OTCXR: Rethinking Self-supervised Alignment using Optimal Transport for Chest X-ray Analysis

Vandan Gorade[1,†], Azad Singh[2,†], Deepak Mishra[2]
[1]Northwestern University
[2]Indian Institute of Technology Jodhpur
vandan.gorade@northwestern.edu, singh.63@iitj.ac.in, dmishra@iitj.ac.in
[†]These authors contributed equally to this work.

1. Failure Case.

Fig. 1 presents the failure case where all the baseline approaches and OTCXR failed, including the proposed one.

2. Visualization of the Transport Plan

Fig. 2 presents the visualization of the transport plan and the cost matrix for OTCXR.

3. Motivation for the R_s and R_t.

The motivation for the R_s and R_t is to learn μ and ν over the visual feature space across diverse viewpoints. Without R_s and R_t, μ and ν are traditionally initialized as uniform distributions (L284-298), which treat all pixels equally, including irrelevant background pixels. However, using a cross-view attention mechanism, the proposed CV-SIM module (sec.3(2)) addresses this limitation by capturing intricate dependencies across different viewpoints. This enables dynamic focusing on discriminative pixels, ensuring that important regions are aligned effectively (sec.3(3)). Discarding the CV-SIM module (Rs and Rt) would cause the model to over-represent irrelevant features(sec.4.1(Fig-2)), leading to poor performance (sec.4.1(Table-3)). Further, Table 3 presents the results with uniform μ and ν (without CV-SIM module), and we observe a considerable degradation in the overall performance. Therefore, the CV-SIM module combined with OT is clinically relevant as it aids in capturing clinically significant information from medical images, enhancing diagnostic accuracy.

4. Superiority over naively integrating OT in SSL framework.

Simply integrating OT and contrastive learning involves aligning probits and logits simultaneously. In contrast, we propose an effective reformulation of the OT problem within the context of SSL to achieve dense semantic invariance by introducing the novel CV-SIM module that utilizes a multi-head cross-view attention mechanism to extract subtle relationships and dependencies across different viewpoints, leading to the initialization of μ and ν distributions (Section 3(3)). Furthermore, unlike existing methods such as DenseCL, SimCLR, and PCRLv2, we shift the focus from traditional pixel-wise differences to dense feature maps, thereby capturing more meaningful semantic relationships and spatial information. Finally, the transport plan is computed using Sinkhorn's al- algorithm, which is easily adaptable to batches of samples with varying lengths, enabling GPU-friendly computations. Importantly, all of Sinkhorn's operations are differentiable, optimizing the embedding with SGD, making it compatible with deep learning frameworks, and reducing computational complexity. However, we agree that OT introduces some computational overhead, but this is during pre-training only. Meanwhile, there is no additional computational overhead in the downstream phase and for inference.

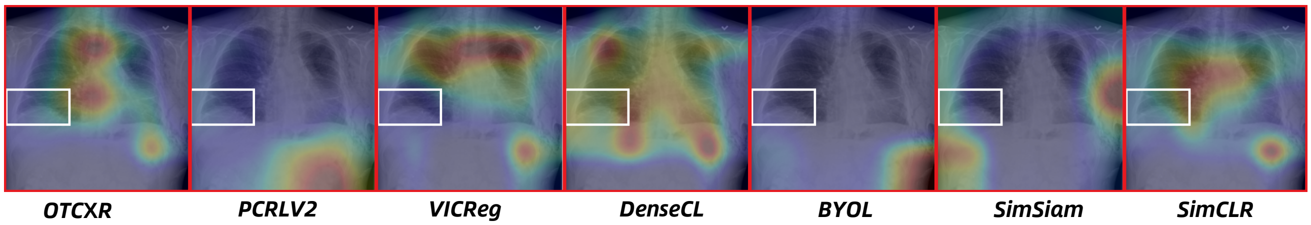

Figure 1. Diagnostic heatmaps for OTCXR and the baseline methods in addition to that in Figure 2 of the manuscript.

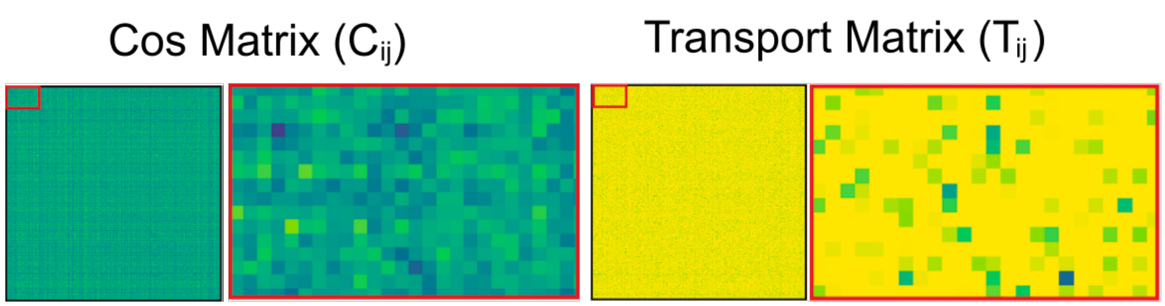

Figure 2. Aquired mean transport plan and cost matrix correspond to the samples in a batch after pre-training of 100 epochs.