Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology

Supplementary Material

1. Additional Method Detail

1.1. Attention Matrix

Here, we further formulate the e_{ij}^l of Equation (6) in the manuscript as,

$$e_{ij}^l = \frac{\left(H_i^l W^Q\right)\left(H_j^l W^K\right)^T}{\sqrt{D}}, \tag{1}$$

where H_i^l is the i-th instance features in l-the region H^l. With EPEG, the R-MSA can be further represented as,

$$\begin{aligned} \text{R-SA} &= \sum_{j=1}^{M\times M} \alpha_{ij}^l \left(H_j^l W^V\right), \\ \text{R-MSA} &= \text{Concat}\left(\text{R-SA}_1, \cdots, \text{R-SA}_{N_{head}}\right) W^O \end{aligned} \tag{2}$$

where the projections W^Q, W^K, W^V, and W^O are parameter matrices sharing across the regions. The N_{head} denotes the number of heads.

1.2. Pseudocodes of CR-MSA

Algorithm 1 gives the details about CR-MSA.

2. Dataset Description

CAMELYON-16 [1] is a WSI dataset proposed for metastasis diagnosis in breast cancer. The dataset contains a total of 400 WSIs, which are officially split into 270 for training and 130 [1] for testing, and the testing sample ratio is 13/40≈1/3. Following [2, 10, 14], we adopt 3-times three-fold cross-validation on this dataset to ensure that each slide is used in training and testing, which can alleviate the impact of data split and random seed on the model evaluation. Each fold has approximately 133 slides. Although CAMELYON-16 provides pixel-level annotations of tumor regions, for weakly supervised learning, we only utilize slide-level annotations.

TCGA NSCLC includes two sub-type of cancers, Lung Adenocarcinoma (**LUAD**) and Lung Squamous Cell Carcinoma (**LUSC**). There are diagnostic slides, LUAD with 541 slides from 478 cases, and LUSC with 512 slides from 478 cases. There are only slide-level labels available for this dataset. Compared to CAMELYON-16, tumor regions in tumor slides are significantly larger in this dataset.

[1]Two slides in the test set are officially considered to be mislabeled, so they are not included in the experiment.

TCGA-BRCA includes two sub-types of cancers, Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC). There are 779 IDC slides and 198 ILC slides. **TCGA-BLCA** contains 376 cases of Bladder Urothelial Carcinoma.

Following prior works [10, 11, 13], we crop each WSI into a series of 256×256 non-overlapping patches at 20X magnification. The background region, including holes, is discarded as in CLAM [10].

3. Implementation Details

Following [10, 11, 13], we use the ResNet-50 model [5] pretrained with ImageNet [3] as the backbone network to extract an initial feature vector from each patch, which has a dimension of 1024. The last convolutional module of the ResNet-50 is removed, and a global average pooling is applied to the final feature maps to generate the initial feature vector. The initial feature vector is then reduced to a 512-dimensional feature vector by one fully-connected layer. As for PLIP [6] features, we also use a fully-connected layer to map 512-dimensional features to 512 dimensions. The head number of R-MSA is 8. An Adam optimizer [7] with learning rate of 2×10^{-4} and weight decay of 1×10^{-5} is used for the model training. The Cosine strategy is adopted to adjust the learning rate. All the models are trained for 200 epochs with an early-stopping strategy. The patience of CAMELYON-16 and TCGA are 30 and 20, respectively. We do not use any trick to improve the model performance, such as gradient cropping or gradient accumulation. The batch size is set to 1. All the experiments are conducted with NVIDIA GPUs. Section 7 gives all codes and weights of the pre-trained PLIP model.

4. Additional Quantitative Experiments

4.1. More on Foundation Model Features

In this section, we evaluate the improvement of R^2T on foundation model features with more MIL models. Figure 1 shows the advantage of R^2T by online fine-tuning, which can further enhance the discriminability of foundation model features on multiple tasks and models. Surprisingly, we find that this improvement does not decrease with more advanced MIL models, such as R^2T+CLAM often outperforms R^2T+ABMIL.

Algorithm 1: PyTorch-style pseudocode for CR-MSA

```
# x:  input instance features
# phi:  learnable parameters Φ 2 ℝ^{c×k}
# msa:  native MSA function

# initialize
r, p, c = x.shape
logits = einsum("r p c, c k -> r p k", x, phi).transpose(1,2)
# compute softmax weights
combine_weights = logits.softmax(dim=-1)
dispatch_weights = logits.softmax(dim=1)
# compute minmax weights
logits_min,_ = logits.min(dim=-1)
logits_max,_ = logits.max(dim=-1)
dispatch_weights_mm = (logits - logits_min) / (logits_max - logits_min + 1e-8)
# get representative features of each region
x_region = einsum("r p c, r k p -> r k p c", x,combine_weights).sum(dim=-2)
# perform native msa
z = msa(x_region)
# distribution of representative features
z = einsum("r k c, r k p -> r k p c", z, dispatch_weights_mm)
# combination k of Φ
z = einsum("r k p c, r k p -> r k p c", z, dispatch_weights).sum(dim=1)
```

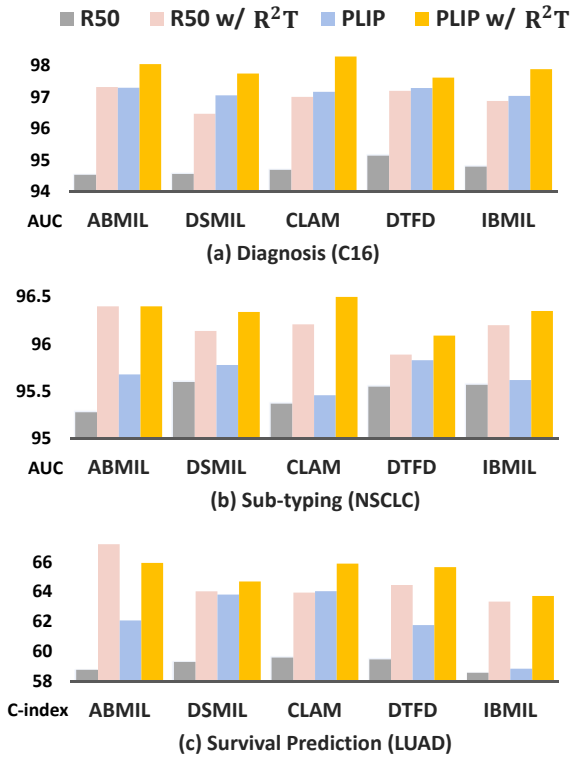

Figure 1. Performance improvement by adding R^2T on different offline features.

4.2. More on EPEG

Different Convolution Kernel. Here, we discuss the impact of different convolution kernels on EPEG. The k is the optimal kernel size, and the upper part of Figure 5 discusses more on different datasets. First, we find that most 1-D convolutional kernels enhance the re-embedding ability of local Transformers. Second, larger convolution kernels typically perform worse. We attribute this to the excessive parameters that tend to overfit on a limited number of slides.

Type	Kernel	C16	NSCLC	LUAD
w/o	none	96.82	96.01	65.45
2-D	3 3	96.73	95.93	66.70
2-D	7 7	96.60	95.71	64.59
1-D	**k 1**	**97.32**	**96.40**	**67.19**

Different Embedded Position. Here, we discuss another variant of EPEG. We place the convolution module after the "value" matrix instead of the default "attn" matrix. Figure 3 shows the specific structures of the two variants. The results in Table 1 demonstrate the feasibility of the "value" variant, but its performance is significantly lower than the original version, especially on the more challenging C16 dataset. We note this is because the original version can incorporate positional information into the core attention matrix to model positional information more directly.

Type	Kernel	C16	NSCLC	LUAD
w/o	none	96.82	96.01	65.45
value	3×3	96.78	96.07	64.47
value	k×1	96.90	96.10	65.76
attn	**k×1**	**97.32**	**96.40**	**67.19**

Table 1. Comparison results of variants of EPEG.

Figure 2. The performances under different region partition strategies on two datasets.

Figure 3. Illustration of variants of EPEG. The left one is the default.

4.3. Region Partition Strategy

Here, we systematically investigate the impacts of different region partition strategies in our method. Figure 2 reports the computational pathology performances of our method with different region partition settings. From observations, we find that the 2-D region partition fashion is superior to the 1-D ones because it preserves more original image structure information. Another phenomenon is that employing a too-small or too-large region will degenerate the performances. We attribute this to the fact that the small region sharply reduces the module receptive field, while the large region damages the diversity of the re-embedded features. Therefore, a moderate region is optimal for R-MSA. Furthermore, even with a larger number of region partitions (e.g., 512 or 32 32), the proposed model still maintains a high level of performance. This reflects that our model can achieve a good trade-off between performance and efficiency (more regions resulting in lower spatial and temporal costs). These findings demonstrate the good scalability of our model to the datasets that have longer input sequences.

4.4. More Parameters are not Always Better

Generally, a Transformer is a multi-layer structure that stacks several blocks with the same structure [4, 8, 9, 12]. However, due to the task specificity, R^2T only contains a few blocks. Here, we systematically investigate the impact of different numbers of layers and different blocks on the performance and computational cost of R^2T. First, Figure 4 shows two different blocks: (a) is used by default in R^2T, (b) introduces a feed-forward network (FFN), which plays an indispensable role in Transformer for NLP or natural image computer vision tasks. From Table 2, we can find that FFN introduces a large number of parameters and computation, but the more expensive computation cost does not bring performance improvement.

Moreover, we can summarize that: 1) Transformer-based methods, represented by TransMIL, bring better long-sequence modeling ability, but also introduce several times more parameters. R^2T-MIL, as one of them, has equal or less parameter size compared to TransMIL and better performance. 2) As Transformer-based methods, Re-embedding paradigm and R^2T have Higher parameter efficiency. "+N-MSA" has the same model structure as TransMIL (N-MSA 2), but thanks to the Re-embedding, it can achieve higher performance with lower cost. Moreover, R^2T (w/o CR-MSA) leverages a more excellent design to further improve performance and training time. This indicates that R^2T-MIL has a good parameter compression space, and can achieve significant improvement with parameters compared to DTFD, reaching the level of the foundation model. 3) In computational pathology, limited by the number of slides, the models face the problem of overfitting[13], and higher parameter size does not imply better performance. Not only does TransMIL perform poorly on some tasks, we also add FFN and increase the number of layers on the R^2T-MIL to increase the parameters of R^2T-MIL (+7.35M), but the performance drops significantly (-4.12% on LUAD).

	PretrainingData	Para. of ine #	TrainTime#	Memo#	FPS"	C16"	NSCLC"	LUAD"
ABMIL	ImageNet-1K	0.65M+26 M	3.1s	2.3G	1250	94.54	95.28	58.78
ABMIL+PLIP [11]	OpenPath-200K	0.65M+151M	1.7s	2.2G	2273	97.30	95.68	62.09
DTFD [41]	ImageNet-1K	0.79M+26 M	5.1s	2.1G	325	95.15	95.55	59.48
TransMIL [22]	ImageNet-1K	2.67M+26 M	13.2s	10.6G	76	93.51	94.97	64.11
Re-embedding								
ABMIL+N-MSA [36]	ImageNet-1K	1.64M+26 M	7.7s	7.2G	158	96.20	95.51	63.99
R^{2}T-MIL(w/o CR-MSA)	ImageNet-1K	1.64M+26 M	6.1s	10.0G	272	96.89	96.24	63.03
R^{2}T-MIL	ImageNet-1K	2.70M+26 M	6.5s	10.1G	236	97.32	96.40	67.19
R^{2}T-MIL [w/ FFN]$_{x2}$	ImageNet-1K	6.90M+26 M	9.9s	12.0G	114	96.23	95.58	64.89
R^{2}T-MIL [w/ FFN]$_{x3}$	ImageNet-1K	10.05M+26 M	14.4s	16.3G	71	96.57	95.70	63.07

Table 2. More about the ef ciency analysis of R^{2}T. FFN denotes the feed-forward network. We use subscripts to indicate the number of layers. It is worth noting that the feature dimensions of the remaining terms except PLIP features are 1024, while PLIP is 512, which explains its rise in ef ciency compared to ABMIL. Other than that, the PLIP feature does not have any computational cost impact on the original method.

case	C16	NSCLC	LUAD
0	97.32	96.40	67.19
500	96.99 (-0.33)	96.23 (-0.17)	62.35 (-4.84)
1000	97.21 (-0.11)	95.98 (-0.42)	62.15 (-5.04)

Table 3. Comparison results between global MSA and local MSA. We perform global MSA computation instead of regional MSA for bags with instance numbers less than the threshold.

Figure 4. Illustration of different blocks of R^{2}T. The (a) is the default.

4.5. More on Local Transformer

Table 3 further explores the impact of local Transformer on computational pathology performance. We set a threshold and perform global MSA computation instead of regional MSA for bags with instance numbers less than that threshold. First, we can nd that more use of global MSA leads to worse performance on both datasets. The characteristic of small tumor areas on the C16 dataset exacerbates the performance degradation caused by global MSA. In addition, this strategy introduces extra hyperparameters, reducing the generalization ability of the model. Overall, our experiments prove that local Transformers can better adapt to the inherent characteristics of WSI such as huge size and small tumor areas than traditional global Transformers.

4.6. Discussion of Hyper-parameter in CR-MSA

The bottom part of Figure 5 shows the results. We can nd that R^{2}T is not sensitive to this parameter, and different values can achieve high-level performance. In addition, different of ine features show similar consistency. This re ects the generality of the preset optimal parameter in different scenarios.

5. Additional Visualization

Figure 6 presents more comprehensive feature visualizations, including cases where the original features have high and extremely low discriminativeness. We use the attention score after softmax normalization to label instances for demonstrating the updated features. Moreover, when the tumor prediction con dence is too low, we assume that the attention score cannot directly indicate the instance tumor probability. In this case, we still use original instance labels to colorize the visualization.

From Figure 6, we can draw the following conclusions: 1) Although the nal MIL model can correctly classify the slide with high discriminator original features, the feature visualization after linear projection (row 1) is still unsatisfactory (high coupling and unclear cohesion). 2) Features

with low discriminativeness (rows 2 and 3) impair the judgment of the MIL model, and the re-embedding module can effectively enhance feature discriminativeness. 3) In features re-embedded by global MSA, the distribution area of tumor instances is linearly correlated with the final tumor prediction score. The larger the distribution area of tumor instances, usually higher its tumor prediction score. However, too many instance numbers and an extremely low tumor instance ratio make it difficult for the module to re-embed all instances as tumor instances, which ultimately affects the performance of the MIL model. We attribute this to a lack of diversity in features re-embedded by global MSA. 4) In contrast, regional MSA addresses this problem well. Because features among different regions are distinct from each other, even if the proportion of re-embedded tumor instances is still low, their discriminativeness is very high (high cohesion and low coupling), which is more favorable for the classification of the final MIL model.

6. Limitation

Although the Transformer-based re-embedding module can effectively improve the discriminativeness of instance features and facilitate classification, we find that the re-embedded features lose their original label information due to the self-attention update of the original features. For example, an original non-tumor patch may be re-embedded as a tumor patch to benefit slide classification. This characteristic of the re-embedding module prevents it from accurately performing weakly supervised localization or segmentation of tumor areas through the final aggregation module. However, the outstanding weakly supervised localization and segmentation capability is one of the features of attention-based MIL models. Therefore, how to use the re-embedding module to improve detection or segmentation performance is our future work.

7. Code and Data Availability

The source code of our project will be uploaded at https://github.com/DearCaat/RRT-MIL.

CAMELYON-16 dataset can be found at https://camelyon16.grand-challenge.org.

All TCGA datasets can be found at https://portal.gdc.cancer.gov.

The script of slide pre-processing and patching can be found at https://github.com/mahmoodlab/CLAM.

The code and weights of PLIP can be found at https://github.com/PathologyFoundation/plip.

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA, 318(22):2199–2210, 2017. 1

[2] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16144–16155, 2022. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 1

[6] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine, pages 1–10, 2023. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 1

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019. 3

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021. 3

[10] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 5(6):555–570, 2021. 1

[11] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. NeurIPS, 34, 2021. 1

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 3

[13] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In Proceedings of the IEEE/CVF Conference on Computer Vi-

sion and Pattern Recognition, pages 18802–18812, 2022. 1, 3

[14] Xiaoxian Zhang, Sheng Huang, Yi Zhang, Xiaohong Zhang, Mingchen Gao, and Liu Chen. Dual space multiple instance representative learning for medical image classification. In 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press, 2022. 1

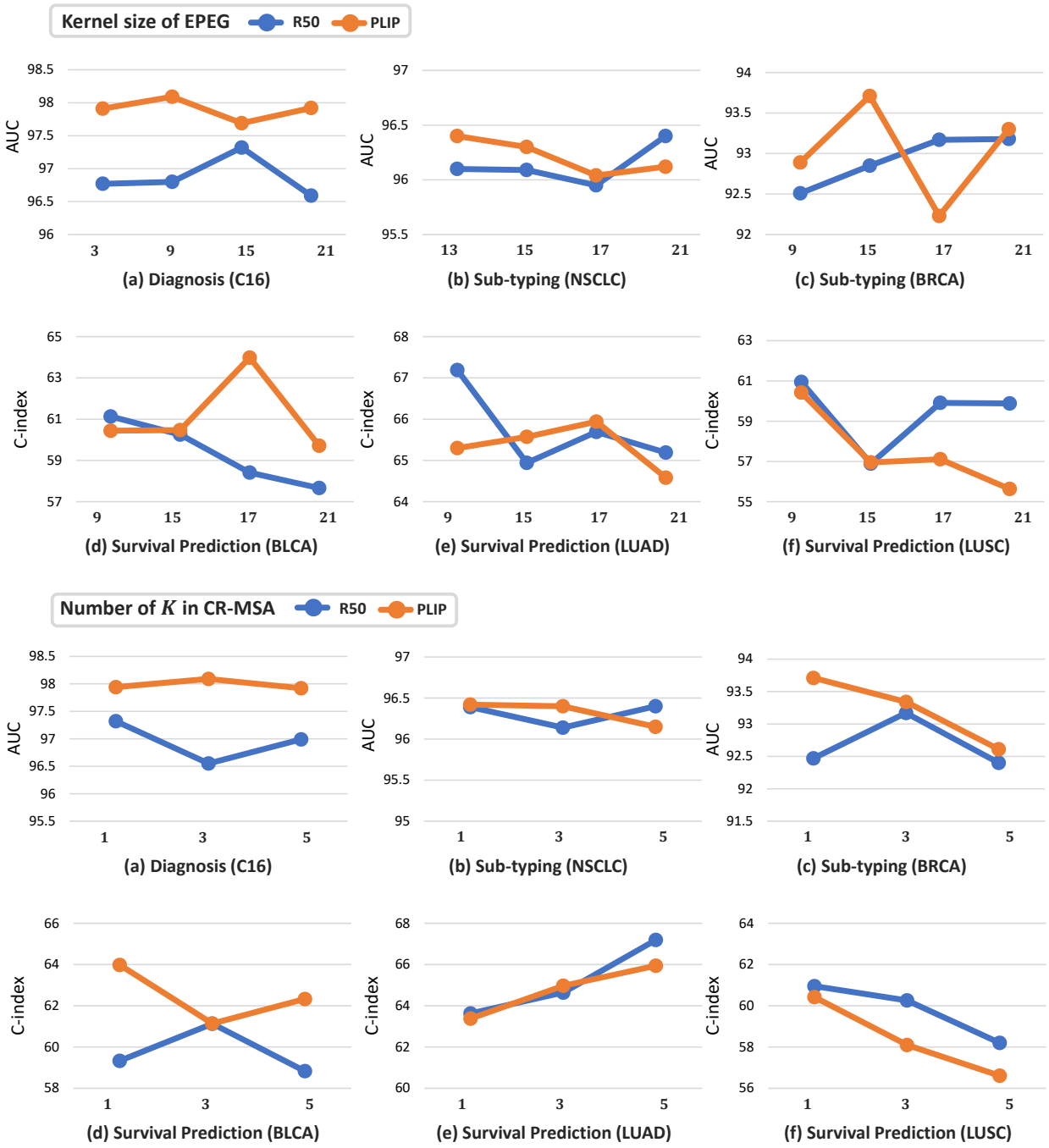

Figure 5. Discussion of important hyper-parameters.

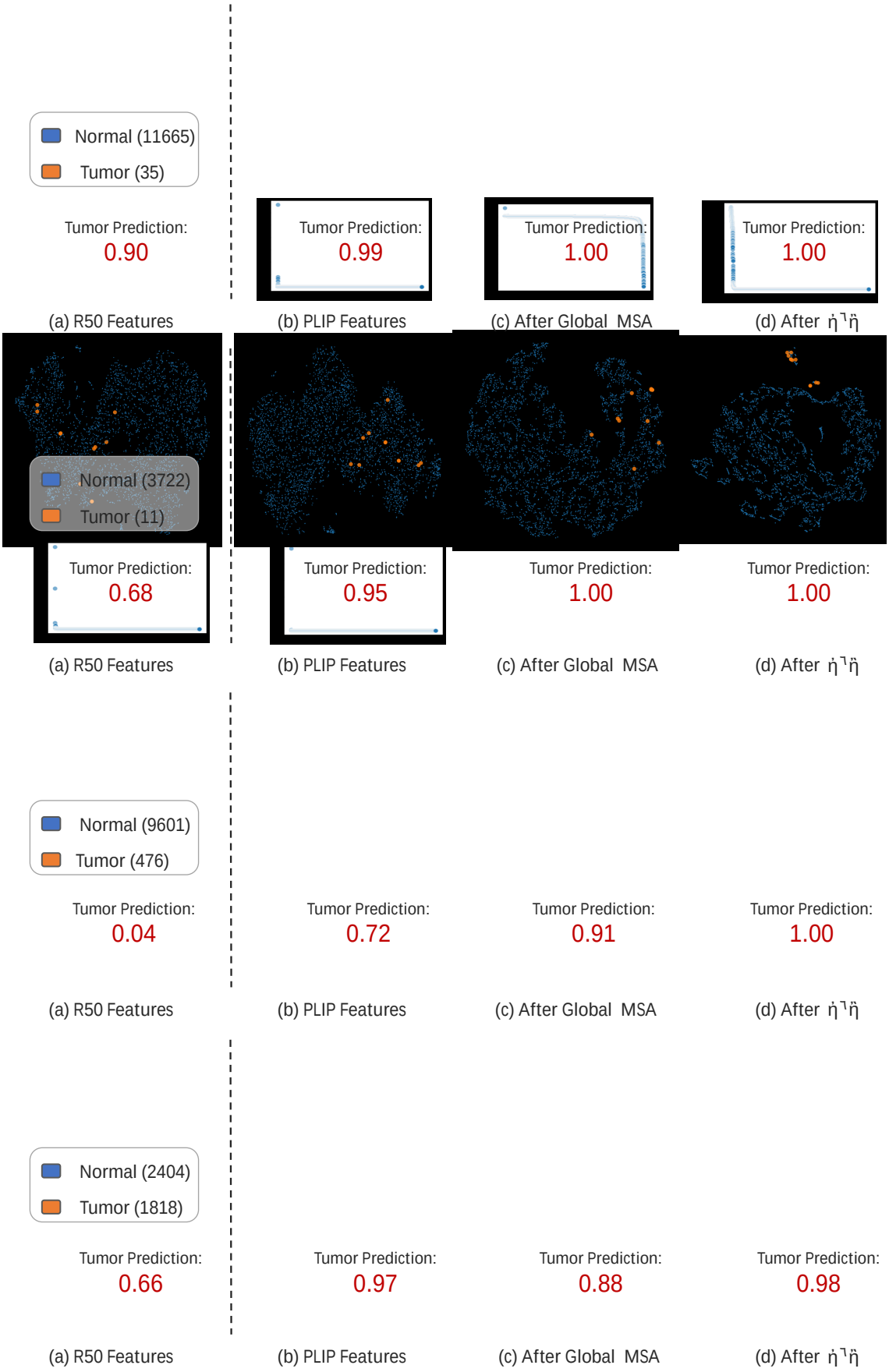

Figure 6. More comparison of t-SNE visualization of instance features. Best viewed in scale.