Hierarchical Text-to-Vision Self Supervised Alignment for Improved Histopathology Representation Learning

Hasindri Watawana[1], Kanchana Ranasinghe[2], Tariq Mahmood[5], Muzammal Naseer[1], Salman Khan[1,4], and Fahad Shahbaz Khan[1,3]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
Hasindri.Watawana@mbzuai.ac.ae
[2] Stony Brook University, New York, USA
[3] Linköping University, Sweden
[4] Australian National University, Australia
[5] Shaukat Khanum Cancer Hospital, Pakistan

Abstract. Self-supervised representation learning has been highly promising for histopathology image analysis with numerous approaches leveraging their patient-slide-patch hierarchy to learn better representations. In this paper, we explore how the combination of domain specific natural language information with such hierarchical visual representations can benefit rich representation learning for medical image tasks. Building on automated language description generation for features visible in histopathology images, we present a novel language-tied self-supervised learning framework, Hierarchical Language-tied Self-Supervision (HLSS) for histopathology images. We explore contrastive objectives and granular language description based text alignment at multiple hierarchies to inject language modality information into the visual representations. Our resulting model achieves state-of-the-art performance on two medical imaging benchmarks, OpenSRH and TCGA datasets. Our framework also provides better interpretability with our language aligned representation space. The code is available at https://github.com/Hasindri/HLSS.

Keywords: Vision-Language Alignment · Self-Supervised Learning

1 Introduction

Self-supervised learning (SSL) has showcased remarkable outcomes for vision tasks [9,3], with recent extensions to medical imaging tasks proving highly successful [13], especially given the expensive and difficult nature of medical image annotation due to necessity of domain-specific expertise knowledge.

In constrast to natural images, medical images contain unique imaging patterns. In clinical studies, it is common to sample multiple gigapixel range image *slides* from a single *patient*, followed by analysis of smaller sub regions of slides, referred as *patches*. This creates a patient, slide, patch hierarchy in captured data where all samples from a single patient correspond to a common diagnosis.

Multiple existing work leverage this hierarchy in the visual modality to learn self-supervised representations of histopathology images [13,4].

With the recent advancement of vision-text alignment research and highly transferable vision language models (VLMs) [18,15], multiple histopathology SSL work go beyond raw visual information, leveraging natural language to learn more generic representations, zero-shot capabilities, and improved interpretability [16,20,17]. However, utilising hierarchy in terms of language has not yet been done in any recent vision-language SSL work in histopathology, especially given most of these work are built upon image-text paired datasets manually annotated or automatically captioned at only patch level [11,7].

In our work, we bridge this gap by exploring hierarchy in both vision and language modalities. We propose a novel framework for hierarchical text-to-vision alignment for language guided visual representation learning named **HLSS (Hierarchical Language-tied Self Supervision)** which extends the self supervised learning objectives across three levels of hierarchy: patient, slide and patch. In contrast to existing histopathology VLM approaches that require a sample specific description per each image, we are the first to use a fixed set of text descriptions depicting dataset specific characteristics, for language guidance. First, we utilise pre-trained LLMs containing extensive world knowledge to extract a fixed set of visual characteristics flagged as useful for a diagnosis in a given dataset type, for each level of the hierarchy. Then a text description is generated per each attribute. Since the three sets of descriptions describe the images at three different granularities; patch level describes more fine grained features while patient level describes features related to overall diagnosis; we refer to them as granular language descriptions. This entire process is automated and is followed by verification from a human expert in histopathology. Collected text descriptions are encoded using a CLIP text encoder and the resulting text vectors are used to construct a hierarchical text-to-vision alignment objective which is combined with a hierarchical vision contrastive objective. The resulting framework, which we named HLSS, learns representations achieving state-of-the-art performance on downstream medical image classification tasks.

In summary, our contributions can be categorized into three parts: **1)** Automated Generation of dataset specific granular *characteristic-description* text pairs that can describe histopathology images at a multitude of levels, **2)** Hierarchical text-to-vision alignment for self-supervised representation learning on histopathology images, **3)** A language guided framework that utilise dataset specific characteristic descriptions instead of sample specific captions. Evaluations on two downstream histopathology benchmarks, OpenSRH and TCGA, demonstrate state-of-the-art performance of HLSS.

2 Related Work

Vision Language Models in Histopathology Numerous recent work in histopathology combine language with self-supervised objectives, mostly basing off image-text contrastive objectives applied on paired image-caption data

[11,17,16,20]. This demands sample-specific text captions that are expensive for medical domains. At the same time, these methods do not explore the hierarchical relations inherent to the data. To the best of our knowledge, our work is a first to introduce a hierarchical vision-language representation learning approach.

3 Methodology

In this section, we present our approach, HLSS, that learns language-guided self-supervised representations for histopathology images that are also interpretable for clinical decision-making. We leverage the inherent hierarchical structure of histopathology data, across both vision and language modalities, and construct two novel language-tied self-supervision objectives, Hierarchical Vision Contrastive (HVC) Loss and Hierarchical Text-to-Vision Alignment (HA) Loss. Additionally, we propose two architectural components, Positive Pairing Module (PPM) and Cross-Modal Alignment Module (CAM) for efficient implementation of our objectives. In the following sections, we discuss some key characteristics of histopathology data, layout the architecture of our framework, introduce our two proposed learning objectives in detail, and describe our overall framework. To the best of our knowledge, in the histopathology imaging domain, we are the first to connect natural language with hierarchy aware self-supervised learning.

3.1 Background

Histopathology domain vision tasks generally involve hierarchical data extraction [13], resulting in an inherent hierarchy for the visual information. A single Whole Slide Image (WSI), referred as a *slide*, could span gigapixel scales, motivating most computer vision approaches to use sub-sampled fields-of-view (e.g., 256×256 pixels region), referred as *patches*. These patches could belong to a single slide or different slides from the same patient. Overall, this results in an inherent three-tier data hierarchy. Interestingly, each level contains some unique visual characteristics which can also be described sufficiently using natural language.

3.2 Architecture

Our HLSS processes patches, $x \in \mathbf{R}^{(H,W,C)}$ where $H = W = 224$ and $C = 3$, to output representations, $z \in \mathbf{R}^D$ where $D = 1024$, which are used in downstream tasks. The inherent data hierarchy allows sampling patches belonging to individual levels. Therein, our setup processes $n_s * n_p * n_a$ patches at each iteration, where n_s slides belong to a common patient, n_p patches are sampled from each slide, and data augmentations provide n_a views per patch. A view here refers to a visually augmented version of a patch. We use a ResNet-50 [10] as our visual encoder with CLIP pretraining [18], $\mathcal{F}_V$ to obtain $z \in \mathbf{R}^D$. Our proposed positive pairing module (PPM) projects z to language guided representation spaces motivated from [19] obtaining features specific to each level as

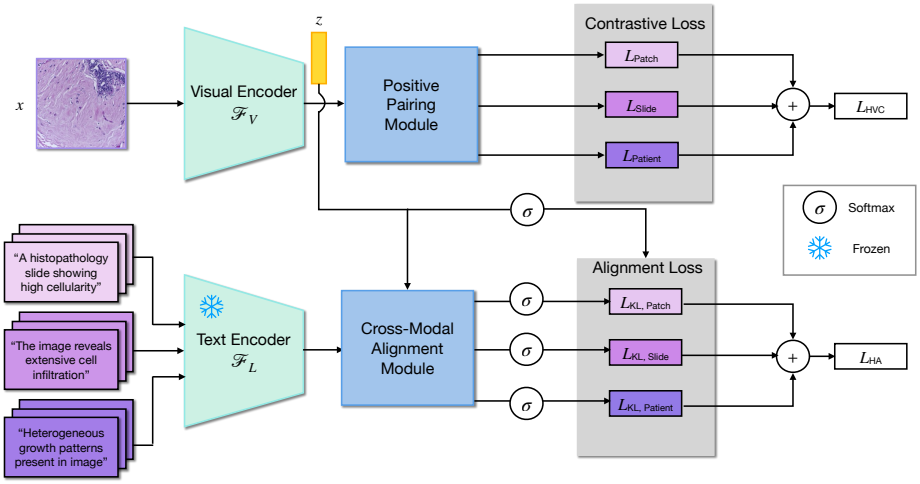

Fig. 1. Overview of HLSS Architecture.

$z'_{\text{patient}}, z'_{\text{slide}}, z'_{\text{patch}}$. These features are processed by our proposed hierarchical vision contrastive (HVC) objective to provide the first learning signal. Additionally, granular textual descriptions relevant to each hierarchy are processed through a language encoder from CLIP, $\mathcal{F}_L$ to obtain language modality features, $t_{\text{patient}}, t_{\text{slide}}, t_{\text{patch}}$. Our cross-modal alignment (CAM) module outputs modified features $t'_{\text{patient}}, t'_{\text{slide}}, t'_{\text{patch}}$ which processed by our Hierarchical Text-to-Vision Alignment (HA) objective, providing a second learning signal. We illustrate this overall architecture in Fig. 1.

3.3 Hierarchical Vision Contrastive Objective

Self-supervised contrastive objectives commonly operate on features that are projected to a secondary feature space, which is considered to learn more generic representations [1]. Given the hierarchy of our data, we propose distinct secondary feature spaces for each level. Additionally, motivated by [19], we propose language-guided construction of these feature spaces. This combined setup is implemented in our proposed Positive Pairing Module (PPM). The resulting hierarchical representations are processed separately using our hierarchical vision contrastive (HVC) loss to provide a suitable learning signal.

Positive Pairing Module We first separate visual encoder outputs, z, to individual levels and process each through level-specific projection layers. These projection layers are implemented as linear layers and output $z'_{\text{patient}}, z'_{\text{slide}}, z'_{\text{patch}}$ where each $z' \in \mathbf{R}^{(m,128)}$ with m equal to $(n_s \cdot n_p \cdot n_a)$, $(n_p \cdot n_a)$, and n_a respectively. The projection layers are initialized with textual vectors (extracted from $\mathcal{F}_L$) such that each axis of the 128-dimensional feature space corresponds to some text description of a histopathology characteristic (details in Sec 3.4).

Motivation behind using a different projection layer per each level is to learn a separate secondary feature space.

Loss Definition Note how each z' above corresponds to more than one patch. Let us refer to features of a single patch as z'_i. Given these projected z'_i, we consider z'_i of common levels positives and others negatives, and we apply a contrastive objective from [14],

$$L_{\text{con}}(z_i, Z_{pi}, Z) = \frac{1}{|Z_{pi}|} \sum_{z_k \in Z_{pi}} \log \frac{\exp(z_i \cdot z_k / \tau)}{\sum_{z_j \in Z \setminus \{z_i\}} \exp(z_i \cdot z_j / \tau)} \tag{1}$$

that supports multiple positives, where Z_{pi} is the set of positives of z_i (excluding z_i) and Z is the set of all z' respectively. Here each of $z'_{\text{patient}}, z'_{\text{slide}}, z'_{\text{patch}}$ corresponds to more than one patch (i.e. $n_s \cdot n_p \cdot n_a$, $n_p \cdot n_a$, and n_a patches respectively) while z_i in Eq. 1 refers to features of a single patch. Applying this loss at each hierarchical level we obtain three different objectives,

$$L_{\text{Patch}} = \sum\nolimits_{z_i \in z'_{\text{patch}}} L_{\text{con}}(z_i, z'_{\text{patch}} \setminus \{z_i\}, Z) \tag{2}$$

$$L_{\text{Slide}} = \sum\nolimits_{z_i \in z'_{\text{slide}}} L_{\text{con}}(z_i, z'_{\text{slide}} \setminus \{z_i\}, Z) \tag{3}$$

$$L_{\text{Patient}} = \sum\nolimits_{z_i \in z'_{\text{patient}}} L_{\text{con}}(z_i, z'_{\text{patient}} \setminus \{z_i\}, Z) \tag{4}$$

We combine these terms to define our first training objective HVC loss as,

$$L_{\text{HVC}} = L_{\text{Patch}} + L_{\text{Slide}} + L_{\text{Patient}} \tag{5}$$

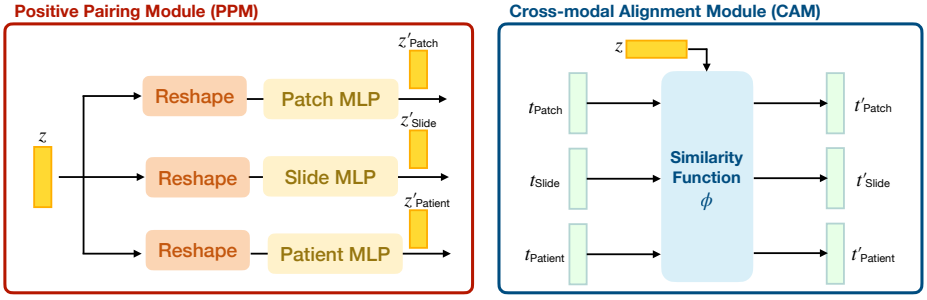

Fig. 2. We illustrate the operations within our proposed Positive Pairing Module (left) and Cross-Modal Alignment Module (right).

3.4 Hierarchical Text-to-Vision Alignment

Our second training objective focuses on explicit alignment of cross-modal representations. We utilize a pretrained large language model (LLM) that contains both world knowledge and domain specific awareness to generate descriptions of

visual characteristics corresponding to each level of our hierarchy. These generated descriptions are additionally verified by human experts (histopathology specialists) to ensure both their meaningfulness and interpretability when proposing diagnoses in downstream tasks.

Granular Language Descriptions We prompt ChatGPT in a multi-stage manner to generate a large set of visual descriptions which is reduced to 128 characteristics per level by eliminating redundancy and invalid responses manually (with human expert supervision). Each characteristic would be represented by 4 natural language descriptions.

Cross-Modal Alignment Given our language dataset of medical characteristic descriptions, we utilize our text encoder $\mathcal{F}_L$ to process these descriptions and extract language embeddings.The averaged embedding (across four descriptions) of each characteristic, $t_{\text{patient}}, t_{\text{slide}}, t_{\text{patch}}$ is calculated, where each $t \in \mathbf{R}^{(128,D)}$ with $D = 1024$. For a given patch, using its visual embedding z we select a best matching text embedding $t'_i = \arg\max_{t_k \in t_i} \langle t_k, z \rangle$ where $\langle \cdot, \cdot \rangle$ is cosine-similarity and $t_k \in t_i$ indexes along the characteristic (128) dimension with $t_k \in \mathcal{R}^D$.

Hierarchical Alignment Objective We utilize the language embeddings t'_i output from CAM to enforce distillation into our visual embeddings, z. Therein, our hierarchical alignment (HA) loss is defined using a KL-divergence loss,

$$L_{\text{HA}} = L_{\text{KL}}(z, t'_{\text{patient}}) + L_{\text{KL}}(z, t'_{\text{slide}}) + L_{\text{KL}}(z, t'_{\text{patch}}) \tag{6}$$

where L_{KL} applies KL-divergence loss with suitable softmax normalization. Here, the KL divergence between the most aligning text vector and the visual representation is calculated using the same z in all levels, as unlike the text hierarchy which is formed of separately curated granular descriptions per each level, the visual hierarchy is always formed on patch sized views. We traverse through visual hierarchy by altering the count of positively paired patches based on a common origin at each level.Therefore, our overall self supervised learning objective can be summarised as,

$$L_{\text{HLSS}} = L_{\text{HVC}} + L_{\text{HA}} \tag{7}$$

4 Experiments

4.1 Experimental Setup

Datasets We perform experiments on two benchmark histopathology image datasets: OpenSRH [12] and TCGA. OpenSRH (Stimulated Raman Histology) is a public dataset of clinical SRH images of 300+ brain tumor patients with classes consisting of normal brain tissue and 6 different brain tumor diagnoses. In TCGA brain cancer dataset, we utilise TCGA-LGG and TCGA-GBM subsets containing brain tumor samples.

Training We train for 40000 epochs at batch size 32 (patient count) using AdamW optimizer with a learning rate of 0.001 and a cosine decay scheduler after warmup in the first 10% iterations. We use $n_s = n_p = n_a = 2$ during

Table 1. Evaluation on SRH Dataset: Baseline results are reproduced on the publicly available OpenSRH dataset under conditions similar to HLSS.

Method	Patch Classification	Slide Classification	Patient Classification
Supervised [13]	88.9	89.0	93.9
SimCLR [5]	81.0	82.1	87.8
SimSiam [6]	80.3	81.4	86.0
BYOL [8]	83.5	84.3	90.5
VICReg [2]	82.1	83.4	87.4
HiDisc [13]	82.2	87.6	88.3
HLSS (ours)	**84.1**	**89.5**	**91.7**

Table 2. Evaluation on TCGA Dataset.

Method	Patch Classification	Slide Classification	Patient Classification
Supervised [13]	85.1	88.3	88.3
SimCLR [5]	77.8	83.0	80.7
SimSiam [6]	68.4	77.2	76.6
BYOL [8]	80.0	84.1	83.1
VICReg [2]	75.5	80.8	77.0
HiDisc [13]	83.1	85.1	83.6
HLSS (ours)	**89.7**	**92.9**	**87.9**

training. Temperature τ is set to 0.7. Other hyperparameters follow standard settings from [13].

Evaluation K Nearest Neighbor (kNN) classification is used to evaluate downstream performance. During evaluation, the pretrained visual backbone is frozen and representations for train and test splits are computed. The class labels of k Nearest Neighbors of the training data is used to predict the class of a given validation patch. Slide and patient level metrics are reported by average pooling the patch level prediction scores of component patches of the given slide or patient. We use all patches from OpenSRH and only 400 randomly loaded patches per slide from TCGA for evaluation. The latter limitation is due to the large slides in TCGA data. We report kNN classification accuracy metric per each task.

4.2 Results Comparison

We first present our results on the OpenSRH dataset in Table 1. Our approach outperforms all prior work by a clear margin while even surpassing supervised pretraining settings for Slide Classification. We next evaluate on TCGA dataset and report these results in Table 2. HLSS outperforms all prior work achieving state-of-the-art results. We take these results as indication for strong representation learning ability of HLSS.

4.3 Ablations

We perform all ablations on OpenSRH dataset following the same self supervised pretraining stage followed by kNN evaluation.

Table 3. Ablation on SSL Objectives.

Method	Loss	Patch Accuracy	Slide Accuracy	Patient Accuracy
HiDisc	HiDisc	82.2	87.6	88.3
Ours	HVC	82.6	89.5	91.7
Ours	HLSS	**84.1**	**89.5**	**91.7**

Table 4. Ablation on Hierarchical Text Integration.

Method	Loss	Vision Hierarchy	Text Hierarchy	Patch Accuracy	Slide Accuracy	Patient Accuracy
HiDisc	HiDisc	patient	none	82.20	87.55	88.33
Ours	HVC	patient	patient	**81.9**	**89.11**	**90.0**
Ours	HVC	patient	slide	81.23	86.53	86.67
Ours	HVC	patient	patch	78.34	84.97	85.0

Table 5. Ablation on Granular Language Descriptions.

Method	Loss	Text Hierarchy	Patch Accuracy	Slide Accuracy	Patient Accuracy
Ours	HVC	non-granular	81.9	89.11	90.0
Ours	HVC	granular	**82.55**	**89.49**	**91.67**

SSL Objectives In the first ablation study, we analyse the contribution of each loss component to the performance of the model. Refer Table 3. HLSS with only HVC loss surpass HiDisc performace. Hierarchical projection layers initialised with granular text vectors, injecting language information to the lower dimensional spaces where hierarchical contrastive loss operates, is the contributing factor in this situation. When HVC loss is combined with a vision-text alignment loss for patch representations, it furthur improves the patch representations as observed in the results.

Hierarchical Integration of Text Integration of text hierarchically to the self-supervised setup improves performance when used at all levels as illustrated by results in Table 4. This study was conducted by using only the HVC loss component.

Granular Descriptions Here, we study the effect of using granular descriptions separately generated for each hierarchical level against using a generic set of textual descriptions repeated across all three levels of hierarchy. Refer Table 5.

4.4 Interpretability

Representations learnt by a language guided approach are more interpretable. We obtained a set of cancer markers specific for each tumor grading in OpenSRH from a histopathologist. These markers were not used during training. A text embedding was derived by averaging the CLIP text embeddings of 4 language descriptions generated via an LLM per each marker. Results demonstrate the close alignment between an image representation with the descriptions of the markers from the ground truth class. More details are included in supplementary.

5 Conclusion

We introduce a novel hierarchical language-tied SSL framework for histopathology by proposing two hierarchical SSL objectives. Our approach compliments a hierarchical vision approach by additionally exploring language hierarchy. In contrast to prior vision-text alignment SSL work that require sample-specific image captions, our model achieves state-of-the-art performance using a set of dataset-specific text descriptions.

Acknowledgments. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partially funded by the Swedish Research Council through grant agreement no. 2022-06725, the LUMI supercomputer hosted by CSC (Finland) and the LUMI consortium, and by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Balestriero, R., et al.: A cookbook of self-supervised learning. ArXiv (2023)
2. Bardes, A., Ponce, J., LeCun, Y.: Variance-invariance-covariance regularization for self-supervised learning. ICLR, Vicreg (2022)
3. 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 (2021)
4. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning (2020)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
7. Gamper, J., Rajpoot, N.: Multiple instance captioning: Learning representations from histopathology textbooks and articles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
8. 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 (2020)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)

11. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., et al.: Quilt-1m: One million image-text pairs for histopathology. Advances in Neural Information Processing Systems (2024)
12. Jiang, C., Chowdury, A., Hou, X., et al.: Opensrh: optimizing brain tumor surgery using intraoperative stimulated raman histology. Advances in neural information processing systems (2022)
13. Jiang, C., Hou, X., Kondepudi, A., Chowdury, A., Freudiger, C.W., Orringer, D.A., Lee, H., Hollon, T.C.: Hierarchical discriminative learning improves visual representations of biomedical microscopy. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. NeurIPS (2020)
15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022)
16. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
17. Lu, M.Y., Chen, B., et al.: Towards a visual-language foundation model for computational pathology. arXiv preprint arXiv:2307.12914 (2023)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
19. Ranasinghe, K., Ryoo, M.S.: Language-based action concept spaces improve video self-supervised learning. NeurIPS 2023 (2023)
20. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)