforms CheXpert classification and on all but two tasks it outperforms ImageNet initialization. When using 100% of the pre-training data LoVT is the best pre-training method on 11 tasks (better by at least the confidence interval on 5 tasks) and when using 30% on 11 tasks (significantly the best on 4 tasks). LoVT performs

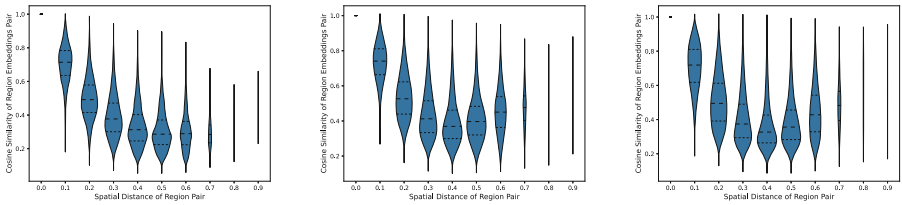

Fig. 3. Spatial smoothness and sensitivity of image region representations. **Left**: LoVT (Ours). **Middle**: No local losses. **Right**: No local losses and no attention pooling. Cosine similarities of image region pairs $\boldsymbol{y}^{\mathcal{I}}_{i,k}, \boldsymbol{y}^{\mathcal{I}}_{i,k'}$ (each from the same sample) plotted as violin plots (with their width representing the number of pairs and quartiles shown as dashed lines) over their spatial distance in the 7×7 image space (normalized and rounded to one decimal digit). We trained all models on 30% of the data and computed the representations on the test set.

best on all COVID Rural tasks, best on most *Linear* tasks, and quite well on the *Frozen* protocol, but does not perform well on the NIH CXR dataset and when finetuned on the RSNA dataset. As there is no single method performing best on all tasks and LoVT performs best in the majority of tasks, this makes LoVT the default method of choice for localized downstream tasks.

Relevance of Pre-Training Dataset Size. We do not observe a consistent benefit of using roughly 210000 pre-training samples (i.e. 100% of the data) over using roughly 63000 samples (i.e. 30%). While on some datasets like RSNA and Object CXR many methods often perform better when pre-trained on 210000 samples (100%), on other datasets like COVID Rural, methods often perform better when pre-trained on 63000 samples (30%). When comparing LoVT pre-trained on 30% of the data with other methods pre-trained in both settings (i.e. 30% and 100%), we observe that LoVT outperforms image-only methods (i.e. BYOL, SimCLR, and PixelPro) on 12 tasks, other text-supervised methods (i.e. ConVIRT and CLIP) on 7 tasks and CheXpert classification on 12 tasks, showing that LoVT effectively reduces the number of required pre-training samples.

Relevance of Downstream Dataset Size. The results shown in Table 1 suggest that, as expected, larger downstream training sets lead to better results. However, we observe that for text-supervised methods (i.e. LoVT, CLIP, and ConVIRT), the downstream training set size is often less relevant compared to other methods. On the *RSAN YOLOv3 Frozen* tasks, LoVT (100%) outperforms ImageNet initialization by 31% when using 100% of the downstream samples, while it outperforms ImageNet initialization by even 167% when only using 1% of the samples.

Spatial Smoothness and Sensitivity. We analyze the influence of the local losses and attention pooling on the spatial smoothness and sensitivity of image region representations and therefore plot in Fig. 3 the distributions of the cosine similarity of image region pairs over their spatial distances. For our LoVT model spatial smoothness and sensitivity can be observed as the quartiles and extreme

points of the cosine similarity distributions decrease monotonously with increasing spatial distance, except for a few very distant region pairs with distances larger than 0.6. Note that these spatially very distant region pairs very likely represent opposite borders (or corners) of the image such that they both very likely contain background, explaining that they have more similar representations. Without local losses $\mathcal{L}_{\text{local-image}}$ and $\mathcal{L}_{\text{local-report}}$, the quartiles and extreme points decrease only for small spatial distances while increasing again for points further away, showing that spatial smoothness is only present for nearby regions and spatial sensitivity of more distant region is not optimal. When additionally replacing attention pooling with average (for image regions) and max (for sentences) pooling, similar results can be observed except that the quartiles are decreasing faster and the maximum points do not decrease for nearby regions. We can therefore deduce that the local losses effectively encourage spatial smoothness and sensitivity while attention pooling alone has only little effect.

Analysis of LoVT and Ablation Study. We refer to the supplementary material for a detailed analysis of our method LoVT, including an ablation study (focusing on local weighting, global and local losses, and attention pooling), an analysis of the distribution and alignment of learned representations, and an analysis of the region weights $w_{i,k}^{\mathcal{I}}$.

5 Discussion

Limitations of Our Evaluation Procedure. In the evaluation procedure, we did not apply extensive hyperparameter tuning, resized all inputs to a resolution of only 224×224, and applied no data augmentation. The presented downstream results are therefore below results typically reported on these datasets. We followed [30] and kept the evaluation procedure simple to limit computational resources and avoid bias induced by tuning to allow for a fair comparison of our method with the baselines.

Limitations of LoVT. LoVT learns its alignment model implicitly based only on latent representations and instance-level pairing information. This makes the model sensitive to hyperparameters and hard to train. Also, it only uses local negatives from the same sample which restricts the number of negatives and may therefore limit its performance. Additionally, the alignment model is restricted to a simple attention mechanism and the regions are based on fixed patches that are not adaptive to the contents of the image. This may restrict the capabilities of the model and therefore of the pre-training method. For a detailed discussion of these limitations as well as of the potential negative societal impact we refer to the supplementary material.

Conclusion. We study pre-training for localized medical imaging on chest X-rays and propose a novel text-supervised method called LoVT, that combines instance-level contrastive learning with local contrastive learning. We evaluate our method on 18 localized tasks on chest X-rays and compare it with typically used pre-training and initialization methods. While there is no single best

method for all tasks, our method LoVT is the best method on 10 out of 18 studied tasks making it the method of choice for localized tasks.

We hope that our work provides valuable insights that encourage using pre-training for localized medical imaging and that our method inspires future work on localized text-supervised pre-training.

References

1. Bachman, P., Hjelm, R., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: NeurIPS (2019)
2. Bardes, A., Ponce, J., LeCun, Y.: VICReg: variance-invariance-covariance regularization for self-supervised learning. In: ICLR (2022)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
4. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: ICCV, pp. 9630–9640 (2021). https://doi.org/10.1109/ICCV48922.2021.00951
5. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. In: NeurIPS (2020)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
7. Chen, X., He, K.: Exploring simple Siamese representation learning. In: CVPR, pp. 15745–15753 (2021). https://doi.org/10.1109/CVPR46437.2021.01549
8. Desai, K., Johnson, J.: VirTex: learning visual representations from textual annotations. In: CVPR, pp. 11157–11168 (2021). https://doi.org/10.1109/CVPR46437.2021.01101
9. Desai, S., et al.: Data from chest imaging with clinical and genomic correlates representing a rural COVID-19 positive population [data set]. The Cancer Imaging Archive (2020). https://doi.org/10.7937/tcia.2020.py71-5978
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
11. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: ICML, pp. 3015–3024 (2021)
12. Gazda, M., Plavka, J., Gazda, J., Drotár, P.: Self-supervised deep convolutional neural network for chest x-ray classification. IEEE Access, 151972–151982 (2021). https://doi.org/10.1109/ACCESS.2021.3125324
13. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation [Online] **101**(23), 215–220 (2000)
14. Grill, J.B., et al.: Bootstrap your own latent - a new approach to self-supervised learning. In: NeurIPS (2020)
15. He, K., Fan, H., Wu, Y., et al.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
17. Hjelm, R.D., et al.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019)

18. Hénaff, O.J., Srinivas, A., et al.: Data-efficient image recognition with contrastive predictive coding. In: ICML, pp. 4182–4192 (2019)
19. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI, pp. 590–597 (2019)
20. JF-Healthcare: object-CXR - automatic detection of foreign objects on chest x-rays. MIDL (2020). https://jfhealthcare.github.io/object-CXR/
21. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
22. Johnson, A., Lungren, M., Peng, Y., et al.: MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). PhysioNet (2019). https://doi.org/10.13026/8360-t248
23. Johnson, A., Pollard, T., Berkowitz, S., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(317) (2019). https://doi.org/10.1038/s41597-019-0322-0
24. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR database (version 2.0.0). PhysioNet (2019). https://doi.org/10.13026/C2JT1Q
25. Li, J., Zhou, P., Xiong, C., Hoi, S.C.H.: Prototypical contrastive learning of unsupervised representations. In: ICLR (2021)
26. Liao, R., et al.: Multimodal representation learning via maximization of local mutual information. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 273–283. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_26
27. Liu, Z., Stent, S., Li, J., Gideon, J., Han, S.: LocTex: learning data-efficient visual representations from localized textual supervision. In: ICCV, pp. 2147–2156 (2021). https://doi.org/10.1109/ICCV48922.2021.00217
28. Mahendran, A., Thewlis, J., Vedaldi, A.: Cross pixel optical-flow similarity for self-supervised learning. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11365, pp. 99–116. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20873-8_7
29. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. In: CVPR, pp. 6706–6716 (2020). https://doi.org/10.1109/CVPR42600.2020.00674
30. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Radiological reports improve pre-training for localized imaging tasks on chest x-rays. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 647–657. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_62
31. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748 (2019)
32. Pinheiro, P.O., Almahairi, A., Benmalek, R.Y., Golemo, F., Courville, A.: Unsupervised learning of dense visual representations. In: NeurIPS (2020)
33. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021)
34. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv: 1804.02767 (2018)
35. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
36. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

37. Saraf, V., Chavan, P., Jadhav, A.: Deep learning challenges in medical imaging. In: Vasudevan, H., Michalas, A., Shekokar, N., Narvekar, M. (eds.) Advanced Computing Technologies and Applications. AIS, pp. 293–301. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-3242-9_28
38. Sariyildiz, M.B., Perez, J., Larlus, D.: Learning visual representations with caption annotations. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12353, pp. 153–170. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58598-3_10
39. Shih, G., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiol. Artif. Intell. **1** (2019). https://doi.org/10.1148/ryai.2019180041
40. Society for Imaging Informatics in Medicine: SIIM-ACR pneumothorax segmentation (2019). https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation
41. Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P.: MoCo pretraining improves representation and transferability of chest x-ray models. In: MIDL (2021)
42. Sriram, A., et al.: COVID-19 prognosis via self-supervised representation learning and multi-image prediction. arXiv preprint arXiv: 2101.04909 (2021)
43. Tang, H., Sun, N., Li, Y.: Segmentation model of the opacity regions in the chest X-rays of the COVID-19 patients in the us rural areas and the application to the disease severity. medRxiv (2020). https://doi.org/10.1101/2020.10.19.20215483
44. Vaswani, A., et al.: Attention is all you need. In: NIPS (2017)
45. Wang, X., Peng, Y., Lu, L., et al.: ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR, pp. 3462–3471 (2017). https://doi.org/10.1109/CVPR.2017.369
46. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: CVPR, pp. 3023–3032 (2021). https://doi.org/10.1109/CVPR46437.2021.00304
47. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR, pp. 3733–3742 (2018). https://doi.org/10.1109/CVPR.2018.00393
48. Xie, E., et al.: DetCo: unsupervised contrastive learning for object detection. In: ICCV, pp. 8372–8381 (2021). https://doi.org/10.1109/ICCV48922.2021.00828
49. Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., Hu, H.: Propagate yourself: exploring pixel-level consistency for unsupervised visual representation learning. In: CVPR, pp. 16679–16688 (2021). https://doi.org/10.1109/CVPR46437.2021.01641
50. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: self-supervised learning via redundancy reduction. In: ICML (2021)
51. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv: 2010.00747 (2020)