Knowledge Boosting: Rethinking Medical Contrastive Vision-Language Pre-training

Xiaofei Chen[1], Yuting He[1], Cheng Xue[1], Rongjun Ge[2], Shuo Li[3], and Guanyu Yang[1,4,5](✉)

[1] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Dhaka, Bangladesh
`yang.list@seu.edu.cn`

[2] Nanjing University of Aeronautics and Astronautics, Nanjing, China

[3] Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA

[4] Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing 210096, China

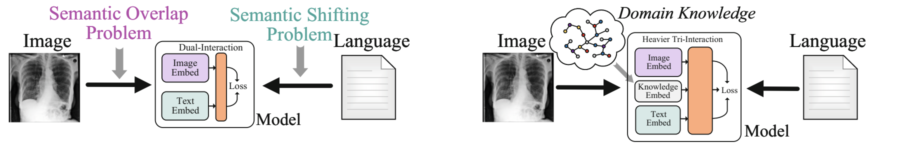
[5] Centre de Recherche en Information Biomédicale Sino-Français (CRIBs), Nanjing, China

Abstract. The foundation models based on pre-training technology have significantly advanced artificial intelligence from theoretical to practical applications. These models have facilitated the feasibility of computer-aided diagnosis for widespread use. Medical contrastive vision-language pre-training, which does not require human annotations, is an effective approach for guiding representation learning using description information in diagnostic reports. However, the effectiveness of pre-training is limited by the large-scale semantic overlap and shifting problems in medical field. To address these issues, we propose the **Kno**wledge-**Bo**osting Contrastive Vision-Language Pre-training framework (KoBo), which integrates clinical knowledge into the learning of vision-language semantic consistency. The framework uses an unbiased, open-set sample-wise knowledge representation to measure negative sample noise and supplement the correspondence between vision-language mutual information and clinical knowledge. Extensive experiments validate the effect of our framework on eight tasks including classification, segmentation, retrieval, and semantic relatedness, achieving comparable or better performance with the zero-shot or few-shot settings. Our code is open on https://github.com/ChenXiaoFei-CS/KoBo.

1 Introduction

Foundation models have become a significant milestone in artificial intelligence, from theoretical research to practical applications [2], like world-impacting large language model ChatGPT [5] and art-history-defining large generative model

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 405–415, 2023.
https://doi.org/10.1007/978-3-031-43907-0_39



(a) Existing Constative Vision-Language Paradigm (b) Knowledge-Boosting Constative Vision-Language Paradigm

Fig. 1. Our knowledge boosting innovates the paradigm of medical vision-language contrastive learning, inspired by two problems in the existing architecture.

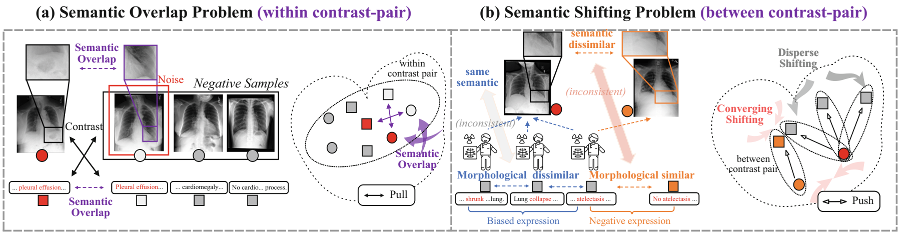


Fig. 2. Two key challenges in medical contrastive vision-language pre-training: **(a)** Semantic overlap exists between negative samples, falsely pulling apart samples with similar semantics. **(b)** Biased expression and negative expression of radiologists cause the inconsistency of semantics and text morphology between sample pairs, causing disperse and converging semantic shifting.

DALL-E [20]. In medical image analysis, foundation models are showing promising future, and pre-training technologies [3,4,8], as the cornerstone of foundation models, facilitated feasibility of computer-aided diagnosis for widespread use.

Medical contrastive vision-language pre-training [10,15,21,23,25] has shown great superiority in medical image analysis, because it utilizes easy-accessible expert interpretation from reports to precisely guide the understanding of image semantics. Therefore, contrastive vision-language pre-training will break through the bottleneck of time-consuming and expensive expert annotation [26] and difficulty in learning fine-grained clinical features with pure-image self-supervised methods [28]. It will improve data efficiency, and achieve comparable or better performance when transferred with the zero-shot or few-shot setting, demonstrating the potential of promoting the ecology of medical artificial intelligence.

However, semantic overlap and semantic shifting are two significant challenges in medical vision-language contrastive learning (Fig. 2). **(a) Semantic Overlap Problem:** There is overlapping semantics between negative samples which should be semantic-distinct, e.g. two medical images sharing the same disease are contrasted which brings noise [25]. Once directly learning, cross-modal representations of the same disease are falsely pulled apart, making the model unable to capture the disease-corresponding image feature. **(b) Semantic Shifting Problem:** Radiologists have writing preferences, e.g. biased for their own familiar concepts and observation view towards similar visual features, and inclined for negation expression towards opposite visual features. Distinct

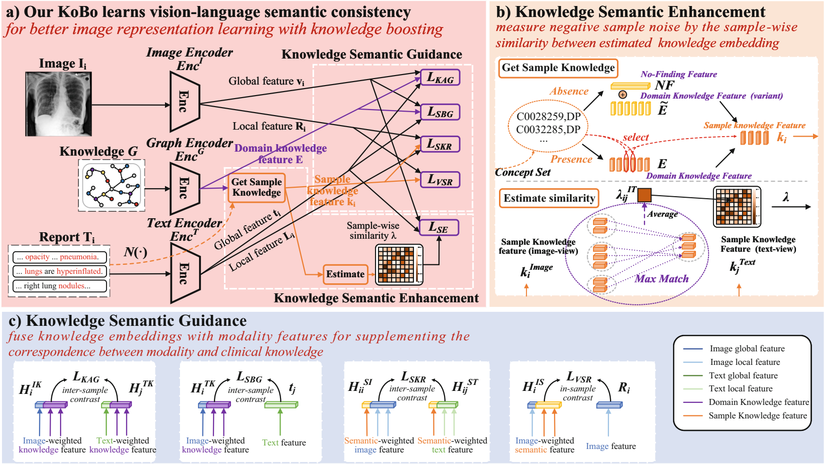

a) Our KoBo learns vision-language semantic consistency
for better image representation learning with knowledge boosting
Image I_i
Image Encoder Enc^I
Knowledge Semantic Guidance
Enc
Global feature v_i
Local feature R_i
Knowledge G
Graph Encoder Enc^G
Domain knowledge feature E
Get Sample Knowledge
Sample knowledge feature k_i
Text Encoder Enc^T
Report T_i
... opacity ... pneumonia.
... lungs are hyperinflated.
... right lung nodules...
$N(\cdot)$
Global feature t_i
Local feature L_i
Sample-wise similarity λ
Estimate
L_{KAG}
L_{SBG}
L_{SKR}
L_{VSR}
L_{SE}
Knowledge Semantic Enhancement
b) Knowledge Semantic Enhancement
measure negative sample noise by the sample-wise similarity between estimated knowledge embedding
Get Sample Knowledge
No-Finding Feature
NF
Domain Knowledge Feature (variant)
$\widetilde{E}$
Absence
C0028259,DP
C0032285,DP
...
Presence
Concept Set
select
E
Domain Knowledge Feature
Sample knowledge Feature
k_i
Estimate similarity
λ_{ij}^{IT}
Average
λ
Sample Knowledge feature (image-view)
k_i^{Image}
Max Match
Sample Knowledge Feature (text-view)
k_j^{Text}
c) Knowledge Semantic Guidance
fuse knowledge embeddings with modality features for supplementing the correspondence between modality and clinical knowledge
H_i^{IK}
L_{KAG}
inter-sample contrast
H_j^{TK}
Image-weighted knowledge feature
Text-weighted knowledge feature
H_i^{TK}
L_{SBG}
inter-sample contrast
t_j
Image-weighted knowledge feature
Text feature
H_{ii}^{SI}
L_{SKR}
inter-sample contrast
H_{ij}^{ST}
Semantic-weighted image feature
Semantic-weighted text feature
H_i^{IS}
L_{VSR}
in-sample contrast
R_i
Image-weighted semantic feature
Image feature
Image global feature
Image local feature
Text global feature
Text local feature
Domain Knowledge feature
Sample Knowledge feature

Fig. 3. Overview of our proposed architecture, where additional clinical knowledge is embedded in. Image encoder, text encoder, graph encoder, knowledge semantic enhancement module, and knowledge semantic guidance module are presented.

concepts describing the same image are morphologically dissimilar for text encoder, while the negation expression of concepts is morphologically similar [17]. Once lack of concept correlation and negation identification, representations with similar semantics are falsely pushed apart and those with opposite semantics are falsely pushed together, interfering with the learning of significant representation [7].

Rethinking the existing methods and challenges of medical contrastive vision-language pre-training [10,21,23,25,26], the lack of clinical knowledge constraints in dual-free-encoding contrastive learning structure is the key problem. Existing methods utilize sample-wise differences to learn mutual information between modalities, improving the representation quality based on the correspondence of learned mutual information and clinical knowledge. However, semantic overlap reduces the learning efficiency of mutual information with the noisy difference, and the mentioned correspondence is vulnerable to semantic shifting. Therefore, if we are able to embed an unbiased, comprehensive representation as knowledge boosting, it will reduce the negative noise and supplement the lacking correspondence. It motivates us to measure the noise with similarities between knowledge representation, and fuse the correspondence between knowledge and modality.

In this paper, we propose a novel knowledge-boosting medical contrastive vision-language pre-training framework (KoBo). Our contributions are as followed. **1)** Our KoBo pre-trains a powerful image encoder including visual information corresponding with the disease described in texts, where knowledge is embedded in our paradigm (Fig. 1) to boost the learning of vision-language consistency. **2)** We propose Knowledge Semantic Enhancement (KSE) module to reduce the negative sample noise with the similarity between open-set sample-wise knowledge embeddings. **3)** We propose Knowledge Semantic Guidance (KSG) module to adjust the semantic shifting during pre-training, fusing the modality feature with unbiased knowledge embeddings for supplementing the correspondence between modality mutual information and clinical knowledge.

2 Methodology

Our Knowledge-Boosting Contrastive Vision-Language Pre-training framework (Fig. 3) boosts vision-language learning with additional clinical knowledge. It contains two modules: KSE for reducing the negative effect of semantic overlap, and KSG for adjusting semantic shifting, aimed at learning effective representation by maximizing semantic consistency between paired image and text features.

2.1 Framework Formulation

In the framework, a powerful image encoder Enc^I and text encoder Enc^T is pre-trained, alongside a graph encoder Enc^G. Given a pair of medical image and diagnostic report $\{I_i, T_i^{Report}\}, I_i \in \mathbb{R}^{H \times W \times C}$, a sentence T_i^{Sent} is randomly selected from T_i^{Report} as a caption comprised of several tokens $\{w_1, w_2, ..., w_{N_L}\}$. Enc^I outputs global feature $z_i^{I,G}$ and local feature $z_i^{I,L}$ for N_I sub-regions, which is from the intermediate feature map. T_i^{Sent} is fed into Enc^T, obtaining global sentence feature $z_i^{T,G}$, and local token feature $z_i^{T,L}$. Distinct projectors are applied to map features into embeddings with lower semantic dim D_S, finally getting global and local image embeddings $v_i \in \mathbb{R}^{D_S}, R_i = \{r_{i1}, r_{i2}, ..., r_{iN_I}\} \in \mathbb{R}^{N_I \times D_S}$, and text embedding $t_i \in \mathbb{R}^{D_S}, L_i = \{l_{i1}, l_{i2}, ..., l_{iN_L}\} \in \mathbb{R}^{N_L \times D_S}$.

Besides using reports and images as the input for our pre-training network, we also input an external knowledge graph to the whole framework for improving the correspondence of modality features and clinical knowledge. The knowledge refers to relations between clinical pathology concepts in the radiology domain in the format of triplet $\mathcal{G} = \{(c_{h_k}, r_k, c_{t_k})\}_{k=1}^{N_G}$, such as UMLS [14]. Domain knowledge embedding for each concept $E = \{e_s\}_{s=1}^{N_E} \in \mathbb{R}^{N_E \times D_S}$ is the output of $Enc^G(\mathcal{G})$.

2.2 Knowledge Semantic Enhancement

To relieve the semantic overlap problem, where negative sample noise harms the effective learning of vision-language mutual information, we propose a semantic enhancement module to identify the noise using sample-wise similarities. The similarity is estimated upon sample knowledge k_i, calculated from domain knowledge embedding E and concept set from texts with negation marker.

Getting Sample knowledge: Firstly, we acquire a concept set that contains pathology concepts extracted from texts with Negbio $\mathcal{N}(\cdot)$ [17]. The image-view concept set which involves the overall observation is from the whole report, while the text-view set only covers the chosen sentence. Secondly, the image and text sample knowledge, as an auxiliary semantic estimation, is selected from domain knowledge embedding E according to the corresponding concept set from the report and sentence respectively, if not considering the negation problem.

Furthermore, considering the challenge that negation expression of concepts commonly exists in radiology reports, which has opposite semantics with similar morphology for text encoder (converging shifting), we randomly generate a *No Finding* embedding $\mathcal{NF}$ and a variant of domain knowledge embedding $\widetilde{E} = \{\widetilde{e}_1, \widetilde{e}_2, ..., \widetilde{e}_{N_E}\}$ of the same size as E with Xavier distribution. Upon the negation mark of concept, sample knowledge embedding $k_i = \{k_{i,s}\}_{s=1}^{N_{ES}}$ is denoted below:

$$k_{i,s} = \begin{cases} e_{i,s} & c_{i,s} \in \mathcal{N}(T_i), P(c_{i,s}) \neq Neg \\ \epsilon \cdot \mathcal{NF} + (1-\epsilon)\widetilde{e}_{i,s} & c_{i,s} \in \mathcal{N}(T_i), P(c_{i,s}) = Neg \end{cases} \tag{1}$$

where P is the negation mark of concepts, and $e_{i,s}, \widetilde{e}_{i,s}$ is the corresponding position of $c_{i,s}$ in E and $\widetilde{E}$. ϵ tunes the variance of negative sample knowledge. $k_{i,s}^{Image}$ and $k_{i,s}^{Text}$ are k_i from the image-view and text-view concept set.

Estimation of Similarities: The semantic similarity is calculated upon sample knowledge. For each image-text pair, a max-match strategy is adopted to match each two sample knowledge embedding with the most similar one for calculating cosine similarities. Sample-wise similarities are aggregated with averages.

$$\lambda_{ij}^{IT} = \frac{1}{N_{ES'}} \sum_{s=1}^{N_{ES'}} \max_{s'=1}^{N_{ES}} (k_{i,s}^{Image})^T k_{j,s'}^{Text}, \lambda_{ij}^{TI} = \frac{1}{N_{ES}} \sum_{s=1}^{N_{ES}} \max_{s'=1}^{N_{ES'}} (k_{i,s}^{Text})^T k_{j,s'}^{Image} \tag{2}$$

where N_{ES} is the number of concepts in T_i^{Sent}, while $N_{ES'}$ is that in T_i^{Report}.

Knowledge Semantic Enhancement Loss: We utilize the sample-wise semantic similarity to estimate negative sample noise, placed in the sample weight of the contrastive loss [18,26], where paired cross-modal embedding are pushed together and unpaired ones are pulled apart. The importance of estimated noisy negative samples is relatively smaller for a subtle pulling between cross-modal embeddings. The semantic enhancement loss is below:

$$\mathcal{L}_{SE} = -\frac{1}{N} \sum_{i=1}^{N} (\log \frac{\exp(v_i^T t_i / \tau_G)}{\sum_{j=1}^{N} (1-\lambda_{ij}^{IT}) \exp(v_i^T t_j / \tau_G)} + \log \frac{\exp(t_i^T v_i / \tau_G)}{\sum_{j=1}^{N} (1-\lambda_{ij}^{TI}) \exp(t_i^T v_j / \tau_G)}) \tag{3}$$

where τ_G is the global temperature, and λ^{IT}, λ^{TI} is the sample similarity measurement. specifically, $\lambda_{i,i}$ is fixed to zero to persist the positive sample weight.

2.3 Knowledge Semantic Guidance

In this section, we propose a semantic guidance module to solve the semantic shifting problem. Utilizing sample knowledge from Sect. 2.2 which contains concept correlation and negation information, the adverse effects of both disperse and converging shifting are alleviated by fusing domain-sample knowledge with global-local modality embeddings. We design four contrast schemes: knowledge anchor guidance for adjusting disperse shifting, semantic knowledge refinement for filtering converging shifting, vision semantic response for consolidating knowledge fusion, and semantic bridge guidance for narrowing the modality gap.

Knowledge Anchor Guidance: Disperse shifting will be adjusted if there are unbiased anchors in semantic space as priors to attract modality embeddings towards clinical semantics, and domain knowledge embedding does a good job. We define knowledge fused embeddings $H_i^{IK} = ATTN(v_i, E, E)$ and $H_i^{TK} = ATTN(t_i, E, E)$, and $ATTN(Q, K, V)$ means the attention function [10]:

$$\mathcal{L}_{KAG} = -\frac{1}{N}\sum_{i=1}^{N}(\log \frac{exp(H_i^{IK} \cdot H_i^{TK}/\tau_G)}{\sum_{j=1}^{N} exp(H_i^{IK} \cdot H_j^{TK}/\tau_G)} + \log \frac{exp(H_i^{TK} \cdot H_i^{IK}/\tau_G)}{\sum_{j=1}^{N} exp(H_i^{TK} \cdot H_j^{IK}/\tau_G)}) \quad (4)$$

where image-weighted and text-weighted knowledge is globally contrasted.

Semantic Knowledge Refinement: Wrong-converging pairs have distinct intrinsic responses on sample knowledge from image and text. Hence, we propose to utilize sample knowledge to refine these falsely gathered dissimilar pairs. We define $H_{ij}^{SI} = ATTN(k_i^{Text}, R_j, R_j)$ and $H_{ij}^{ST} = ATTN(k_i^{Text}, L_j, L_j)$:

$$\mathcal{L}_{SKR} = -\frac{1}{N}\sum_{i=1}^{N}\log \frac{exp(\frac{1}{N_{ES}\cdot\tau_L}\sum_{k=1}^{N_{ES}} H_{iik}^{SI} \cdot H_{iik}^{ST})}{\sum_{j=1}^{N} exp(\frac{1}{N_{ES}\cdot\tau_L}\sum_{k=1}^{N_{ES}} H_{ijk}^{SI} \cdot H_{ijk}^{ST})} \quad (5)$$

where local semantic-weighted image and text embeddings are contrasted.

Vision Semantic Response: Instead of matching single token with image subregions in [10], we propose to match the concept with sub-regions. As the concept is a more complete and atomic semantic unit, local response upon concept will better guide the representation learning with a fine-grained semantic match through an in-sample contrast. We define $H_i^{IS} = ATTN(R_i, k_i^{Text}, k_i^{Text})$, and the fusion of knowledge will be consolidated as below:

$$\mathcal{L}_{VSR} = -\frac{1}{N \cdot N_I}\sum_{i=1}^{N}\sum_{k=1}^{N_I}\log \frac{exp(H_{ik}^{IS} \cdot r_{ik}/\tau_L)}{\sum_{k'=1}^{N_I} exp(H_{ik}^{IS} \cdot r_{ik'}/\tau_L)} \quad (6)$$

where there is an in-sample local contrast between H_i^{IS} and vision features.

Semantic Bridge Guidance: We propose to narrow disperse shifting enlarged by the modality gap between vision and language. Specifically, the gap is bridged by the fusion of domain knowledge which is better compatible with text:

$$\mathcal{L}_{SBG} = -\frac{1}{N}\sum_{i=1}^{N}(\log \frac{exp(H_i^{IK} \cdot t_i/\tau_G)}{\sum_{j=1}^{N} exp(H_i^{IK} \cdot t_j/\tau_G)} + \log \frac{exp(t_i \cdot H_i^{IK}/\tau_G)}{\sum_{j=1}^{N} exp(t_i \cdot H_j^{IK}/\tau_G)}) \quad (7)$$

where the image-weighted domain knowledge is contrasted with text features between samples. Finally, $\mathcal{L}_{SG}$ is aggregated by these four parts as below:

$$\mathcal{L}_{SG} = \lambda_1\mathcal{L}_{KAG} + \lambda_2\mathcal{L}_{SKR} + \lambda_3\mathcal{L}_{VSR} + \lambda_4\mathcal{L}_{SBG} \quad (8)$$

3 Experiment

Experiment Protocol: Pre-training performs on MIMIC-CXR [12] following the pre-process style of [9]. The impression section of reports and frontal view of images are selected to generate 203k image-report pair. Five downstream task datasets (CheXpert [11], Covidx [24], MIMIC-CXR, UMNSRS [16] and SIIM [22]) are applied on eight tasks. Semantic relatedness is to verify the text understanding of radiology concepts, where text embedding with certain prompts predicts the relatedness. A new semantic relatedness benchmark is generated from MIMIC-CXR, adding in the extra negation discriminating. CheXpert5X200 [10](Multi-classification) is from CheXpert, and CheXpert-labeller[11] generates retrieval labels in MIMIC-CXR. More details are in appendix.

Table 1. Comparison results in eight downstream tasks. (*) defines that official pre-trained weight is used, and the remaining methods are reproduced with the same batch size, pre-processing and the evaluation. CLS, RR, SR, and SEG mean classification, retrieval, semantic relatedness and semantic segmentation, V or L means vision and language tasks. Few-shot-Frozen means the frozen encoder of the backbone and only 1% of total training data. ResNet-50 is the equal-comparing backbone except for KoBo-Vit. The best two results are highlighted in underlined red and violet.

Method	**Zero-shot**					**Few-shot-Frozen**		
	CLS(V+L)	**RR(V)**	**RR(V+L)**	**SR(L)**	**SR(L)**	**CLS(V)**	**SEG(V)**	**CLS(V)**
	CheXpert	CheXpert5X200	MIMIC	UMNSRS	MIMIC	CheXpert	SIIM	Covidx
	(Auroc)	(mAP)	(mAP)	(Pearson)	(Pearson)	(Auroc)	(Dice)	(Acc)
CLIP [18](*)	0.4702	0.2544	0.7577	0.1985	-0.2879	0.5748	/	0.8975
ConVIRT [26]	0.8252	0.3808	**0.8482**	**0.2506**	0.1429	0.8548	0.4992	0.9475
Gloria [10]	0.8257	0.3875	0.8390	0.2294	0.1100	0.8492	0.5479	0.9250
MGCA [23]	0.8496	0.3906	0.8428	0.1889	0.1809	0.8616	0.5696	0.9375
MedCLIP [25](*)	0.7805	**0.4298**	0.7258	0.2032	-0.1321	0.8214	0.5619	0.9325
KoBo	**0.8590**	0.3918	**0.8467**	**0.2563**	**0.3712**	**0.8628**	**0.6393**	**0.9550**
KoBo-Vit	**0.8635**	**0.4123**	0.8455	0.1824	**0.4229**	**0.8660**	**0.6554**	**0.9525**

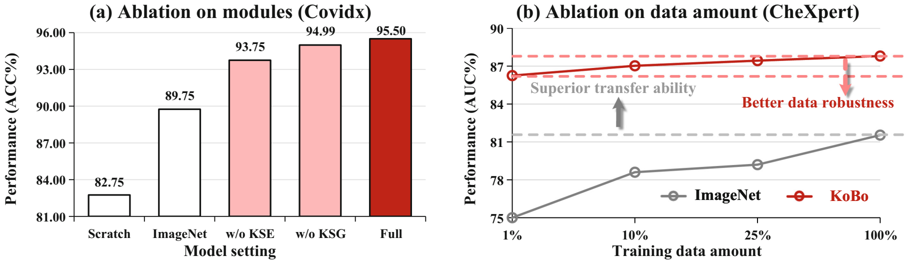


Fig. 4. **(a)** Module ablation study of our KoBo framework is performed on Covidx dataset compared with ImageNet and random initialization, upon few-shot frozen setting. **(b)** Data ablation study is performed on CheXpert with frozen setting when classification training data amount reduces to 25%, 10% and 1%.

For implementation, ResNet50 [6] and Vit [13] are image encoder, and BioClinicalBERT [1] is the text encoder. CompGCN with LTE [27] is our graph encoder, and domain knowledge contains 10,244 concepts in UMLS which exist in MIMIC-CXR. Negbio [17] combined with UMLS disambiguation tool [14] serves as $\mathcal{N}(\cdot)$. Embeddings are projected into the dim of 256. Pre-training has the batch size of 100 and max epochs of 50 based on Pytorch on two RTX3090 GPUs. Adam optimizer with the learning rate of 5e-5 and ReduceLR scheduler are applied. τ_G is 0.07 and τ_L is 0.1. λ in KSG loss are all 0.25, while ϵ in KSE loss is 0.1.

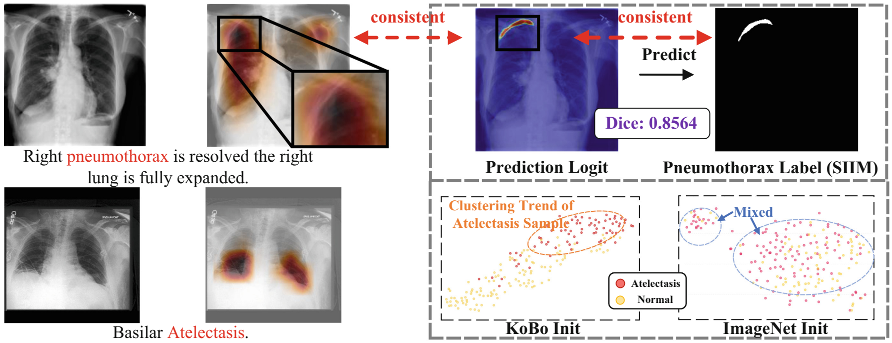

consistent
consistent
Predict
Dice: 0.8564
Prediction Logit
Pneumothorax Label (SIIM)
Right pneumothorax is resolved the right lung is fully expanded.
Clustering Trend of Atelectasis Sample
Mixed
Atelectasis
Normal
KoBo Init
ImageNet Init
Basilar Atelectasis.

Fig. 5. Visualization of pneumothorax and atelectasis. AblationCAM [19] generates the class activate map (CAM) upon last layer of Kobo-ResNet. There is strong consistency between CAM, prediction logits and segmentation label. t-SNE [23] is applied on image embedding from CheXpert-valid, showing gathering cluster trend of disease samples.

Comparison Study: Table 1 verifies our powerful representation ability, reaching state-of-art in classification, segmentation, and semantic relatedness compared with existing vision-language pre-training tasks, while our method is also top two for retrieval. In zero-shot classification tasks, our KoBo outperforms MGCA and ConVIRT 0.94% and 3.38% respectively, exceeding most methods even in their training setting. For CheXpert5X200, our framework is second only to MedCLIP which presents a superior performance in this dataset. In three few-shot setting task, our KoBo has an absolute leading position.

Ablation Study: As is demonstrated in Fig. 4, we perform module ablation and data amount ablation. **(a)** For module ablation, both modules bring benefits in representation learning and are respectively effective. When KSG module is removed, our KoBo also extracts effective feature related to pneumonia with a subtle decrease of 0.51%. When KSE is removed, there is a reduction of 1.25% accuracy. **(b)** For data amount ablation, KoBo has better data robustness with a subtle decrease when training data reduce to 1%. KoBo also has a superior transfer ability with an absolutely better AUC with 1% data than ImageNet with all training data than ImageNet with all training data.

Qualitative Analysis: In Fig. 5, our Kobo has learned fine-grained and effective image feature with the fusion of knowledge modeling. The deepest region in the first image gathered on the top left side, showing an obvious expansion on the right lung. There is consistency with the expert annotation and our output logit. The precise location of atelectasis region in CAM of second image and clustering trend interpret for the increase in zero-shot classification.

4 Conclusion

In our paper, we propose a Knowledge-Boosting Contrastive Vision-Language Pre-traing framework (KoBo). Sample and domain knowledge are used to differentiate noisy negative samples and supplement the correspondence between modality and clinical knowledge. Our experiments on eight tasks verify the effectiveness of our framework. We hope that our work will encourage more research on knowledge-granularity alignment in medical vision-language learning.

Acknowledgements. This research was supported by the Intergovernmental Cooperation Project of the National Key Research and Development Program of China(2022YFE0116700). We thank the Big Data Computing Center of Southeast University for providing the facility support.

References

1. Alsentzer, E., et al.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78 (2019)
2. Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
3. Chen, Z., et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part V, pp. 679–689. Springer (2022). https://doi.org/10.1007/978-3-031-16443-9_65
4. Chen, Z., Li, G., Wan, X.: Align, reason and learn: enhancing medical vision-and-language pre-training with knowledge. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 5152–5161 (2022)
5. van Dis, E.A., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L.: Chatgpt: five priorities for research. Nature **614**(7947), 224–226 (2023)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. He, Y., et al.: Learning better registration to learn better few-shot medical image segmentation: Authenticity, diversity, and robustness. IEEE Trans. Neural Netw. Learn. Syst. (2022)
8. He, Y., et al.: Geometric visual similarity learning in 3d medical image self-supervised pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9538–9547 (2023)

9. Hou, B., Kaissis, G., Summers, R.M., Kainz, B.: RATCHET: medical transformer for Chest X-ray diagnosis and reporting. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 293–303. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_28
10. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951 (2021)
11. Irvin, J., 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)
12. Johnson, A.E., et al.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
13. Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
14. Mao, Y., Fung, K.W.: Use of word and graph embedding to measure semantic relatedness between unified medical language system concepts. J. Am. Med. Inform. Assoc. **27**(10), 1538–1546 (2020)
15. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Radiological reports improve pre-training for localized imaging tasks on chest x-rays. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part V, pp. 647–657. Springer (2022). https://doi.org/10.1007/978-3-031-16443-9_62
16. Pakhomov, S.: Semantic relatedness and similarity reference standards for medical terms (2018)
17. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z.: Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits Trans. Sci. Proc. **2018**, 188 (2018)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
19. Ramaswamy, H.G., et al.: Ablation-cam: visual explanations for deep convolutional network via gradient-free localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 983–991 (2020)
20. Reddy, M.D.M., Basha, M.S.M., Hari, M.M.C., Penchalaiah, M.N.: Dall-e: creating images from text. UGC Care Group I J. **8**(14), 71–75 (2021)
21. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part V, pp. 690–700. Springer (2022). https://doi.org/10.1007/978-3-031-16443-9_66
22. Viniavskyi, O., Dobko, M., Dobosevych, O.: Weakly-supervised segmentation for disease localization in Chest X-Ray images. In: Michalowski, M., Moskovitch, R. (eds.) AIME 2020. LNCS (LNAI), vol. 12299, pp. 249–259. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59137-3_23
23. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Advances in Neural Information Processing Systems

24. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: a tailored deep convolutional neural network design for detection of Covid-19 cases from chest x-ray images. Sci. Rep. **10**(1), 1–12 (2020)
25. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3876–3887 (2022)
26. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference, pp. 2–25. PMLR (2022)
27. Zhang, Z., Wang, J., Ye, J., Wu, F.: Rethinking graph convolutional networks in knowledge graph completion. In: Proceedings of the ACM Web Conference 2022, pp. 798–807 (2022)
28. Zhou, Z.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42