

Hierarchical Generalized Category Discovery for Brain Tumor Classification in Digital Pathology

Matthias Perkonigg^{*1}, Patrick Rockenschaub¹, Georg Göbel¹, and Adelheid Wöhrer²

¹Institute of Clinical Epidemiology, Public Health, Health Economics, Medical Statistics and Informatics, Medical University of Innsbruck, Austria

²Institute of Neuropathology and Neuromolecular Pathology, Medical University of Innsbruck, Austria

Abstract

Accurate brain tumor classification is critical for intra-operative decision making in neuro-oncological surgery. However, existing approaches are restricted to a fixed set of predefined classes and are therefore unable to capture patterns of tumor types not available during training. Unsupervised learning can extract general-purpose features, but it lacks the ability to incorporate prior knowledge from labelled data, and semi-supervised methods often assume that all potential classes are represented in the labelled data. Generalized Category Discovery (GCD) aims to bridge this gap by categorizing both known and unknown classes within unlabelled data. To reflect the hierarchical structure of brain tumor taxonomies, in this work, we introduce Hierarchical Generalized Category Discovery for Brain Tumor Classification (HGCD-BT), a novel approach that integrates hierarchical clustering with contrastive learning. Our method extends contrastive learning based GCD by incorporating a novel semi-supervised hierarchical clustering loss. We evaluate HGCD-BT on OpenSRH, a dataset of stimulated Raman histology brain tumor images, achieving a +28% improvement in accuracy over state-of-the-art GCD methods for patch-level classification, particularly in identifying previously unseen tumor categories. Furthermore, we demonstrate the generalizability of HGCD-BT on slide-level classification of hematoxylin and eosin stained whole-slide images from the Digital Brain Tumor Atlas, confirming its utility across imaging modalities.

Keywords: Brain Tumor Classification, Generalized Category Discovery, Semi-supervised learning

1 Introduction

For brain tumor patients, surgical intervention is often a critical component of treatment. During surgery, the intra-operative classification of brain tumors is essential to guide personalized decision-making. Depending on the underlying pathological classification the surgical radicality varies between biopsy and supramaximal resection. However, the large diversity and complexity of brain tumor types make it difficult to assemble representative datasets that span the full spectrum of pathologies for training machine learning models. Current classification methods are typically constrained to a set of pathological patterns defined prior to training of a model. In practice, however, datasets often lack pathological patterns of certain subtypes, the number of different subtypes may be unknown (e.g. when sourced from a large-scale, unlabelled database), or include only partially annotated subsets, typically biased toward the most prevalent tumor types. Unsupervised and semi-supervised methods can achieve remarkable success in extracting general-purpose feature representations that could be leveraged for a wide-range of downstream tasks, as demonstrated by foundation models in digital pathology (e.g. [3, 33, 28]). However, unsupervised learning approaches are inherently unable to incorporate prior knowledge in the form of labelled data, while semi-supervised methods typically assume that all potential classes are present in the labelled subset of the data. These limitations hinder their applicability to neuro-oncology, where datasets

^{*}corresponding author: matthias.perkonigg@i-med.ac.at

are both incomplete and imbalanced. Generalized Category Discovery (GCD) addresses these limitations by aiming to categorize unlabelled data through a combination of labelled and unlabelled samples. The unlabelled samples may belong to a known category represented in the labelled set or to novel categories that need to be discovered [27]. Different GCD approaches [27, 20, 35] have demonstrated strong performance in image analysis tasks using benchmark datasets of natural images. When extended to brain tumor classification based on imaging, GCD approaches have the potential to reduce the need for extensive labelling, especially for rare or difficult-to-annotate tumor subtypes. Moreover, by revealing previously unrecognized subgroups, these approaches could contribute to a finer-grained understanding of tumor biology and support precision oncology. Brain tumor classification is inherently hierarchical, as reflected in the WHO classification system [15]. This hierarchical structure encodes clinically meaningful relationships that directly influence diagnosis and treatment. Designing a GCD approach to leverage hierarchical information offers the opportunity to align computational discovery with underlying biological organization, improving robustness and interpretability.

Contribution In this work, we propose a novel approach to Generalized Category Discovery (GCD) tailored for the specific challenges of brain tumor classification. Since brain tumor taxonomies naturally follow a hierarchical structure, therefore the key idea of our approach is to explicitly model this hierarchy to capture the underlying pathological concepts better than existing GCD methods focusing mainly on contrastive learning. To this end we introduce Hierarchical Generalized Category Discovery for Brain Tumor Classification (HGCD-BT), a method that incorporates this hierarchical organization directly during the training process. This is achieved by extending the unsupervised hierarchical loss proposed by [37] to a novel semi-supervised hierarchical clustering loss function specifically tailored to the setting of GCD. We conduct experiments on a dataset of Stimulated Raman Histology (SRH) data of brain tumors [12] demonstrating that HGCD-BT shows promising performance for GCD for brain tumor classification and outperforms GCD methods relying solely on contrastive learning. Furthermore, we validate its generalizability on hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) [26], achieving accurate slide-level classification across 12 tumor types. These results highlight the robustness of HGCD-BT across both imaging modalities and classification granularities.

2 Related work

GCD is a setting in ML in which the goal is to classify images in a dataset, a subset of which has known labels. GCD aims at assigning labels to all remaining images, using class labels that may or may not have been observed in the labeled subset. The setting of GCD has been primarily explored in natural image classification [27, 20, 35]. In addition, GCD has been combined with active learning [17] or continual learning [31, 36]. Particularly relevant to our work are GCD methods considering the hierarchical nature of categories. To leverage this inherent hierarchy, a variety of techniques have been proposed. These include self-coding to implicitly learn a category tree [21], hierarchical pseudo-labeling [22], semi-supervised hierarchical clustering [9] or hierarchical prototyping [30]. [8, 14] use the properties of the hyperbolic space to model hierarchical relationships and improve category discovery. GCD is rarely applied to medical imaging, one approach [8] explores discovery of concepts via probabilistic modeling in diverse medical imaging datasets.

Artificial Intelligence for Stimulated Raman Histology SRH, a label-free optical imaging technique, able to produce virtual histology images in 2-3 minutes was introduced in the clinic [19, 10] and has the potential to revolutionize intraoperative diagnosis. SRH utilizes laser-based imaging to identify variations in macromolecule concentration within biomedical specimens, and creates imaging contrast based on these differences. It has been demonstrated that SRH can be used to provide non-inferior human diagnosis compared to frozen section diagnosis in a significantly shorter timeframe [7]. The inherently digital nature of SRH images and their real-time application during surgery led to the development of various ML methods, particularly for the classification of different brain tumor types, primarily covering the more common tumor types. Convolutional Neural Networks (CNNs) were used in fully-supervised algorithms to classify SRH images in low quality, tumor and non-tumor regions [25], to differentiate between patches of tumor recurrence and pseudoprogression [11] or to classify 13 diagnostic tumor types [11]. Self-supervised representation

learning was applied to the differentiation between primary Central nervous system (CNS) lymphoma (PCNSL) and other CNS entities [24].

Artificial Intelligence for FFPE brain tumor classification Artificial Intelligence has been extensively applied to hematoxylin and eosin (H&E)-stained slides of formalin-fixed and paraffin-embedded (FFPE) for brain tumor subtyping. Specifically, [23] identified 23 studies dealing with glioma subtyping. Those studies focus on classifying a limited set of two to five common subtypes. Besides studies specifically build for brain tumor classification, different foundation models in pathology [3, 29] use brain tumor subtyping as a downstream task to demonstrate the model. They demonstrate promising performance by applying linear probing on tasks such as predicting IDH status [3, 29, 32] or common tumor subtyping [3]. However, different from GCD methods, linear probing requires annotation for all tumor subtypes.

3 Method

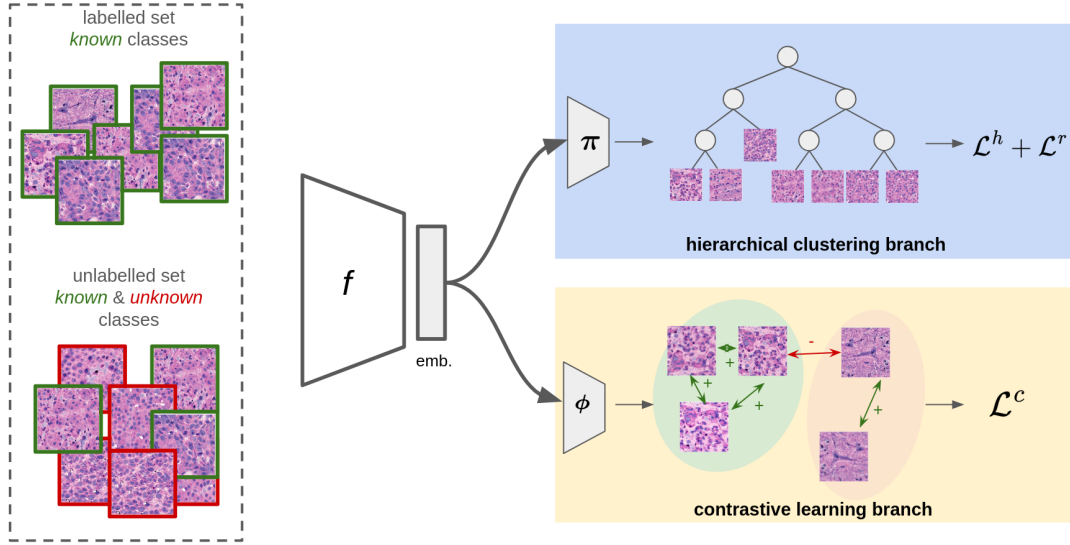


Figure 1: Overview of our proposed method HCGD-BT consisting of a hierarchical clustering branch and a contrastive learning branch.

GCD is a framework for classifying images when only a subset of the dataset has known class labels. The core goal is to assign appropriate labels to the remaining unlabelled images, which may belong to both known and previously unseen classes. Formally, let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label space. The dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ consists of a labelled subset $\mathcal{D}_l = \{x_i, y_i\}_{i=0}^N \in \mathcal{X} \times \mathcal{Y}_L$ and an unlabelled subset $\mathcal{D}_u = \{x_i, y_i\}_{i=0}^M \in \mathcal{X} \times \mathcal{Y}_U$, where the label space of the labelled dataset is a subset of the label space of the unlabelled dataset $\mathcal{Y}_L \subset \mathcal{Y}_U$. During training the labels in $\mathcal{Y}_U$ are not accessible. After training labels should be estimated for every sample in $\mathcal{D}_u$.

Figure 1 illustrates the architecture of our proposed method. HCGD-BT utilizes a (pre-trained) feature encoder f , and two separate branches, namely a hierarchical clustering branch using a novel loss formulation (see Section 3.2) and a contrastive learning branch (see Section 3.1). The feature encoder and the two branches are trained jointly to achieve category discovery. The contrastive learning branch structures the feature space such that data points of the same class are close together. The hierarchical learning branch enforces a hierarchy which clusters data points by a sequence of decisions.

3.1 Contrastive learning loss

For the contrastive loss we follow the concept proposed in [27], which combines an unsupervised contrastive loss and a supervised contrastive loss to utilize both the labelled $\mathcal{D}_l$ and unlabelled $\mathcal{D}_u$ data. The unsupervised loss [4] is computed from two randomly augmented views x_i and x'_i of the

same input in a mini-batch $\mathcal{B}$. First, both views are embedded by using the feature extractor f followed by a projection head ϕ into an embedding vector $\mathbf{z}_i = \phi(f(x_i))$. Then, the loss is given as:

$$\mathcal{L}_i^u = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}'_i)}{\sum_{n=1}^{|\mathcal{B}|} \mathbb{1}_{[n \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_n)}, \quad (1)$$

where $\mathbb{1}_{[n \neq i]}$ is the indicator function, equal to 1 when $n \neq i$ and 0 otherwise. For the supervised loss we define $\mathcal{P}(i)$ as the indices of those images with the same label as x_i in $\mathcal{B}$. Furthermore, $\mathcal{B}_l$ denotes the labeled subset within the mini-batch $\mathcal{B}$. We can then write the supervised loss as:

$$\mathcal{L}_i^s = -\frac{1}{|\mathcal{P}(i)|} \sum_{q \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_q)}{\sum_{n=1}^{|\mathcal{B}_l|} \mathbb{1}_{[n \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_n)} \quad (2)$$

Finally the loss for a whole mini-batch is given as:

$$\mathcal{L}^c = (1 - \alpha) \sum_{i \in \mathcal{B}} \mathcal{L}_i^u + \alpha \sum_{j \in \mathcal{B}_l} \mathcal{L}_j^s \quad (3)$$

with a weight coefficient α to balance the unsupervised and supervised loss.

3.2 Semi-supervised hierarchical clustering loss

To reflect the hierarchical structure of brain tumor diagnostics, inspired by CoHiClust [37] we propose to use contrastive hierarchical clustering in the training phase of the model to design a loss that takes the hierarchical nature of pathological patterns into account. To utilize the prior information in form of the labelled dataset we extend the unsupervised loss in [37] to a semi-supervised formulation. To formulate the hierarchical tree loss we follow [37] to construct a soft binary decision tree. Different from hard decision trees, within a soft decision tree every internal node defines a probability of taking the right or left branch. The idea is that similar data points will be routed through the same nodes of the tree. We parametrize this tree by adding a projection head π after the encoder f . π maps the features extracted by f to a K -dimensional vector where $K = 2^T - 1$ with T being the height of the binary decision tree. The parameter K denotes the number of internal nodes in the tree. The projection head π is modeled as a simple feed-forward network that uses the sigmoid function σ to generate outputs that can be interpreted as the probabilities of taking the left branch of an internal node. Based on these calculated probabilities and the observation that similar data points should follow the same path through the tree, we can formulate a loss function.

Formally, for two inputs x_1, x_2 we can consider their posterior probabilities $P_t(x_1), P_t(x_2)$ at tree level t and compute the probability that they reach the same node at level t as the scalar product $P_t(x_1) \cdot P_t(x_2)$, to avoid trivial solutions [37] we further formulate this probability using the Bhattacharyya coefficient [1]:

$$s_t(x_1, x_2) = \sqrt{P_t(x_1) \cdot P_t(x_2)} = \sum_{i=0}^{2^t-1} \sqrt{P_t^i(x_1) P_t^i(x_2)} \quad (4)$$

For a final similarity function over the whole tree those level-wise similarities are summed: $s(x_1, x_2) = \sum_{t=0}^{T-1} s_t(x_1, x_2)$.

For an input x_i and its augmented version x'_i in a batch B our contrastive hierarchical loss is written as:

$$\mathcal{L}_i^h = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} s(x_i, x_j) - \frac{1}{|\mathcal{P}(i)| + 1} \left[\left(\sum_{j \in \mathcal{P}(i)} s(x_i, x_j) \right) + s(x_i, x'_i) \right] \quad (5)$$

where $\mathcal{N}(i)$ are the indices of inputs with different labels than x_i , and $\mathcal{P}(i)$ are those that share a label with x_i . Note that, if $x_i \in B_l$, $\mathcal{N}(i)$ contains all samples in B that have a different label than x_i and those for which no label is provided. If $x_i \notin B_l$ $\mathcal{N}(i)$ contains all samples from B except x_i .

To avoid solutions where only parts of the tree are used for clustering, we employ the regularization loss $\mathcal{L}^r$ proposed in [37], which encourages the model to use both the left and the right sub-tree.

Tumor type	WSIs	known/novel
Glioblastoma, IDH-wildtype	378	known
Pilocytic astrocytoma	138	novel
Meningothelial meningioma	84	known
Pituitary adenoma	79	known
Ganglioglioma	70	novel
Grade 3, Oligodendroglioma, IDH-mut. and 1p/19q cod.	70	known
Haemangioblastoma	69	known
Grade 2, Oligodendroglioma, IDH-mut. and 1p/19q cod.	66	known
Atypical meningioma	66	novel
Adamantinomatous craniopharyngioma	66	novel
Schwannoma	65	known
Diffuse astrocytoma, IDH-mut.	58	novel

Table 1: Overview of the DBTA dataset for tumor types with more than 50 WSIs.

This is achieved by minimizing the cross-entropy between the desired balanced distribution $[0.5, 0.5]$ and the mean probabilities of all samples in B at each internal node.

The overall loss to train our proposed HGCD-BT method is written as:

$$\mathcal{L} = \mathcal{L}^h + \beta \mathcal{L}^r + \gamma \mathcal{L}^c \quad (6)$$

where β and γ are weighting factors to balance the influence of the loss functions.

4 Experiments and Results

4.1 Data

We used two datasets for evaluating the proposed approach, a dataset of Stimulated Raman histology *OpenSRH* [12] for patch-level classification, and the Digital Brain Tumor Atlas (DBTA) for slide-level classification of H&E-stained WSIs.

OpenSRH

OpenSRH [12] is a publicly available image classification dataset consisting of SRH data from brain tumor patients. The dataset comprises 1348 SRH slides from 307 patients, annotated with pathological labels for six brain tumor types — high-grade glioma (HGG), low-grade glioma (LGG), metastases, meningioma, schwannoma, pituitary adenoma — along with a normal tissue category. The slides are divided into non-overlapping 300x300 pixel patches resulting in 282,931 patches, which serve as inputs for our category discovery model. For evaluation of the category discovery we designated four classes as known (HGG, meningioma, metastases, normal) and three as unknown or novel classes (LGG, schwannoma, pituitary adenoma). Among the known classes, we set a proportion of 50% as labelled in the training data. Image processing follows the methodology in [12] to generate three-channel image patches.

Digital Brain Tumor Atlas (DBTA)

The DBTA dataset [26] is an openly available dataset of digitized H&E- stained brain tumour slides from FFPE tissue. It contains over 126 different brain tumor subtypes. For this study, we selected a subset of the dataset containing tumor types with at least 50 samples each, yielding 12 classes and a total of 1,209 whole-slide images (WSIs). From these we randomly sampled 7 classes as known and the remaining 5 as novel classes. An overview of the selected subset including number of WSIs and split into known and novel classes is provided in Table 1. From this we can observe that DBTA dataset reflects real-world prevalence, resulting in a highly imbalanced distribution across classes. As for SRH imaging, we set 50% of the known class WSIs as labelled during training.

4.2 Implementation details

All implementation were performed using Python (v3.10 Python Software Foundation) and PyTorch (v 2.5.1). The source code is available at: https://github.com/mperkonigg/HGCD_BT.

SRH setting As feature extractor f we adopt the Vision Transformer (ViT-B-16) model [6] pre-trained on ImageNet using DINO [2], where we use the output [CLS] token as feature representations. Since the pre-training was performed on natural images, we choose to fine-tune the whole model during our training. The contrastive projection head ϕ is modeled as a four-layer multi-layer perceptron (MLP) with GeLU activation following the setup in [27, 20]. The projection head for hierarchical clustering π is a two-layer MLP with ReLU activation after the first and sigmoid activation after the second layer with an output dimension of 7, corresponding to a tree level $T = 3$, which is chosen based on the number of classes in the dataset. To generate different input views (x_i, x'_i) we applied data augmentation including horizontal and vertical flipping, randomized contrast adjustments, and Gaussian sharpening. We used a batch size of 64, a learning rate of 0.01 and trained for 200 epochs using stochastic gradient descent. For HGCD-BT we incorporated a 50-epoch warm-up phase before applying the hierarchical clustering loss. During this warm-up, we only use the contrastive loss (Eq. 3) to adapt the pre-trained features of f to the SRH imaging domain. Following [27] we set $\alpha = 0.35$ in Equation 3. We set $\beta = 2^{-T} = 2^{-3} = 0.125$ in Equation 6, following the recommendations in [37]. To have equal contribution of contrastive learning and hierarchical clustering we set $\gamma = 1$.

DBTA setting For DBTA we built on a domain-specific multimodal whole slide foundation model *TITAN* as a slide-encoder [5] with a *CONCH* patch encoder [16]. For processing we utilized the *Trident* toolkit [34]. Both *TITAN* and *CONCH* are using self-supervised visual-language alignment to train the models. *TITAN* operates on a slide-level and is built on diverse large-scale set of WSIs spanning 20 different organs and including neoplastic and non-neoplastic tissue. During training it extracts patch-level features (using *CONCHv1.5*) that are aggregated to task-agnostic slide-level encodings. Due to the size and computational needs for inference and training on WSIs we extract slide encodings once prior to training. During GCD training we used a three layer MLP as feature encoder f to transform those slide encodings to GCD embeddings. For contrastive learning the same projection head ϕ as for SRH patches is used. For hierarchical clustering the output dimension of π is adapted to 31, corresponding to a tree level of $T = 5$. Data augmentation is applied to the *TITAN* slide encodings, including Gaussian noise, scaling, shifting, and feature dropout, prior to transformation by f . A batch size of 16 was applied, the learning rate was set to 0.001 and the model was trained for 200 epochs using stochastic gradient descent. As for SRH we used a warm-up phase of 50 epochs where only the contrastive learning loss is applied. As for SRH we set $\alpha = 0.35$ and $\gamma = 1$. β is adapted to the tree level of T and set to $2^{-5} = 0.03125$.

4.2.1 Baselines

We compare our approach to several baselines. We evaluate two state-of-the-art category discovery methods developed and tested for natural images GCD [27] and Dynamic Conceptual Contrastive Learning (DCCL) [20]. In addition, we use the features extracted from the pretrained models with DINO [2] for SRH and *TITAN* for DBTA without any special attention to category discovery. GCD [27] introduced the setting of generalized category discovery and proposed a simple semi-supervised contrastive learning approach. DCCL [20] is alternating between estimating visual conceptions, using a semi-supervised Infomap clustering, and learning representations during training. This setup allows DCCL to place classes (e.g. cats and dogs) from the same conception (e.g. animal) close together.

4.2.2 Evaluation protocol

As evaluation we follow the standard protocol in the GCD setting [27]. First, SemiKMeans clustering is performed on the predicted embeddings [27]. Next, the Hungarian assignment algorithm [13] is used to optimally align the predicted clusters and ground truth. Finally, accuracy scores are calculated on all unlabelled samples and reported for all classes, as well as separately known classes and novel classes.

4.3 Results

4.3.1 Patch-level classification - OpenSRH

Table 2 shows a comparison of the accuracy of HGCD-BT and the baseline methods. The results show that the hierarchical approach of HGCD-BT is more suitable for patch-level brain tumor

classification, as seen in the increase in performance between HGCD-BT and GCD (+38.2%) and DCCL (+28.1%). Especially for novel classes the difference between accuracies is high. Both GCD and DCCL show significant increase in performance compared to using pre-trained DINO features only. However, GCD is increasing the performance in novel categories the least (+7.0%). Both DCCL and GCD are performing significantly better for tumor types partially labelled in the training set, whereas HGCD-BT shows a balanced, high accuracy.

Method	All Acc.	Known Acc.	Novel Acc.
ViT (ImageNet pretrained DINO)	35.7	35.5	36.0
GCD [27]	55.5	70.9	43.0
DCCL [20]	65.6	73.0	59.6
HGCD-BT (ours)	93.7	93.1	94.2

Table 2: Results for HGCD-BT compared to baseline methods reporting overall accuracy and accuracies for known and novel classes separately.

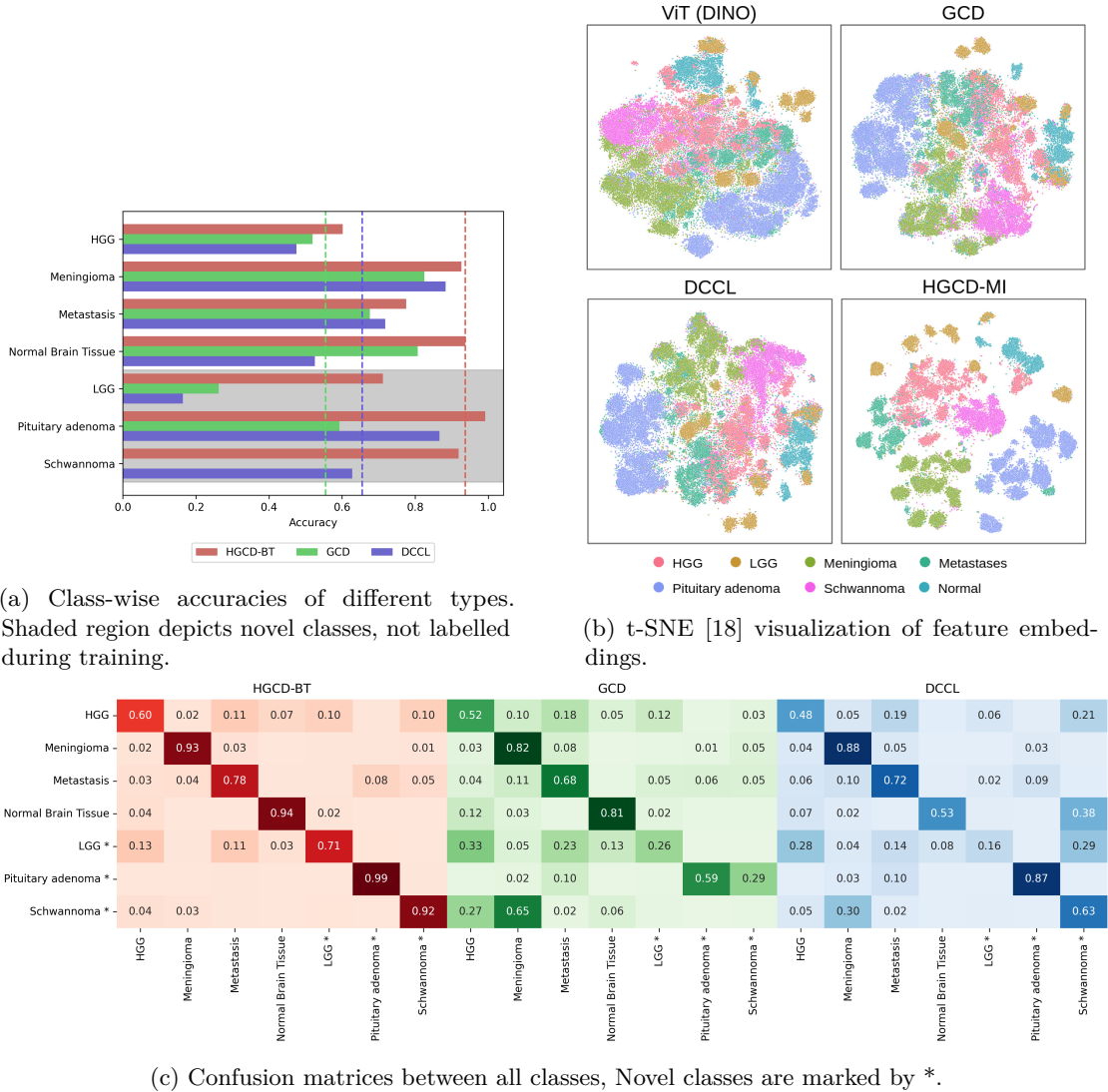


Figure 2: Results for GCD, DCCL and HGCD-BT for patch-level classification on OpenSRH.

We further analyse the performance differences by examining class-wise accuracies, as shown in Figure 2a. In particular, GCD fails to discover the category of schwannoma, a benign nerve sheath tumor, and it is mainly confused with meningioma and HGG. HGCD-BT achieves significantly better accuracy for the novel brain tumor classes - LGG, pituitary adenoma, and schwannoma - compared to the other methods. Overall, HGG (in known classes) and LGG (in novel classes) exhibit

lower accuracy compared to other categories. This is primarily due to high pairwise confusion between these classes and metastases. Biologically, this is expected, as both LGG and HGG belong to the glioma category, making their distinction particularly challenging. The hierarchical nature of the features extracted by the proposed method appears better suited for capturing subtle features such as increased cellularity, anaplasia and mitotic activity, differentiating HGG and LGG.

Comparing the t-sne [18] feature embedding visualization for the pre-trained ViT [2], GCD [27], DCCL [20] and our HGCD-BT method in Figure 2b shows a better distinction of categories for HGCD-BT. While ViT exhibits a general trend toward separation, it lacks well-defined class boundaries. GCD and DCCL improve separation, but HGCD-BT achieves the most distinct class boundaries. Additionally, HGCD-BT reveals small, well-defined subgroups within categories, which can be further investigated.

4.3.2 Slide-level classification - DBTA

For WSI analysis the results in terms of accuracy are shown in Table 3. Given that all compared methods are using a pathology foundation model as slide-level feature encoder the differences in accuracy is not as high as for SRH. Nevertheless, HGCD-BT reaches the best overall accuracy of 79.2%, when compared to the pretrained TITAN feature encoder (+5.4%), GCD (+7.2%) and DCCL (+3.5%). Especially for novel classes HGCD-BT shows strong performance (+8%) against DCCL and the TITAN feature encoder.

Method	All Acc.	Known Acc.	Novel Acc.
TITAN [5]/Conch[16]	73.8	77.0	70.1
GCD [27]	72.0	79.3	63.8
DCCL [20]	75.7	80.6	70.1
HGCD-BT (ours)	79.2	80.2	78.1

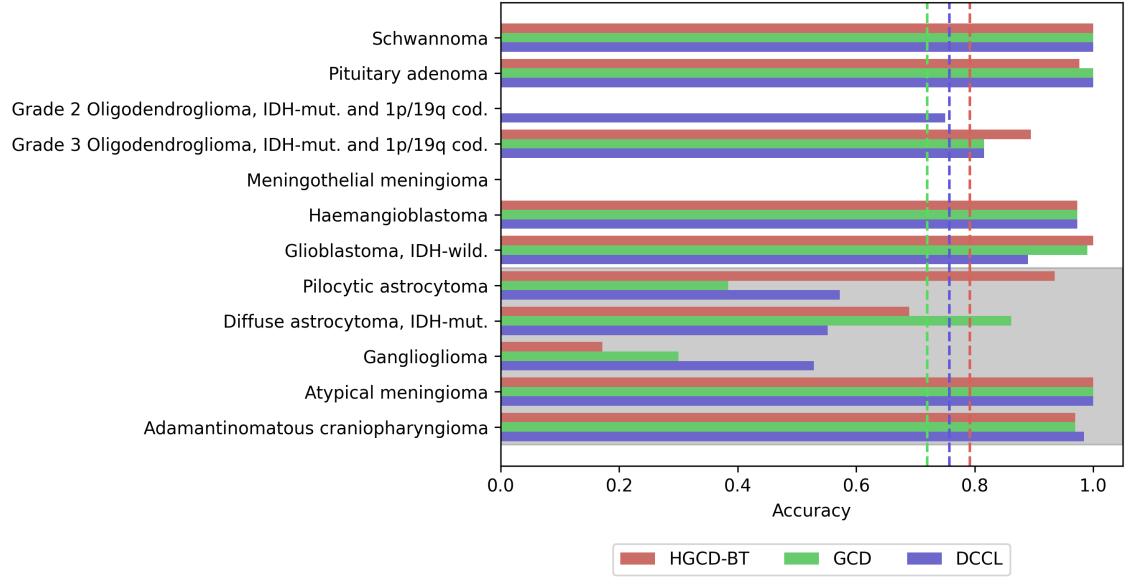
Table 3: Results for HGCD-BT compared to baseline methods reporting overall accuracy and accuracies for known and novel classes separately for DBTA.

For a more detailed analysis Figure 3b shows the performance for type-wise accuracy. We observe some tumor type confusions that are common across all methods and are explainable from a neuropathological standpoint. Meningothelial meningioma and atypical meningioma are merged into one class by all approaches, which is expected since both share cytological features and arrangements but differ mostly in mitotic frequency. A frequent confusion, particularly for HGCD-BT and GCD, involves Grade 2 oligodendroglioma IDH-mut., 1p/19q cod., which is misclassified as either Grade 3 oligodendroglioma IDH-mut., 1p/19q cod. or diffuse astrocytoma. The distinction between Grade 2 and 3 oligodendroglioma is also visually difficult and lies along a morphological spectrum, therefore, it cannot be separated in the feature space. Grade 2 oligodendroglioma IDH-mut., 1p/19q cod. and diffuse astrocytoma both fall within the family of low-grade gliomas and share histopathological features. HGCD-BT and GCD struggle to capture the fine-grained cues needed for accurate discrimination. Finally, ganglioglioma is often confused with pilocytic and diffuse astrocytoma. Ganglioglioma is a mixed glio-neuronal tumour, comprised of a glial and a ganglion component, where the glial component is often astrocytic.

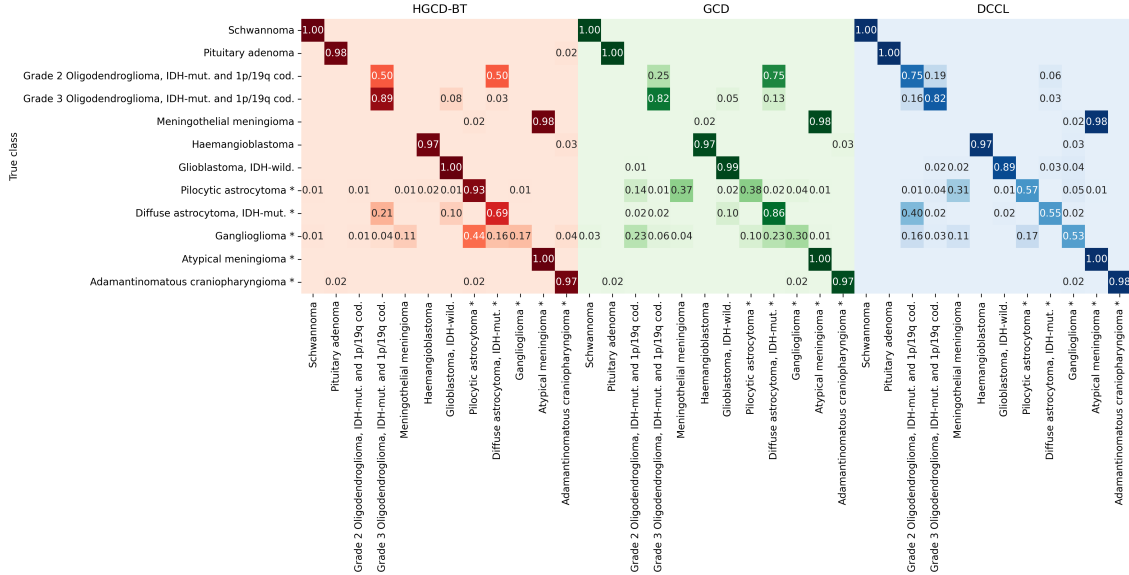
4.3.3 Ablation study

We analyse the contributions of various elements of our proposed approach for patch-level on SRH imaging data.

Influence of key components Different key components of our method are studied in Table 4. First, we can assess the importance of the inclusion of both the hierarchical loss $\mathcal{L}^h$ and the contrastive loss $\mathcal{L}^c$. Using only the components related to $\mathcal{L}^h$ (row 2) results in a drop of -2.3% in overall accuracy. While using only $\mathcal{L}^c$ defaults to GCD [27]. Second, the novel extension of the hierarchical loss to a semi-supervised loss shows a increased performance (row 3 vs. row 5), especially for novel categories (+5.1%). Finally, the effect of warm-up phase used during training to adapt the features to the SRH imaging domain prior to performing hierarchical clustering is compared, while not using a warm-up phase before hierarchical clustering results in a drop of 5.3% (9.2% for novel classes) (row 4).



(a) Class-wise accuracies of different types. Shaded region depicts novel classes, not labelled during training.



(b) Confusion matrices between all classes, Novel classes are marked by *.

Figure 3: Results for GCD, DCCL and HGCD-BT for slide-level classification on DBTA.

Influence of tree levels T We test different values for the number of levels of the decision tree T used during the calculation of $\mathcal{L}^h$ (see Figure 4). For T values between 2 and 4, we observe a stable overall accuracy ranging between 91.7% ($T = 4$) and 93.7% ($T = 3$). Adding more tree levels resulted in a performance loss, with $T = 5$ accuracy on novel classes drops, while accuracy on known classes is stable. For an increase to $T = 6$ the overall accuracy drops to 79.5%, with both known and novel accuracy dropping significantly. This drop might be explained by a more challenging optimization process by adding more tree levels. In addition, tree level 3 results in eight leave nodes, close to the number of classes of seven which might support training. This observation is similar to the findings in [37].

5 Discussion and Conclusion

In this work, we introduced HGCD-BT, a novel approach that leverages hierarchical clustering and contrastive learning to improve category discovery in histopathological imaging. By incorporating a

	$\mathcal{L}^c$	su. $\mathcal{L}^h$	un. $\mathcal{L}^h$	wu.	All Acc.	Known Acc.	Novel Acc.
1 (GCD) [27]	✓	-	-	-	55.5	70.9	43.0
2	-	✓	✓	✓	91.4	90.1	92.0
3	✓	-	✓	✓	90.1	92.8	89.1
4	✓	✓	✓	-	88.4	92.8	85.0
5 (HGCD-BT)	✓	✓	✓	✓	93.7	93.1	94.2

Table 4: Analysis of different components of our method. su. $\mathcal{L}^h$ / un. $\mathcal{L}^h$ = supervised and unsupervised hierarchical clustering loss respectively, wu. = warm-up phase

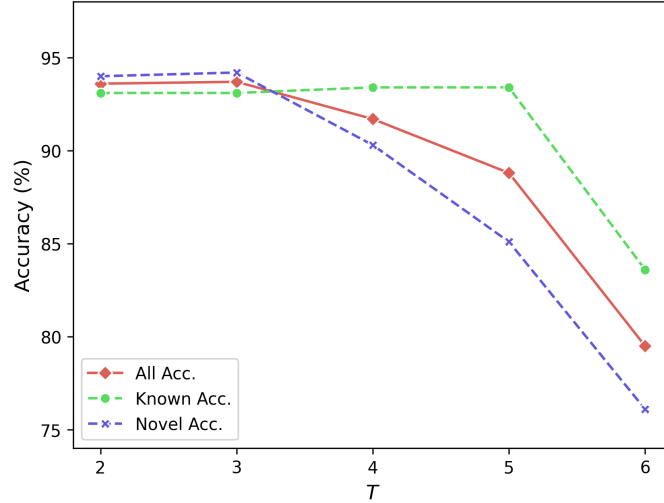


Figure 4: Different settings for the tree level T for patch-level classification on OpenSRH

semi-supervised hierarchical clustering loss, our method effectively captures the structured nature of pathological patterns, outperforming existing GCD techniques. Through experiments on the OpenSRH and DBTA dataset, we demonstrated that HGCD-BT achieves superior classification performance, particularly in identifying novel classes not represented in the labelled part of the data. The results demonstrate generalizability across modalities (SRH and H&E-staining) and level of classification (patch- and slide-level). From a clinical perspective, more accurate and rapid identification of brain tumor types shortens the time of surgical interventions and may reduce complications and side-effects due to anesthesia and surgery. By discovering novel subtypes without requiring costly and labor-intensive annotations, HGCD-BT has the potential to support diagnostic workflows. In addition, using the hierarchical nature of HGCD-BT has the potential to improve interpretability of the decisions made by the model. Our results demonstrate that concept discovery is a promising approach for brain tumor classification. The ability to uncover previously unrecognized categories could contribute to the development of more personalized treatment strategies, aligning with the broader goal of tailoring interventions to specific molecular or morphological subtypes. In addition, the newly identified classes might be incorporated in a specialized classification model using continual learning to enable an adaptive system without the need for retraining from scratch. While we focused here on brain tumors, the methodology is broadly applicable to any disease domain with a hierarchical taxonomy. Potential applications include other tumor types (e.g., skin or thoracic cancers) as well as other disease areas such as lung infections, where hierarchical relationships naturally exist (e.g., lung infections $\rightarrow$ pneumonia $\rightarrow$ fungal pneumonia).

Despite the potential of HGCD-BT, there are a few limitations that need consideration. First, the number of tumor subtypes is restricted to the most common types. Data of rare tumor types is challenging to collect and fine-grained features are needed to distinguish the types. To fully realize its potential, future work should explore datasets containing rare diseases and address challenges related to long-tailed distributions. In HGCD-BT we need to set the number of T manually, and our experiments confirm that performance is sensitive to this choice. Future work should explore strategies how to estimate T dynamically to allow for a flexible extension of the classes represented

in the data. Similarly, for evaluation we use the standard approach for GCD evaluation, which relies on a semi-supervised version of k-means clustering. This requires setting the number of clusters. While this enables direct comparison to other methods, future work should explore the use of the learned hierarchy in the hierarchical clustering branch directly to make a classification. To construct the tree hierarchy, we are following [37] in constructing a binary tree to facilitate training. Nevertheless, this restricts the flexibility of the approach, hierarchies for brain tumor subtypes do not follow a strictly binary decision tree. Extending HGCD-BT to allow non-binary hierarchies would more accurately reflect true pathological relationships. Finally, our work considered imaging data alone, extending the method to multimodal data, by integrating genomic, radiological, or clinical data, could offer new opportunities for discovering disease trajectories and a more comprehensive modelling of diseases.

Acknowledgments

The computational results have been achieved in part using the Austrian Scientific Computing (ASC) infrastructure.

References

- [1] Bhattacharyya, A.: On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics* **7**(4), 401–406 (Jul 1946)
- [2] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9630–9640 (Oct 2021). <https://doi.org/10.1109/ICCV48922.2021.00951>, <https://ieeexplore.ieee.org/document/9709990>, iSSN: 2380-7504
- [3] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **30**(3), 850–862 (Mar 2024). <https://doi.org/10.1038/s41591-024-02857-3>, <https://www.nature.com/articles/s41591-024-02857-3>
- [4] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations (Jul 2020). <https://doi.org/10.48550/arXiv.2002.05709>, <http://arxiv.org/abs/2002.05709>, arXiv:2002.05709 [cs]
- [5] Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., Williamson, D.F.K., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Komura, D., Kawabe, A., Ishikawa, S., Gerber, G., Peng, T., Le, L.P., Mahmood, F.: Multimodal Whole Slide Foundation Model for Pathology (Nov 2024). <https://doi.org/10.48550/arXiv.2411.19666>, <http://arxiv.org/abs/2411.19666>, arXiv:2411.19666 [eess]
- [6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Oct 2020), <https://openreview.net/forum?id=YicbFdNTTy>
- [7] Einstein, E.H., Ablyazova, F., Rosenberg, A., Harshan, M., Wahl, S., Har-El, G., Constantino, P.D., Ellis, J.A., Boockvar, J.A., Langer, D.J., D’Amico, R.S.: Stimulated Raman histology facilitates accurate diagnosis in neurosurgical patients: a one-to-one noninferiority study. *Journal of Neuro-Oncology* **159**(2), 369–375 (Sep 2022). <https://doi.org/10.1007/s11060-022-04071-y>, <https://doi.org/10.1007/s11060-022-04071-y>
- [8] Fan, J., Liu, D., Chang, H., Huang, H., Chen, M., Cai, W.: Seeing Unseen: Discover Novel Biomedical Concepts via Geometry-Constrained Probabilistic Modeling. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11524–11534. IEEE,

Seattle, WA, USA (Jun 2024). <https://doi.org/10.1109/CVPR52733.2024.01095>, <https://ieeexplore.ieee.org/document/10657229/>

- [9] Hao, S., Han, K., Wong, K.Y.K.: CiPR: An Efficient Framework with Cross-instance Positive Relations for Generalized Category Discovery (Mar 2024). <https://doi.org/10.48550/arXiv.2304.06928>, <http://arxiv.org/abs/2304.06928>, arXiv:2304.06928 [cs]
- [10] Hollon, T.C., Pandian, B., Adapa, A.R., Urias, E., Save, A.V., Khalsa, S.S.S., Eichberg, D.G., D’Amico, R.S., Farooq, Z.U., Lewis, S., Petridis, P.D., Marie, T., Shah, A.H., Garton, H.J.L., Maher, C.O., Heth, J.A., McKean, E.L., Sullivan, S.E., Hervey-Jumper, S.L., Patil, P.G., Thompson, B.G., Sagher, O., McKhann, G.M., Komotar, R.J., Ivan, M.E., Snuderl, M., Otten, M.L., Johnson, T.D., Sisti, M.B., Bruce, J.N., Muraszko, K.M., Trautman, J., Freudiger, C.W., Canoll, P., Lee, H., Camelo-Piragua, S., Orringer, D.A.: Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature Medicine* **26**(1), 52–58 (Jan 2020). <https://doi.org/10.1038/s41591-019-0715-9>, <https://www.nature.com/articles/s41591-019-0715-9>, publisher: Nature Publishing Group
- [11] Hollon, T.C., Pandian, B., Urias, E., Save, A.V., Adapa, A.R., Srinivasan, S., Jairath, N.K., Farooq, Z., Marie, T., Al-Holou, W.N., Eddy, K., Heth, J.A., Khalsa, S.S.S., Conway, K., Sagher, O., Bruce, J.N., Canoll, P., Freudiger, C.W., Camelo-Piragua, S., Lee, H., Orringer, D.A.: Rapid, label-free detection of diffuse glioma recurrence using intraoperative stimulated Raman histology and deep neural networks. *Neuro-Oncology* **23**(1), 144–155 (Jan 2021). <https://doi.org/10.1093/neuonc/noaa162>
- [12] Jiang, C., Chowdury, A., Hou, X., Kondepudi, A., Freudiger, C.W., Conway, K., Camelo-Piragua, S., Orringer, D.A., Lee, H., Hollon, T.C.: OpenSRH: optimizing brain tumor surgery using intraoperative stimulated Raman histology. *Advances in neural information processing systems* **35**(DB), 28502–28516 (Dec 2022), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10114931/>
- [13] Kuhn, H.W.: The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**(1-2), 83–97 (1955). <https://doi.org/10.1002/nav.3800020109>
- [14] Liu, Y., Han, K.: DebGCD: Debiased Learning with Distribution Guidance for Generalized Category Discovery (2025), <https://openreview.net/forum?id=9B8o9AxSyb>
- [15] Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., Soffietti, R., von Deimling, A., Ellison, D.W.: The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* **23**(8), 1231–1251 (Jun 2021). <https://doi.org/10.1093/neuonc/noab106>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8328013/>
- [16] Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A visual-language foundation model for computational pathology. *Nature Medicine* **30**(3), 863–874 (Mar 2024). <https://doi.org/10.1038/s41591-024-02856-4>, <https://www.nature.com/articles/s41591-024-02856-4>, publisher: Nature Publishing Group
- [17] Ma, S., Zhu, F., Zhong, Z., Zhang, X.Y., Liu, C.L.: Active Generalized Category Discovery. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16890–16900 (Jun 2024). <https://doi.org/10.1109/CVPR52733.2024.01598>
- [18] Maaten, L.v.d., Hinton, G.: Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**(86), 2579–2605 (2008), <http://jmlr.org/papers/v9/vandermaaten08a.html>
- [19] Orringer, D.A., Pandian, B., Niknafs, Y.S., Hollon, T.C., Boyle, J., Lewis, S., Garrard, M., Hervey-Jumper, S.L., Garton, H.J.L., Maher, C.O., Heth, J.A., Sagher, O., Wilkinson, D.A., Snuderl, M., Venneti, S., Ramkissoon, S.H., McFadden, K.A., Fisher-Hubbard, A., Lieberman, A.P., Johnson, T.D., Xie, X.S., Trautman, J.K., Freudiger, C.W., Camelo-Piragua, S.: Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based

stimulated Raman scattering microscopy. *Nature Biomedical Engineering* **1**, 0027 (2017). <https://doi.org/10.1038/s41551-016-0027>

- [20] Pu, N., Zhong, Z., Sebe, N.: Dynamic Conceptual Contrastive Learning for Generalized Category Discovery. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7579–7588 (Jun 2023). <https://doi.org/10.1109/CVPR52729.2023.00732>, iSSN: 2575-7075
- [21] Rastegar, S., Doughty, H., Snoek, C.: Learn to Categorize or Categorize to Learn? Self-Coding for Generalized Category Discovery. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 72794–72818. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/e6789e468c65a7816760a00a487d3c4e-Paper-Conference.pdf
- [22] Rastegar, S., Salehi, M., Asano, Y.M., Doughty, H., Snoek, C.G.M.: Se-IEEx: Self-Expertise in Fine-Grained Generalized Category Discovery (Nov 2024). <https://doi.org/10.48550/arXiv.2408.14371>, <http://arxiv.org/abs/2408.14371>, arXiv:2408.14371 [cs]
- [23] Redlich, J.P., Feuerhake, F., Weis, J., Schaadt, N.S., Teuber-Hanselmann, S., Buck, C., Luttmann, S., Eberle, A., Nikolin, S., Appenzeller, A., Portmann, A., Homeyer, A.: Applications of artificial intelligence in the analysis of histopathology images of gliomas: a review. *npj Imaging* **2**(1), 16 (Jul 2024). <https://doi.org/10.1038/s44303-024-00020-8>, <https://www.nature.com/articles/s44303-024-00020-8>
- [24] Reinecke, D., Maroouf, N., Smith, A., Alber, D., Markert, J., Goff, N.K., Hollon, T.C., Chowdury, A., Jiang, C., Hou, X., Meissner, A.K., Fürtjes, G., Ruge, M.I., Ruess, D., Stehle, T., Al-Shughri, A., Körner, L.I., Widhalm, G., Roetzer-Pejrimovsky, T., Golfinos, J.G., Snuderl, M., Neuschmelting, V., Orringer, D.A.: Fast intraoperative detection of primary CNS lymphoma and differentiation from common CNS tumors using stimulated Raman histology and deep learning. *medRxiv: The Preprint Server for Health Sciences* p. 2024.08.25.24312509 (Aug 2024). <https://doi.org/10.1101/2024.08.25.24312509>
- [25] Reinecke, D., von Spreckelsen, N., Mawrin, C., Ion-Margineanu, A., Fürtjes, G., Jünger, S.T., Khalid, F., Freudiger, C.W., Timmer, M., Ruge, M.I., Goldbrunner, R., Neuschmelting, V.: Novel rapid intraoperative qualitative tumor detection by a residual convolutional neural network using label-free stimulated Raman scattering microscopy. *Acta Neuropathologica Communications* **10**(1), 109 (Aug 2022). <https://doi.org/10.1186/s40478-022-01411-x>, <https://doi.org/10.1186/s40478-022-01411-x>
- [26] Roetzer-Pejrimovsky, T., Moser, A.C., Atli, B., Vogel, C.C., Mercea, P.A., Prihoda, R., Gelpi, E., Haberler, C., Höftberger, R., Hainfellner, J.A., Baumann, B., Langs, G., Woehrer, A.: The Digital Brain Tumour Atlas, an open histopathology resource. *Scientific Data* **9**(1), 55 (Feb 2022). <https://doi.org/10.1038/s41597-022-01157-0>, <https://www.nature.com/articles/s41597-022-01157-0>, publisher: Nature Publishing Group
- [27] Vaze, S., Hant, K., Vedaldi, A., Zisserman, A.: Generalized Category Discovery. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7482–7491. IEEE, New Orleans, LA, USA (Jun 2022). <https://doi.org/10.1109/CVPR52688.2022.00734>
- [28] Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Yang, E., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y.K., Kunz, J.D., Lee, M.C.H., Bernhard, J.H., Godrich, R.A., Oakley, G., Millar, E., Hanna, M., Wen, H., Retamero, J.A., Moye, W.A., Yousfi, R., Kanan, C., Klimstra, D.S., Rothrock, B., Liu, S., Fuchs, T.J.: A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine* pp. 1–12 (Jul 2024). <https://doi.org/10.1038/s41591-024-03141-0>
- [29] Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., Wang, F., Peng, Y., Zhu, J., Zhang, J., Jackson, C.R., Zhang, J., Dillon, D., Lin, N.U., Sholl, L., Denize, T., Meredith, D., Ligon, K.L., Signoretti, S., Ogino, S., Golden, J.A., Nasrallah, M.P., Han, X., Yang, S., Yu, K.H.: A pathology foundation model for cancer diagnosis and

prognosis prediction. *Nature* **634**(8035), 970–978 (Oct 2024). <https://doi.org/10.1038/s41586-024-07894-z>, <https://www.nature.com/articles/s41586-024-07894-z>, publisher: Nature Publishing Group

- [30] Wang, Y., Zhong, Z., Qiao, P., Cheng, X., Zheng, X., Liu, C., Sebe, N., Ji, R., Chen, J.: Discover and Align Taxonomic Context Priors for Open-world Semi-Supervised Learning. *Advances in Neural Information Processing Systems* **36**, 19015–19028 (Dec 2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/3c646b713f5de2cf1ab1939d49a4036d-Abstract-Conference.html
- [31] Wu, Y., Chi, Z., Wang, Y., Feng, S.: MetaGCD: Learning to Continually Learn in Generalized Category Discovery. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1655–1665. IEEE, Paris, France (Oct 2023). <https://doi.org/10.1109/ICCV51070.2023.00159>
- [32] Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.H., Willens, S., Olguin, F.M., Nirschl, J.J., Neal, J., Diehn, M., Yang, S., Li, R.: A vision–language foundation model for precision oncology. *Nature* **638**(8051), 769–778 (Feb 2025). <https://doi.org/10.1038/s41586-024-08378-w>, <https://www.nature.com/articles/s41586-024-08378-w>
- [33] Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**(8015), 181–188 (Jun 2024). <https://doi.org/10.1038/s41586-024-07441-w>, <https://www.nature.com/articles/s41586-024-07441-w>, publisher: Nature Publishing Group
- [34] Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating Data Processing and Benchmarking of AI Models for Pathology (Feb 2025). <https://doi.org/10.48550/arXiv.2502.06750>, <http://arxiv.org/abs/2502.06750>, arXiv:2502.06750 [cs]
- [35] Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., Khan, F.S.: PromptCAL: Contrastive Affinity Learning via Auxiliary Prompts for Generalized Novel Category Discovery. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3479–3488 (Jun 2023). <https://doi.org/10.1109/CVPR52729.2023.00339>, iSSN: 2575-7075
- [36] Zhao, B., Aodha, O.M.: Incremental Generalized Category Discovery. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19080–19090. IEEE, Paris, France (Oct 2023). <https://doi.org/10.1109/ICCV51070.2023.01753>
- [37] Znalezniak, M., Rola, P., Kaszuba, P., Tabor, J., Śmieja, M.: Contrastive Hierarchical Clustering. In: Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., Bonchi, F. (eds.) *Machine Learning and Knowledge Discovery in Databases: Research Track*. pp. 627–643. Springer Nature Switzerland, Cham (2023). <https://doi.org/10.1007/978-3-031-43412-9-37>