VII. Online Appendix

A. Fine-tuning Datasets for ROI Classification

In the finetuning stage for ROI classification, four diverse datasets are utilized to evaluate the classification performance across multiple pathological scales from cellular to tissue level. These datasets are split in a manner reflective of their content: CAM16 [26], pRCC [27], and ROSE [7] are patient-wise divided into training, validation, and test sets with a 7:1:2 ratio, while the WBC [28] dataset, following its original separation, which is a 10:2:5 ratio. To ensure data integrity and prevent leakage, the ROSE and the test sets for other datasets are independent of CPIA-Mini. In addition to Table II, more details of these datasets are provided.

Camelyon16 (CAM16): From the Cancer Metastases in Lymph Nodes challenge, this dataset consists of 540 tumors and 541 normal histopathology images, each cropped to 8000x8000 dimensions.

Papillary Renal Cell Carcinoma Dataset (pRCC): It includes 870 type I and 547 type II histopathology images, with an average size of 2000x2000. Type I images present small cells with clear cytoplasm, while Type II display cells with voluminous cytoplasm and high-grade nuclei.

ROSE Dataset: A private collection from Peking Union Medical College Hospital, this dataset contains cytopathology images from pancreatic liquid samples, including 1,773 pancreatic cancer and 3,315 normal images.

Raabin-WBC Dataset (WBC): Comprising cytopathology images of five blood cell types, it includes 301 basophil, 1,066 eosinophil, 3,461 lymphocyte, 795 monocyte, and 8,891 neutrophil images.

B. Effectiveness on Cell Segmentation

In previous sections, we have evaluated the PuzzleTuning pre-trained models on high-level classification tasks. In this section, we further explore downstream low-level tasks on weakly supervised semantic segmentation (WSSS). Specifically, we adopted CellViT [33], a recent WSSS framework designed for cell segmentation, with ViT as its backbone. A U-Net-like encoder-decoder structure is used with skip connections concatenating each pair of ViT encoder and CNN decoder layers. We have modified the input image size of CellViT from the original 256*256 to 224*224 to match our pre-training ViT. Additionally, we designed CellVPT, which changes the backbone from ViT to VPT by utilizing additional prompt tokens. In both ViT and VPT designs, we evaluate the effectiveness of PuzzleTuning pre-trained models against others.

Following CellViT [33], the PanNuke dataset [34] is applied, which contains 7,904 images in size of 256*256. Specifically, 189,744 annotated nuclei from 19 different tissue types and 5 distinct cell categories are included. With pre-trained weights, the CellViT and CellVPT models are trained for 160 epochs, with a learning rate of 0.00005 for CellViT and 0.0002 for CellVPT. Each comparison experiment is trained under the same or higher-performed settings. In the ViT experiments, the pre-trained ViT are loaded as the backbone while their VPT loads additional empty prompt tokens. In the PuzzleTuning VPT experiments, baseline ViT is ImageNet-trained ViT, and PuzzleTuning pre-trained VPT prompt tokens are loaded. Prompting strategies are explored in the downstream WSSS, where only the prompt tokens are updated in the '+pt' process while '+ft' has all parameters trained.

TABLE VI

THE RESULTS OF PRE-TRAINED WEIGHTS ON THE CELLVIT AND CELLVPT FOR WSSS NUCLEI SEGMENTATION, WHERE VIT/VPT DENOTE DIFFERENT BACKBONE MODELS AND PT/FT DENOTE UPDATING THE PROMPT TOKENS OR UPDATING ALL THE PARAMETERS IN DOWNSTREAM TRAINING.

Initialization	ViT+ft		ViT+pt		VPT+ft		VPT+pt	
	DICE	Jacard	DICE	Jacard	DICE	Jacard	DICE	Jacard
SimCLR	79.37	70.97	78.99	70.88	78.80	70.29	78.99	70.88
MoCo	76.60	66.97	78.13	69.09	75.81	65.72	78.11	69.08
BYOL	76.33	65.87	60.42	48.13	62.11	49.84	63.21	50.61
DINO	77.01	67.51	77.32	67.79	75.73	65.54	77.32	67.79
MAE	79.29	71.05	79.91	72.04	79.60	71.55	79.84	71.98
CAE	74.97	64.58	70.37	57.98	70.60	58.39	70.35	57.96
GCMAE	79.70	71.72	78.16	69.18	79.41	71.37	78.16	69.18
MaskFeat	80.01	72.13	79.81	71.85	79.35	71.19	79.83	71.92
DropPos	78.94	70.45	74.94	64.26	77.53	68.36	75.02	64.40
Jigsaw	77.30	67.63	75.75	65.42	75.02	64.20	75.81	65.47
SimMIM	79.06	70.48	79.49	71.00	77.92	68.77	79.49	70.95
BeyondMask	79.56	71.46	76.25	66.10	74.34	63.52	76.29	66.07
PuzzleTuning	**79.80**	**71.98**	**80.05**	**72.43**	**79.80**	**71.96**	**79.97**	**72.18**

Explicitly enhanced by the pre-training with explicit tasks focusing on appearance consistency, spatial consistency, and restoration understanding, PuzzleTuning further improved the low-level vision tasks for the ViT backbone. As shown in Table VI, PuzzleTuning pre-trained ViT and VPT achieve the DICE score of 79.80, 80.05, 79.80, 79.97, through the finetuning and prompting process of WSSS. The improved results of PuzzleTuning regarding other SSL pre-training methods support the effectiveness of our explicit task design in low-level downstream tasks.

TABLE VII

THE NUCLEI SEGMENTATION PERFORMANCE WITH DIFFERENT PRE-KNOWLEDGE IN PUZZLETUNING PRE-TRAINING, WHERE THE CELLVIT AND CELLVPT ARE USED. THE VIT/VPT DENOTES DIFFERENT BACKBONE MODELS AND PT/FT DENOTES UPDATING THE PROMPT TOKENS OR UPDATING ALL THE PARAMETERS IN DOWNSTREAM TRAINING. THE KNOWLEDGE APPLIED IS DENOTED AS: RANDOM: RANDOMLY SET WEIGHT OF VIT; SEMANTIC: MAE-TRAINED VIT WITH IMAGENET; ABSTRACT: SUPERVISED IMAGENET-TRAINED VIT. TWO PUZZLETUNING CURRICULUM ARE EXPLORED WITH P16-RD: PATCH SIZE IS FIXED AT 16 AND FIX-POSITION RATIO DECAYS FROM 90% TO 20%, P16-R25: PATCH SIZE AND FIX-POSITION RATIO ARE FIXED AT 16, 25%.

Initialization	ViT+ft		ViT+pt		VPT+ft		VPT+pt	
	DICE	Jacard	DICE	Jacard	DICE	Jacard	DICE	Jacard
Random	77.18	67.76	77.18	67.74	75.55	65.23	77.15	67.64
Semantic	77.95	68.97	79.92	71.86	79.56	71.45	79.92	71.86
Abstract	79.78	71.54	79.95	72.14	79.71	71.83	79.82	71.93
PuzzleTuning-p16r25	79.14	70.81	77.21	67.75	79.69	71.69	79.87	71.97
PuzzleTuning-p16rd	78.74	70.33	78.73	69.88	79.87	71.90	79.87	71.91
PuzzleTuning-Abstract	79.80	71.98	**80.05**	**72.43**	79.80	71.96	**79.97**	**72.18**
PuzzleTuning-Semantic	**80.18**	**72.64**	79.96	72.09	**80.00**	**72.14**	79.93	71.99

In low-level vision ablation tasks at Table VII, comparing abstract (ImageNet) with semantic (MAE-ImageNet) knowledge initialization, they generally yield similar DICE performance (+1.83%, +0.03%, +0.25% and -0.1%) while abstract initialization still perform slightly better. Consistently, both are significantly better performance than the random knowledge initialization. Encompassed with PuzzleTuning, bridging the

abstract general knowledge yields consistent improvement in downstream tasks. Specifically, when bridging the pathological images with semantic (MAE-ImageNet) initialization, the performances gain further improvement in the low-level WSSS tasks (ViT+ft and VPT+ft).

Moreover, in downstream WSSS at Table VII, most experiments indicate that updating only the prompt tokens (+pt) achieves better performance than updating the backbone ViT and prompts (+ft). While in the high-level abstract vision tasks, updating both backbone and prompt tokens yields better performance. These findings may related to the task complexity of WSSS, where updating more backbone parameters may encounter slightly more overfitting.

C. Effectiveness on Whole Slide Images Classification

To further evaluate the effectiveness of PuzzleTuning pre-training in pathological image analysis, we explored a lung cancer sub-typing task using whole slide images (WSIs) for slide-level classification. Similar to the region-of-interest (ROI) images in other sections, WSIs are commonly used diagnostic tools [30]. They are of gigapixel scale and cannot be directly processed by GPUs. Most studies crop each WSI into thousands of patches. However, the scarcity of labels results in only having sample-level labels instead of patch-level ones. Given these characteristics, SSL pre-training markedly enhances the performance of WSI models.

Specifically, most WSI studies involve a two-stage process: a feature extraction stage, where the cropped patches are embedded into features, and a feature modeling stage, where a model is employed to analyze these extracted features. Accordingly, we employed the pre-trained ViT into two distinct roles with two widely applied WSI analysis frameworks: CLAM [31] and Graph-Transformer (GTP) [32]. Additionally, we encompass the prompt tokens in several explorations, where the VPT is applied with the pre-trained ViT backbone and prompt tokens.

In this study, 789 lung adenocarcinoma (LUAD), 707 lung squamous cell carcinoma (LUSC), and 567 non-cancerous (normal) tissue WSI samples from TCGA-LUAD and TCGA-LUSC [29] datasets are used. The WSIs are separated into training, validation, and test sets based on patient ID with a ratio of 7:1:2. All models are trained 50 epochs with the same learning rate of 0.00005 or hyper-parameters optimized for higher performance.

1) In WSI feature extraction: Applied in the feature extraction stage, PuzzleTuning-trained ViTs are employed in CLAM [31] to embed cropped images into feature vectors of 768 dimensions. Subsequently, the CLAM framework is trained with the embedded feature vectors for slide-level classification. As depicted in the first two columns of Table VIII, the performance of the comparisons varies significantly due to the fixed feature extractor, which is not updated during downstream training. Among them, the PuzzleTuning-trained VPT achieves the highest performance with an accuracy of 87.42%. Furthermore, compared to most other pre-trained models, PuzzleTuning pre-trained ViT models demonstrate superior performance.

TABLE VIII

THE ACCURACY OF PUZZLETUNING PRE-TRAINED MODEL IN WSI SLIDE-LEVEL CLASSIFICATION. SERVING IN THE FEATURE EXTRACTION STAGE OF CLAM [31], PRE-TRAINED MODEL EMBEDS PATCHES INTO FEATURES; SERVING IN THE FEATURE MODELING STAGE OF GTP [32], PRE-TRAINED MODELS PREDICT SLIDE-LEVEL CATEGORY WITH THE EMBEDDED WSI FEATURES. VIT: VIT-BASE MODEL, VPT: PROMPTING WITH ADDITIONAL PROMPT TOKENS OF VIT. IN THE DOWNSTREAM TASKS: FT: FINETUNING ALL PARAMETERS, PT: PROMPTING ONLY THE PROMPT TOKENS. TWO PUZZLETUNING CURRICULUM DESIGNS ARE EXPLORED WITH P16-RD: PATCH SIZE IS FIXED AT 16 AND FIX-POSITION RATIO DECAYS FROM 90% TO 20%, P16-R25: PATCH SIZE AND FIX-POSITION RATIO ARE FIXED AT 16, 25%.

Initialization	Feature Extraction		Feature Modeling			
	ViT	VPT	ViT+ft	ViT+pt	VPT+ft	VPT+pt
Random	53.99	–	73.93	76.69	72.70	76.69
ImageNet	**82.82**	–	76.07	77.91	76.69	77.91
SimCLR	65.34	–	76.07	74.54	77.30	74.54
MoCo	53.68	–	74.85	77.31	76.07	77.31
BYOL	55.21	–	73.93	73.62	75.77	73.62
DINO	54.60	–	76.07	76.69	76.07	76.69
MAE	57.06	–	75.77	75.46	77.91	75.46
CAE	53.68	–	73.62	74.85	75.77	74.85
GCMAE	61.35	–	74.23	75.46	76.69	75.46
MaskFeat	64.72	–	73.01	74.54	76.07	74.54
DropPos	58.90	–	73.62	74.23	75.46	74.23
Jigsaw	52.15	–	75.46	77.30	75.46	77.30
SimMIM	55.83	–	74.85	77.91	76.69	77.91
BeyondMask	64.42	–	75.15	77.30	76.99	77.30
PuzzleTuning-p16r25	64.42	77.91	76.07	76.69	74.23	78.83
PuzzleTuning-p16rd	**74.85**	80.67	**76.69**	77.61	76.69	75.15
PuzzleTuning	66.56	**87.42**	76.38	**78.22**	**79.14**	**79.45**

However, the highest performance with ViT backbone alone is observed with ImageNet's natural knowledge initialization (accuracy of 82.82%). As the pre-trained ViTs underperform compared to the ImageNet baseline without pathological image pre-training, the drop in embedding performance (ranging from -16.26 to -30.67%) further underscores the importance of general vision knowledge. Nevertheless, through prompt tuning in PuzzleTuning, additional domain-bridging knowledge can be effectively encoded, resulting in a performance increase (82.82 to 87.42%). Moreover, explicit bridging yields significant performance improvements (87.42 vs. 66.56%) over implicit finetuning of all parameters. Its efficacy at the WSI feature extraction stage supports our objective of adapting natural to the pathological images by learning the bridging prompts.

2) In WSI feature modeling: Applied in the feature modeling stage, GTP [32] is built with the ViT/VPT backbone for slide-level classification. In their training process, the GTP models are initialized with ViT weights from the pre-training stage, and if the VPT is used, the empty additional prompt tokens are attached. For the PuzzleTuning experiments, the ViT+pt loads pre-trained ViT and empty prompt tokens, while the VPT+pt loads baseline ViT and PuzzleTuning pre-trained prompt tokens. In the '+ft' experiments, the ViT and VPT have all parameters trained, while only the prompt tokens are updated in the '+pt' process.

In the right four columns of Table VIII, the PuzzleTuning pre-trained models achieve the slide-level classification accuracy of 76.38%, 78.22%, 79.14%, and 79.45%. Consistent with other comparison experiments against the SOTAs, applying PuzzleTuning-trained models achieve the highest performance in the WSI feature modeling stage. In the slide-level down-

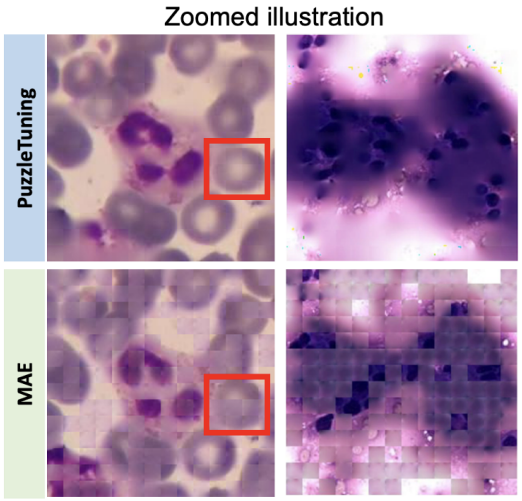

Fig. 9. Illustrations of the reconstructed images (by MAE and PuzzleTuning) with the visible patches (the unmasked patches or hint patches) replaced by the original patches. The combined image of PuzzleTuning closely resembles the original image with a low-saturation purple color feature, presenting a smooth and coherent appearance. In contrast, the MAE-generated patches display significant color differences from the original patches.

stream task, the relationships of distinct features in the WSI are effectively modeled. This further proves the effectiveness of multiple puzzle restoration tasks for building relationships based on appearance and spatial consistency. Additionally, the findings in WSI applications further enhance its ability as a general approach in pathological image analysis.

D. Reconstruction Performance

We illustrate the original images, puzzles, and reconstructions in PuzzleTuning training in Fig. 2. Regarding the SSL task to understand pathological images, both MAE and PuzzleTuning are trained through reconstruction. However, the task focuses differ in their learning and restoring approaches, where PuzzleTuning employs multiple puzzles restoration to explicitly learn the grouping, junction, and semantic alignment relationship. In contrast, MAE mainly focuses on interpreting the junction relationship of remaining patches with MIM. With the explicit task design on appearance consistency, spatial consistency, and restoration understanding of Fig. 1, we visually compare PuzzleTuning with MAE. In Fig. 8, guided by 25% of visible original patches, a batch of 4 images and their puzzles and masked samples are generated.

Regarding appearance consistency, PuzzleTuning demonstrates a superior capability to reconstruct homogeneous global and local features compared to MAE. Globally, on the first image set of blood cells (sample 1) in Fig. 8, the combined image of PuzzleTuning closely resembles the original image with a low-saturation purple color feature, presenting a smooth and coherent appearance. In contrast, the MAE-generated patches display significant color differences from the original patches. Locally, highlighted in 9 with red boxes, PuzzleTuning consistently imparts the concave shape in the center of all surrounding red blood cells. In contrast, the MAE-generated images fail to preserve the concave shape of most red blood cells. Similarly, in the second set (sample 2), PuzzleTuning generates cells based on limited visible original patches, restoring cell clusters from the original image. However, MAE-generated images do not effectively restore cell morphology around visible patches, thus failing to achieve the desired consistency between cells.

Regarding spatial consistency presented with junction patches, PuzzleTuning achieves smooth image reconstruction, preserving the advanced texture alignment toward visible original patches. On the contrary, the images generated by MAE exhibit relatively coarse connections between patches. Shown explicitly in the second and the third image sets (sample 2 and 3) in Fig. 8, the edges between different patches in MAE-generated images become notably prominent. Their colors and textures present misalignment to the visible original patches. Additionally, the expected junction textures, such as rounded cell shape, which should be interpreted through the junction patches, are fuzzy in MAE-generated patches in Fig. 8. In contrast, PuzzleTuning reconstructs junction patches more clearly with continuous cell texture and color. Aligned closely with visible patches, PuzzleTuning clearly maintains spatial consistency through smooth transitions along patch edges.