Supplementary Material of Submission 2187

A Implementation of Non-parametric Baselines

In this part, we introduce the detailed implementations of baselines non-parametric classification methods. Specifically, for the KNN methods, we apply the global average pooling and max pooling to get global features for both query and gallery. And the ProtoType baseline uses the global averaged feature over positive patches and negative patches, guided by the mask. The SimpleShot [23] baseline (also MI-SimpleShot [5] using fair backbone in this paper) is similar to ProtoType, while it applies extra centering and normalization. Note, ProtoType and SimpleShot use the best query patch as ours. While KNN methods use only one global feature vector for the query, thus the top-k operation is conducted on the global gallery features following the common number 5. Note, ProtoType and SimpleShot use the same mask label as ours. To ensure a fair comparison with these baselines, we improve the performance of baseline methods as much as possible. The main improvement is about the logit design. Specifically, previous KNN methods only return integer predictions, which lost confidence information and lead to suboptimal AUC results. Thus, we improve the baselines via converting their integer predictions to float probabilities to better evaluate AUC, and use a threshold searched on val set to decide the label for accuracy evaluation. Note, we randomly take 100 validation WSIs and 8 high-quality positive gallery WSIs (tumor patches size from 1000 to 3000) for 5 repeat experiments. The data split and evaluation protocol are the same as ours for fair comparison.

B Ablation Study on Hyperparameters

Param.	v1	v2	v3	v4	v5
N_k	93.5	91.78	94.52	96.20	96.55
t	96.55	96.33	96.21	96.21	95.75
τ	96.54	96.55	96.33	96.32	95.39

Table A. Grid search for hyperparameters via AUC (%) on CAMELYON16 [1] val split. Number of gallery N_k in Eq. 4 and softmax temperature τ in Eq. 6 use values: 5, 10, 20, 30, 40 (v1 - v5), while related threshold t in Eq. 5 is searched among 0.88, 0.89, 0.9, 0.91, 0.92.

In this paper, we conduct grid search on the CAMELYON16 validation split to set hyperparameters. The involved parameters are listed in Tab. A, with the searched values in the caption. We set initial N_k, t, τ as 20, 0.9, 10, respectively. Then, we search N_k in the first, and set the best N_k before searching the next t. In this way, we search τ after setting N_k and t. Experimentally, and we find the related threshold t and softmax temperature τ are robust hyperparameters with little fluctuations. From this table, we set N_k, t, τ to 40, 0.88, 10, respectively. This setting is also applied to CAMELYON17 and CAMELYON16-C.

C Visualization to Understand INC Process

To better understand how our method works, we depict the retrieval processes of INC in Fig. A. The most representative query patch $\hat{\mathbf{s}}$ in Eq. 4 are depicted in the left with black borders. We find our method searching right positives patches (blue borders) and negative patches (red borders) in general, which visualizes the top positive gallery patches and negatives in Eq. 4. These results suggest we can generate reasonable logit in Eq. 4, which is close to positive local features and far from negative local features, as much as possible. However, we also find there are some cancer cells in the negative patches, which introduce noises and interference to the classification. Therefore, taking sufficient gallery patches are important to obtain robust logits (*e.g.*, 40 for CAMELYON16 dataset).

Another solution is to retrieve more representative queries to smooth the logits, which is the motivation of our retrieval aggregation. As shown in this figure, the patches with gray borders are the related queries searched in the second query-to-query retrieval. The retrieved instances are quite similar in visual patterns, which ensures the correctness to mean their underlying logits. Overall, our INC methods are visually explainable with high robustness.

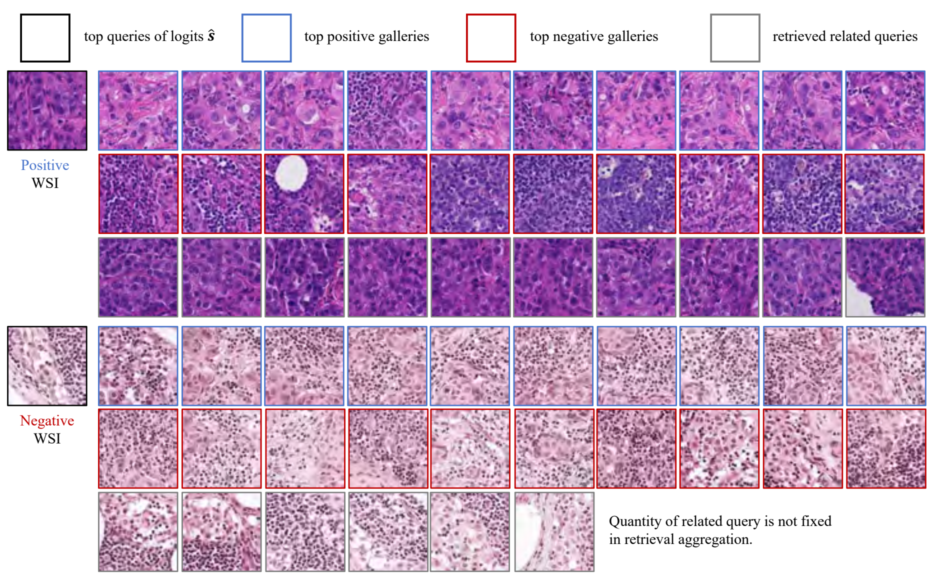

Fig. A. Visualization of top positive galleries, negative galleries in Eq. 4, and retrieved query patches in Eq. 5. These results help to understand the process of INC method.