Appendix D. Qualitative Results

The quantitative metrics evaluated in our study work on the assumption that labels are good indicators of how visually similar patches are to one another. It assumes that patches belonging to the same class should be more like each other than patches of different classes. But this is not always the case, and it depends a lot on how the dataset was built, how patches were obtained from WSI, especially if the dataset contains very large classes and has a lot of variability. In particular, in the ULiege dataset, there are several classes containing patches that are visually dissimilar and seem to match better to samples from another class than their own. Figure D.1 illustrates such situation.

To complement the quantitative results, we present a study on the embedding vectors obtained by each model (Section D.1), as well as qualitative retrieval results (Sections D.2-D.4). For each model, several queries and their top-1 results are shown and discussed. The queries were selected at random from some classes. For the ULiege dataset, only 6 classes out of the 67 have been selected based on the number of samples they contain (2 of the smallest classes, 2 of the biggest and 2 of medium size). The results of our, non-expert, qualitative evaluations are however not fully exhaustive or objective, as it was not realistic to present them for every query of every dataset. Ideally, a formal reader study involving multiple pathologists should be conducted to derive less biased qualitative results. However, this was beyond the scope of this work due to the quantity of models being evaluated and the variability of the dataset in terms of pathology, requiring different expertise in pathology.

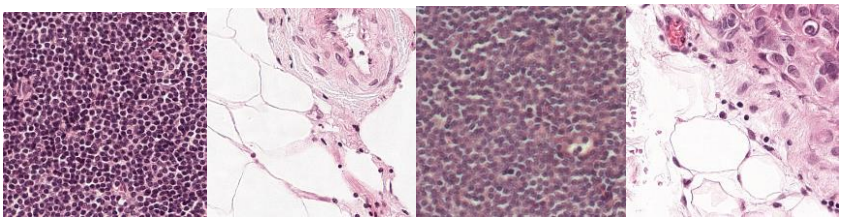

Figure D.1 This figure shows four patches from the ULiege dataset. The first two belong to the Camelyon16 0 class. The third patch resembles the first one but belongs to the Janowczyk7 0 class. The fourth one resembles the second patch but belongs to the Camelyon16 1 class.

Appendix D.1: Embeddings

In Figure D.2, the image representations of the ULiege indexing set are projected into a two-dimensional space using t-SNE [72]. The 67 classes are distinguished by different colors, as indicated by the color scale displayed alongside each graph.

Overall, most models produce well-separated clusters in the embedding space. Notable exceptions include the pre-trained variants of iBOT and BYOL, as well as cdpath21-gan, which fail to generate clusters that clearly discriminate between classes.

The effect of the training strategy is further reflected in the definition of cluster boundaries. For out-of-domain models trained in a self-supervised manner (bottom-left of the figure), the pre-trained variants yield clusters with less well-defined boundaries compared to models trained from scratch or subsequently fine-tuned. This observation is linked to the fact that the representations learned by pre-trained models are less discriminative, as they rely exclusively on features derived from natural image data without prior exposure to the target domain. Interestingly, self-supervised models exhibit more sharply defined cluster boundaries than supervised models (top-left).

Finally, as a byproduct, these visualizations highlight the class imbalance present in the ULiege indexing set. Certain classes contain substantially more samples than others, resulting in two consequences: (i) a higher likelihood of retrieving samples from majority classes, and (ii) improved representation learning for these classes, compared to underrepresented ones, in models trained on the ULiege data.

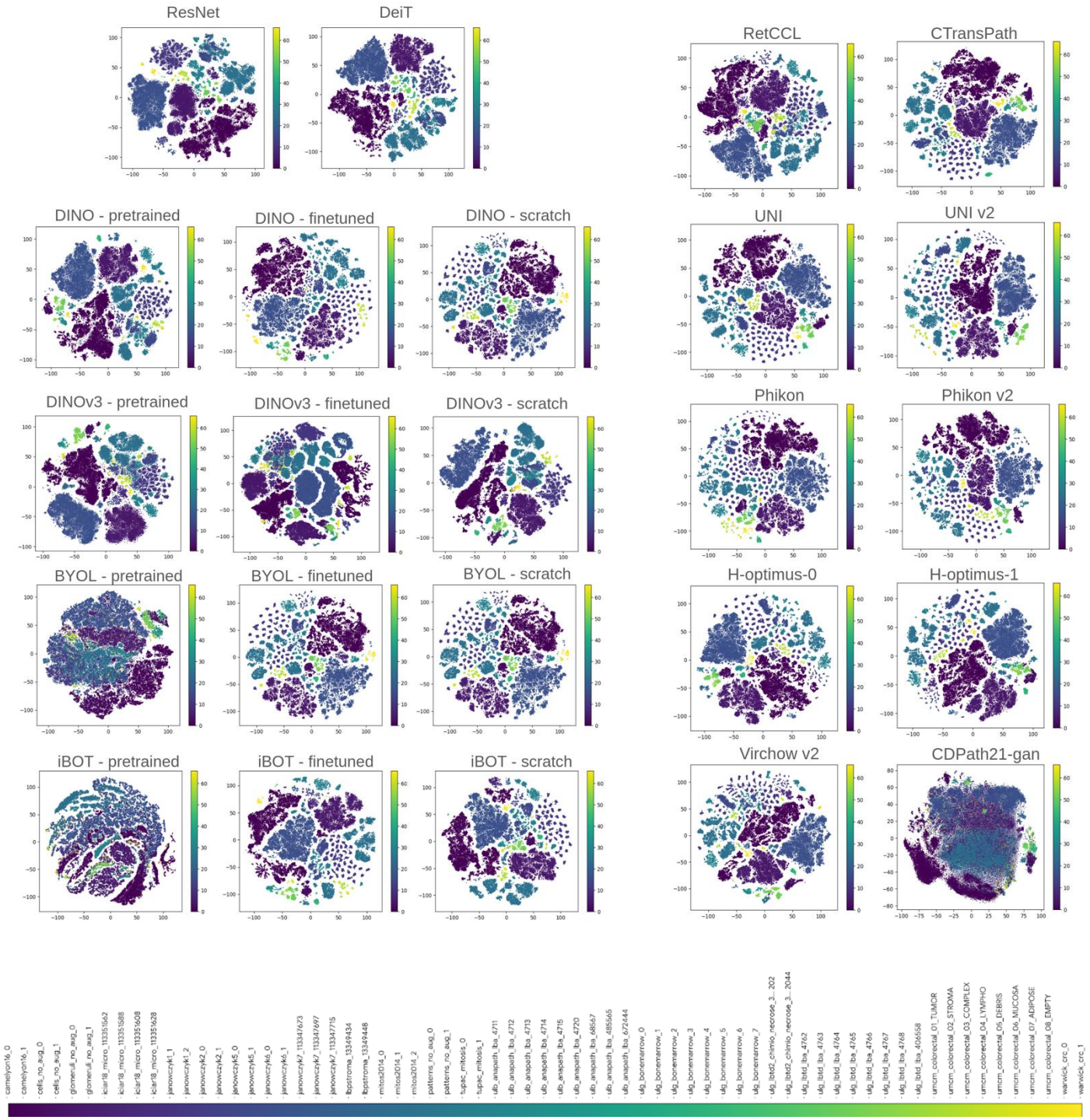

Figure D.1: 2D representations of the embeddings of the indexing set of the ULiege dataset, using the t-SNE method computed by the supervised models (top left), the out-of-domain SSL models (bottom left) and the in-domain SSL models (right)

Appendix D.2. Supervised models

Figure D3 shows the Top-1 results retrieved using the ResNet and DeiT models for queries selected from six classes of the ULiege dataset. The queries were selected such that two belong to the smallest dataset classes (d - e), two to the dominant classes (c - f) and two to intermediate classes (a - b). At first glance, the images retrieved by both ResNet and DeiT appear visually similar to their corresponding queries. However, examining the labels reveals that four of the ResNet's retrieved images are from a different class than the query, while DeiT has one mismatch. This further illustrates intra/extra class diversity, whereby images from classes different from the query can nevertheless be visually similar, and, according to the models, sometimes more so than images from the correct class. In this case, the class size does not

appear to impact the results, as all retrieved images are visually similar to the queries regardless of class, and label mismatches occur in both a dominant class (f) and a minority one (e).

The same tests were carried out on the other two datasets and the results are shown in Figure D4. The results for NCT-CRC-HE-100K are consistent across both models, with the retrieved images matching the query and only one label mismatch. This is less the case for Camelyon17-WILDS. Both models retrieve similar images, but they are slightly less similar to the query. However, it is important to note that the labels still match. This discrepancy can be attributed to the dataset characteristics: the images in the validation dataset (used as queries) differ in tone from those in the test set (indexed in the database).

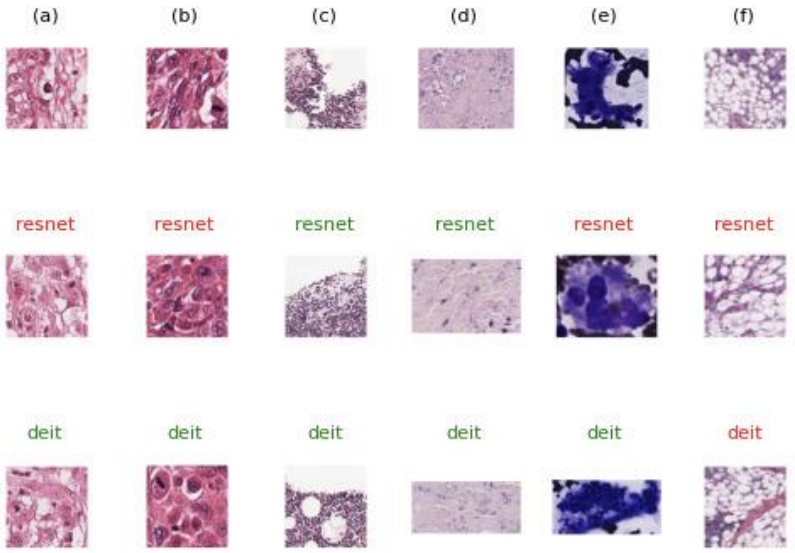

Figure D3: Qualitative results with the two supervised models for query images sampled from the following classes of the ULiege dataset: (a) mitos2014 1 (b) tupac mitosis 1, (c) Camelyon16 0, (d) ulg lbtd2 chimio necrose 36362044, (e) patterns no aug 0 and (f) janowczyk6 1. The first row shows the query image, the second row shows the top-1 retrieved image obtained with the ResNet model, and the third row shows the top-1 retrieved image obtained with the DeiT model. Extractor names highlighted in green (resp. red) indicate that the retrieved image belongs (resp. does not belong) to the same class as the corresponding query. All retrieved images, even the misclassified ones, are from the same project as the query.

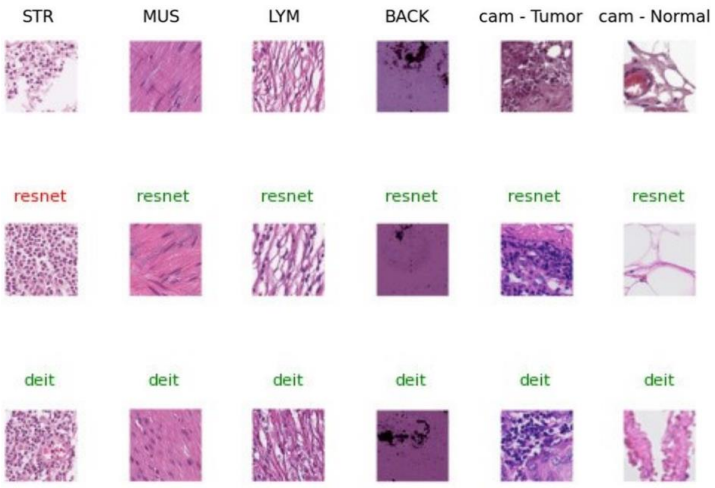

Figure D4: Qualitative results with the two supervised models for queries taken from the Normal and Tumor classes of the Camelyon17-WILDS dataset (first two columns), and from the BACK, LYM, MUS, and STR classes from the NCT-CRC-HE-100K dataset (columns 3 to 6). The first row shows the query image, the second row shows the top-1 retrieved image obtained with the ResNet model, and the third row shows the top-1 retrieved image obtained with the DeiT model. Extractor names highlighted in green (resp. red) indicate that the retrieved image belongs (resp. does not belong) to the same class as the corresponding query.

Appendix D.3. Self-supervised out-of-domain models

Figures D5 and D6 show the Top-1 retrieved images for each model and for queries of the three datasets. All models retrieve images that match the query well for both the ULiege and the NCT-CRC-HE-100K datasets. The pre-trained BYOL model that struggles slightly more, retrieving images that are less visually similar than the other models for queries a, b and d in the ULiege dataset, and queries a and c in the NCT-CRC-HE-100K dataset. The models trained from scratch using Dinov1 and v3 exhibit a slight visual mismatch for the MUS class query of the NCT-CRC-HE_100K dataset, though this does not extend to the other classes of that dataset. These results are consistent with the quantitative results shown in Section 6. For the Camelyon17-WILDS dataset, the results are consistent across models, with similarities observed between the retrieved images.

Examining the labels of the retrieved images in comparison to the query again reveals the influence of intra-class versus extra-class diversity. For instance, in case (a) of the ULiege dataset (Figure D5), almost no models retrieved an image from the same class as the query, while all retrieved images nevertheless appear visually similar to the query. However, it is important to note that despite the label mismatches, all retrieved images come from the same project as the query (except one, with BYOL). This suggests that while the models may not have focused on small details, they are capable of recognizing the organ/tissue type.

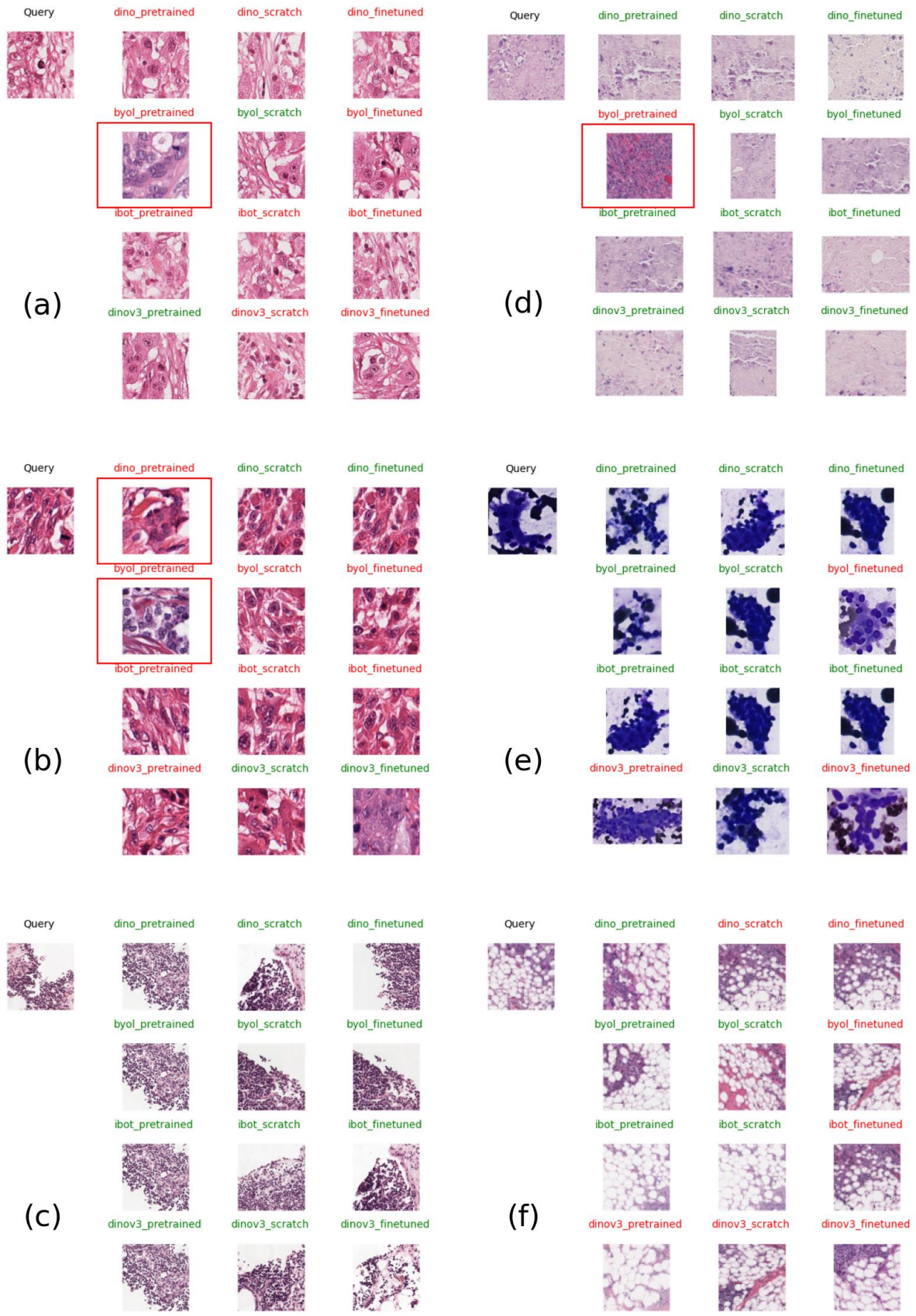

Figure D5: Qualitative results with all out-of-domain self-supervised models for queries taken from the following classes from the ULiege dataset: (a) mitos2014 1 (b) tupac mitosis 1, (c) Camelyon16 0, (d) ulg lbtd2 chimio necrose 36362044, (e) patterns no aug 0 and (f) janowczyk6 1. Extractor names highlighted in green (resp. red) indicate that the retrieved image belongs (resp. does not belong) to the same class as the corresponding query. Retrieved images surrounded by a red box do not belong the same project as the query.

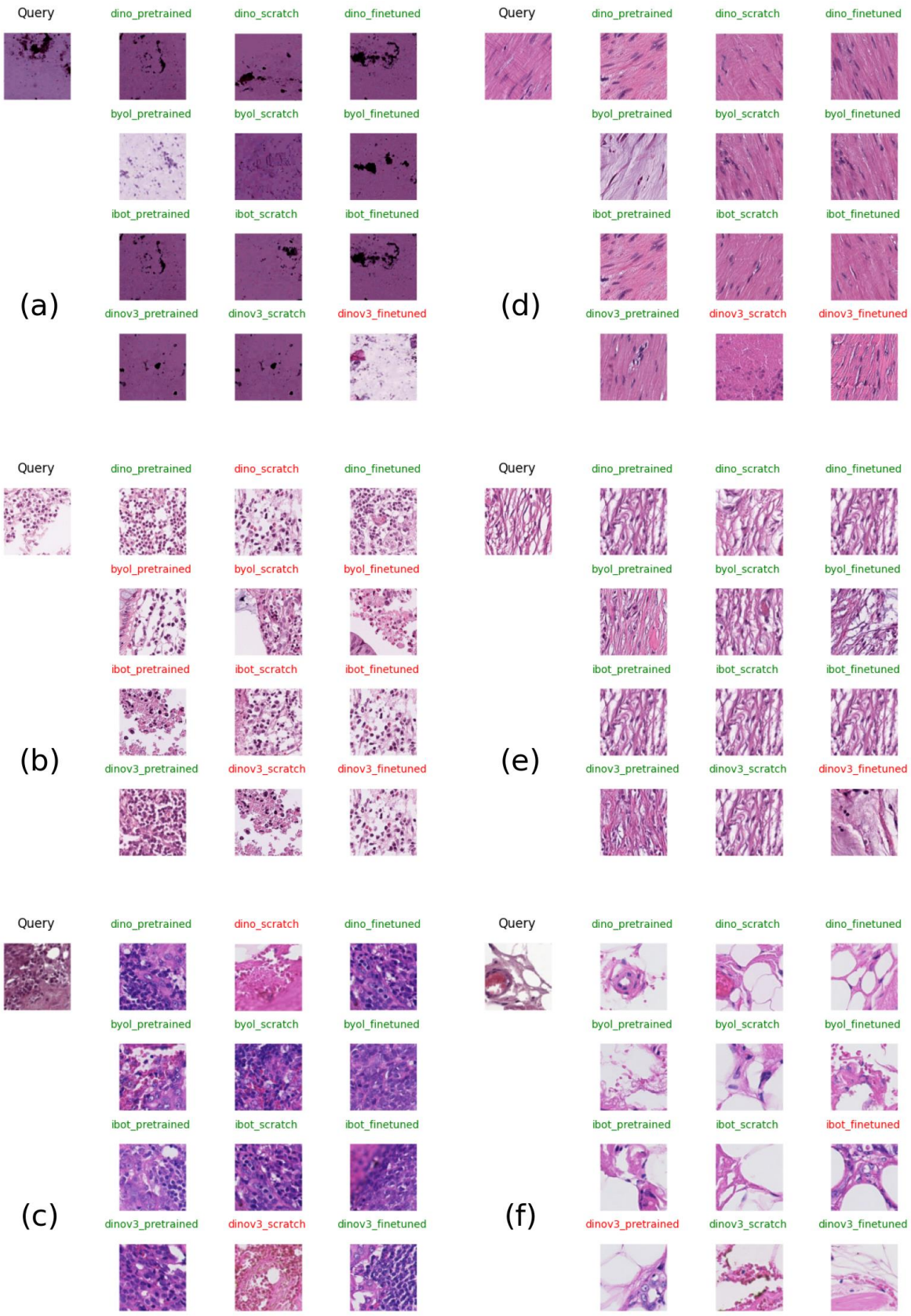

Figure D6: Qualitative results with all out-of-domain self-supervised models for queries taken from the BACK (a), LYM (b), MUS (d), and STR (e) classes of the NCT-CRC-HE-100K dataset and from the Tumor (c) and Normal (f) classes of the Camelyon17-WILDS dataset. Extractor names highlighted in green (resp. red) indicate that the retrieved image belongs (resp. does not belong) to the same class as the corresponding query.

Appendix D.4. Self-supervised in-domain models

Figures D7 and D8 show the Top-1 results of each model for different queries. All models retrieve images that match the corresponding query on the ULiege dataset. On the NCT-CRC-HE-100K dataset, the retrieved images also closely match the queries, showing good performances across all models. For the Camelyon17-WILDS dataset, however, we can again observe that all retrieved images are similar to one another but match the query less well due to the discrepancy between the test and validation sets.

The cdpath21-gan model presents better qualitative results than expected given its poor quantitative performance. However, a closer inspection of the results on the ULiege dataset reveals that the Top-1 images consistently differ slightly from both the query and the Top-1 images of the other models. In particular, in case (b) (Figure D7), there is a slight difference in tone, while in case (c), (e) and (f), the Top-1 image contains additional visual elements not found in the other models' results (bottom left in case (c), bottom right in cases (e) and (f)). This observation seems to be consistent with the quantitative results. The other models confirm their quantitative performance, as all retrieved images remain highly similar to the queries. UNI v1 and H-optimus-0 provide slightly less similar images for query (a) of the ULiege dataset, but several elements of their retrieved images can still be directly matched to the queries.

Regarding the labels, we again observe that some Top-1 results that look similar to their queries are not of the same label. It confirms however the quantitative results of cdpath21-gan as we can see that it frequently retrieved a ‘wrong’ query compared to the other models that are much more consistent. It is also the only model that returns a result from a different project on the ULiege dataset, rather than just mistakenly taking an image from a different class but similar project.

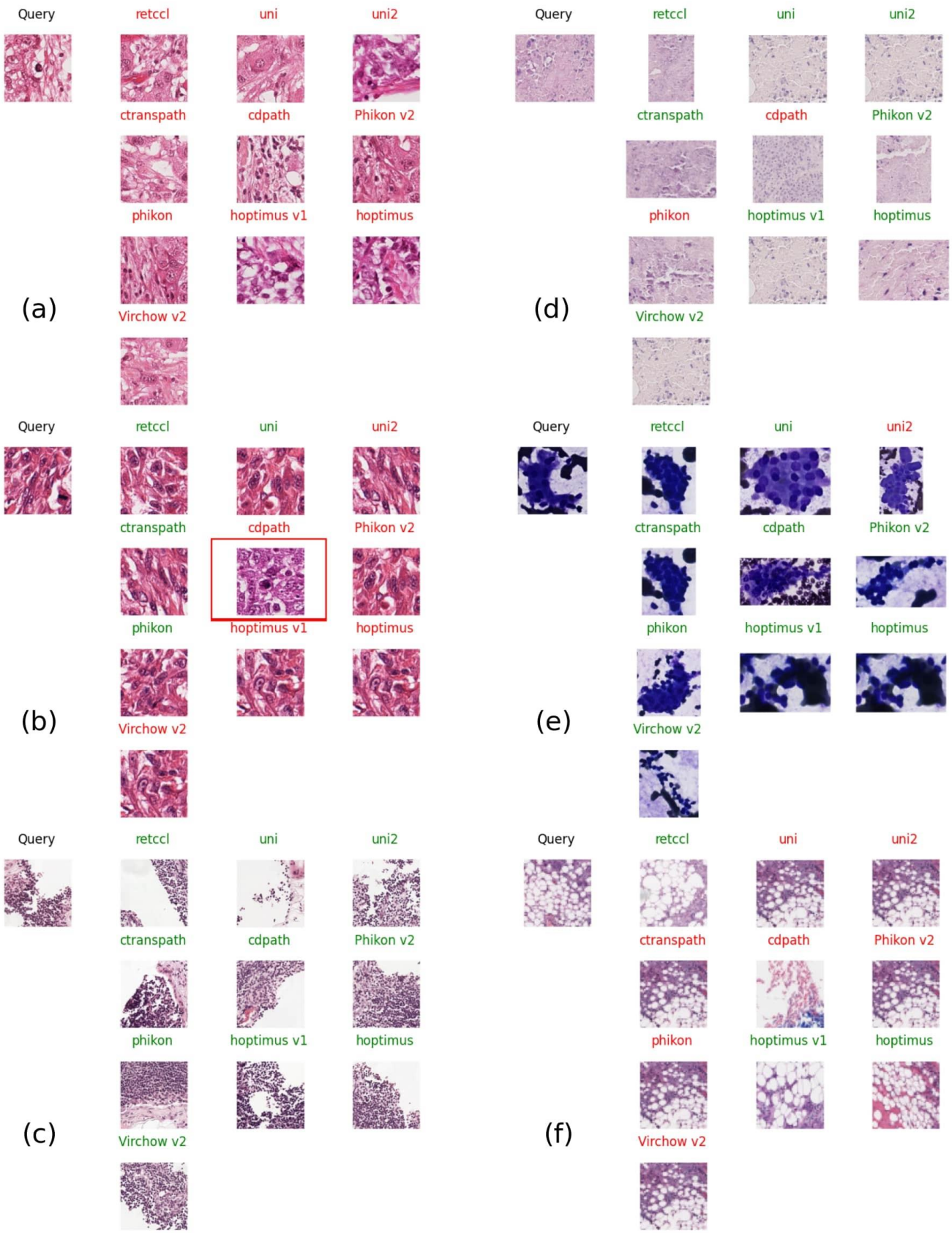

Figure D7: Qualitative results with all in-domain self-supervised models for queries taken from the following classes of the ULiege dataset: (a) mitos2014 1 (b) tupac mitosis 1, (c) Camelyon16 0, (d) ulg lbtd2 chimio necrose 36362044, (e) patterns no aug 0 and (f) janowczyk6 1. Extractor names highlighted in green (resp. red) indicate that the retrieved image belongs (resp. does not belong) to the same class as the corresponding query. Retrieved images surrounded by a red box do not belong the same project as the query.

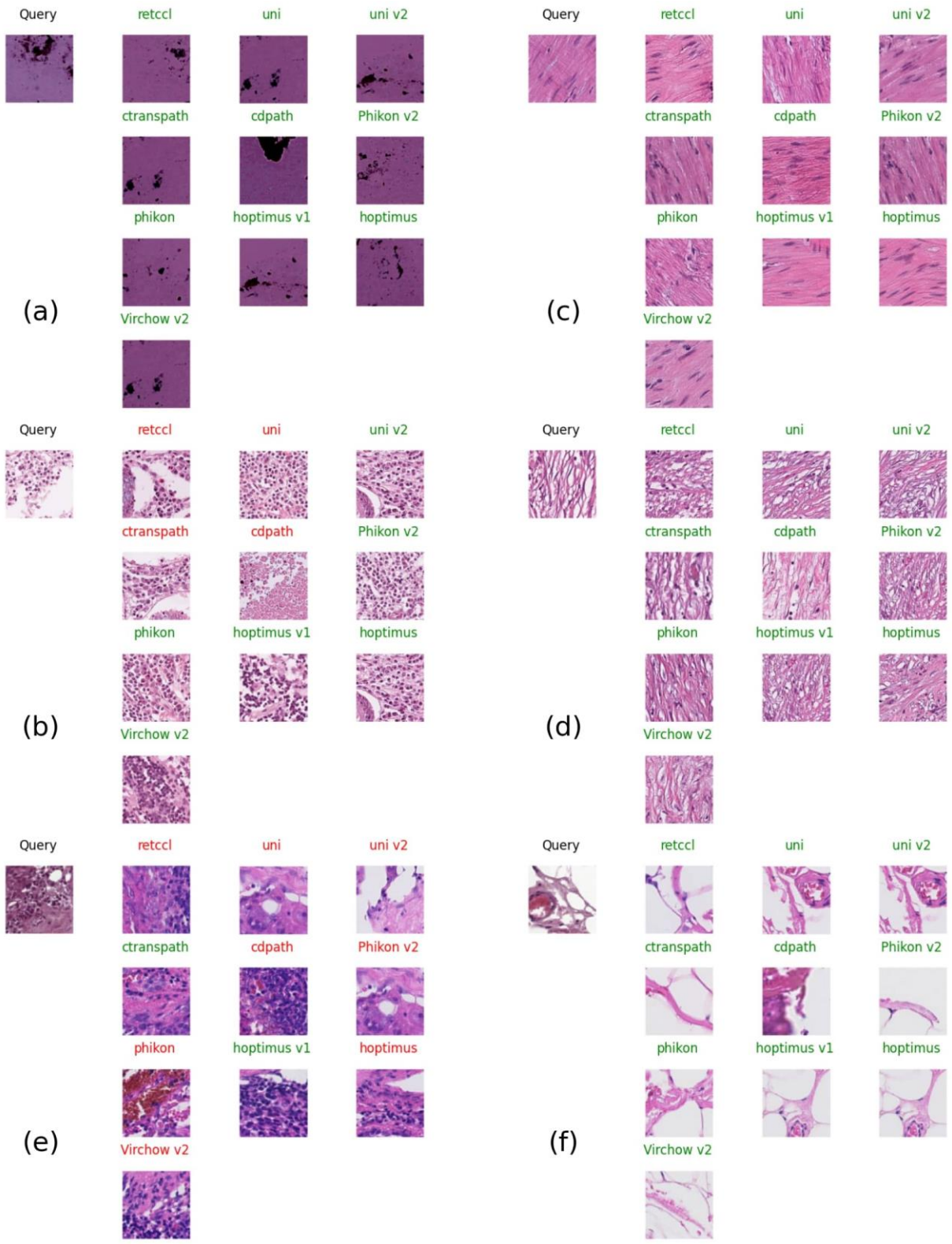

Figure D8: Qualitative results with all in-domain self-supervised models for queries taken from the BACK (a), LYM (b), MUS (c), and STR (d) classes of the NCT-CRC-HE-100K dataset and from the Tumor (e) and Normal (f) classes of the Camelyon17-WILDS dataset. Extractor names highlighted in green (resp. red) indicate that the retrieved image belongs (resp. does not belong) to the same class as the corresponding query.