SLPD: Slide-Level Prototypical Distillation for WSIs

Zhimiao Yu, Tiancheng Lin, and Yi Xu$^{(\boxtimes)}$

MoE Key Lab of Artificial Intelligence, AI Institute,
Shanghai Jiao Tong University, Shanghai, China
`xuyi@sjtu.edu.cn`

Abstract. Improving the feature representation ability is the foundation of many whole slide pathological image (WSIs) tasks. Recent works have achieved great success in pathological-specific self-supervised learning (SSL). However, most of them only focus on learning patch-level representations, thus there is still a gap between pretext and slide-level downstream tasks, *e.g.*, subtyping, grading and staging. Aiming towards slide-level representations, we propose Slide-Level Prototypical Distillation (SLPD) to explore intra- and inter-slide semantic structures for context modeling on WSIs. Specifically, we iteratively perform intra-slide clustering for the regions (4096×4096 patches) within each WSI to yield the prototypes and encourage the region representations to be closer to the assigned prototypes. By representing each slide with its prototypes, we further select similar slides by the set distance of prototypes and assign the regions by cross-slide prototypes for distillation. SLPD achieves state-of-the-art results on multiple slide-level benchmarks and demonstrates that representation learning of semantic structures of slides can make a suitable proxy task for WSI analysis. Code will be available at https://github.com/Carboxy/SLPD.

Keywords: Computational pathology · Whole slide images(WSIs) · Self-supervised learning

1 Introduction

In computational histopathology, visual representation extraction is a fundamental problem [14], serving as a cornerstone of the (downstream) task-specific learning on whole slide pathological images (WSIs). Our community has witnessed the progress of the *de facto* representation learning paradigm from the supervised ImageNet pre-training to self-supervised learning (SSL) [15,36]. Numerous

---

Z. Yu and T. Lin—Equal contribution.

pathological applications benefit from SSL, including classification of glioma [7], breast carcinoma [1], and non-small-cell lung carcinoma [25], mutation prediction [32], microsatellite instability prediction [31], and survival prediction from WSIs [2,16]. Among them, pioneering works [12,22,27] directly apply the SSL algorithms developed for natural images (*e.g.*, SimCLR [10], CPC [30] and MoCo [11]) to WSI analysis tasks, and the improved performance proves the effectiveness of SSL. However, WSI is quite different from natural images in that it exhibits a hierarchical structure with giga-pixel resolution. Following works turn to *designing pathological-specific tasks* to explore the inherent characteristics of WSIs for representation learning, *e.g.*, resolution-aware tasks [18,34,37] and color-aware tasks [2,38]. Since the pretext tasks encourage to mine the pathologically relevant patterns, the learned representations are expected to be more suitable for WSI analysis. Nevertheless, these works only consider learning the representations at the patch level, *i.e*, the cellular organization, but neglecting macro-scale morphological features, *e.g.*, tissue phenotypes and intra-tumoral heterogeneity. As a result, there is still a gap between the pre-trained representations and downstream tasks, as the latter is mainly at the slide level, *e.g.*, subtyping, grading and staging.

More recently, some works propose to close the gap via *directly learning slide-level representations in pre-training.* For instance, HIPT [8], a milestone work, introduces hierarchical pre-training (DINO [6]) for the patch-level (256×256) and region-level (4096×4096) in a two-stage manner, achieving superior performance on slide-level tasks. SS-CAMIL [13] uses EfficientNet-B0 for image compression in the first stage and then derives multi-task learning on the compressed WSIs, which assumes the primary site information, *e.g.*, the organ type, is always available and can be used as pseudo labels. SS-MIL [35] also proposes a two-stage pre-training framework for WSIs using contrastive learning (SimCLR [10]), where the differently subsampled bags[1] from the same WSI are positive pairs in the second stage. A similar idea can be found in Giga-SSL [20] with delicate patch- and WSI-level augmentations. The aforementioned methods share the same two-stage pre-training paradigm, *i.e.*, patch-to-region/slide. Thus broader context information is preserved to close the gap between pretext and downstream tasks. However, they are essentially *instance discrimination* where only the self-invariance of region/slide is considered, leaving the *intra- and inter-slide semantic structures* unexplored.

In this paper, we propose to encode the intra- and inter-slide semantic structures by modeling the mutual-region/slide relations, which is called SLPD: Slide-Level Prototypical Distillation for WSIs. Specifically, we perform the slide-level clustering for the 4096×4096 regions within each WSI to yield the prototypes, which characterize the medically representative patterns of the tumor (*e.g.*, morphological phenotypes). In order to learn this intra-slide semantic structure, we encourage the region representations to be closer to the assigned prototypes. By representing each slide with its prototypes, we further select semantically simi-

---

[1] By formulating WSI tasks as a multi-instance learning problem, the WSI is treated as a bag with corresponding patches as instances.
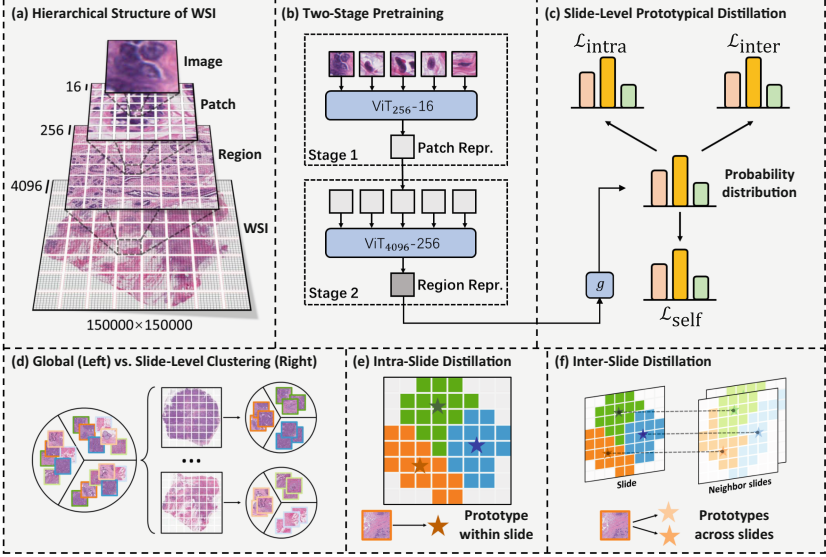
Fig. 1. (a) A WSI possesses the hierarchical structure of WSI-region-patch-image, from coarse to fine. (b) Two-stage pre-training paradigm successively performs the image-to-patch and patch-to-region aggregations. (c-e) The proposed SLPD. SLPD explores the semantic structure by slide-level clustering. Besides self-distillation, region representations are associated with the prototypes within and across slides to comprehensively understand WSIs.

lar slides by the set-to-set distance of prototypes. Then, we learn the inter-slide semantic structure by building correspondences between region representations and cross-slide prototypes. We conduct experiments on two benchmarks, NSCLC subtyping and BRCA subtyping. SLPD achieves state-of-the-art results on multiple slide-level tasks, demonstrating that representation learning of semantic structures of slides can make a suitable proxy task for WSI analysis. We also perform extensive ablation studies to verify the effectiveness of crucial model components.

2   Method

2.1   Overview

As shown in Fig. 1(a), a WSI exhibits hierarchical structure at varying resolutions under $20\times$ magnification: 1) the 4096×4096 regions describing macro-scale organizations of cells, 2) the 256×256 patches capturing local clusters of cells, 3) and the 16×16 images characterizing the fine-grained features at the cell-level. Given N unlabeled WSIs $\{w_1, w_2, \cdots, w_N\}$, consisting of numerous regions $\{\{x_n^l\}_{l=1}^{L_n}\}_{n=1}^N$, where L_n denotes the number of regions of WSI w_n, we aim to learn a powerful encoder that maps each x_n^l to an embedding $z_n^l \in \mathbb{R}^D$. SLPD

is built upon the two-stage pre-training paradigm proposed by HIPT, which will be described in Sect. 2.2. Fig 1(c-d) illustrates the pipeline of SLPD. We characterize the semantic structure of slides in Sect. 2.2, which is leveraged to establish the relationship within and across slides, leading to the proposed intra- and inter-slide distillation in Sect. 2.4 and Sect. 2.5.

2.2   Preliminaries

We revisit Hierarchical Image Pyramid Transformer (HIPT) [8], a cutting-edge method for learning representations of WSIs via self-supervised vision transformers. As shown in Fig. 1(b), HIPT proposes a two-stage pre-training paradigm considering the hierarchical structure of WSIs. In stage one, a patch-level vision transformer, denoted as ViT_{256}-16, aggregates non-overlapping 16×16 images within 256×256 patches to form patch-level representations. In stage two, the pre-trained ViT_{256}-16 is freezed and leveraged to tokenize the patches within 4096×4096 regions. Then a region-level vision transformer ViT_{4096}-256 aggregates these tokens to obtain region-level representations. With this hierarchical aggregation strategy, a WSI can be represented as a bag of region-level representations, which are then aggregated with another vision transformer, ViT_{WSI}-4096, to perform slide-level prediction tasks.

HIPT leverages DINO [6] to pre-train ViT_{256}-16 and ViT_{4096}-256, respectively. The learning objective of DINO is self-distillation. Taking stage two as an example, DINO distills the knowledge from teacher to student by minimizing the cross-entropy between the probability distributions of two views at region-level:

$$\mathcal{L}_{\text{self}} = \mathbb{E}_{x \sim p_d} H(g_t(\hat{z}), g_s(z)), \tag{1}$$

where $H(a, b) = -a \log b$, and p_d is the data distribution that all regions are drawn from. The teacher and the student share the same architecture consisting of an encoder (*e.g.*, ViT) and a projection head g_t/g_s. $\hat{z}$ and z are the embeddings of two views at region-level yielded by the encoder. The parameters of the student are exponentially moving averaged to the parameters of the teacher.

2.3   Slide-Level Clustering

Many histopathologic features have been established based on the morphologic phenotypes of the tumor, such as tumor invasion, anaplasia, necrosis and mitoses, which are then used for cancer diagnosis, prognosis and the estimation of response-to-treatment in patients [3,9]. To obtain meaningful representations of slides, we aim to explore and maintain such histopathologic features in the latent space. Clustering can reveal the representative patterns in the data and has achieved success in the area of unsupervised representation learning [4,5,24,26]. To characterize the histopathologic features underlying the slides, a straightforward practice is the global clustering, *i.e.*, clustering the region embeddings from all the WSIs, as shown in the left of Fig. 1(d). However, the obtained clustering centers, *i.e.*, the prototypes, are inclined to represent the visual bias

related to staining or scanning procedure rather than medically relevant features [33]. Meanwhile, this clustering strategy ignores the hierarchical structure "region→WSI→whole dataset" underlying the data, where the ID of the WSI can be served as an extra learning signal. Therefore, we first consider the slide-level clustering that clusters the embeddings within each WSI, which is shown in the right of Fig. 1(d). Specifically, we conduct k-means algorithm before the start of each epoch over L_n region embeddings $\{z_n^l\}_{l=1}^{L_n}$ of w_n to obtain M prototypes $\{c_n^m \in \mathbb{R}^D\}_{m=1}^M$. Similar operations are applied across other slides, and then we acquire N groups of prototypes $\{\{c_n^m\}_{m=1}^M\}_{n=1}^N$. Each group of prototypes is expected to encode the semantic structure ($e.g.$, the combination of histopathologic features) of the WSI.

2.4 Intra-Slide Distillation

The self-distillation utilized by HIPT in stage two encourages the correspondence between two views of a region at the macro-scale because the organizations of cells share mutual information spatially. However, the self-distillation, which solely mines the spatial correspondences inside the 4096×4096 region, cannot comprehensively understand the histopathologic consistency at the slide-level. In order to achieve better representations, the histopathologic connections between the WSI and its regions should be modeled and learned, which is called intra-slide correspondences. With the proposed slide-level clustering in Sect. 2.3, a slide can be abstracted by a group of prototypes, which capture the semantic structure of the WSI. As shown in Fig. 1(e), we assume that the representation z and its assigned prototype c also share mutual information and encourage z to be closer to c with the intra-slide distillation:

$$\mathcal{L}_{\text{intra}} = \mathbb{E}_{x \sim p_d} H\left(g_t(c), g_s(z)\right), \tag{2}$$

We omit super-/sub-scripts of z for brevity. Through Eq. 2, we can leverage more intra-slide correspondences to guide the learning process. For further understanding, a prototype can be viewed as an augmented representation aggregating the slide-level information. Thus this distillation objective is encoding such information into the corresponding region embedding, which makes the learning process semantic structure-aware at the slide-level.

2.5 Inter-Slide Distillation

Tumors of different patients can exhibit morphological similarities in some respects [17,21], so the correspondences across slides should be characterized during learning. Previous self-supervised learning methods applied to histopathologic images only capture such correspondences with positive pairs at the patch-level [22,23], which overlooks the semantic structure of the WSI. We rethink this problem from the perspective how to measure the similarity between two slides accurately. Due to the heterogeneity of the slides, comparing them with the local crops or the averaged global features are both susceptible to being one-sided. To

address this, we bridge the slides with their semantic structures and define the semantic similarity between two slides w_i and w_j through an optimal bipartite matching between two sets of prototypes:

$$D(w_i, w_j) = \max\{\frac{1}{M} \sum_{m=1}^{M} \cos(c_i^m, c_j^{\sigma(m)}) \mid \sigma \in \mathfrak{S}_M\}, \ D(w_i, w_j) \in [-1, 1], \quad (3)$$

where $\cos(\cdot, \cdot)$ measures the cosine similarity between two vectors, and $\mathfrak{S}_M$ enumerates the permutations of M elements. The optimal permutation σ^* can be computed efficiently with the Hungarian algorithm [19]. With the proposed set-to-set distance, we can model the inter-slide correspondences conveniently and accurately. Specifically, for a region embedding z belonging to the slide w and assigned to the prototype c, we first search the top-K nearest neighbors of w in the dataset based on the semantic similarity, denoted as $\{\hat{w}_k\}_{k=1}^{K}$. Second, we also obtain the matched prototype pairs $\{(c, \hat{c}_k)\}_{k=1}^{K}$ determined by the optimal permutation, where $\hat{c}_k$ is the prototype of $\hat{w}_k$. Finally, we encourage z to be closer to $\hat{c}_k$ with the inter-slide distillation:

$$\mathcal{L}_{\text{inter}} = \mathbb{E}_{x \sim p_d}[\frac{1}{K} \sum_{k=1}^{K} H\left(g_t(\hat{c}_k), g_s(z)\right)]. \quad (4)$$

The inter-slide distillation can encode the sldie-level information complementary to that of intra-slide distillation into the region embeddings.

The overall learning objective of the proposed SLPD is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{self}} + \alpha_1 \mathcal{L}_{\text{intra}} + \alpha_2 \mathcal{L}_{\text{inter}}, \quad (5)$$

where the loss scale is simply set to $\alpha_1 = \alpha_2 = 1$. We believe the performance can be further improved by tuning this.

3   Experimental Results

Datasets. We conduct experiments on two public WSI datasets[2]. *TCGA-NSCLC* dataset includes two subtypes in lung cancer, Lung Squamous Cell Carcinoma and Lung Adenocarcinoma, with a total of 1,054 WSIs. *TCGA-BRCA* dataset includes two subtypes in breast cancer, Invasive Ductal and Invasive Lobular Carcinoma, with a total of 1,134 WSIs.

Pre-training. We extract 62,852 and 60,153 regions at $20\times$ magnification from TCGA-NSCLC and TCGA-BRCA for pre-training ViT$_{4096}$-256 in stage two. We leverage the pre-trained ViT$_{256}$-16 in stage one provided by HIPT to tokenize the patches within each region. Following the official code of HIPT, ViT$_{4096}$-256 is optimized for 100 epochs with optimizer of AdamW, base learning rate of 5e-4 and batch size of 256 on 4 GTX3090 GPUs.

---

[2] The data is released under a CC-BY-NC 4.0 international license.

Fine-tuning. We use the pre-trained ViT_{256}-16 and ViT_{4096}-256 to extract embeddings at the *patch-level* (256×256) and the *region-level* (4096×4096) for downstream tasks. With the pre-extracted embeddings, we fine-tune three aggregators (*i.e.*, MIL [28], DS-MIL [22] and ViT_{WSI}-4096 [8]) for 20 epochs and follow other settings in the official code of HIPT.

Evaluation Metrics. We adopt the 10-fold cross validated Accuracy (Acc.) and area under the curve (AUC) to evaluate the weakly-supervised classification performance. The data splitting scheme is kept consistent with HIPT.

Table 1. Slide-level classification. "Mean" leverages the averaged pre-extracted embeddings to evaluate KNN performance. Bold and underlined numbers highlight the best and second best performance

#	Feature Aggragtor	Feature Extraction	Pretrain Method	NSCLC		BRCA	
				Acc.	AUC	Acc.	AUC
Weakly supervised classification							
1	MIL [28]	patch-level	DINO	$0.780_{\pm 0.126}$	$0.864_{\pm 0.089}$	$0.822_{\pm 0.047}$	$0.783_{\pm 0.056}$
2		region-level	SLPD	$0.856_{\pm 0.025}$	$0.926_{\pm 0.017}$	$\mathbf{0.879}_{\pm 0.035}$	$0.863_{\pm 0.076}$
3	DS-MIL [22]	patch-level	DINO	$0.825_{\pm 0.054}$	$0.905_{\pm 0.059}$	$0.847_{\pm 0.032}$	$0.848_{\pm 0.075}$
4		region-level	DINO	$0.841_{\pm 0.036}$	$0.917_{\pm 0.035}$	$0.854_{\pm 0.032}$	$0.848_{\pm 0.075}$
5		region-level	SLPD	$\underline{0.858}_{\pm 0.040}$	$\underline{0.938}_{\pm 0.026}$	$0.854_{\pm 0.039}$	$0.876_{\pm 0.050}$
6	ViT$_{WSI}$-4096 [8]	region-level	DINO	$0.843_{\pm 0.044}$	$0.926_{\pm 0.032}$	$0.849_{\pm 0.037}$	$0.854_{\pm 0.069}$
7		region-level	DINO+$\mathcal{L}_{intra}$	$0.850_{\pm 0.042}$	$0.931_{\pm 0.041}$	$0.866_{\pm 0.030}$	$\underline{0.881}_{\pm 0.069}$
8		region-level	DINO+$\mathcal{L}_{inter}$	$0.850_{\pm 0.043}$	$\underline{0.938}_{\pm 0.028}$	$0.860_{\pm 0.030}$	$0.874_{\pm 0.059}$
9		region-level	SLPD	$\mathbf{0.864}_{\pm 0.042}$	$\mathbf{0.939}_{\pm 0.022}$	$\underline{0.869}_{\pm 0.039}$	$\mathbf{0.886}_{\pm 0.057}$
K-nearest neighbors (KNN) evaluation							
10	Mean	region-level	DINO	$0.770_{\pm 0.031}$	$0.840_{\pm 0.038}$	$0.837_{\pm 0.014}$	$0.724_{\pm 0.055}$
11		region-level	DINO+$\mathcal{L}_{intra}$	$0.776_{\pm 0.039}$	$0.850_{\pm 0.023}$	$0.841_{\pm 0.012}$	$0.731_{\pm 0.064}$
12		region-level	DINO+$\mathcal{L}_{inter}$	$\underline{0.782}_{\pm 0.027}$	$\underline{0.854}_{\pm 0.025}$	$\underline{0.845}_{\pm 0.014}$	$\underline{0.738}_{\pm 0.080}$
13		region-level	SLPD	$\mathbf{0.792}_{\pm 0.035}$	$\mathbf{0.863}_{\pm 0.024}$	$\mathbf{0.849}_{\pm 0.014}$	$\mathbf{0.751}_{\pm 0.079}$

3.1 Weakly-Supervised Classification

We conduct experiments on two slide-level classification tasks, NSCLC subtyping and BRCA subtyping, and report the results in Table 1. The region-level embeddings generated by SLPD outperform the patch-level embeddings across two aggregators[3] and two tasks (#1∼ 5). This illustrates that learning representations with broader image contexts is more suitable for WSI analysis. Compared with the strong baseline, *i.e.*, the two-stage pre-training method proposed by HIPT (#6), SLPD achieves performance increases of 1.3% and 3.2% AUC on NSCLC and BRCA (#9). Nontrivial performance improvements are also observed under KNN evaluation (#10 vs.#13): +2.3% and +3.1% AUC on NSCLC and BRCA. The superior performance of SLPD demonstrates that learning representations with slide-level semantic structure appropriately can significantly narrow the gap between pre-training and downstream slide-level

---

[3] The feature extraction of the patch-level is impracticable for the ViT-based model due to its quadratic complexity in memory usage.

tasks. Moreover, intra-slide and inter-slide distillation show consistent performance over the baseline, corroborating the effectiveness of these critical components of SLPD.

3.2   Ablation Study

Different Clustering Methods. As discussed in Sect. 2.3, we can alternatively use the global clustering to obtain prototypes and then optimize the network with a similar distillation objective as Eq. 2. For a fair comparison, the total number of prototypes of the two clustering methods is approximately the same. Table 2(#1,2) reports the comparative results, where the slide-level clustering surpasses the global clustering by 0.6% and 1.8% of AUC on NSCLC and BRCA, which verifies the effectiveness of the former. The inferior performance of the global clustering is due to the visual bias underlying the whole dataset.

Table 2. Ablation studies of SLPD. ViT$_{\text{WSI}}$-4096 is the aggregator with region-level embeddings.

#	Ablation	Method	NSCLC		BRCA	
			Acc.	AUC	Acc.	AUC
1	Different cluster-	global	$0.848_{\pm0.045}$	$0.925_{\pm0.033}$	$0.842_{\pm0.048}$	$0.863_{\pm0.060}$
2	ing methods	slide-level	$0.850_{\pm0.042}$	$0.931_{\pm0.041}$	$0.866_{\pm0.030}$	$0.881_{\pm0.069}$
3	Different inter-	region	$0.828_{\pm0.040}$	$0.915_{\pm0.025}$	$0.843_{\pm0.024}$	$0.849_{\pm0.067}$
4	slide distillations	prototype	$0.850_{\pm0.043}$	$0.938_{\pm0.028}$	$0.860_{\pm0.030}$	$0.874_{\pm0.059}$
5		$M=2$	$0.859_{\pm0.036}$	$0.936_{\pm0.021}$	$0.869_{\pm0.039}$	$0.886_{\pm0.057}$
6	Number of prototypes	$M=3$	$0.864_{\pm0.035}$	$0.938_{\pm0.022}$	$0.861_{\pm0.056}$	$0.878_{\pm0.069}$
7		$M=4$	$0.864_{\pm0.042}$	$0.939_{\pm0.022}$	$0.860_{\pm0.031}$	$0.872_{\pm0.060}$
8		$K=1$	$0.864_{\pm0.042}$	$0.939_{\pm0.022}$	$0.869_{\pm0.039}$	$0.886_{\pm0.057}$
9	Number of slide neighbors	$K=2$	$0.862_{\pm0.039}$	$0.938_{\pm0.029}$	$0.875_{\pm0.038}$	$0.889_{\pm0.057}$
10		$K=3$	$0.869_{\pm0.034}$	$0.936_{\pm0.024}$	$0.873_{\pm0.051}$	$0.880_{\pm0.058}$

Different Inter-slide Distillations. The proposed inter-slide distillation is semantic structure-aware at the slide-level, since we build the correspondence between the region embedding and the matched prototype (#4 in Table 2). To verify the necessity of this distillation method, we turn to another design where the inter-slide correspondence is explored through two nearest region embeddings across slides (#3 in Table 2). As can be seen, the region-level correspondences lead to inferior performances, even worse than the baseline (#5 in Table 1), because the learning process is not guided by the slide-level information.

Number of Prototypes. As shown in Table 2(#5∼7), the performance of SLPD is relatively robust to the number of prototypes on NSCLC, but is somewhat affected by it on BRCA. One possible reason is that the heterogeneity of invasive breast carcinoma is low [29], and thus the excessive number of prototypes cannot obtain medically meaningful clustering results. Empirically, we set $M=4$ on NSCLC and $M=2$ on BRCA as the default configuration. We

suggest the optimal number of prototypes should refer to clinical practice, by considering tissue types, cell morphology, gene expression and other factors.

Number of Slide Neighbors. As demonstrated in Table 2($\#5{\sim}7$), the performance of SLPD is robust to the number of slide neighbors. Considering that more slide neighbors require more computation resources, we set $K = 1$ as the default configuration. For more results, please refer to the Supplementary.

4   Conclusion

This paper reflects on slide-level representation learning from a novel perspective by considering the intra- and inter-slide semantic structures. This leads to the proposed Slide-Level Prototypical Distillation (SLPD), a new self-supervised learning approach achieving the more comprehensive understanding of WSIs. SLPD leverages the slide-level clustering to characterize semantic structures of slides. By representing slides as prototypes, the mutual-region/slide relations are further established and learned with the proposed intra- and inter-slide distillation. Extensive experiments have been conducted on multiple WSI benchmarks and SLPD achieves state-of-the-art results. Though SLPD is distillation-based, we plan to apply our idea to other pre-training methods in the future, *e.g.*, contrastive learning [10,11].

References

1. Abbasi-Sureshjani, S., Yüce, A., Schönenberger, et al.: Molecular subtype prediction for breast cancer using H&E specialized backbone. In: MICCAI, pp. 1–9. PMLR (2021)
2. Abbet, C., Zlobec, I., Bozorgtabar, B., Thiran, J.-P.: Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 480–489. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_46
3. Amin, M.B., Greene, F.L., Edge, S.B., et al.: The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more personalized approach to cancer staging. CA Cancer J. Clin. **67**(2), 93–99 (2017)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV, pp. 132–149 (2018)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. **33**, 9912–9924 (2020)
6. Caron, M., Touvron, H., Misra, I., et al.: Emerging properties in self-supervised vision transformers. In: ICCV, pp. 9650–9660 (2021)

7. Chen, L., Bentley, P., et al.: Self-supervised learning for media using image context restoration. MedIA **58**, 101539 (2019)
8. Chen, R.J., et al.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: CVPR, pp. 16144–16155 (2022)
9. Chen, R.J., Lu, M.Y., Williamson, D.F.K., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878 (2022)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML, pp. 1597–1607. PMLR (2020)
11. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
12. Dehaene, O., Camara, A., Moindrot, O., et al.: Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583 (2020)
13. Fashi, P.A., Hemati, S., Babaie, M., Gonzalez, R., Tizhoosh, H.: A self-supervised contrastive learning approach for whole slide image representation in digital pathology. J. Pathol. Inform. **13**, 100133 (2022)
14. Gurcan, M.N., Boucheron, L.E., Can, A., et al.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. **2**, 147–171 (2009)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9729–9738 (2020)
16. Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., Wu, H.: Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 561–570. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_54
17. Jass, J.R.: HNPCC and sporadic MSI-H colorectal cancer: a review of the morphological similarities and differences. Fam. Cancer **3**, 93–100 (2004)
18. Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.: Self-path: self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging **40**(10), 2845–2856 (2021)
19. Kuhn, H.W.: The Hungarian method for the assignment problem. Nav. Res. Logist. Q. **2**(1–2), 83–97 (1955)
20. Lazard, T., Lerousseau, M., Decencière, E., Walter, T.: Self-supervised extreme compression of gigapixel images
21. Levy-Jurgenson, A., et al.: Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Sci. Rep. **10**(1), 1–11 (2020)
22. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR, pp. 14318–14328 (2021)
23. Li, J., Lin, T., Xu, Y.: SSLP: spatial guided self-supervised learning on pathological images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 3–12. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_1
24. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966 (2020)
25. Li, L., Liang, Y., Shao, M., et al.: Self-supervised learning-based multi-scale feature fusion network for survival analysis from whole slide images. Comput. Biol. Med. **153**, 106482 (2023)
26. Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering. In: AAAI, vol. 35, pp. 8547–8555 (2021)

27. Lu, M.Y., Chen, R.J., Mahmood, F.: Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (conference presentation). In: Medical Imaging 2020: Digital Pathology, vol. 11320, p. 113200J. International Society for Optics and Photonics (2020)

28. Lu, M.Y., Williamson, D.F., Chen, T.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**(6), 555–570 (2021)

29. Öhlschlegel, C., Zahel, K., Kradolfer, D., Hell, M., Jochum, W.: Her2 genetic heterogeneity in breast carcinoma. J. Clin. Pathol. **64**(12), 1112–1116 (2011)

30. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

31. Saillard, C., Dehaene, et al.: Self supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. arXiv preprint arXiv:2109.05819 (2021)

32. Saldanha, O.L., Loeffler, et al.: Self-supervised deep learning for pan-cancer mutation prediction from histopathology. bioRxiv, pp. 2022–09 (2022)

33. Sharma, Y., Shrivastava, A., Ehsan, L., et al.: Cluster-to-conquer: a framework for end-to-end multi-instance learning for whole slide image classification. In: Medical Imaging with Deep Learning, pp. 682–698. PMLR (2021)

34. Srinidhi, C.L., Kim, S.W., Chen, F.D., Martel, A.L.: Self-supervised driven consistency training for annotation efficient histopathology image analysis. Media **75**, 102256 (2022)

35. Tavolara, T.E., Gurcan, M.N., Niazi, M.K.K.: Contrastive multiple instance learning: an unsupervised framework for learning slide-level representations of whole slide histopathology images without labels. Cancers **14**(23), 5778 (2022)

36. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via nonparametric instance discrimination. In: CVPR, pp. 3733–3742 (2018)

37. Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., Zheng, Y.: Instance-aware self-supervised learning for nuclei segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 341–350. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_33

38. Yang, P., et al.: CS-CO: a hybrid self-supervised visual representation learning method for H&E-stained histopathological images. Media **81**, 102539 (2022)