Medical Image Analysis 94 (2024) 103143

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

CellViT: Vision Transformers for precise cell segmentation and classification

Fabian Hörst [a,b,*], Moritz Rempe [a,b], Lukas Heine [a,b], Constantin Seibold [a,c], Julius Keyl [a,d], Giulia Baldini [a,e], Selma Ugurel [f,g], Jens Siveke [h,i], Barbara Grünwald [j,k], Jan Egger [a,b], Jens Kleesiek [a,b,g,l]

[a] Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), 45131 Essen, Germany
[b] Cancer Research Center Cologne Essen (CCCE), West German Cancer Center Essen, University Hospital Essen (AöR), 45147 Essen, Germany
[c] Clinic for Nuclear Medicine, University Hospital Essen (AöR), 45147 Essen, Germany
[d] Institute of Pathology, University Hospital Essen (AöR), 45147 Essen, Germany
[e] Institute of Interventional and Diagnostic Radiology and Neuroradiology, University Hospital Essen (AöR), 45147 Essen, Germany
[f] Department of Dermatology, University Hospital Essen (AöR), 45147 Essen, Germany
[g] German Cancer Consortium (DKTK, Partner site Essen), 69120 Heidelberg, Germany
[h] West German Cancer Center, partner site Essen, a partnership between German Cancer Research Center (DKFZ) and University Hospital Essen, University Hospital Essen (AöR), 45147 Essen, Germany
[i] Bridge Institute of Experimental Tumor Therapy (BIT) and Division of Solid Tumor Translational Oncology (DKTK), West German Cancer Center Essen, University Hospital Essen (AöR), University of Duisburg-Essen, 45147 Essen, Germany
[j] Department of Urology, West German Cancer Center, 45147 University Hospital Essen (AöR), Germany
[k] Princess Margaret Cancer Centre, M5G 2M9 Toronto, Ontario, Canada
[l] Department of Physics, TU Dortmund University, 44227 Dortmund, Germany

ARTICLE INFO

Dataset link: https://github.com/TIO-IKIM/CellViT

MSC:
68T45
68T10
68U07
92C55

Keywords:
Cell segmentation
Digital pathology
Deep learning
Vision transformer

ABSTRACT

Nuclei detection and segmentation in hematoxylin and eosin-stained (H&E) tissue images are important clinical tasks and crucial for a wide range of applications. However, it is a challenging task due to nuclei variances in staining and size, overlapping boundaries, and nuclei clustering. While convolutional neural networks have been extensively used for this task, we explore the potential of Transformer-based networks in combination with large scale pre-training in this domain. Therefore, we introduce a new method for automated instance segmentation of cell nuclei in digitized tissue samples using a deep learning architecture based on Vision Transformer called CellViT. CellViT is trained and evaluated on the PanNuke dataset, which is one of the most challenging nuclei instance segmentation datasets, consisting of nearly 200,000 annotated nuclei into 5 clinically important classes in 19 tissue types. We demonstrate the superiority of large-scale in-domain and out-of-domain pre-trained Vision Transformers by leveraging the recently published *Segment Anything Model* and a ViT-encoder pre-trained on 104 million histological image patches — achieving state-of-the-art nuclei detection and instance segmentation performance on the PanNuke dataset with a mean panoptic quality of 0.50 and an F_1-detection score of 0.83. The code is publicly available at https://github.com/TIO-IKIM/CellViT.

1. Introduction

Cancer is a severe disease burden worldwide, with millions of new cases yearly and ranking as the second leading cause of death after cardiovascular diseases (Tran et al., 2022). Despite novel and powerful non-invasive radiological imaging modalities, collecting tissue samples and evaluating them with a microscope remains a standard procedure for diagnostic evaluation. A pathologist can draw conclusions about potential therapeutic approaches or use them as a starting point for further investigations by identifying abnormalities within the tissue. One crucial component is the analysis of the cells and their distribution within the tissue, such as detecting tumor-infiltrating lymphocytes (Stanton and Disis, 2016) or inflammatory cells in the tumor microenvironment (Greten and Grivennikov, 2019; Grünwald et al., 2021). However, large-scale analysis on the cell level is time-consuming and suffers from a high intra- and inter-observer variability.

Due to the development of high-throughput scanners for pathology, it is now possible to create digitized tissue samples (whole-slide images, WSI), enabling the application of computer vision (CV) algorithms. CV facilitates automated slide analysis, for example, to create tissue segmentation (Ester et al., 2023), detect tumors (Lu et al., 2021),

* Correspondence to: Insitute for Artificial Intelligence in Medicine, Girardetstraße 2, 45131 Essen, Germany.
E-mail address: fabian.hoerst@uk-essen.de (F. Hörst).

https://doi.org/10.1016/j.media.2024.103143
Received 30 June 2023; Received in revised form 14 February 2024; Accepted 12 March 2024
Available online 16 March 2024

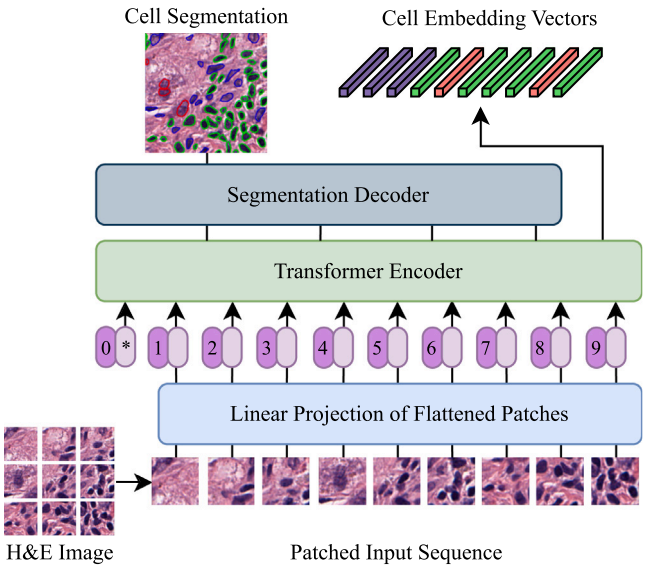

Fig. 1. Network structure of CellViT. An input image is transformed into a sequence of tokens (flattened input sections). By using skip connections at multiple encoder depth levels and a dedicated upsampling decoder network, precise nuclei instance segmentations are derived. Nuclei embeddings are extracted from the Transformer encoder.

evaluate therapy response (Hörst et al., 2023), and the computer-aided detection and segmentation of cells (Graham et al., 2019; Ilyas et al., 2022). In addition to the clinical applications mentioned above, cell instance segmentation can be leveraged for downstream deep learning tasks, as each WSI contains numerous nuclei of diverse types, fostering systematic analysis and predictive insights (Graham et al., 2023). Sirinukunwattana et al. (2018) showed that cell analysis supports the creation of high-level tissue segmentation based on cell composition. Corredor et al. (2017) used hand-crafted features extracted from cells to detect tumor regions in a slide.

Existing algorithms for analyzing WSI (Lu et al., 2021; Campanella et al., 2019; Hörst et al., 2023) are often based on convolutional neural networks (CNNs) used as feature extractors for image regions. The algorithms, despite achieving clinical-grade performance (Campanella et al., 2019), face limitations in interpretability, which in turn poses challenges in defining novel human-interpretable biomarkers. However, accurate cell analysis within these slides presents an opportunity to construct explainable pipelines, incorporating human-interpretable features effectively in downstream tasks (Graham et al., 2023, 2021). Nevertheless, since subtask WSI analysis models (Lu et al., 2021; Campanella et al., 2019; Hörst et al., 2023) rely on abstract entity embeddings, features must be extracted from the detected cells. One approach is to generate hand-crafted features, such as morphological attributes, from the segmentation (Kothari et al., 2009). In the radiology setting, this is referred to as Radiomics (Murray et al., 2019). Alternatively, employing a CNN on image sections of single cells can derive deep learning features. While hand-crafted features may have limited performance, using CNNs for each cell is computationally complex. Thus, the need for automated and reliable detection and segmentation of cells in conjunction with cell-feature extraction in WSI is evident.

We developed a novel deep learning architecture based on Vision Transformer for automated instance segmentation of cell nuclei in digitized tissue samples (**CellViT**). Our approach eliminates the need for additional computational effort for deriving cell features via parallel feature extraction during runtime. The CellViT model proves to be highly effective in collecting nuclei information within patient cohorts and could serve as a reliable nucleus feature extractor for downstream algorithms. Our solution demonstrates exceptional performance on the PanNuke (Gamper et al., 2020) dataset by leveraging transfer learning and pre-trained models (Chen et al., 2022; Kirillov et al., 2023). The PanNuke dataset contains 189,744 segmented nuclei and includes 19 different types of tissues. Among these tissues, there are five clinically important nuclei classes: Neoplastic, inflammatory, epithelial, dead, and connective/soft cells. In addition to the high number of tissue classes and nuclei types, the dataset is highly imbalanced, creating additional complexity. Besides class imbalance, segmenting cell nuclei itself is a difficult task. The cell nuclei may overlap, have a high level of heterogeneity and inter- or intra-instance variability in shape, size, and staining (Ilyas et al., 2022). Sophisticated training methods such as transfer learning, data augmentation, and specific training sampling strategies next to postprocessing algorithms are necessary to achieve satisfactory results.

The proposed network architecture is based on a U-Net-shaped encoder–decoder architecture similar to HoVer-Net (Graham et al., 2019), one of the leading models for nuclei segmentation. Notably, we replace the traditional CNN-based encoder network with a Vision Transformer, inspired by the UNETR architecture (Hatamizadeh et al., 2022b). This approach is depicted in Fig. 1. Vision Transformers are token-based neural networks that use the attention mechanism to capture both local and global context information. This ability enables ViTs to understand relationships among all cells in an image, leveraging long-range dependencies and substantially improving their segmentation. Moreover, when using the common token size of 16 pixels (px) and pixel-resolutions such as 0.25 µm/px (commonly ×40 magnification) or 0.50 µm/px (commonly ×20 magnification) of the images, the token size of ViTs is approximately equivalent to that of a cell, enabling a direct association between a detected cell and its corresponding token embedding from the ViT encoder. As a result, we directly obtain a localizable feature vector during our cell detection that we can extract simultaneously within one forward pass, unlike CNN networks.

Given the limited amount of available data in the medical domain, pre-trained models are an essential requirement as ViTs have increased data requirements compared to CNNs. Chen et al. (2022) recently published a ViT pre-trained on 104 million histological images (ViT_{256}). Their network outperformed current state-of-the-art (SOTA) cancer subtyping and survival prediction methods. Another important contribution is the *Segment Anything Model (SAM)*, proposed by Kirillov et al. (2023). They developed a generic segmentation network for various image types, whose zero-shot performance is almost equivalent to many supervised trained networks. In our work, we compare the performance of pre-trained ViT_{256} (Chen et al., 2022) and SAM (Kirillov et al., 2023) models as building blocks of our architecture for nuclei segmentation and classification. We demonstrate superior performance over existing nuclei instance segmentation models. We summarize our contributions as follows:

1. We present a novel U-Net-shaped encoder–decoder network for nuclei instance segmentation, levering Vision Transformers as encoder networks. Our approach surpasses existing methods for nuclei detection by a substantial margin and achieves competitive segmentation results with other SOTA methods on the PanNuke dataset. We demonstrate the generalizability of CellViT by applying it to the MoNuSeg dataset without finetuning. We showcase that the performance increase can mainly be attributed to the use of large-scale in-domain and out-of-domain pre-trained encoders, by comparing a ViT-encoder pre-trained on 104 million histological patches and to the *Segment Anything Model* and a random initialized model.
2. We provide a framework that enables fast inference results applied on Gigapixel WSI by using a large inference patch size of 1024×1024 px in contrast to conventional 256 px-sized patches. Compared to HoVer-Net, our inference pipeline runs 1.85 times faster.

3. Our new architecture inherently facilitates the extraction of cell embedding vectors, which correspond to the ViT token of the last transformer layer containing the detected cell. Experiments on the CoNSeP dataset demonstrate the utility of these embeddings for adapting the model to new classes, achieving a linear classification AUROC of 0.963 for the CellViT-SAM-H model.

2. Related work

2.1. Instance segmentation of nuclei

Numerous methods have been developed to solve the challenging task of cell nuclei instance segmentation in WSIs. Previous works have explored diverse approaches, ranging from traditional image processing techniques to deep learning (DL) methods. Commonly used image processing techniques involve the design and extraction of domain-specific features. These features encompass characteristics such as intensity, texture, shape, and morphological properties of the nuclei. The primary challenge is separating overlapping nuclei, and different techniques have been devised to do this (Yang et al., 2006; Malpica et al., 1998; Tareef et al., 2018; Cheng and Rajapakse, 2009; Veta et al., 2013; Ali and Madabhushi, 2012; Wienert et al., 2012; Liao et al., 2016). For instance, the works of Cheng and Rajapakse (2009), Veta et al. (2013), and Ali and Madabhushi (2012) rely on a predefined nuclei geometry and the watershed algorithm to separate clustered nuclei, while Wienert et al. (2012) used morphological operations without watershed and Liao et al. (2016) utilized eclipse-fitting for cluster separation. A common drawback of these techniques is their dependency on hand-crafted features, which require expert-level domain knowledge, have limited representative power, and are sensitive to hyperparameter selection (Graham et al., 2019; Chen et al., 2023). The complexity of extracting meaningful features increases when cell nuclei classification is added to the segmentation task. Consequently, their performance is insufficient for our needs to classify and segment nuclei in various tissue types (Chen et al., 2023).

To overcome the limitations of traditional image processing techniques, DL has emerged as a powerful approach for nuclei instance segmentation. An inherent advantage of DL networks is their automatic extraction of relevant features for the given task, surpassing the need for expert-level domain knowledge to generate hand-crafted features. DL algorithms, particularly CNNs (Minaee et al., 2022; Esteva et al., 2019), have shown remarkable success in various computer vision tasks (LeCun et al., 2015). Especially the invention of the U-Net architecture by Ronneberger et al. (2015) has significantly impacted medical image analysis by enabling accurate and efficient segmentation of complex structures, contributing to advancements in various medical domains such as radiology (Isensee et al., 2020; Kelly et al., 2022) and digital pathology (Siddique et al., 2021). It consists of a U-shaped encoder–decoder structure with skip connections at multiple network depths to preserve fine-grained details in the decoder. However, the original U-Net implementation is not able to separate clustered nuclei (Graham et al., 2019). Therefore, specialized network architectures are necessary to separate clustered and overlapping cell nuclei. In the current literature, DL algorithms for nuclei instance segmentation are further divided into two-stage and one-stage methods (Ilyas et al., 2022).

Two-stage methods incorporate a cell detection network in the first stage to localize cell nuclei within an image, generating bounding box predictions of nuclei. These detected nuclei are then passed on to a subsequent segmentation stage to retrieve a fine-grained nucleus segmentation. Mask-RCNN (He et al., 2017) is one of the leading two-stage models built on top of the object detection model Fast-RCNN (Girshick, 2015). Koohbanani et al. (2019) utilized Mask-RCNN networks for nuclei instance segmentation. Based on the proposed nuclei detections in the first stage, the model incorporates a segmentation branch for the fine-grained nucleus segmentations in the second stage. A rectangular image section of the detected nuclei is used as input for the segmentation stage, which causes the problem that overlapping neighboring nuclei may be segmented as well and need to be cleaned up by an additional postprocessing algorithm. Another two-stage method for nuclei segmentation is BRP-Net (Song et al., 2017), which creates nuclei proposals in the first place, then refines the boundary, and finally creates a segmentation out of this. However, this network structure is computationally complex and not designed for end-to-end training due to three independent stages. Additionally, the network requires a considerable time of 12 min to segment a 1360×1024 px image, making its practical application nearly impossible (Song et al., 2017). While two-stage systems offer advantages in localizing cells and improving individual nucleus detection, they often require additional postprocessing for segmentation and suffer from time and computational complexity.

In comparison, one-stage methods combine a single DL network with postprocessing operations. Micro-Net (Raza et al., 2019) extends the U-Net by using multiple resolution input images to be invariant against nuclei of varying sizes. The DIST model by Naylor et al. (2019) adds an additional decoder branch next to the segmentation branch to detect nuclei markers for a watershed postprocessing algorithm. For this, they predict distance maps from the nucleus boundary to the center of mass of the nuclei. Distance maps are regression maps indicating the distance of a pixel to a reference point, e.g., from a nuclei pixel to the center of mass. HoVer-Net (Graham et al., 2019), one of the current SOTA methods for automatic nuclei instance segmentation, uses horizontal and vertical distances of nuclei pixels to their center of mass and separates the nuclei by using the gradient of the horizontal and vertical distance maps as an input to an edge detection filter (Sobel operator). The models STARDIST (Weigert and Schmidt, 2022; Schmidt et al., 2018) and its extension CPP-Net (Chen et al., 2023) generate polygons defining the nuclei boundaries over a set of predicted distances. For this, STARDIST utilizes a star-convex polygon representation to approximate the shape of nuclei. Whereas in STARDIST, the polygons are derived just by features of the centroid pixel, CPP-Net uses context information from sampled points within a nucleus and proposes a shape-aware perceptual loss to constrain the polygon shape. STARDIST demonstrates comparable segmentation performance to HoVer-Net, while CPP-Net exhibits slightly superior results.

In contrast, boundary-based methods such as DCAN (Chen et al., 2016) and TSFD-Net (Ilyas et al., 2022) adopt a different approach, where instead of using distance maps, watershed markers, or polygon predictions, they directly predict the nuclear contour using a prediction map. While DCAN is based on the U-Net architecture, TSFD-Net utilizes a Feature Pyramid Network (FPN) (Lin et al., 2017a) to leverage multiple scales of features. Additionally, the authors of TSFD-Net introduce a tissue-classifier branch to learn tissue-specific features and guide the learning process. To address the class imbalance across nuclei and tissue types, they employ the focal loss (Lin et al., 2017b) for the tissue detection branch, a modified cross-entropy loss with dynamic scaling, and the Focal Tversky loss (Abraham and Khan, 2019) for the segmentation branch, which enlarges the contribution of challenging regions. While TSFD-Net shows promising results, its comparability to other methods is limited due to the lack of a standardized evaluation procedure.

2.2. Vision transformer

All promising DL models (He et al., 2017; Song et al., 2017; Raza et al., 2019; Naylor et al., 2019; Weigert and Schmidt, 2022; Graham et al., 2019; Chen et al., 2023, 2016; Ilyas et al., 2022) for nuclei instance segmentation mentioned previously are based on CNNs. Even though CNN models have demonstrated their effectiveness in image processing, they are bound to local receptive fields and may struggle to capture spatial long-range relationships (Ester et al., 2023). Inspired by the Transformer architecture in NLP (Vaswani et al., 2017), Vision Transformers (Dosovitskiy et al., 2021) have recently emerged as an

alternative to CNNs for CV (Caron et al., 2021). Their architecture is based on the self-attention mechanism (Vaswani et al., 2017), allowing the model to attend to any region within an image to capture long-range dependencies. Unlike CNNs, they are also not bound to fixed input sizes and can process images of arbitrary sizes depending on computational capacity. Vision Transformers have shown promising results not only in image classification (Dosovitskiy et al., 2021; Caron et al., 2021; Raghu et al., 2021), but also in other vision tasks such as object detection (Zhang et al., 2021) and semantic segmentation (Hatamizadeh et al., 2022b; Ester et al., 2023).

Vision transformers for instance segmentation. In recent years, various ideas to use the Transformer architecture for instance segmentation have been developed (Chen and Yu, 2021; Li et al., 2021; Hatamizadeh et al., 2022b,a; Xie et al., 2021; Zheng et al., 2021). Primarily, these methods integrate Transformer models into encoder–decoder architectures by exchanging or extending the encoder network of existing U-Net-based solutions. Chen and Yu (2021) used a Transformer in their TransUNet network to encode tokenized patches from a CNN feature map as the input sequence to derive global context within the CNN network. Li et al. (2021) applied a squeeze-and-expansion Transformer as a variant of the original Vision Transformer by Dosovitskiy et al. (2021) for medical segmentation. The Segformer model by Xie et al. (2021) incorporates an adapted Transformer as an image encoder connected to a lightweight MLP decoder segmentation head. In contrast to these methods, the SETR model (Zheng et al., 2021), used the original ViT as encoder and a fully convolution network as decoder, both connected without intermediate skip connections. Building upon these advancements, the UNETR model (Hatamizadeh et al., 2022b) combined a standard ViT connected to a U-Net-like decoder with skip connections, outperforming TransUNet and the SETR model on three medical image segmentation datasets. The integration of the original ViT implementation without adaptions into the powerful U-Net framework allows the use of pre-trained ViT-networks, which is an important property exploited in our work.

Large-scale pre-training. Pre-training a Vision Transformer on a large amount of data serves as a crucial step to initialize the model's parameters with meaningful representations. Dosovitskiy et al. (2021) demonstrated that ViTs require a larger amount of data compared to CNNs to learn meaningful representations. This is attributed to the inductive biases of the receptive fields of CNNs that are useful for smaller datasets. In contrast, ViTs need to learn relevant patterns, but when provided with sufficiently large datasets, these patterns are more meaningful (Raghu et al., 2021). In the medical domain, where annotated data is often limited, pre-trained ViT-based networks become even more critical. By utilizing self-supervised pre-training approaches (Chen et al., 2020; He et al., 2020; Caron et al., 2020; Grill et al., 2020; Chen and He, 2021; Caron et al., 2021), available unlabeled data can be facilitated effectively to initialize network weights before fine-tuning the network on the target domain. One popular self-supervised pre-training approach, specifically adapted for Vision Transformers, is DINO (knowledge distillation with no labels) (Caron et al., 2021). Vision Transformers trained with this method contain features that explicitly include information about the semantic segmentation of images, which does not emerge as clearly with CNNs (Caron et al., 2021).

In the histopathological domain, Chen et al. (2022) developed a hierarchical network for slide-level representation by stacking multiple ViT blocks. Their approach involves a three-stage hierarchical architecture performing a bottom-up aggregation, with each stage pre-trained independently with DINO. The first stage focuses on processing 16×16 px-sized visual tokens out of 256×256 px patches to create a local cell-cluster token. This first stage ViT, which we refer to as $\mathbf{ViT_{256}}$ (ViT-Small, 21.7 M parameters), is particularly relevant for semantic segmentation. The authors pre-trained the ViT_{256} on 104 million 256×256 px-sized histological image patches from The Cancer Genome Atlas (TCGA) and made the network weights publicly available. It was demonstrated that the ViT_{256} network successfully learned visual concepts specific to histopathological tissue images, including fine-grained cell locations, stroma, and tumor regions, making the model a powerful pre-trained backbone network for histological image analysis.

As for the "natural image"-domain, Kirillov et al. (2023) recently published a promptable open-source segmentation model as a "foundation model" (Bommasani et al., 2021) for semantic segmentation, also known as **Segment Anything (SAM)**. The SAM framework comprises an image encoder (ViT) and a lightweight mask decoder network. The final backbone (ViT-H) of SAM was trained supervised on 1.1 billion segmentation masks from 11 million images. A three-stage data engine consisting of assisted manual, semi-automatic, and automatic mask generation acquired this extensively annotated dataset. Pre-trained weights for three different ViT-scales (ViT-Base with 86 M parameters, denoted as SAM-B, ViT-Large with 307 M parameters, denoted as SAM-L, and ViT-Huge with 632 M parameters, denoted as SAM-H) are publicly available.

3. Methods

Our architecture is inspired by the UNETR model (Hatamizadeh et al., 2022b) for 3D volumetric images, but we adapt its architecture for processing 2D images as shown in Fig. 2. Unlike traditional segmentation networks that employ a single decoder branch for computing the segmentation map, our network employs three distinct multi-task output branches inspired by the approach of HoVer-Net (Graham et al., 2019). The first branch predicts the binary segmentation map of all nuclei (nuclei prediction, NP), capturing their boundaries and shapes. The second branch generates horizontal and vertical distance maps (horizontal–vertical prediction, HV), providing crucial spatial information for precise localization and delineation. Lastly, the third branch predicts the nuclei type map (NT), enabling the classification of different nucleus types. In summary, our network has the following multi-task branches for instance segmentation:

- NP-branch: Predicts binary nuclei map
- HV-branch: Predicts the horizontal and vertical distances of nuclear pixels to their center of mass, normalized between −1 and 1 for each nuclei
- NT-branch: Predicts the nuclei types as instance segmentation maps

To integrate these outputs, we utilize additional postprocessing steps. These steps involve merging the information from the different branches, separating overlapping nuclei to ensure accurate individual segmentation, and determining the nuclei class based on the nuclei type map.

In our experiments, we also evaluated the effectiveness of the STARDIST decoder method and its extension, CPP-Net. We integrate their techniques into the proposed UNETR-HoVer-Net architecture with modifications. Instead of the NP-branch, an object probability branch *PD* is used to predict whether a pixel belongs to an object by predicting the Euclidean distance to the nearest background pixel. The HV-branch is replaced by a branch *RD* to predict the radial distances of an object pixel to the boundary of the nucleus (star-convex representation) (Weigert and Schmidt, 2022). The NT-branch remains unchanged. For the CPP-Net decoder, an additional refinement step is added for the radial distances (Chen et al., 2023).

3.1. Network structure

In our network, we integrate a Vision Transformer as an image encoder that is connected to an upsampling decoder network via skip connections. This architecture allows us to leverage the strengths of a Vision Transformer as an image encoder for instance segmentation without losing fine-grained information. Even though many other adaptations of the U-Net structure for Vision Transformers have been

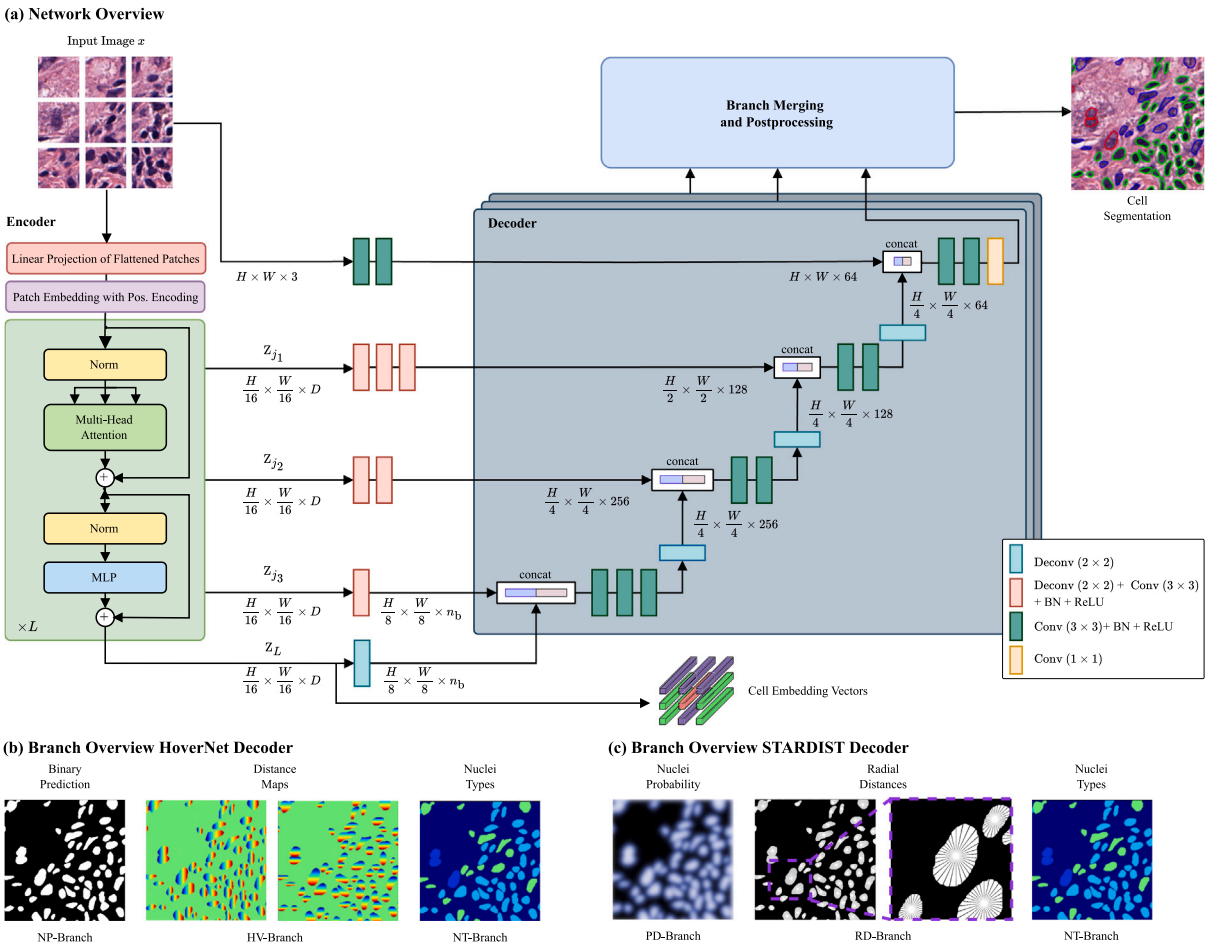

Fig. 2. Network structure of our proposed CellViT-network consisting of a ViT encoder connected to multiple decoders via skip connections (a). We employ pre-trained ViT_{256} and SAM models as encoder networks. Both decoder structures, HoVer-Net and STARDIST, consist of multiple segmentation branches with identical architectures, differing only in the number of output channels. For visualization purposes, the tissue classification branch is not illustrated. In the lower part, we illustrate the output of the HoVer-Net decoder (b) with three branches: NP (binary nuclei prediction), HV (distance maps), and NT (nuclei type). For the STARDIST decoder structure (c), the branches include PD (nuclei probability), RD (radial distances), and again NT.

proposed (e.g., SwinUNETR (Hatamizadeh et al., 2022a)), it was important for us to choose a network structure that incorporates the original ViT structure by Dosovitskiy et al. (2021) without modifications such that we can make use of the large-scale pre-trained ViTs, namely ViT_{256} and SAM.

As in NLP (Vaswani et al., 2017), Vision Transformers take as input a 1D sequence of tokens embeddings (Dosovitskiy et al., 2021; Vaswani et al., 2017). Therefore we need to divide an input image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ with height H, width W and C input channels into a sequence of flattened tokens $\boldsymbol{x}_\text{p} \in \mathbb{R}^{N \times (P^2 \cdot C)}$. Each token is a squared image section with the dimension $P \times P$. The number of tokens N can be calculated via $N = HW/P^2$, which is the effective input sequence length (Hatamizadeh et al., 2022b). Accordingly, a linear projection layer $\boldsymbol{E} \in \mathbb{R}^{N \times D}$ is used to map the flattened tokens $\boldsymbol{x}_\text{p}$ into a D-dimensional latent space. The latent vector size D remains constant through all of the Transformer layers. In contrast to the UNETR-network, we incorporate a learnable class token $\boldsymbol{x}_\text{class}$ (Dosovitskiy et al., 2021), which we can use for classification tasks and append it to the token sequence.

Unlike CNNs, which inherently capture spatial relationships through their local receptive fields, Transformers are permutation invariant and, therefore, cannot capture spatial relationships. Thus, a learnable 1D positional embedding $\boldsymbol{E}_\text{pos} \in \mathbb{R}^{(N+1) \times D}$ is added to the projected token embeddings to preserve spatial context (Hatamizadeh et al., 2022b). In summary, the final input sequence $\boldsymbol{z}_0$ for the Transformer encoder is:

$$\boldsymbol{z}_0 = \left[\boldsymbol{x}_\text{class}; \boldsymbol{x}_\text{p}^1 \boldsymbol{E}; \boldsymbol{x}_\text{p}^2 \boldsymbol{E}; \ldots; \boldsymbol{x}_\text{p}^N \boldsymbol{E}\right] + \boldsymbol{E}_\text{pos}. \tag{1}$$

The Transformer encoder comprises alternating layers of multiheaded self-attention (MHA) (Dosovitskiy et al., 2021) and multilayer perceptrons (MLP), assembled in one Transformer block. A ViT is composed of several stacked Transformer blocks such that the latent tokens $\boldsymbol{z}_i$ are calculated by

$$\boldsymbol{z}'_i = \text{MHA}(\text{Norm}(z_{i-1})) + \boldsymbol{z}_{i-1}, \; i = 1 \ldots L \tag{2}$$

$$\boldsymbol{z}_i = \text{MLP}(\text{Norm}(z_{i-1})) + \boldsymbol{z}_{i-1}, \; i = 1 \ldots L, \tag{3}$$

with L denoting the number of Transformer blocks, Norm$(\cdot)$ denoting layer normalization, and i is the intermediate block identifier (Hatamizadeh et al., 2022b). Inspired by the U-Net and UNETR architectures, we add skip connections to leverage information at multiple encoder depths in the decoder. In total, we use five skip connections. The first skip connection takes $\boldsymbol{x}$ as input and processes it by two convolution layers (3×3 kernel size) with batch-normalization and ReLU activation functions. For the remaining four skip connections, the intermediate and bottleneck latent tokens $\boldsymbol{z}_j$, $j \in \left\{\frac{L}{4}, \frac{2L}{4}, \frac{3L}{4}, L\right\}$ are extracted without the class token and reshaped to a 2D tensor $\boldsymbol{Z}_j \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times D}$. This is only valid if $4 \mid L$ holds, which is commonly satisfied for common ViT implementations (Dosovitskiy et al., 2021; Chen et al., 2022; Kirillov et al., 2023). Each of the feature maps $\boldsymbol{Z}_j$ is transformed by a combination of deconvolutional layers that increase the resolution in both directions by a factor of two and convolutions to adjust the latent dimension. Subsequently, the transformed feature maps are successively processed in each decoder, beginning

with Z_L, and fused with the corresponding skip connection at each stage. This iterative fusion ensures the effective incorporation of multi-scale information, enhancing the overall performance of the decoder. Our network is designed in such a way that the output resolution of the segmentation results exactly matches the input image resolution.

As denoted in Fig. 2, our three segmentation branches (NP, HV, NT) share the same image encoder with the same skip connections and their transformations. The only difference lies in the isolated upsampling pathways of the decoders specific to each branch.

To leverage the additional tissue type information available in the PanNuke dataset, we introduce a tissue classification branch (TC) to guide the learning process of the encoder. For this, we use the class token $z_{L,\text{class}}$ as input to a linear layer with softmax activation function to predict the tissue class.

3.2. Target and losses

For faster training and better convergence of the network, we employ a combination of different loss functions for each network branch, aligned with TSFD-Net (Ilyas et al., 2022). The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NP}} + \mathcal{L}_{\text{HV}} + \mathcal{L}_{\text{NT}} + \mathcal{L}_{\text{TC}} \tag{4}$$

where $\mathcal{L}_{\text{NP}}$ denotes the loss for the NP-branch, $\mathcal{L}_{\text{HV}}$ the loss for the HV-branch, $\mathcal{L}_{\text{NT}}$ the loss for the NT-branch, and $\mathcal{L}_{\text{TC}}$ the loss for the TC-branch. Overall, the individual branch losses are composed of the following weighted loss functions:

$$\begin{aligned}
\mathcal{L}_{\text{NP}} &= \lambda_{\text{NP}_{\text{FT}}}\mathcal{L}_{\text{FT}} + \lambda_{\text{NP}_{\text{DICE}}}\mathcal{L}_{\text{DICE}} \\
\mathcal{L}_{\text{HV}} &= \lambda_{\text{HV}_{\text{MSE}}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{HV}_{\text{MSGE}}}\mathcal{L}_{\text{MSGE}} \\
\mathcal{L}_{\text{NT}} &= \lambda_{\text{NT}_{\text{FT}}}\mathcal{L}_{\text{FT}} + \lambda_{\text{NT}_{\text{DICE}}}\mathcal{L}_{\text{DICE}} + \lambda_{\text{NT}_{\text{BCE}}}\mathcal{L}_{\text{BCE}} \\
\mathcal{L}_{\text{TC}} &= \lambda_{\text{TC}_{\text{CE}}}\mathcal{L}_{\text{CE}}
\end{aligned} \tag{5}$$

with the individual segmentation losses

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{n}\sum_{i=1}^{N_{\text{px}}}\sum_{c=1}^{C} y_{i,c}\log\hat{y}_{ic} \tag{6}$$

$$\mathcal{L}_{\text{DICE}} = 1 - \frac{2\times\sum_{i=1}^{N_{\text{px}}} y_{ic}\hat{y}_{ic} + \varepsilon}{\sum_{i=1}^{N_{\text{px}}} y_{ic} + \sum_{i=1}^{N_{\text{px}}}\hat{y}_{ic} + \varepsilon} \tag{7}$$

$$\mathcal{L}_{\text{FT}} = \sum_{c=1}^{C}\left(1 - \frac{\sum_{i=1}^{N_{\text{px}}} y_{ic}\hat{y}_{ic} + \varepsilon}{\sum_{i=1}^{N_{\text{px}}} y_{ic}\hat{y}_{ic} + \alpha_{\text{FT}}\sum_{i=1}^{N_{\text{px}}} y_{ic}\hat{y}_{ic} + \beta_{\text{FT}}\sum_{i=1}^{N_{\text{px}}} y_{ic}\hat{y}_{ic}}\right)^{\frac{1}{\gamma_{\text{FT}}}} \tag{8}$$

and the cross-entropy as tissue classification loss

$$\mathcal{L}_{\text{CE}} = -\sum_{c_{\text{T}}=1}^{C_{\text{T}}} y_{c_{\text{T}}}\log\hat{y}_{c_{\text{T}}}, C_{\text{T}} = 19,$$

with the contribution of each branch loss (5) to the total loss (4) controlled by the ith hyperparameters λ_i. $\mathcal{L}_{\text{MSE}}$ denotes the mean squared error of the horizontal and vertical distance maps and $\mathcal{L}_{\text{MSGE}}$ the mean squared error of the gradients of the horizontal and vertical distance maps, each summarized for both directions separately. In the segmentation losses (6)–(8), y_{ic} is the ground-truth and $\hat{y}_{ic}$ the prediction probability of the ith pixel belonging to the class c, C the total number of nuclei classes, N_{px} the total amount of pixels, ε a smoothness factor and $\alpha_{\text{FT}}, \beta_{\text{FT}}$ and γ_{FT} are hyperparameters of the Focal Tversky loss $\mathcal{L}_{\text{FT}}$. The Cross-Entropy loss (6) and Dice loss (7) are commonly used in semantic segmentation. To address the challenge of underrepresented instance classes, the Focal Tversky loss (8), a generalization of the Tversky loss, is used. The Focal Tversky loss places greater emphasis on accurately classifying underrepresented instances by assigning higher weights to those samples. This weighting enhances the model's capacity to handle class imbalance and focuses its learning on the more challenging regions of the segmentation task. More information about the selection of the weights for Eq. (5) are given in the Appendix A.3.

3.3. Postprocessing

As the network does not directly provide a semantic instance segmentation with separated nuclei, postprocessing is necessary to obtain accurate results.

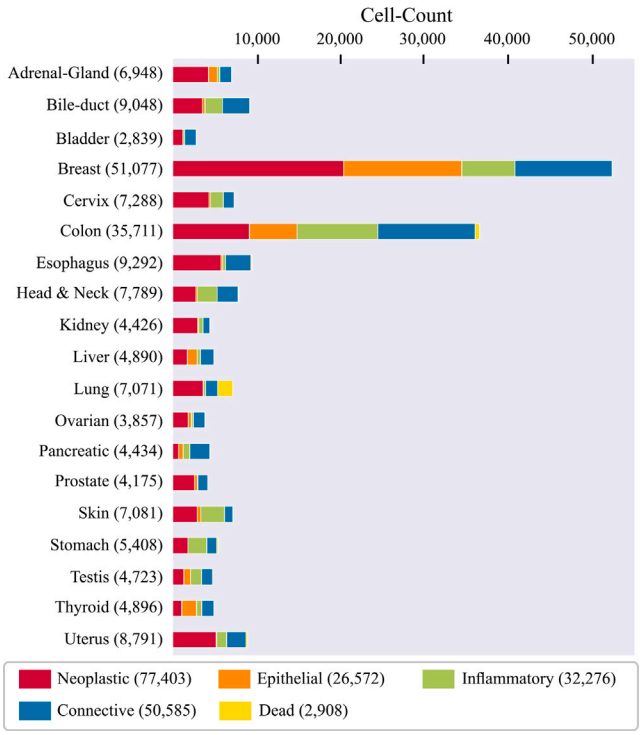

Fig. 3. PanNuke nuclei distribution overview for each of the nineteen tissue types, sorted by the total number of nuclei inside the tissue. The total number of nuclei within a tissue type is given in parentheses.
Source: Adapted from (Gamper et al., 2020).

This involves several steps, including merging the information from the different branches, separating overlapping nuclei to ensure accurate individual segmentation, and determining the nuclei class based on the nuclei type map. Moreover, when performing inference on whole gigapixel WSI, a fusion mechanism is necessary. Due to the significant size of WSIs, inference needs to be performed on image patches extracted from them using a sliding-window approach. The segmentation results obtained from these patches must be assembled to generate a segmentation map of the entire WSI. The postprocessing methods are therefore explained in the following two paragraphs, starting with the segmentation of a single patch followed by its composition into a segmentation output for the entire WSI.

Nuclei separation and classification. To separate adjacent and overlapping nuclei from each other, we utilize HoVer-Net's validated postprocessing pipeline. This involves computing the gradients of the horizontal and vertical distance maps to capture transitions between nuclei boundaries and the boundary between nuclei and the background. At these transition points significant value changes occur in the gradient. The Sobel operator (edge detection filter) is then applied to identify regions with substantial differences in neighboring pixels within the distance maps. Finally, a marker-controlled watershed algorithm is employed to generate the final boundaries.

To calculate the nuclei class, the output of the separated nuclei is merged with the nuclei type predictions. For this purpose, majority voting is performed in the nuclei region using the NT prediction map with the majority class assigned to all nuclei pixels (Graham et al., 2019).

The STARDIST and CPP-Net decoder methods, on the other hand, use non-maximum suppression (NMS) to prune redundant polygons that likely represent the same object (Weigert and Schmidt, 2022; Schmidt et al., 2018). We use this approach when testing CellViT with STARDIST and CPP-Net decoders. In difference to STARDIST, the CPP-Net approach uses the refined radial distances as input for the NMS. The nuclei classes are then again assigned to the resulting binary polygons via majority voting.

Inference. The encoder ViT offers a significant advantage for performing inference on gigapixel WSI over CNNs based U-Nets. Its capability to process input sequences of arbitrary length, constrained only by memory consumption and positional embedding interpolation, allows for increased input image sizes during inference. It is important to note that positional embedding

interpolation must be considered when scaling the input images. In preliminary experiments on the MoNuSeg dataset (see Section 5.3), we found that our network achieves equal performance when inferring on a single 1024×1024 px patch compared to cutting the same patch into 256×256 px sub-patches with an overlap of 64 px. Based on these findings, we have chosen to perform WSI inference using 1024×1024 px large patches with a 64 px overlap. Due to the high computational overhead, it is not feasible to keep the segmentation results of the entire WSI in memory. Consequently, we process and merge only the overlapping nuclei during postprocessing. By utilizing just a small overlap in the inference patches relative to the patch size, the postprocessing effort is reduced. To efficiently store the results in a structured and readable format, as well as for compatibility with software such as QuPath (Bankhead et al., 2017), the nuclei predictions for an entire WSI are exported in a JSON file. Each nucleus is represented by several parameters, including the nuclei class, bounding-box coordinates, shape polygon of the boundaries, and the center of mass for detection location. In the Appendix, we provide example visualizations of the prediction results from an internal esophageal adenocarcinoma and melanoma cohort, imported into QuPath (see Fig. A.2)). This approach ensures the accessibility of the instance segmentation results for further analysis and visualization.

Moreover, for each detected nuclei $\hat{y}$, we store the corresponding embedding token $\boldsymbol{z}_L^{\hat{y}} \in \mathbb{R}^D$. Importantly, as the cell embedding vectors can be directly extracted during the forward pass and are spatially linked to each nuclei $\hat{y}$, there is no need for an additional forward pass on cropped image patches of the detected cells, again saving inference time. If a nucleus is associated with multiple tokens, we average over all token embeddings in which the nucleus is located. The cell embedding can be used as extracted cell-features for downstream DL algorithms addressing problems such as disease prediction, treatment response, and survival prediction.

4. Experimental setup

4.1. Datasets

PanNuke. We use the PanNuke dataset as the main dataset to train and evaluate our model. The dataset contains 189,744 annotated nuclei in 7904 256×256 px images of 19 different tissue types and 5 distinct cell categories, as depicted in Fig. 3. Cell-images were captured at a magnification of ×40 with a resolution of 0.25 µm/px. The dataset is highly imbalanced, especially the nuclei class of dead cells is severely underrepresented, as apparent in the nuclei and tissue class statistics (see Fig. 3). PanNuke is regarded as one of the most challenging datasets to perform the simultaneous nuclei instance segmentation task (Ilyas et al., 2022).

MoNuSeg. The MoNuSeg (Kumar et al., 2020, 2017) dataset serves as an additional dataset for nuclei segmentation. In contrast to PanNuke, the dataset is much smaller and does not divide the nuclei into different classes. For this work, we only use the test dataset of MoNuSeg to evaluate our model. The test dataset consists of 14 images with a resolution of 1000×1000 px, acquired at ×40 magnification with 0.25 µm/px. In total, the test dataset contains more than 7000 annotated nuclei across the seven organ types kidney, lung, colon, breast, bladder, prostate, and brain at several disease states (benign and tumors at different stages). Since no nuclei labels are included, the dataset cannot be used for evaluating classification performance. To process the dataset more effectively with our ViT-based networks with a token size of 16 px, we resized the data to a size of 1024×1024 px. Due to the sufficient patch size of the original data, we also created a ×20 dataset with 0.50 µm/px resolution, where the patch size is 512×512 px accordingly.

CoNSeP. We utilized the colorectal nuclear segmentation and phenotypes (CoNSeP) dataset by Graham et al. (2019) to analyze extracted cell embeddings (see Section 3.3) of detected cells on an external validation dataset. This dataset comprises 41 H&E-stained colorectal adenocarcinoma WSI at a resolution of 0.25 µm/px and an image size of 1000×1000 px, which we rescale to 1024×1024 px similar to the MoNuSeg data. The dataset exhibits significant diversity, encompassing stromal, glandular, muscular, collagen, adipose, and tumorous regions, along with various types of nuclei derived from originating cells: normal epithelial, dysplastic epithelial, inflammatory, necrotic, muscular, fibroblast, and miscellaneous nuclei, including necrotic and mitotic cells.

4.2. Experiments

In this study, we conducted two experiments on the PanNuke dataset and one on the MoNuSeg dataset to assess algorithms performance. We additionally used an internal dataset for comparing inference speed time. Given the higher clinical relevance of the detection task over achieving the optimal segmentation quality, we **(1)** performed an ablation study on PanNuke to determine the most suitable network architecture for nuclei detection. We compared the performance of pre-trained models (see Section 4.4) against randomly initialized models and explored the impact of regularization techniques such as data augmentation, loss functions, and customized oversampling, as well as comparing the HoVer-Net decoder method to the STARDIST and CPP-Net decoder methods in our UNETR-structure. Based on these investigations, we identified the best models, which were **(2)** subsequently evaluated for segmentation quality. To assess both detection and segmentation performance, we compared our models with multiple baseline architectures, namely DIST (Naylor et al., 2019), Mask-RCNN (He et al., 2017), Micro-Net (Raza et al., 2019), HoVer-Net (Graham et al., 2019), TSFD-Net (Ilyas et al., 2022), and CPP-Net (Chen et al., 2023). We also re-trained the STARDIST model with a ResNet50 (He et al., 2016) backbone and the hyperparameters of Chen et al. (2023) to retrieve comparable detection results. For comparison, we conducted our experiments using the same three-fold cross-validation splits provided by the PanNuke dataset organizers and report the averaged results over all three splits. It is worth mentioning that all the comparison models we evaluate in this study adhere to the same evaluation scheme for the PanNuke dataset, with one exception. The TSFD-Net publication reports results based on an 80-20 train-test split, making their results more optimistic. Nevertheless, we include their results for the purpose of comparison.

As a third experiment **(3)**, we evaluated our models trained on PanNuke on the publicly available 14 test images of the MoNuSeg dataset to test generalizability. The dataset serves a second purpose next to generalization: We compare various input image sizes and assess the performance of our inference pipeline outlined in Section 3.3. In this context, we evaluate the performance using two scenarios — one involving an uncropped MoNuSeg slide with 1024 px input patch size and the other using cropped 256 px input images. Additionally, we investigate the impact of our overlapping strategy with a 64-pixel overlap, focusing on the 256 px input size.

To analyze the cell embeddings $\boldsymbol{z}_L^{\hat{y}}$ for detected nuclei with our CellViT models, we utilize the CoNSeP dataset **(4)**. To achieve this, we perform inference with the pre-trained PanNuke models on the CoNSeP images (1024 px input patch size) and extract the token embeddings $\boldsymbol{z}_L^{\hat{y}}$ for each nuclei $\hat{y}$ from the last Transformer block that are spatially associated with $\hat{y}$. Subsequently, we employ the Uniform Manifold Approximation and Projection (UMAP) method for dimension reduction to transform the cell embedding vectors (of the 27 training images) into a two-dimensional representation, which can be visualized in a two-dimensional scatter plot. We additionally trained a linear classifier on top of the cell embeddings (extracted from the 27 training images) to classify the detected cells into the CoNSeP nuclei classes and tested the classifier on the cell embeddings of the cells from the 14 test images.

Finally, to compare the inference runtime **(5)**, we collected a diverse dataset of 10 esophageal WSIs with tissue areas ranging from 2.79 mm^2 to 74.07 mm^2. We measured the inference runtime for the HoVer-Net model, as well as for the $CellViT_{256}$ and CellViT-SAM-H models with 256 px and 1024 px patch input size and an overlap of 64 px. For each WSI, we repeated the process three times and averaged the runtime results.

4.3. Evaluation metrics

Nuclear instance segmentation evaluation. Usually, the Dice coefficient (DICE) or the Jaccard index are used as evaluation metrics for semantic segmentation. However, as Graham et al. (2019) have already shown, these two metrics are insufficient for evaluating nuclear instance segmentation as they did not account for the detection quality of the nuclei. Therefore, a metric is needed that assess the following three requirements (see Graham et al. (2019)):

1. Separate the nuclei from the background
2. Detect individual nuclei instances and separate overlapping nuclei
3. Segment each instance

These three requirements cannot be evaluated with the Jaccard index and the DICE score, as they just satisfy requirement 1. In line with (Graham et al., 2019) and the PanNuke dataset evaluation recommendations (Gamper et al., 2020), we use the panoptic quality (PQ) (Kirillov et al., 2019) to quantify the instance segmentation performance. The PQ us defined as

$$PQ = \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{Detection Quality (DQ)}} \times \underbrace{\frac{\sum_{(y,\hat{y})\in TP} IoU(y,\hat{y})}{|TP|}}_{\text{Segmentation Quality (SQ)}}, \tag{9}$$

with $IoU(y,\hat{y})$ denoting the intersection-over-union (Kirillov et al., 2019). In this equation, y denotes a ground-truth (GT) segment, and $\hat{y}$ denotes a predicted segment, with the pair $(y,\hat{y})$ being a unique matching set of one ground-truth segment and one predicted segment. As Kirillov et al. (2019) proved, each pair of segments $(y,\hat{y})$, i.e., each pair of true and predicted nuclei, in an image is unique if $IoU(y,\hat{y}) > 0.5$ is satisfied. For each class, the unique matching of $(y,\hat{y})$ splits the predicted and the GT segments into three sets:

- True Positives (TP): Matched pairs of segments, i.e., correctly detected instances
- False Positives (FP): Unmatched predicted segments, i.e., predicted instances without matching GT instance
- False negatives (FN): Unmatched GT segments, i.e., GT instances without matching predicted instance

The PQ score can be intuitively decomposed into two parts, the detection quality similar to the F_1 score commonly used in classification and detection scenarios, and the segmentation quality as the average IoU of matched segments (Graham et al., 2019; Kirillov et al., 2019). To ensure a fair comparison, we use binary PQ (bPQ) pretending that all nuclei belong to one class (nuclei vs. background) and the more challenging multi-class PQ (mPQ), taking the nuclei class into account. In doing so for mPQ, we calculate the PQ independently for each nuclei class and subsequently average the results over all classes (Gamper et al., 2020).

Nuclear classification evaluation. To evaluate the detection quality of our model, we employ commonly used detection metrics. Similar to the approach used in the PQ-score for nuclear instance segmentation evaluation, we split GT and predicted instances into TPs, FPs, and FNs. We use the conventional detection metrics precision (P_d), recall (P_d) and the ($F_{1,\mathrm{d}}$)-score as a harmonic mean between precision and recall. The index 'd' indicates that these are the scores for the entire binary nuclei detection over all classes c. Thus, the binary detection scores are defined as follows:

$$F_{1,\mathrm{d}} = \frac{2TP_\mathrm{d}}{2TP_\mathrm{d} + FP_\mathrm{d} + FN_\mathrm{d}}$$
$$P_\mathrm{d} = \frac{TP_\mathrm{d}}{TP_\mathrm{d} + FP_\mathrm{d}}$$
$$R_\mathrm{d} = \frac{TP_\mathrm{d}}{TP_\mathrm{d} + FN_\mathrm{d}}$$

We further break down TP_d into correctly classified instances of class c (TP_c), false positives of class c (FP_c) and false negatives of class c (FN_c) to derive cell-type specific scores. We then define the $F_{1,c}$-score, precision (P_c) and recall (R_c) of each nuclei class c as

$$F_{1,c} = \frac{2(TP_c + TN_c)}{2(TP_c + TN_c) + 2FP_c + 2FN_c + FP_\mathrm{d} + FN_\mathrm{d}},$$
$$P_c = \frac{TP_c + TN_c}{TP_c + TN_c + 2FP_c + FP_\mathrm{d}},$$
$$R_c = \frac{TP_c + TN_c}{TP_c + TN_c + 2FN_c + FN_\mathrm{d}}.$$

In order to prioritize the classification of different nuclear types, we incorporated an additional weighting factor for the nuclei classes, as suggested in the official PanNuke evaluation metrics (Gamper et al., 2020; Graham et al., 2019), Since we cannot use the $IoU(y,\hat{y}) > 0.5$ criterion to find matching instances $(y,\hat{y})$ between GT-instances and predictions for the detection task, we use the methodology of Sirinukunwattana et al. (2016) and define a match $(y,\hat{y})$ if both centers of mass are within a radius of 6 px (0.50 µm/px) and 12 px (0.25 µm/px), respectively.

Statistical analysis. To test if the performance of one network architecture yields significantly better results than another network architecture, we employed the one-tailed paired t-test with a significance threshold set at $\alpha = .05$. We test the hypothesis that network A_1 yields higher or equal mean performance to network A_2. Our alternative hypothesis proposes that network A_1 achieves a statistically lower mean performance than network A_2. More information about the validity and limitation of the test selection is given in the Appendix A.1.2. We used SciPy $v1.11.3$ for calculation (Virtanen et al., 2020).

4.4. Model training

Oversampling. Even though the PanNuke dataset has around 200,000 annotated nuclei, they are distributed just across a limited number of 8000 patches with 256×256 px patch size. Furthermore, there is a substantial class imbalance among tissue types and nuclei classes (see Fig. 3). Thus, we developed a new oversampling strategy based on class weightings to balance both tissue classes and nuclei classes. For each patch i in the training dataset with N_Train training samples, we calculate the sampling weights for the tissue class and the cell class with

$$p_i(\gamma_\mathrm{s}) = \frac{w_\mathrm{Tissue}(i,\gamma_\mathrm{s})}{\max\limits_{j\in[1,N_\mathrm{Train}]} w_\mathrm{Tissue}(j,\gamma_\mathrm{s})} + \frac{w_\mathrm{Cell}(i,\gamma_\mathrm{s})}{\max\limits_{j\in[1,N_\mathrm{Train}]} w_\mathrm{Cell}(j,\gamma_\mathrm{s})}, \tag{10}$$

where $w_\mathrm{Tissue}(i,\gamma_\mathrm{s})$ is a weight factor for the tissue class and $w_\mathrm{Cell}(i,\gamma_\mathrm{s})$ for the nuclei class. The parameter $\gamma_\mathrm{s} \in [0,1]$ is a weighting factor that determines the strength of the oversampling. A γ_s value of 0 indicates no oversampling, while $\gamma_\mathrm{s} = 1$ corresponds to maximum balancing. To ensure neither $w_\mathrm{Tissue}(i,\gamma_\mathrm{s})$ nor $w_\mathrm{Cell}(i,\gamma_\mathrm{s})$ dominates the sampling, normalization is applied to both summands in Eq. (10). The calculation of the weighting factor of the tissue class can be calculated directly via

$$w_\mathrm{Tissue}(i,\gamma_\mathrm{s}) = \frac{N_\mathrm{Train}}{\gamma_\mathrm{s}\left(\sum\limits_{j\in[1,N_\mathrm{Train}]|c_{\mathrm{T},j}=c_{\mathrm{T},i}} 1\right) + (1-\gamma_\mathrm{s})N_\mathrm{Train}} \tag{11}$$

as each patch can only belong to one tissue class denoted by $c_{\mathrm{T},i}$. For cell weighting, it must be considered that each patch can contain multiple nuclei from different cell classes. Therefore, we create a binary vector $\boldsymbol{c}_i \in \{0,1\}^C$, where each entry is set to 1 for each existing nuclei type c in the patch. To get a reference value for scaling similar to Eq. (11), we calculate $N_\mathrm{Cell} = \sum_{i=1}^{N_\mathrm{Train}} \|\boldsymbol{c}_i\|_1$. The cell weighting for each training image i is then calculated by

$$w_\mathrm{Cell}(i,\gamma_\mathrm{s}) = (1-\gamma_\mathrm{s}) + \gamma_\mathrm{s}\sum_{j=1}^{C} c_{ij}\frac{N_\mathrm{Cell}}{\gamma_\mathrm{s}\sum_{k=1}^{N_\mathrm{Train}} c_{kj} + (1-\gamma_\mathrm{s})N_\mathrm{Cell}},$$

with c_{ij} the vector entry of $\boldsymbol{c}_i$ at position j. The training images are randomly sampled in a training epoch with replacement based on their sampling weights $p_i(\gamma_\mathrm{s})$.

Data augmentation. In addition to our customized oversampling strategy, we extensively employ data augmentation techniques to enhance data variety and discourage overfitting. We use a combination of the following geometrical and noisy/intensity-based augmentation methods: random 90-degree rotation, horizontal flipping, vertical flipping, downscaling, blurring, gaussian noise, color jittering, superpixel representation of image sections (SLIC), zoom blur, random cropping with resizing and elastic transformations. These augmentation techniques were selected to introduce variations in the shape, orientation, texture, and appearance of the nuclei, enhancing the robustness and generalization capabilities of the model. For detailed information on the augmentation methods utilized, including the selected probabilities and corresponding hyperparameters, please refer to the Appendix.

Optimization and training strategy. We train all our models for 130 epochs and incorporate exponential learning rate scheduling with a scheduling factor of 0.85 to gradually reduce the learning rate during training (denoted as CellViT hyperparameters). To balance our training, we use our modified oversampling strategy with $\gamma_\mathrm{s} = 0.85$. For the STARDIST and CPP-Net models, we also conducted experiments using the proposed CPP-Net hyperparameters by Chen et al. (2023). A complete overview of all hyperparameters, including optimizer, data augmentation, and weighting factors of the loss functions in Eq. (5) is provided in the Appendix. As for the encoder models, we leverage the ViT_{256}-model (ViT-S, $D = 384$, $L = 12$), which has been pre-trained on histological data (see Section 2.2). Additionally, we compare the performance with the three pre-trained SAM checkpoints: SAM-B (ViT-B, $D = 768$, $L = 12$), SAM-L (ViT-L, $D = 1024$, $L = 24$) and SAM-H (ViT-H, $D = 1280$, $L = 32$). These checkpoints provide different model sizes and complexities, allowing us to evaluate their respective performance and choose the most suitable one for our task. During training, we initially freeze the encoder weights for the first 25 epochs.

Table 1
Precision (P), Recall (R) and F_1-score (F_1) for detection and classification across the three PanNuke splits for each nuclei type. The centroid of each nucleus was used for computing detection metrics for segmentation networks. *TSFD-Net was not evaluated on the official three-fold splits of the PanNuke dataset and left out by the comparison **Model re-trained by ourselves ***Models trained on downscaled 0.50 µm/px PanNuke images.

Model	Decoder	Hyperparameters	Detection			Classification														
						Neoplastic			Epithelial			Inflammatory			Connective			Dead		
			P_d	R_d	$F_{1,d}$	P_{Neo}	R_{Neo}	$F_{1,Neo}$	P_{Epi}	R_{Epi}	$F_{1,Epi}$	P_{Inf}	R_{Inf}	$F_{1,Inf}$	P_{Con}	R_{Con}	$F_{1,Con}$	P_{Dead}	R_{Dead}	$F_{1,Dead}$
DIST			0.74	0.71	0.73	0.49	0.55	0.50	0.38	0.33	0.35	0.42	0.45	0.42	0.42	0.37	0.39	0.00	0.00	0.00
Mask-RCNN			0.76	0.68	0.72	0.55	0.63	0.59	0.52	0.52	0.52	0.46	0.54	0.50	0.42	0.43	0.42	0.17	0.30	0.22
Micro-Net			0.78	0.82	0.80	0.59	0.66	0.62	0.63	0.54	0.58	0.59	0.46	0.52	0.50	0.45	0.47	0.23	0.17	0.19
HoVer-Net			0.82	0.79	0.80	0.58	0.67	0.62	0.54	0.60	0.56	0.56	0.51	0.54	0.52	0.47	0.49	0.28	0.35	0.31
TSFD-Net*			0.84	0.87	0.85	0.60	0.71	0.65	0.56	0.58	0.57	0.59	0.58	0.57	0.55	0.49	0.53	0.33	0.40	0.43
STARDIST (ResNet50) **	STARDIST	CPP-Net	0.85	0.80	0.82	0.69	0.69	0.69	0.73	0.68	0.70	0.62	0.53	0.57	0.54	0.49	0.51	0.39	0.09	0.10
STARDIST (ResNet50) **	STARDIST	CellViT	0.85	0.79	0.82	0.70	0.66	0.68	0.71	0.66	0.68	0.58	0.58	0.58	0.54	0.49	0.51	0.39	0.34	0.36
$CellViT_{256}$ – Raw	HoVer-Net	CellViT	0.80	0.77	0.78	0.61	0.64	0.63	0.63	0.59	0.61	0.55	0.46	0.50	0.45	0.43	0.44	0.43	0.16	0.23
$CellViT_{256}$ – Over	HoVer-Net	CellViT	0.79	0.78	0.78	0.62	0.63	0.62	0.65	0.59	0.62	0.54	0.47	0.50	0.44	0.45	0.44	**0.46**	0.16	0.24
$CellViT_{256}$ – Aug	HoVer-Net	CellViT	0.83	0.82	0.82	0.70	0.69	0.69	0.68	0.71	0.69	0.58	**0.59**	0.58	0.54	0.51	0.52	0.38	0.35	0.36
$CellViT_{256}$ – No-FC	HoVer-Net	CellViT	0.82	**0.83**	0.82	0.69	**0.70**	0.69	0.70	0.69	0.70	0.58	0.58	0.58	0.53	0.51	0.52	0.40	0.33	0.36
CellViT-Random (no pre-train)	HoVer-Net	CellViT	0.79	0.81	0.80	0.63	0.65	0.64	0.63	0.62	0.72	0.54	0.57	0.55	0.49	0.46	0.48	0.30	0.34	0.31
$CellViT_{256}$	HoVer-Net	CellViT	0.83	0.82	0.82	0.69	**0.70**	0.69	0.68	0.71	0.70	0.59	0.58	0.58	0.53	0.51	0.52	0.39	0.35	0.37
CellViT-SAM-B	HoVer-Net	CellViT	0.83	0.82	**0.83**	0.70	**0.70**	0.70	0.70	0.72	0.71	0.59	0.58	**0.59**	0.54	**0.52**	**0.53**	**0.46**	0.29	0.36
CellViT-SAM-L	HoVer-Net	CellViT	0.84	0.82	**0.83**	0.71	**0.70**	0.70	0.71	0.72	0.72	0.59	0.58	0.58	0.54	**0.52**	**0.53**	0.42	0.36	**0.39**
CellViT-SAM-H	HoVer-Net	CellViT	0.84	0.81	**0.83**	0.72	0.69	**0.71**	0.72	**0.73**	**0.73**	0.59	0.57	0.58	0.55	**0.52**	**0.53**	0.43	0.32	0.36
$CellViT_{256}$	STARDIST	CPP-Net	0.84	0.75	0.79	0.64	0.60	0.62	0.65	0.56	0.60	**0.64**	0.45	0.52	**0.58**	0.47	0.47	0.30	0.27	0.28
$CellViT_{256}$	STARDIST	CellViT	0.83	0.79	0.81	0.71	0.65	0.68	0.68	0.68	0.68	0.59	0.57	0.58	0.52	0.49	0.50	0.37	**0.38**	0.37
CellViT-SAM-H	STARDIST	CPP-Net	0.84	0.78	0.81	0.68	0.66	0.67	0.71	0.62	0.66	0.57	0.57	0.57	0.54	0.45	0.49	0.36	0.32	0.32
CellViT-SAM-H	STARDIST	CellViT	0.84	0.80	0.82	0.72	0.68	0.70	**0.74**	0.71	0.72	0.60	0.57	0.58	0.53	0.51	0.52	0.44	0.34	0.38
$CellViT_{256}$	CPP-Net	CPP-Net	0.85	0.76	0.80	0.69	0.62	0.65	0.70	0.62	0.65	0.57	0.55	0.56	0.53	0.46	0.49	0.32	**0.38**	0.33
$CellViT_{256}$	CPP-Net	CellViT	**0.87**	0.76	0.81	0.73	0.64	0.68	0.71	0.65	0.68	0.58	0.57	0.58	0.55	0.47	0.51	0.37	0.37	0.37
CellViT-SAM-H	CPP-Net	CPP-Net	0.86	0.78	0.82	0.72	0.67	0.70	0.73	0.68	0.70	0.62	0.55	0.58	0.55	0.50	0.52	0.27	0.14	0.18
CellViT-SAM-H	CPP-Net	CellViT	**0.87**	0.78	0.82	**0.74**	0.67	0.70	**0.74**	0.70	0.72	0.60	0.57	0.58	0.57	0.49	**0.53**	0.41	0.36	0.38
$CellViT_{256}$ (0.50 µm/px)***	HoVer-Net	CellViT	0.86	0.60	0.71	0.72	0.59	0.65	0.71	0.58	0.64	0.60	0.38	0.47	0.53	0.32	0.40	0.43	0.04	0.07
CellViT-SAM-H (0.50 µm/px)***	HoVer-Net	CellViT	0.88	0.63	0.73	0.74	0.62	0.67	0.74	0.61	0.67	0.60	0.42	0.49	0.56	0.34	0.42	0.49	0.04	0.08

After this initial warm-up phase to train the decoder, we proceed to train the entire model including the image encoder.

Implementation. All models are implemented in PyTorch 1.13.1. To augment images and masks, we used the Albumentations library (Buslaev et al., 2020). Other used libraries include the official STARDIST (Schmidt et al., 2018), CPP-Net (Chen et al., 2023) and CellSeg-models implementations (Okunator, 2022). For the pre-trained ViT_{256}-model, we utilized the ViT-S checkpoint[1] provided by Chen et al. (2022). As for the SAM-B, SAM-L, and SAM-H models, we use the encoder backbones of each final training stage of SAM (Kirillov et al., 2023), published on GitHub.[2] All experiments were conducted on an 80 GB NVIDIA A100 GPU with automatic mixed precision. However, it is worth noting that a 48 GB NVIDIA RTX A6000 is also sufficient for the ViT_{256} and SAM-B model training.

5. Results

In the section below, the results for the experiments (1) nuclei detection quality and (2) segmentation quality on PanNuke, (3) generalization performance on the independent MoNuSeg cohort, (4) cell-embedding analysis and (5) inference speed comparisons are given. If not stated otherwise, all models were trained on the PanNuke dataset with a resolution of 0.25 µm/px.

5.1. Detection quality on PanNuke

Considering the clinical importance of nuclei detection and classification over achieving the best possible segmentation quality, our ablation study aimed to determine the best model based on the detection results using the PanNuke dataset. Table 1 presents the precision, recall, and F_1-Score for both detection and classification performance across all nuclei classes, including the binary case. To determine the optimal settings, we evaluated different variations of our network. These include a randomly initialized network (CellViT-Random), networks with pre-trained weights from the ViT_{256} network ($CellViT_{256}$), and networks with different pre-trained SAM backbones (CellViT-SAM-B, CellViT-SAM-L, CellViT-SAM-H). To ensure comparability, the CellViT-Random network shares the same architecture (ViT-S, $D = 384$, $L = 12$) as the $CellViT_{256}$ network. All mentioned CellViT model variants were trained using data augmentation and our customized sampling strategy as regularization methods. The decoder network strategies (HoVer-Net, STARDIST or CPP-Net decoder) and hyperparameter settings are given behind the network name in Table 1.

We first analyze the CellViT models with HoVer-Net decoder. Compared to the baseline models, the randomly initialized CellViT-Random network achieves detection results comparable to the HoVer-Net CNN network. However, when using pre-trained encoder networks, we observe a significant performance increase, reaching state-of-the-art performance ($p < .05$, see Fig. A.3 in the Appendix). We notice a strong increase in F_1-scores compared to all existing solutions, especially for the epithelial nuclei class. Both the ViT_{256} and the three different SAM encoders exhibit significantly better performance, all at a similar level, with the CellViT-SAM-H model as the best solution ($p < .05$, see Fig. A.4). Notably, we even outperform purely detection-based methods like Mask-RCNN and all state-of-the-art approaches by a large margin with up to a 26% increase in the $F_{1,Epi}$-score of epithelial nuclei.

To demonstrate the effect of extensive data augmentation, customized sampling strategy and the Focal Tversky loss, we additionally report the results for a $CellViT_{256}$ model without regularization ($CellViT_{256}$-Raw), with oversampling only ($CellViT_{256}$-Over), with data-augmentation only ($CellViT_{256}$-Aug) and a model trained with oversampling and all augmentations, but without Focal Tversky loss ($CellViT_{256}$-No-FC) in Table 1. Our experiments reveal that data augmentation, in particular, is a crucial regularization method that significantly enhances performance (p-values for ablation study are Fig. A.3 in the Appendix). Specifically, the addition of data augmentation results in a 0.13 increase in the $F_{1,Dead}$ score for the dead nuclei class compared the ($CellViT_{256}$-Raw) model. Oversampling and Focal Tversky loss just lead to minimal improvements in the detection scores.

We also tested the STARDIST and CPP-Net decoder structures with the $CellViT_{256}$ and CellViT-SAM-H model with our hyperparameters and the CPP-Net hyperparameters suggested by Chen et al. (2023). These models usually achieve higher precision values but often a significantly lower recall and a lower F1 score than the models with the HoVer-Net

[1] https://github.com/mahmoodlab/HIPT
[2] https://github.com/facebookresearch/segment-anything

Table 2
Average mPQ and bPQ across the 19 tissue types of the PanNuke dataset for three-fold cross-validation. The standard deviation (STD) of the splits is provided in the final row. STARDIST models with ResNet50 (RN50) encoder were re-trained with CPP-Net hyperparameters (CPP-HP) and CellViT hyperparameters (CellViT-HP) for comparison. For the CellViT models, just the architecture with HoVer-Net decoder (HV-Net) is given. *TSFD-Net was not evaluated on the official three-fold splits of the PanNuke dataset and left out by the comparison **STARDIST trained by Chen et al. (2023) ***Model re-trained by ourselves.

	HoVer-Net		TSFD-Net*		STARDIST**		STARDIST***		STARDIST***		CPP-Net		$CellViT_{256}$		CellViT-SAM-H	
					RN50 encoder		RN50 encoder CPP-HP		RN50 encoder CellViT-HP		RN50 encoder		HV-Net decoder CellViT-HP		HV-Net decoder CellViT-HP	
Tissue	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ	mPQ	bPQ
Adrenal	0.4812	0.6962	0.5223	0.6900	0.4868	0.6972	0.4928	0.6954	0.4834	0.6884	0.4922	0.7031	0.4950	0.7009	**0.5134**	**0.7086**
Bile Duct	0.4714	0.6696	0.5000	0.6284	0.4651	0.6690	0.4632	0.6583	0.4680	0.6564	0.4650	0.6739	0.4721	0.6705	**0.4887**	**0.6784**
Bladder	0.5792	0.7031	0.5738	0.6773	0.5793	0.6986	0.5643	0.6949	0.5730	0.6901	**0.5932**	0.7057	0.5756	0.7056	0.5844	**0.7068**
Breast	0.4902	0.6470	0.5106	0.6245	0.5064	0.6666	0.4948	0.6585	0.4889	0.6497	0.5066	0.6718	0.5089	0.6641	**0.5180**	**0.6748**
Cervix	0.4438	0.6652	0.5204	0.6561	0.4628	0.6690	0.4752	0.6739	0.4781	0.6685	0.4779	**0.6880**	0.4893	0.6862	**0.4984**	0.6872
Colon	0.4095	0.5575	0.4382	0.5370	0.4205	0.5779	0.4230	0.5704	0.4087	0.5555	0.4269	0.5888	0.4245	0.5700	**0.4485**	**0.5921**
Esophagus	0.5085	0.6427	0.5438	0.6306	0.5331	0.6655	0.5200	0.6508	0.5175	0.6446	0.5410	**0.6755**	0.5373	0.6619	**0.5454**	0.6682
Head & Neck	0.4530	0.6331	0.4937	0.6277	0.4768	0.6433	0.4660	0.6305	0.4629	0.6215	0.4667	0.6468	0.4901	0.6472	**0.4913**	**0.6544**
Kidney	0.4424	0.6836	0.5517	0.6824	**0.5880**	0.6998	0.5090	0.6888	0.4750	0.6800	0.5092	0.7001	0.5409	0.6993	0.5366	**0.7092**
Liver	0.4974	0.7248	0.5079	0.6675	0.5145	0.7231	0.4899	0.7106	0.5034	0.7051	0.5099	0.7271	0.5065	0.7160	**0.5224**	**0.7322**
Lung	0.4004	0.6302	0.4274	0.5941	0.4128	0.6362	0.3627	0.6087	0.3931	0.6205	0.4234	0.6364	0.4102	0.6317	**0.4314**	**0.6426**
Ovarian	0.4863	0.6309	0.5253	0.6431	0.5205	0.6668	0.5106	0.6573	0.5204	0.6547	0.5276	**0.6792**	0.5260	0.6596	**0.5390**	0.6722
Pancreatic	0.4600	0.6491	0.4893	0.6241	0.4585	0.6601	0.4548	0.6516	0.4526	0.6439	0.4680	**0.6742**	**0.4769**	0.6643	0.4719	0.6658
Prostate	0.5101	0.6615	0.5431	0.6406	0.5067	0.6748	0.4905	0.6561	0.4812	0.6457	0.5261	**0.6903**	0.5164	0.6695	**0.5321**	0.6821
Skin	0.3429	0.6234	0.4354	0.6074	0.3610	0.6289	0.3826	0.6349	0.3709	0.6197	0.3547	0.6192	0.3661	0.6400	**0.4339**	**0.6565**
Stomach	**0.4726**	0.6886	0.4871	0.6529	0.4477	0.6944	0.4239	0.6769	0.4194	0.6642	0.4553	**0.7043**	0.4475	0.6918	0.4705	0.7022
Testis	0.4754	0.6890	0.4843	0.6435	0.4942	0.6869	0.4819	0.6848	**0.5141**	0.6812	0.4917	**0.7006**	0.5091	0.6883	0.5127	0.6955
Thyroid	0.4315	0.6983	0.5154	0.6692	0.4300	0.6962	0.4246	0.6962	0.4175	0.6921	0.4344	0.7094	0.4412	0.7035	**0.4519**	**0.7151**
Uterus	0.4393	0.6393	0.5068	0.6204	0.4480	0.6599	0.4452	0.6455	0.4683	0.6428	**0.4790**	0.6622	0.4737	0.6516	0.4737	**0.6625**
Average	0.4629	0.6596	0.5040	0.6377	0.4796	0.6692	0.4671	0.6602	0.4682	0.6539	0.4815	0.6767	0.4846	0.6696	**0.4980**	**0.6793**
STD	0.0076	0.0036	–	–	–	–	0.0489	0.0340	0.0496	0.0348	–	–	0.0503	0.0340	0.0413	0.0318

decoder architecture. As an extension of the STARDIST method, the CPP-Net decoder achieves slightly better results. Overall, the models achieve better detection results than comparable CNN-based SOTA networks and outperform the ResNet50-based STARDIST model, but are inferior to our suggested models with HoVer-Net decoder architecture. The results also reveal that our hyperparameters provide better detection performance.

In addition to the provided dataset resolution of 0.25 µm/px, we performed training and evaluation for the two best model variants $CellViT_{256}$ and CellViT-SAM-H on downscaled PanNuke data (from 256×256 to 128×128 px patch size), resulting in 0.50 µm/px resolution. The results are presented in the last two rows of Table 1. The downsizing leads to a substantial drop in performance compared to the 0.25 µm/px networks, with detection results approaching the baseline models. Notably, the recall of individual classes significantly decreases (by an average of −0.20). In particular, the recall for the dead nuclei class drops to 0.04, indicating that this class is almost not detected at all. Interestingly, the precision increases minimally or remains almost the same compared to our best 0.25 µm/px models. We conclude that despite detecting significantly fewer nuclei, when a nucleus is identified and classified correctly, it corresponds to the true nucleus class with high accuracy for most classes.

For subsequent investigations, we decided to further just consider the $CellViT_{256}$ and CellViT-SAM-H models to enable a comparison between in-domain and out-of-domain pre-training. All p-values for the detection metrics are in the Appendix (Fig. A.3, Fig. A.4)

5.2. Segmentation quality on PanNuke

To assess the segmentation quality, the panoptic quality is used. Table 3 presents the PQ values for each nuclei type, averaged over all tissue types. Among all settings, $CellViT_{256}$ and CellViT-SAM-H networks with HoVer-Net decoder excel in neoplastic, connective, and epithelial nuclei. However, in the case of inflammatory and connective nuclei, they are outperformed by TSFD-Net due to its larger training dataset (80/20 split vs. 33/67 split). Notably, all models consistently yield the lowest results for dead cells, attributed to class imbalance and the small size of dead cells. To further analyze the influence of Focal Tversky loss and our custom oversampling strategy, we included PQ values for the $CellViT_{256}$ model (HoVer-Net decoder) with different regularization techniques in Table 3. It is observed that the segmentation quality is improved by oversampling ($CellViT_{256}$-Over) for almost all nuclei classes except neoplastic nuclei. The deterioration of neoplastic nuclei is attributed to the class rebalancing, as neoplastic nuclei constitute the majority class in the dataset. Removing the Focal Tversky loss ($CellViT_{256}$-No-FC), leads to a decrease in panoptic quality for all classes, except neoplastic nuclei again, but compared to the $CellViT_{256}$ model with all regularization techniques, this decline is not statistically significant ($p = 0.51$, see Fig. A.3)). Models employing STARDIST and CPP-Net decoders achieve lower panoptic quality than HoVer-Net decoder models but surpass baseline networks. The results for the 0.50 µm/px models reveal a significant drop in performance when using the downscaled data.

Finally, we evaluate the segmentation performance of the $CellViT_{256}$ and CellViT-SAM-H models with HoVer-Net decoder against the best baseline models by computing the binary PQ (bPQ) and the more challenging multi-class PQ (mPQ) for each of the 19 tissue types in PanNuke, providing an assessment of both instance segmentation qualities. As baseline experiments, we just include the best HoVer-Net model by Graham et al. (2019), TSFD-Net and the original STARDIST and CPP-Net models with ResNet50 encoder Chen et al. (2023). For our detection experiments in Section 5.1, we retrained the baseline STARDIST model with the ResNet50 encoder. Even though we are not able to reproduce segmentation results reported by Chen et al. (2023) with CPP-Net and our hyperparameter settings, we include all three results in Table 2 for a fair comparison (STARDIST ResNet50 (Chen et al., 2023), STARDIST ResNet50 re-trained with CPP-Net hyperparameters, and STARDIST ResNet50 with our hyperparameters). Our experimental results demonstrate that CPP-Net and STARDIST (both with ResNet50 encoder) exhibit comparable bPQ values, whereas our CellViT models achieve superior mPQ, which is statistically significant ($p < .05$ for all CellViT-models with HoVer-Net decoder, see Fig. A.5 and A.6 in the Appendix). This is primarily attributed to the superior detection capabilities of our models, which impacts the mPQ value. The best average model is CellViT-SAM-H with the HoVer-Net decoder architecture trained with our hyperparameter settings. Segmentation results per tissue for 0.50 µm/px are given in the Appendix A.1.

To provide a visual representation of the segmentations, we include tissue-wise comparisons between ground-truth and segmentation predictions of the CellViT-SAM-H model in Fig. 4. As observed in the lung example, the instance segmentation of dead cells poses a significant challenge due to their small size. Furthermore, detecting and segmenting dead nuclei becomes even more difficult when these images are scaled down from 0.25 µm/px to 0.50 µm/px resolution.

Table 3
Average *PQ* across the three PanNuke splits for each nuclear category on the PanNuke dataset. *TSFD-Net was not evaluated on the official three-fold splits of the PanNuke dataset and left out by the comparison. **Model re-trained by ourselves ***Models trained on downscaled 0.50 µm/px PanNuke images.
Abbreviations: Decoder (Dec.), Hyperparameter (HP.), HoVer-Net (HV), STARDIST (SD), CPP-Net (CPP).

Model	Dec.	HP.	Neoplastic	Epithelial	Inflammatory	Connective	Dead
DIST			0.439	0.290	0.343	0.275	0.000
Mask-RCNN			0.472	0.403	0.290	0.300	0.069
Micro-Net			0.504	0.442	0.333	0.334	0.051
HoVer-Net			0.551	0.491	0.417	0.388	0.139
TSFD-Net*			0.572	0.566	0.453	0.423	0.214
STARDIST (RN50)**	SD	CPP	0.564	0.543	0.398	0.388	0.024
STARDIST (RN50)**	SD	CellViT	0.547	0.532	**0.424**	0.380	0.123
CellViT-SAM-H	HV	CellViT	**0.581**	**0.583**	0.417	**0.423**	0.149
$CellViT_{256}$	HV	CellViT	0.567	0.559	0.405	0.405	0.144
$CellViT_{256}$ – Raw	HV	CellViT	0.495	0.465	0.344	0.335	0.067
$CellViT_{256}$ – Over	HV	CellViT	0.494	0.467	0.349	0.339	0.071
$CellViT_{256}$ – Aug	HV	CellViT	0.565	0.558	0.419	0.403	**0.156**
$CellViT_{256}$ – No-FC	HV	CellViT	0.567	0.548	0.416	0.404	0.141
$CellViT_{256}$	SD	CellViT	0.516	0.507	0.400	0.331	0.128
CellViT-SAM-H	SD	CellViT	0.548	0.544	0.400	0.347	0.132
$CellViT_{256}$	CPP	CellViT	0.540	0.524	0.414	0.369	0.133
CellViT-SAM-H	CPP	CellViT	0.571	0.565	0.405	0.395	0.131
$CellViT_{256}$*** (0.50 µm/px)	HV	CellViT	0.497	0.467	0.292	0.285	0.021
CellViT-SAM-H*** (0.50 µm/px)	HV	CellViT	0.528	0.502	0.315	0.311	0.031

Table 4
MoNuSeg validation result for $CellViT_{256}$ and CellViT-SAM-H models with HoVer-Net decoder and trained with CellViT hyperparameters on different dataset resolutions and inference patch sizes averaged over all three PanNuke training folds. The original image size for 0.25 µm/px resolution with ×40 magnification (mag.) is 1024 px, and 512 px for 0.25 µm/px (×20 mag.). *Models trained on downscaled 0.50 µm/px PanNuke images.

Dataset resolution	Inference patch size	256 px with 64 px overlap				256 px without overlap				1024 px (no patching)			
	Metric	*bPQ*	P_d	R_d	F_1,d	*bPQ*	P_d	R_d	F_1,d	*bPQ*	P_d	R_d	F_1,d
0.25 µm/px (×40 mag.)	$CellViT_{256}$	0.660	0.841	0.886	0.863	0.621	0.814	0.897	0.853	0.661	0.838	0.859	0.848
	CellViT-SAM-H	0.671	0.846	0.893	**0.868**	0.631	0.814	0.906	0.857	**0.672**	**0.847**	0.885	0.865
	$CellViT_{256}$ (0.50 µm/px)*	0.509	0.748	0.893	0.804	0.491	0.728	0.895	0.792	0.515	0.759	0.905	0.813
	CellViT-SAM-H (0.50 µm/px)*	0.524	0.746	0.963	0.840	0.514	0.729	0.963	0.829	0.540	0.749	**0.966**	0.842
		256 px with 64 px overlap				256 px without overlap				512 px (no patching)			
0.50 µm/px (×20 mag.)	$CellViT_{256}$	0.588	0.918	0.766	0.834	0.586	0.902	0.759	0.824	0.593	0.919	0.771	0.837
	CellViT-SAM-H	0.627	**0.922**	0.791	**0.851**	0.620	0.908	0.784	0.841	0.627	0.909	0.792	0.846
	$CellViT_{256}$ (0.50 µm/px)*	0.643	0.874	0.803	0.836	0.640	0.867	0.797	0.830	0.644	0.873	0.810	0.840
	CellViT-SAM-H (0.50 µm/px)*	0.649	0.835	0.814	0.824	0.648	0.841	0.820	0.830	**0.655**	0.840	**0.829**	0.834

5.3. MoNuSeg test performance

In this experiment, we focused on instance segmentation without classification on the MoNuSeg dataset to assess the generalizability of our models (just with HoVer-Net decoder) at resolutions of 0.25 µm/px and 0.50 µm/px. Additionally, we aim to evaluate the impact of changing the input sequence size by performing inference on large-scale tiles of size 1024 px (0.25 µm/px) and 512 px (0.50 µm/px), respectively, comparing the results to non-overlapping 256 px patches and 256 px patches with an overlap of 64 px derived by a shifting window approach. We utilized the three final models of the PanNuke training folds for each architecture and conducted inference on the MoNuSeg data without retraining.

The evaluation results are presented in Table 4. Consistent with the previous experiments, the CellViT-SAM-H model is the best-performing model. It achieves a *bPQ*-score of 0.672 on 1024 px tiles when no patching was applied and of 0.671 for 256 px tiles with an overlap of 64 px. However, when using 256 px patches without overlap, the *bPQ*-score decreases to 0.631, likely due to the absence of merging overlapping nuclei at cell borders and double detected cells (higher recall). Importantly, the overall comparison between larger tiles and smaller tiles with overlapping indicates that inference on larger tiles did not lead to a degradation in performance. This justifies our inference pipeline for large-scale WSI, in which we are using 1024 px sized patches with an overlap of 64 px and overlapping merging strategies. The $CellViT_{256}$ model yields slightly inferior results compared to the CellViT-SAM-H model.

Using the models trained on 0.50 µm/px data on the 0.25 µm/px data and vice versa, the 0.50 µm/px trained models exhibit poor performance on 0.25 µm/px data, while the 0.25 µm/px trained models experience a less severe performance drop on the 0.50 µm/px data. Nevertheless, networks trained and evaluated on the same WSI resolution achieved the best performance, thus it is advisable to align image resolution between different dataset and use the appropriate model. Consistently, the best results are achieved for WSI acquired with a resolution of 0.25 µm/px.

We include a visual demonstration presenting a tissue tile from the MoNuSeg test set along with binary segmentation masks generated by the CellViT-SAM-H and CellViT-SAM-H (0.50 µm/px) models in the Appendix.

5.4. Token analysis

In Fig. 5, we present the two-dimensional UMAP embeddings of cell tokens from the CoNSeP dataset. The CellViT-SAM-H and $CellViT_{256}$

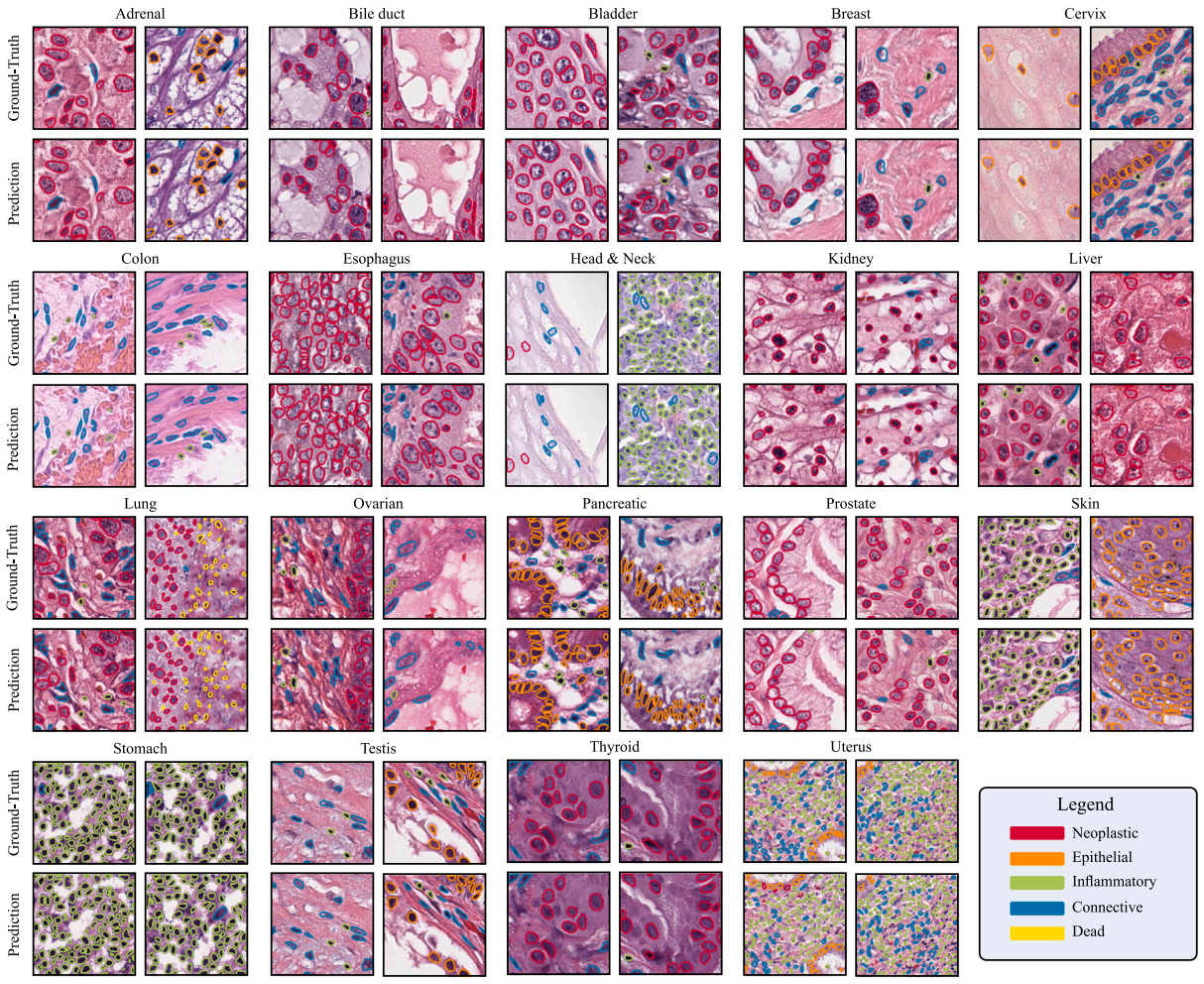

Fig. 4. Example of PanNuke patches with ground-truth annotations and CellViT-SAM-H predictions overlaid for each tissue type.

models with HoVer-Net decoder, trained on the PanNuke dataset, were utilized. The tokens were extracted simultaneously with cell detections in a single inference pass. The color overlay in the scatter plots (left) and tissue images (right) indicates the respective nuclei classes. Consistent with Graham et al. (2019), we grouped normal and malignant/dysplastic epithelial nuclei into an "epithelial" class, while fibroblast, muscle, and endothelial nuclei were grouped into the "spindle-shaped nuclei" class. The global clusters in the scatter plot represent cells from different images, with clusters containing cells from the same tissue phenotype being grouped together. An example of this is cluster 1 for the CellViT-SAM-H model. It comprises cell clusters from two images, both containing multiple glands. Within this cluster, the local spatial arrangement of the cell embeddings allows differentiation of nuclei types (epithelial, spindle-shaped, and inflammatory) despite the model not being explicitly trained for all cell classes (spindle-shaped cells are not explicitly defined in the PanNuke dataset). Cluster 3, which is spatially close to cluster 1, contains even more glands, while the tissue image associated with the distant cluster 2 lacks glands and primarily consists of spindle-shaped and inflammatory nuclei. In summary, the global UMAP arrangement primarily captures differences in the nuclei's tissue environment (e.g., nearby glands, muscles). The local arrangement highlights distinctions between nuclei without the need for fine-tuning the model for specific nuclei types. Notably, in the CellViT$_{256}$ model, there is an increased emphasis on global tissue differences. This, however, results in the generation of new clusters even for subtle tissue variations (e.g., number of glands changed). While clusters 2 and 3 from the CellViT-SAM-H plot were successfully re-identified, cluster 1, encompassing tissue from two different samples but both containing glands, underwent separation into two distinct clusters and therefore could not be clearly recognized as a cluster.

To quantitatively assess the quality of the embeddings, we trained a linear nuclei classifier on the embeddings of the training data (15,548 nuclei) to classify the nuclei into the CoNSeP classes. We evaluated the classifier on the embeddings of the test images (8773 nuclei). The model achieved an area under the receiver operating characteristics curve (AUROC) of 0.963 for the validation data using the CellViT-SAM-H embeddings. When utilizing the CellViT$_{256}$ embeddings, the model achieved an AUROC of 0.960. This demonstrates the effectiveness of our embeddings in classifying unknown nuclei classes, with both CellViT-SAM-H and CellViT$_{256}$ embeddings yielding high AUROC values.

5.5. Inference runtime

Our inference runtime benchmark shows that our inference pipeline is accelerated by a factor of 2.49 (CellViT$_{256}$) and 2.25 (CellViT-SAM-H) when using 1024 px input patches instead of 256 px. The CellViT$_{256}$ model with 1024 px input patches is 1.34 times faster than the CellViT-SAM-H model with 1024 px patches, due to the smaller ViT-Structure of CellViT$_{256}$ (ViT-Small with 21.7 M parameters vs. ViT-Huge with 632 M parameters) and reduced computational complexity. Both CellViT models with their large 1024 px input patch size outperform the HoVerNet model (input 256 px with overlap, output 164 px), with speedups of 1.85 (CellViT$_{256}$) and 1.39 (CellViT-SAM-H), respectively. Further

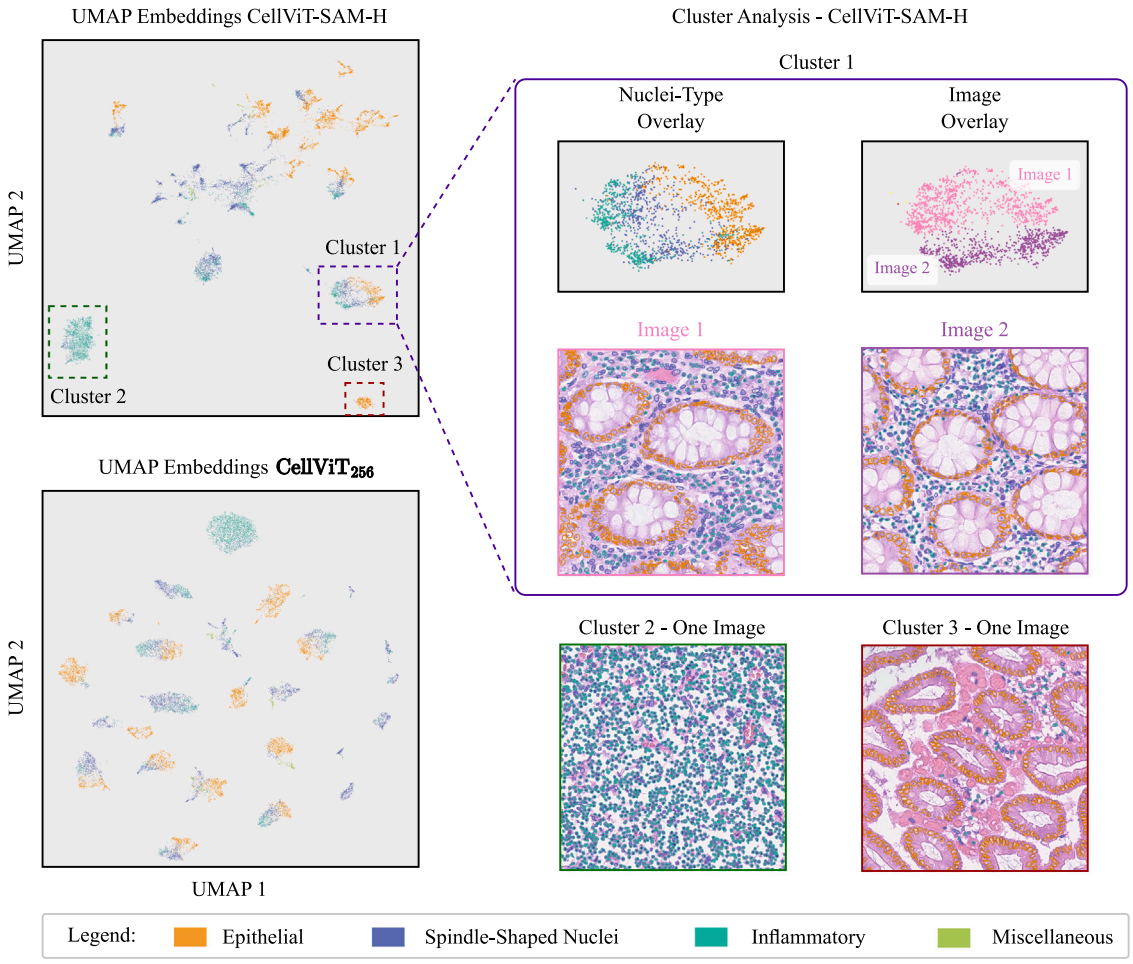

Fig. 5. Two-dimensional UMAP embedding visualization (left) of the CoNSeP dataset with the CellViT-SAM-H and CellViT$_{256}$ (HoVer-Net encoder) models trained on PanNuke. We extract cell-tokens for each detected cell with our model, resulting in one embedding vector per cell. On the right side of the figure, representative clusters derived with the CellViT-SAM-H model are displayed alongside corresponding tissue images. The color overlay illustrates the ground-truth nuclei types within the dataset.

results containing network parameter counts and number of multiply-accumulate operations (MACs) are given in Tab A.5 in the Appendix. The observed speedup between HoVer-net and our CellViT models can be attributed to the interaction of multiple components, namely reduced computational complexity for a 1024px sized Field-of-View for the CellViT model (see Tab A.5, HoVer-Net 5361.6 million MACs, CellViT$_{256}$ 2127.9 million MACs, CellViT-SAM-H 3413.4 million MACs) and reduced patch extraction effort. Simultaneously, a decrease in the number of merging operations for overlapping patches further contributes to the overall realized speed advantage.

6. Discussion and conclusion

Nuclei instance segmentation is crucial for clinical applications, requiring automated tools that offer high robustness and reliability. In the clinical context of performing large-scale analysis on clinical patient cohorts, accurate detection is considered more important than precise segmentation.

In this work, we introduced a novel deep learning-based method for simultaneously segmenting and detecting nuclei in digitized H&E tissue samples. Our work was inspired by the success of previous works using large-scale trained Vision Transformers, particularly by the contributions made by Chen et al. (2022) (ViT$_{256}$) and Kirillov et al. (2023) (SAM). The CellViT network proposed in this study demonstrates state-of-the-art performance for both nuclei instance segmentation and nuclei detection on the PanNuke dataset. Additionally, the results on the MoNuSeg dataset validate the generalizability of our model to previously unseen cohorts. Notably, our model surpasses all other existing methods by a significant margin for nuclei detection and classification, elevating nuclei detection in H&E-slides to a new level. By leveraging the most recent computer vision approaches, we showed that both in-domain pre-training (ViT$_{256}$) and the use of the SAM foundation model yields significantly better results compared to randomly initialized network weights. Our larger inference patch size allows us to be 1.85 times faster than the popular HoVer-Net inference framework by Graham et al. (2019), which could save hours in computational time when dealing with huge gigapixel WSI. Moreover, our framework allows direct assessment of a localizable ViT-token from a detected nucleus that can be further used in downstream tissue analysis tasks. Although an evaluation of this aspect is pending, we anticipate promising prospects based on our first results in Section 5.4. Our work provides the potential to design interpretable algorithms that directly correlate with specific cells or cell patterns. One possible direction for future research involves graph-based networks with attention mechanisms using these embeddings.

Nevertheless, external validation of the results is necessary. Yet, additional datasets are required, especially to verify the detection quality of our model. Furthermore, our models exhibit reliable performance only for WSI acquired at 0.25 μm/px resolution. While the results obtained with 0.50 μm/px images are acceptable in terms of detection, there is room for improvement, as there is a huge performance gap between 0.25 μm/px and 0.50 μm/px-WSI processing. We recommend to scan the tissue samples on a resolution of 0.25 μm/px if technically possible. In the future, we plan to apply the proposed model with extracted nuclei tokens to downstream histological image analysis tasks. This will enable us to validate if simultaneously extracted tokens are an advantage for building interpretable algorithms for computational pathology. Additionally, it will allow us to evaluate which tokens have achieved a more meaningful representation of the tissue and are better suited for downstream tasks, as there are just minimal differences in the segmentation and detection performance between our best-performing CellViT_{256} and CellViT-SAM-H models. To ensure the accessibility of our results, we have made both the code and pre-trained models publicly available under an open-source license for non-commercial use.

CRediT authorship contribution statement

Fabian Hörst: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Moritz Rempe:** Methodology, Writing – original draft, Writing – review & editing. **Lukas Heine:** Methodology, Writing – original draft, Writing – review & editing. **Constantin Seibold:** Conceptualization, Writing – review & editing. **Julius Keyl:** Validation, Writing – review & editing. **Giulia Baldini:** Validation, Writing – review & editing. **Selma Ugurel:** Validation, Writing – review & editing. **Jens Siveke:** Validation, Writing – review & editing. **Barbara Grünwald:** Validation, Writing – review & editing. **Jan Egger:** Supervision, Writing – review & editing. **Jens Kleesiek:** Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Selma Ugurel reports a relationship with Bristol Myers Squibb that includes: consulting or advisory, funding grants, speaking and lecture fees, and travel reimbursement. Selma Ugurel reports a relationship with Merck Serono that includes: consulting or advisory, funding grants, and speaking and lecture fees. Selma Ugurel reports a relationship with Merck Sharp & Dohme that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Selma Ugurel reports a relationship with Novartis that includes: consulting or advisory, speaking and lecture fees, and travel reimbursement. Selma Ugurel reports a relationship with Almirall that includes: travel reimbursement. Selma Ugurel reports a relationship with IGEA Clinical Biophysics that includes: travel reimbursement. Selma Ugurel reports a relationship with Pierre Fabre that includes: travel reimbursement. Selma Ugurel reports a relationship with Sun Pharma that includes: travel reimbursement. Jens T. Siveke reports a relationship with AstraZeneca that includes: consulting or advisory and speaking and lecture fees. Jens T. Siveke reports a relationship with Bayer AG that includes: consulting or advisory and speaking and lecture fees. Jens T. Siveke reports a relationship with Boehringer Ingelheim that includes: consulting or advisory, funding grants, and speaking and lecture fees. Jens T. Siveke reports a relationship with Bristol-Myers Squibb that includes: consulting or advisory, funding grants, and speaking and lecture fees. Jens T. Siveke reports a relationship with Immunocore that includes: consulting or advisory and speaking and lecture fees. Jens T. Siveke reports a relationship with MSD Sharp Dohme that includes: consulting or advisory and speaking and lecture fees. Jens T. Siveke reports a relationship with Novartis that includes: consulting or advisory and speaking and lecture fees. Jens T. Siveke reports a relationship with Roche Genentech that includes: consulting or advisory, funding grants, and speaking and lecture fees. Jens T. Siveke reports a relationship with Servier that includes: consulting or advisory and speaking and lecture fees. Jens T. Siveke reports a relationship with Abalos Therapeutics that includes: funding grants. Jens T. Siveke reports a relationship with Celgene that includes: funding grants. Jens T. Siveke reports a relationship with Eisbach Bio that includes: funding grants. Jens T. Siveke reports a relationship with Pharma15 that includes: board membership and equity or stocks.

Data availability

The used datasets are publicly available. All models and source-code are available online here: https://github.com/TIO-IKIM/CellViT.

Acknowledgments

This work received funding from 'KITE' (Plattform für KI-Translation Essen) from the REACT-EU initiative (https://kite.ikim.nrw/, EFRE-0801977) and the Cancer Research Center Cologne Essen (CCCE). We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103143.

References

Abraham, N., Khan, N.M., 2019. A novel focal tversky loss function with improved attention U-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 683–687. http://dx.doi.org/10.1109/ISBI.2019.8759329.

Ali, S., Madabhushi, A., 2012. An Integrated Region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. IEEE Trans. Med. Imaging 31 (7), 1448–1460. http://dx.doi.org/10.1109/TMI.2012.2190089.

Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., James, J.A., Salto-Tellez, M., Hamilton, P.W., 2017. QuPath: Open source software for digital pathology image analysis. Sci. Rep. 7 (1), http://dx.doi.org/10.1038/s41598-017-17204-5.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2021. On the opportunities and risks of foundation models. http://dx.doi.org/10.48550/arXiv.2108.07258, arXiv.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: fast and flexible image augmentations. Information 11 (2), 125.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309. http://dx.doi.org/10.1038/s41591-019-0508-1.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. 33, 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660. http://dx.doi.org/10.1109/ICCV48922.2021.00951.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16144–16155. http://dx.doi.org/10.1109/CVPR52688.2022.01567.

Chen, S., Ding, C., Liu, M., Cheng, J., Tao, D., 2023. CPP-Net: Context-aware polygon proposal network for nucleus segmentation. IEEE Trans. Image Process. 32, 980–994. http://dx.doi.org/10.1109/TIP.2023.3237013.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758. http://dx.doi.org/10.1109/CVPR46437.2021.01549.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Chen, H., Qi, X., Yu, L., Heng, P., 2016. DCAN: Deep contour-aware networks for accurate gland segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, pp. 2487–2496. http://dx.doi.org/10.1109/CVPR.2016.273.

Chen, L., Yu, Q.T., 2021. Transformers make strong encoders for medical image segmentation. http://dx.doi.org/10.48550/arXiv.2102.04306, arXiv.

Cheng, J., Rajapakse, J.C., 2009. Segmentation of clustered nuclei with shape markers and marking function. IEEE Trans. Biomed. Eng. 56 (3), 741–748. http://dx.doi.org/10.1109/TBME.2008.2008635.

Corredor, G., Whitney, J., Arias, V., Madabhushi, A., Romero, E., 2017. Training a cell-level classifier for detecting basal-cell carcinoma by combining human visual attention maps with low-level handcrafted features. J. Med. Imaging 4 (2), 021105. http://dx.doi.org/10.1117/1.jmi.4.2.021105.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv Preprint.

Ester, O., Hörst, F., Seibold, C., Keyl, J., Ting, S., Vasileiadis, N., Schmitz, J., Ivanyi, P., Grünwald, V., Bräsen, J.H., Egger, J., Kleesiek, J., 2023. Valuing vicinity: Memory attention framework for context-based semantic segmentation in histopathology. Comput. Med. Imaging Graph. 107, 102238. http://dx.doi.org/10.1016/j.compmedimag.2023.102238.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25 (1), 24–29. http://dx.doi.org/10.1038/s41591-018-0316-z.

Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N., 2020. PanNuke dataset extension, insights and baselines. http://dx.doi.org/10.48550/arXiv.2003.10778, arXiv Preprint.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 1440–1448. http://dx.doi.org/10.1109/ICCV.2017.322.

Graham, S., Jahanifar, M., Vu, Q.D., Hadjigeorghiou, G., Leech, T., Snead, D., Raza, S.E.A., Minhas, F., Rajpoot, N., 2021. CoNIC: Colon nuclei identification and counting challenge 2022. http://dx.doi.org/10.48550/arXiv.2111.14485, arXiv Preprint.

Graham, S., Vu, Q.D., Jahanifar, M., Raza, S.E.A., Minhas, F., Snead, D., Rajpoot, N., 2023. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. Med. Image Anal. 83, 102685. http://dx.doi.org/10.1016/j.media.2022.102685.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. 58, 101563. http://dx.doi.org/10.1016/j.media.2019.101563.

Greten, F.R., Grivennikov, S.I., 2019. Inflammation and cancer: Triggers, mechanisms, and consequences. Immunity 51 (1), 27–41. http://dx.doi.org/10.1016/j.immuni.2019.06.025.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284.

Grünwald, B.T., Devisme, A., Andrieux, G., Vyas, F., Aliar, K., McCloskey, C.W., Macklin, A., Jang, G.H., Denroche, R., Romero, J.M., Bavi, P., Bronsert, P., Notta, F., O'Kane, G., Wilson, J., Knox, J., Tamblyn, L., Udaskin, M., Radulovich, N., Fischer, S.E., Boerries, M., Gallinger, S., Kislinger, T., Khokha, R., 2021. Spatially confined sub-tumor microenvironments in pancreatic cancer. Cell 184 (22), 5577–5592.e18. http://dx.doi.org/10.1016/j.cell.2021.09.022.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2022a. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, pp. 272–284. http://dx.doi.org/10.1007/978-3-031-08999-2_22.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. UNETR: Transformers for 3D medical image segmentation. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 1748–1758. http://dx.doi.org/10.1109/WACV51458.2022.00181.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738. http://dx.doi.org/10.1109/CVPR42600.2020.00975.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hörst, F., Ting, S., Liffers, S.-T., Pomykala, K.L., Steiger, K., Albertsmeier, M., Angele, M.K., Lorenzen, S., Quante, M., Weichert, W., Egger, J., Siveke, J.T., Kleesiek, J., 2023. Histology-based prediction of therapy response to neoadjuvant chemotherapy for esophageal and esophagogastric junction adenocarcinomas using deep learning. JCO Clin. Cancer Inform. (7).

Ilyas, T., Mannan, Z.I., Khan, A., Azam, S., Kim, H., De Boer, F., 2022. TSFD-Net: Tissue specific feature distillation network for nuclei segmentation and classification. Neural Netw. 151, 1–15. http://dx.doi.org/10.1016/j.neunet.2022.02.020.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2020. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211. http://dx.doi.org/10.1038/s41592-020-01008-z.

Kelly, B.S., Judge, C., Bollard, S.M., Clifford, S.M., Healy, G.M., Aziz, A., Mathur, P., Islam, S., Yeom, K.W., Lawlor, A., Killeen, R.P., 2022. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur. Radiol. 32 (11), 7998–8007. http://dx.doi.org/10.1007/s00330-022-08784-6.

Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P., 2019. Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9404–9413. http://dx.doi.org/10.1109/CVPR.2019.00963.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. http://dx.doi.org/10.48550/arXiv.2304.02643, arXiv Preprint.

Koohbanani, N.A., Jahanifar, M., Gooya, A., Rajpoot, N., 2019. Nuclear instance segmentation using a proposal-free spatially aware deep learning framework. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 622–630. http://dx.doi.org/10.1007/978-3-030-32239-7_69.

Kothari, S., Chaudry, Q., Wang, M., 2009. Extraction of informative cell features by segmentation of densely clustered tissue images. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 6706–6709. http://dx.doi.org/10.1109/iembs.2009.5333810.

Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.-A., Li, J., et al., 2020. A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging 39 (5), 1380–1391. http://dx.doi.org/10.1109/TMI.2019.2947628.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36 (7), 1550–1560. http://dx.doi.org/10.1109/TMI.2017.2677499.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R., 2021. Medical image segmentation using squeeze-and-expansion transformers. http://dx.doi.org/10.48550/arXiv.2105.09511, arXiv.

Liao, M., qian Zhao, Y., hua Li, X., shan Dai, P., wen Xu, X., kai Zhang, J., ji Zou, B., 2016. Automatic segmentation for cell images based on bottleneck detection and ellipse fitting. Neurocomputing 173, 615–622. http://dx.doi.org/10.1016/j.neucom.2015.08.006.

Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, pp. 936–944. http://dx.doi.org/10.1109/CVPR.2017.106.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988. http://dx.doi.org/10.1109/tpami.2018.2858826.

Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570. http://dx.doi.org/10.1038/s41551-020-00682-w.

Malpica, N., de Solórzano, C.O., Vaquero, J.J., Santos, A., Vallcorba, I., García-Sagredo, J.M., del Pozo, F., 1998. Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry 28 (4), 289–297. http://dx.doi.org/10.1002/(sici)1097-0320(19970801)28:4<289::aid-cyto3>3.0.co;2-7.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2022. Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 44 (7), 3523–3542. http://dx.doi.org/10.1109/TPAMI.2021.3059968.

Murray, J.M., Kaissis, G., Braren, R., Kleesiek, J., 2019. Wie funktioniert radiomics? Der Radiol. 60 (1), 32–41. http://dx.doi.org/10.1007/s00117-019-00617-w.

Naylor, P., Laé, M., Reyal, F., Walter, T., 2019. Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Trans. Med. Imaging 38 (2), 448–459. http://dx.doi.org/10.1109/TMI.2018.2865709.

Okunator, 2022. okunator/cellseg_models.pytorch: v0.1.23. http://dx.doi.org/10.5281/ZENODO.7064617.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A., 2021. Do vision transformers see like convolutional neural networks? Adv. Neural Inf. Process. Syst. 34, 12116–12128.

Raza, S.E.A., Cheung, L., Shaban, M., Graham, S., Epstein, D., Pelengaris, S., Khan, M., Rajpoot, N.M., 2019. Micro-net: A unified model for segmentation of various objects in microscopy images. Med. Image Anal. 52, 160–173. http://dx.doi.org/10.1016/j.media.2018.12.003.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Schmidt, U., Weigert, M., Broaddus, C., Myers, G., 2018. Cell detection with star-convex polygons. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, pp. 265–273. http://dx.doi.org/10.1007/978-3-030-00934-2_30.

Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V., 2021. U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access 9, 82031–82057. http://dx.doi.org/10.1109/ACCESS.2021.3086020.

Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R.J., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1196–1206. http://dx.doi.org/10.1109/TMI.2016.2525803.

Sirinukunwattana, K., Snead, D., Epstein, D., Aftab, Z., Mujeeb, I., Tsang, Y.W., Cree, I., Rajpoot, N., 2018. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Sci. Rep. 8 (1), http://dx.doi.org/10.1038/s41598-018-31799-3.

Song, Y., Tan, E.-L., Jiang, X., Cheng, J.-Z., Ni, D., Chen, S., Lei, B., Wang, T., 2017. Accurate cervical cell segmentation from overlapping clumps in pap smear images. IEEE Trans. Med. Imaging 36 (1), 288–300. http://dx.doi.org/10.1109/TMI.2016.2606380.

Stanton, S.E., Disis, M.L., 2016. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. J. ImmunoTherapy Cancer 4 (1), http://dx.doi.org/10.1186/s40425-016-0165-6.

Tareef, A., Song, Y., Huang, H., Feng, D., Chen, M., Wang, Y., Cai, W., 2018. Multi-pass fast watershed for accurate segmentation of overlapping cervical cells. IEEE Trans. Med. Imaging 37 (9), 2044–2059. http://dx.doi.org/10.1109/TMI.2018.2815013.

Tran, K.B., Lang, J.J., Compton, K., Xu, R., Acheson, A.R., Et. al, H.J.H., 2022. The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 400 (10352), 563–591. http://dx.doi.org/10.1016/s0140-6736(22)01438-6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Veta, M., van Diest, P.J., Kornegoor, R., Huisman, A., Viergever, M.A., Pluim, J.P.W., 2013. Automatic nuclei segmentation in H&E stained breast cancer histopathology images. PLOS ONE 8 (7), null. http://dx.doi.org/10.1371/journal.pone.0070221.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Weigert, M., Schmidt, U., 2022. Nuclei instance segmentation and classification in histopathology images with stardist. In: 2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC). pp. 1–4. http://dx.doi.org/10.1109/ISBIC56247.2022.9854534.

Wienert, S., Heim, D., Saeger, K., Stenzinger, A., Beil, M., Hufnagl, P., Dietel, M., Denkert, C., Klauschen, F., 2012. Detection and segmentation of cell nuclei in virtual microscopy images: A minimum-model approach. Sci. Rep. 2 (1), http://dx.doi.org/10.1038/srep00503.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Yang, X., Li, H., Zhou, X., 2006. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy. IEEE Trans. Circuits Syst. I. Regul. Pap. 53 (11), 2405–2414. http://dx.doi.org/10.1109/TCSI.2006.884469.

Zhang, Z., Lu, X., Cao, G., Yang, Y., Jiao, L., Liu, F., 2021. ViT-YOLO:Transformer-based YOLO for object detection. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops. ICCVW, pp. 2799–2808. http://dx.doi.org/10.1109/ICCVW54120.2021.00314.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890. http://dx.doi.org/10.1109/CVPR46437.2021.00681.

Fabian Hörst is an electrical engineer with a master's in information engineering. He completed his master's in 2022 at the Institute for Artificial Intelligence in Medicine (Essen, Germany) and subsequently joined the group as a Ph.D. student. In his research, he works in medical imaging, especially computational pathology.

Moritz Rempe graduated from the Technical University in Dortmund (Germany) with a master's degree in medical physics. In his master's thesis he worked with complex-valued neural networks to exploit the raw data domain of magnetic resonance imaging data, which he now proceeds in his doctoral research at the Institute for Artificial Intelligence in Medicine (Essen, Germany).

Lukas Heine graduated at the Friedrich-Alexander-University of Erlangen–Nuremberg (Erlangen, Germany) with a degree in medical physics and is currently a Ph.D. student at the Institute for Artificial Intelligence in Medicine (Essen, Germany). His research focuses on clinical data integration, technical infrastructure to deploy deep learning algorithms to the hospital, and computer vision for radiological imaging data.

Constantin Seibold graduated in 2019 with a master's degree in computer science from the Karlsruhe Institute of Technology (Karlsruhe, Germany). Afterward, he pursued a Ph.D. with a focus on medical image analysis in low supervision settings in Computer Science within the Helmholtz Information and Data Science School for Health at the Karlsruhe Institute of Technology, which he completed with distinction in 2023. He is now working as Junior Group Leader in the Clinic for Nuclear Medicine at the University Hospital Essen, focusing on artificial intelligence in medicine.

Julius Keyl After completing his medical studies at the LMU in Munich, Julius Keyl began his professional career as a resident in oncology and is currently a resident in pathology at the University Hospital Essen, Germany. His research focuses on using machine learning to develop personalized treatment strategies for cancer patients.

Giulia Baldini is a computer scientist with a bachelor's degree from the Free University of Bozen-Bolzano (Italy) and a master's degree from the University of Bonn (Germany). Currently, she is working as a Ph.D. student at the Institute for Artificial Intelligence in Medicine (Essen, Germany) and works on classification and segmentation projects in the field of medical imaging, with a focus on integration in the clinical routine.

Selma Ugurel studied medicine at the Heinrich-Heine University Düsseldorf and RWTH Aachen University (Germany). Subsequently, she completed her training as a dermatologist at the Clinic for Dermatology, Venereology, and Allergology at Saarland University (Homburg/Saar, Germany), followed by her habilitation (venia legendi). Since 2014, she has been a senior physician at the Department of Dermatology, Venereology, and Allergology at the University Hospital Essen, Germany. Her research in the field of dermato-oncology and the relevance of prognostic markers for malignant melanomas has received numerous awards.

Jens Siveke is scientific and deputy director of the West German Tumor Center (WTZ) of the University Hospital Essen and director of the Bridge Institute for Experimental Tumor Therapy at the WTZ (Essen, Germany). He is the spokesperson of the German Cancer Consortium (DKTK) partner site Essen/Düsseldorf and heads the German Cancer Research Center (DKFZ) Division of Solid Tumor Translational Oncology. He is board-certified in internal medicine and gastroenterology. The major focus of his research and clinical activities is in gastrointestinal cancers, especially pancreatic and biliary malignancies, and in personalized medicine. He is Principal Investigator of multiple trials.

Barbara Grünwald studied molecular biotechnology in Munich and obtained a Ph.D. in experimental medicine in 2016. After postdoctoral training at the Princess Margaret Cancer Center in Toronto, she recently started her own research group at the University Hospital Essen. Her research focuses on tissue self-organization in solid tumors with a specific focus on how malignant functions are implemented across multicellular interactions.

Jan Egger has over 10 years of experience in developing algorithms in medical image analysis and image-guided therapies, including 5 years of postdoctoral research and development in the healthcare industry. Until now, he has published over 150 peer-reviewed papers, edited several books, and won numerous awards for theses and publications. He holds a Ph.D. and a (German) Habilitation in computer science, and an interdisciplinary Ph.D. in human biology. Currently, Jan has a dual appointment at the Institute of Computer Graphics and Vision (ICG) of the Graz University of Technology (TU Graz) and the Institute for Artificial Intelligence in Medicine (IKIM) at the Essen University Hospital (UKE) in Germany, where he is a Professor and a Privatdozent for AI in Medicine.

Jens Kleesiek studied medicine in Heidelberg and bioinformatics in Hamburg, where he obtained a Ph.D. in computer science in 2012. After training at the University Hospital Heidelberg and the German Cancer Research Center (DKFZ), he received his board certification (Facharzt) in radiology, his habilitation (venia legendi), and specialization (Zusatzweiterbildung) in medical informatics. His research focuses on applying self-supervised and weakly supervised learning paradigms to recognize clinically relevant patterns in large and complex data and integrating multimodal data sources to enhance the decision-making process at the point of care.