Uni-Hema: Unified Model for Digital Hematopathology

Supplementary Material

1. Overview

This supplementary material provides detailed information referenced in the main paper, presenting a comprehensive view of our experimental results and performance of our unified approach. It includes complete descriptions of all datasets used in our paper (Supplementary Section 2), detailing segmentation, detection, classification, and multimodal datasets. The training configurations and loss functions discussed in Section 4.2 of the main paper are presented in Supplementary Section 3, enabling reproducibility. Extended comparisons for detection, segmentation, and classification, originally mentioned in Section 4.3, are provided in Supplementary Section 4. Ablation studies with task-specific analyses (Supplementary Section 5) explore our proposed approach, while qualitative evaluations across detection and segmentation tasks are included in Supplementary Section 6, highlighting strengths, limitations, and insights beyond the main paper.

Table 1. Segmentation datasets: A total of approximately 222k images are aggregated from multiple sources, with around 221k samples utilized for Uni-Hema training and evaluation, and the remainder reserved for validation or excluded due to annotation limitations.

Sr	Dataset	# Images	# Classes	Train	Test
1	Malaria-Detection-2019 [1]	883	2	706	177
2	NuClick [23]	1,470	5	1220	250
3	KRD-WBC [3]	600	5	480	120
4	WBC Image Dataset [51]	400	5	300	–
5	White Blood Cell dataset [29]	367	3	367	–
6	ErythrocytesIDB [14]	50	3	40	10
7	AneRBC-II-Anemic [38]	6,000	5	5,000	1,000
8	AneRBC-II-Healthy [38]	6,000	5	5,000	1,000
9	MP-IDB [25]	281	5	140	–
10	Elsafty-RBC for AI [11]	204,510	9	156,408	48,102
11	BBBC041Seg [9]	1,560	1	1248	312

2. Datasets

We utilize a total of 46 datasets with different annotations, including publicly available and curated datasets, spanning all tasks (cell detection, cell segmentation, cell morphology prediction, cell classification, visual text completion, and visual question answering) in our unified framework. This includes 11 segmentation datasets with a total of 222,121 images (Table 1), 16 object detection datasets comprising 84, 207 images (Table 2), 16 single-cell classification datasets containing 381,931 images (Table 4), and in house cureted two vision languages dataset for the VQA and MLM task containing 27,884 image with QA/ML text pair. Together, these resources amount to nearly 0.7 million images, making our work one of the most extensive in digital hematopathology (per our knowledge). In addition, we incorporate semi-synthetically curated vision–language datasets, including 6,940 masked language modeling (MLM) samples and 21,887 VQA pairs, enabling enriched multimodal pretraining and contextual reasoning.

Table 2. Detection datasets: Approximately 84k images are collected from different datasets. About 66 k images are used for Uni-Hema training and testing, while the remaining images are part of validation or not used for testing or excluded due to annotation errors.

Sr	Dataset	# Images	# Classes	Train	Test
1	LeukemiaAttri [35]	28,000	14	16,770	7260
2	M5 [43]	7,500	4	4980	2250
3	TXL_PBC [13]	1,500	3	1008	144
4	BCCD [41]	364	3	225	36
5	Sickle-cell [46]	413	1	339	74
6	Plasmodium [34]	2,703	4	2418	–
7	Plasmodium_Phonecamera	1,185	1	948	–
8	Tuberculosis_phonecamera	2,503	1	1218	–
9	Vivax [17]	1,328	4	1208	120
10	ThickBloodSmears [48]	1,830	1	1830	–
11	NIH-NLM-Thick PV [21]	3,013	1	3,013	–
12	Parasite[10]	265	2	225	40
13	Acevedo [2]	10,435	6	10,435	-
14	MP-IDB [25]	281	4	–	106
15	Bio-Net [39]	2,005	4	–	401
16	Raabin-M1 [24]	17,769	7	14215	3554
17	Raabin-M2 [24]	3,114	7	2491	623

Table 3. Single-cell classification datasets: In total, about 382k images are compiled across diverse datasets, of which nearly 365k are employed for Uni-Hema model development, while the rest are allocated for validation or omitted due to quality constraints.

Sr	Dataset	# Images	# Classes	Train	Test
1	BMC [28]	171,373	21	137098	34275
2	AML Matek [27]	18,365	15	18365	–
3	Raabin WBC [24]	14,514	5	10175	4339
4	Warty pig [4]	2,871	5	1408	–
5	LISC	2,263	5	241	–
6	KRD-WBC [3]	600	5	480	–
7	BCCD	364	3	225	–
8	HRLS [6]	16,027	9	16,027	–
9	APL_AML [40]	25,915	20	14250	–
10	White-Blood-Cell-dataset [6]	376	1	376	–
11	Acevedo [2]	4446	2	3510	936
12	RV_PBS [33]	752	10	601	151
13	PBC-8-DA [33]	13,042	8	10433	2609
14	C-NMC2019 [30]	12,528	2	10661	1867
15	BloodMNIST [49]	17,092	8	11959	3421
16	AML Hehr [16]	81,403	5	–	–

Table 4. Visual question answering and visual mask language modeling dataset: total images 27,884

Sr	Dataset	Task	Images	Train	Test
1	WBCAtt-VQA	VQA	20,944	17,859	3,085
2	LeukemiaAttri-MLM	MLM	6,940	5,552	1,388

3. Implementation details

Table 5 reports the details of the full configuration for all six training stages, including hyperparameters and the model components trained at each stage.

Table 5. Comprehensive configuration settings for the six-step training procedure. The table expands the implementation details discussed in Section 4.2 and includes all hyperparameters used at each stage.

Parameter	Step-1	Step-2	Step-3	Step-4	Step-5	Step-6
num_classes_detection	0	0	30	30	0	0
num_classes_classification	45	0	0	0	0	0
num_classes_segmentation	0	0	2	2	2	0
lr	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
lr_backbone	1e-5	0	1e-5	0	0	0
batch_size	32	48	6	4	16	16
weight_decay	0.01	0.01	0.01	0.01	0.01	0.01
epochs	24	50	12	12	8	8
lr_drop	22	-	10	10	-	-
backbone_training	True	False	True	False	False	False
Image_Encoder_training	False	False	True	False	False	False
Image_Decoder_training	False	False	True	True	False	False
Text_Encoder_training	False	True	False	False	False	True
Text_Decoder_training	False	True	False	False	False	True
CMF_training	False	False	False	False	False	True
TGVR_training	False	False	True	True	False	False
QGMF_training	False	False	True	False	True	False
SCFE_training	True	False	True	False	False	False
position_embedding	sine	-	sine	sine	sine	-
hidden_dim_text	384	384	384	384	384	384
hidden_dim_vision	256	256	256	256	256	384

3.1. Multi-Task loss formulation

We optimize our unified model using a composite multi-task loss that combines contributions from cell classification, segmentation, cell detection, morphology (multi-label) prediction, and text tasks. Formally, the total loss is expressed as:

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}_{cls}^{cell}}_{\text{Cell Classification}} + \underbrace{\mathcal{L}_{seg}}_{\text{Segmentation}} + \underbrace{\lambda_{det}\,\mathcal{L}_{det}}_{\text{Cell Detection}} + \underbrace{\mathcal{L}_{morph}}_{\text{Morphology (Multi-label)}} + \underbrace{\mathcal{L}_{text}}_{\text{Text / T5}} \quad (1)$$

where:

$$\mathcal{L}_{cls}^{cell} = \mathrm{CE}(\hat{y}_{image}, y_{image}) \quad (2)$$

$$\mathcal{L}_{seg} = \mathcal{L}_{dice}(\hat{M}, M) + \mathcal{L}_{focal}(\hat{M}, M) \quad (3)$$

$$\mathcal{L}_{det} = \mathcal{L}_{set\_cls} + \mathcal{L}_{set\_bbox} + \mathcal{L}_{giou} \quad (4)$$

$$\mathcal{L}_{morph} = \mathrm{AsymmetricLoss}(\hat{y}_{morph}, y_{morph}) \quad (5)$$

$$\mathcal{L}_{text} = \mathrm{CE}(\hat{y}_{text}, y_{text}) \quad (6)$$

For the cell detection loss, we follow the standard λ_{det} weighting scheme used in DINO [50], ensuring consistency with prior transformer-based detectors.

4. Extended results

This section presents extended experimental results for all tasks discussed in the main paper. In the supplementary Tables, the Blue entries indicate baseline results that are also reported in the main paper.

Detailed single cell morphological results: In the main paper, we reported only the overall mean F1-Micro score for the single-cell morphology task. Here in Table 6, we provide a detailed breakdown of the F1-Macro scores for each of the 11 morphological attributes. We compare our model against four strong baselines: DINOv2 [32] (a non-medical vision foundation model), CONCH [26] (pathology-specialized foundation model), DINO-Bloom [22] (hematology foundation model), and ResNet-50 [15] pretrained on ImageNet. Across the 11 attributes, our model achieves the best performance on five attributes and ranks second on four others, demonstrating strong and consistent performance in fine-grained morphological reasoning.

Detailed morphological results for multi-cell scenario: In the main paper (Table 1), we reported only the mean F1 score for the Field-of-View (FoV) morphology prediction task. In this Supplementary Material Table 7, we provide a detailed comparison of the F1 scores for all six morphology labels for the H_100x_C2 [35] dataset. We compare our method against AttriDet [35] and CBM [**?**], two strong task-specific baselines. Our model outperforms both AttriDet and CBM in five of the six morphology categories, achieving a clear performance margin in most attributes. These detailed scores further validate the robustness of our approach, multi-cell morphological predication in microscopy images

Comparison with SOTA object detectors: Expanded detection results comparing our method against additional state-of-the-art detectors. In the main paper, we report on performance against YOLO [20] and DINO [50]. Here, in Supplementary Table 8, we include comparisons with Faster R-CNN [36], FCOS [45], and Sparse R-CNN [44].

Table 6. Morphology results of individual attributes of a single cell image.

Model	Dinov2s	Conch	Dino-bloom	ResNet	Ours
Attribute	**F1 Macro**				
Cell Shape	81.4	85.7	87.6	88.9	**90.8**
Cell Size	79.3	81.3	79.8	**84.4**	**84.4**
Chromatin Density	86.3	**87.0**	85.8	85.7	85.1
Cytoplasm Colour	82.7	86.7	**90.3**	88.4	88.4
Cytoplasm Texture	89.7	93.8	93.6	**93.9**	92.7
Cytoplasm Vacuole	80.8	87.4	83.8	88.9	**90.2**
Granularity	99.7	**99.7**	99.5	99.6	**99.7**
Granule Colour	98.1	**98.9**	98.7	98.8	**98.9**
Granule Type	99.0	99.4	99.1	**99.5**	99.4
N-C Ratio	94.1	**96.9**	95.1	96.3	96.8
Nucleus Shape	59.0	67.7	59.8	**79.1**	78.4

Table 7. Morphology results of individual attributes of the cell from the Complete field of view image

Model	AttriDet	CBM	Ours
Attribute	**F1 Mean**		
N-C Ratio	73.9	21.9	**86.4**
Nucleus Shape	95.9	96.2	**97.4**
Nucleous	54.3	41.8	**73.1**
Cytoplasm	89.7	77.2	**95.9**
Cytoplasmic basophil	83.6	70.2	**90.9**
Cytoplasm Vacuole	**29.1**	3.33	11.0

Methods marked with * indicate results that were not computed for certain datasets. All values correspond to $\text{mAP}_{@50}$. Note that our method is trained on a unified dataset covering all tasks, whereas the other models are trained individually on each dataset. The results demonstrate that our unified approach achieves competitive or superior detection performance across multiple datasets.

Extended segmentation results: Expanded segmentation results comparing our method against additional state-of-the-art models. In the main paper, we presented the performance of our method against U-Net [37] and TransNetR [19]. Here, in Supplementary Table 9, we included the NanoNet [18] and R2U-Net [5] for a more comprehensive comparison. All values correspond to Dice scores. Our method demonstrates competitive performance with the state-of-the-art models, while being trained on a unified dataset covering all tasks, whereas the other methods are trained individually on each dataset.

Extended cell classification results: Single-cell classification results on unseen (Acevedo [2], Medmnist [49], C-NMC [30], and RV-PBC-8 [8]) and seen (Raabin [24] and BMC [28]) datasets. In Supplementary Table 10, we report F1-scores for multiple small and large variants of non-medical vision foundation models (DINOv2 [32], DINOv3 [42]), pathology-specific models (Phikon-v2 [12], UNI [7], CONCH [26], TransPath [47]), and hematology-focused models (DinoBloom [22]). Across all datasets, the mean performance of our unified model is superior to all compared baselines, including both medical and non-medical foundation models, demonstrating strong generalization to diverse cell types and domains.

VQA and V-MLM tasks performance: The table 11 presents the results of our model on the VQA and VLM tasks using standard evaluation metrics, including BLEU-1 to BLEU-4 and ROUGE-1, ROUGE-2, and ROUGE-L scores. These extended results demonstrate that our unified model achieves reasonable and consistent performance across both tasks, complementing the main paper's findings

Fine-tuning results: The table 12 presents the fine-tuning results ($mAP_{@50}$) in the Raabin dataset [24] for the detection task. Our method, initialized with our own pretrained weights, achieves higher performance compared to the DINO [50] detector pretrained weights, demonstrating stronger transferability and improved detection capability.

5. Ablation

In the ablation studies, we analyze the impact of key design choices across tasks.

Effect of different up-sampling techniques on segmentation: Table 13 compares simple resizing (bilinear interpolation) with our learnable two-layer convolutional layer up-sampler. In the baseline, segmentation is performed using plain interpolation during both training and testing. We then introduce a lightweight up-sampler composed of two convolutional layers (with interpolation) and fine-tune only this module while freezing the rest of the model. The learnable up-sampler consistently yields higher Dice scores than simple resizing, demonstrating its effectiveness in improving segmentation accuracy.

Accuracy with and without prompt: Table 14 presents an ablation study evaluating the effect of disease-prompt guidance on detection and segmentation. Incorporating disease prompts consistently improves localization accuracy and mask quality across datasets, whereas removing the prompts leads to noticeable drops in precision.

Effect of different level features for classification: Ablation study on feature selection for single-cell classification. In Table 15, we compare the performance of the features from the backbone only, encoder only, and a combined backbone+encoder representation for the single classification task. The combined features achieve the best overall performance, demonstrating the benefit of multi-level feature fusion.

Cross model fusion (CMF) ablation: Backbone features $\mathcal{E}_{\mathbf{B}}^{\mathbf{I}}$ preserve dense spatial and global scene information, making them more suitable for cross-modal fusion, while encoder features are object-centric (as discussed in Section 3.2 of [50]). Since encoder embeddings are learned with strong object-level supervision, they are less effective for global alignment. An ablation study confirms this: using backbone features $\mathcal{E}_{\mathbf{B}}^{\mathbf{I}}$ achieves a higher BLEU-4 score (56.4) than encoder features $\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$ (51.0) against the WBCAtt-VQA dataset.

Table 8. Cell detection results comparison with state-of-the-art object detectors

Dataset	Sparse R-CNN	FCOS	Faster R-CNN	DINO	YOLO	Ours	Ours (mean)
HCM_40x_C2 [35]	32.7	32.7	36.9	36.9	37.3	**43.6**	43.1 ± 0.70
HCM_100x_C2[35]	36.7	40.6	41.2	43.7	44.2	**49.8**	47.1 ± 2.16
LCM_40x_C2[35]	33.9	28.5	36.3	36.6	34.9	**40.7**	41.7 ± 1.85
LCM_100x_C2[35]	25.9	34.3	38.8	38.2	38.1	**45.6**	43.5 ± 2.20
HCM_1000x [43]	*	77.7	74.7	79.8	77.3	**83.1**	79.8 ± 4.63
LCM_1000x [43]	*	56.6	57.2	**64.2**	56.5	62.4	60.8 ± 0.64
HCM_400x [43]	*	65.3	60.7	**70.4**	66.9	69.0	68.3 ± 0.62
LCM_400x [43]	*	52.3	48.8	58.3	**59.9**	54.5	56.0 ± 2.91
Sickle Cell[46]	*	*	61.8	**73.6**	68.6	67.0	66.2 ± 0.66
Parasites[10]	*	*	18.8	38.6	**46.1**	36.2	36.4 ± 0.6
BCCD [41]	*	*	87.4	**89.5**	88.1	87.8	86.7 ± 1.76
TXL [13]	*	*	94.3	**95.3**	94.9	94.0	92.3 ± 0.47

Table 9. Comparison of segmentation results against state-of-the-art methods.

Model	AneRBC-Anemic	AneRBC-Healthy	Elsafty	IDB2	KRD	MD-2019
Unet [37]	78.3	75.1	93.4	91.5	93.3	75.2
R2U-Net [5]	71.2	72.3	99.2	-	91.6	69.6
Atten_Unet [31]	75.2	68.4	99.6	89.6	92.3	76.6
Nanonet [18]	91.2	90.9	98.3	33.1	86.7	84.7
TransNetR[19]	**93.6**	**95.2**	99.5	**92.1**	**94.9**	**86.7**
Ours	93.4	94.1	**99.9**	90.5	94.5	77.2
Ours (mean)	93.4 ± 0.01	94.3 ± 0.34	99.9 ± 0.05	91.5 ± 1.05	94.7 ± 0.25	80.4 ± 2.85

Table 10. Comparison of single-cell classification performance with foundation models.

Method	Acevedo	Minisit	C-NMC	RV-PBS_8	Raabin	BMC	*mean*	# Params
Dinov2G[32]	96.78	97.75	69.17	91.17	95.35	71.04	86.88	1136M
DinoBloomG[22]	98.80	**98.92**	67.89	94.07	98.76	84.56	90.50	1136M
Dinov3h+ /16 [42]	98.19	98.42	71.40	90.25	96.22	73.41	87.98	840M
Dinov2L[32]	96.81	97.49	69.08	91.17	95.29	71.09	86.22	304M
DinoBloomL [22]	**98.89**	98.77	68.88	**94.35**	98.46	84.52	90.65	304M
Phikon-v2 [12]	96.90	98.39	**73.67**	90.17	94.23	72.11	97.58	300M
Dinovl/16 [42]	97.81	98.13	69.94	91.93	95.93	72.82	87.76	300M
UNI [7]	98.10	98.36	70.45	91.83	95.72	76.36	88.47	300M
CONCH [26]	95.35	97.10	66.03	91.20	94.37	71.56	85.93	200M
Dinov2B[32]	96.22	97.51	70.56	90.22	94.75	70.38	96.60	86M
DinoBloomB [22]	98.66	98.74	70.82	94.10	98.57	84.78	90.94	86M
Dinov3b/16 [42]	96.98	97.28	70.71	91.19	95.28	71.80	87.20	86M
TransPath [47]	94.00	95.67	64.77	88.06	91.75	59.15	82.23	55M
ResNet50 [15]	90.04	95.15	64.95	86.07	88.63	64.71	81.60	25M
Dinov3s+/16 [42]	94.88	97.45	69.57	89.65	93.34	66.89	85.30	29M
DinoBloomS [22]	98.16	98.77	69.60	92.80	98.03	84.95	90.38	22M
Dinov2S[32]	94.50	96.76	71.02	90.69	93.70	68.18	85.80	22M
Dinov3s/16 [42]	96.03	97.14	67.97	90.83	93.81	67.82	85.60	21M
Ours	98.14	98.64	72.83	93.61	**98.83**	**86.23**	**91.38**	31M
Ours (mean)	98.1 ± 0.36	98.18 ± 0.60	71.81 ± 1.18	92.57 ± 1.17	98.53 ± 0.25	86.10 ± 0.55	91.15 ± 0.48	31M

Table 11. Evaluation results for VQA and MLM tasks across multiple metrics, demonstrating the performance of our unified model.

Approach	Task	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Ours (best)	VQA	70.9	63.9	59.8	56.4	72.0	60.0	69.9
Ours (mean)	VQA	70.47 ± 0.51	62.60 ± 1.18	58.03 ± 1.57	54.30 ± 1.85	71.03 ± 0.85	58.23 ± 1.55	68.70 ± 1.04
Ours	V-MLM	92.1	87.8	83.8	79.8	91.7	82.0	91.2
Ours (mean)	V-MLM	94.13 ± 1.76	90.64 ± 2.46	87.44 ± 3.16	84.23 ± 3.84	93.40 ± 1.47	85.76 ± 3.26	93.10 ± 1.65

6. Qualitative results

The qualitative segmentation results are shown in Figure 1a, where our method demonstrates competitive performance with state-of-the-art approaches across diverse datasets. While overall localization is accurate, slight edge softness, common in transformer-based architectures, may appear when compared to TransNetR. Qualitative detection results in Figure 1b further show that our model performs well across multiple datasets and object types, consistently producing robust and reliable detections even under significant domain variation.

Table 12. Performance of Uni-Hema after fine-tuning the cell detection task.

Dataset	Dino	Ours
Raabin [24]	53.7	**55.9**

Table 13. The results demonstrate that including the CNN-based upsampler enhances segmentation accuracy.

Dataset	Resize	Up-sampler (Freeze)
Dice Score		
AneRBC-Anemic[38]	91.8	**93.4**
AneRBC-Healthy[38]	92.8	**94.1**
Elsafty [11]	99.2	**99.9**
IDB2[25]	78.8	**90.5**
KRD[3]	94.0	**94.5**
MD-2019[1]	72.8	**77.6**

Table 14. The results show that the use of disease-specific prompts positively influences the model's performance, enhancing both segmentation and detection accuracy.

Dataset	Without Prompt	With Prompt	*Disease*
Cell Detection (mAP_{50})			
H_40x_C2 [35]	40.3	**43.6**	Leukemia
H_100x_C2 [35]	46.2	**49.8**	Leukemia
L_40x_C2 [35]	41.0	**40.7**	Leukemia
L_100x_C2 [35]	44.3	**45.6**	Leukemia
H_1000x [43]	76.8	**83.1**	Malaria
L_1000x [43]	61.6	**62.4**	Malaria
H_400x [43]	66.0	**69.0**	Malaria
L_400x [43]	51.8	**54.5**	Malaria
Sickle Cell [46]	66.3	**67.0**	Sickle Cell
Parasites [10]	34.1	**36.2**	Parasites
BCCD [41]	77.7	**87.8**	Normal
TXL [13]	88.7	**94.0**	Normal
Unseen			
Bio-Net [39]	37.5	**54.7**	Normal
Malaria [25]	74.6	**78.5**	Malaria
Segmentation (Dice Score)			
AneRBC-Anemic[38]	92.1	**93.4**	Anemia
AneRBC-Healthy[38]	93.0	**94.1**	Normal
Elsafty [11]	99.2	**99.9**	Anemia
IDB2[25]	86.6	**90.5**	Leukemia
KRD[3]	94.3	**94.5**	Unknown
MD-2019[1]	**79.0**	77.6	Malaria
Unseen			
BBBC041Seg[9]	90.9	86.2	Normal

7. Medical impact and practical advantages

Our proposed method offers a unified paradigm capable of performing multiple hematology tasks, including detection, segmentation, morphology prediction, visual interaction, and classification, within a single framework. This unified approach allows the model to be further fine-tuned for additional hematology tasks as new datasets become available, providing flexibility and adaptability for evolving clinical needs. By consolidating multiple tasks into one model, our method reduces the overhead of installing, maintaining, and synchronizing several specialized models, which is particularly beneficial in resource-constrained and resource-efficient healthcare environments.

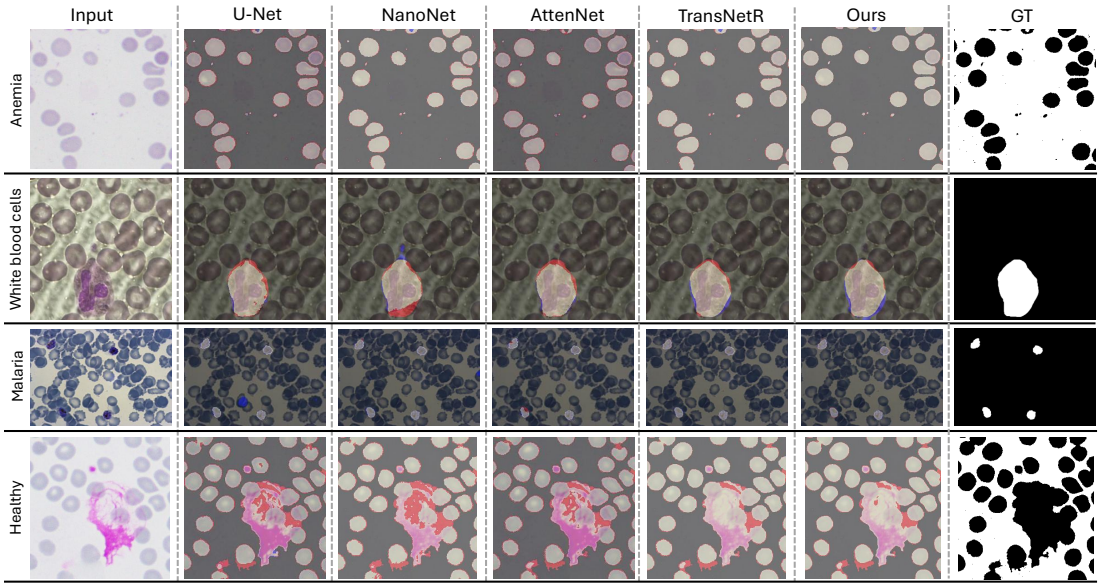
. Moreover, the dataset, code, and pretrained weights will be made publicly available upon acceptance, ensuring reproducibility and enabling the broader research and medical community to leverage and extend our work for practical clinical applications. This approach ultimately facilitates efficient, accurate analysis across diverse blood-related tasks, supporting faster diagnosis, better patient monitoring, and more accessible computational hematopathology.

Table 15. The results demonstrate how leveraging features at multiple hierarchical levels enhances the model's classification accuracy.

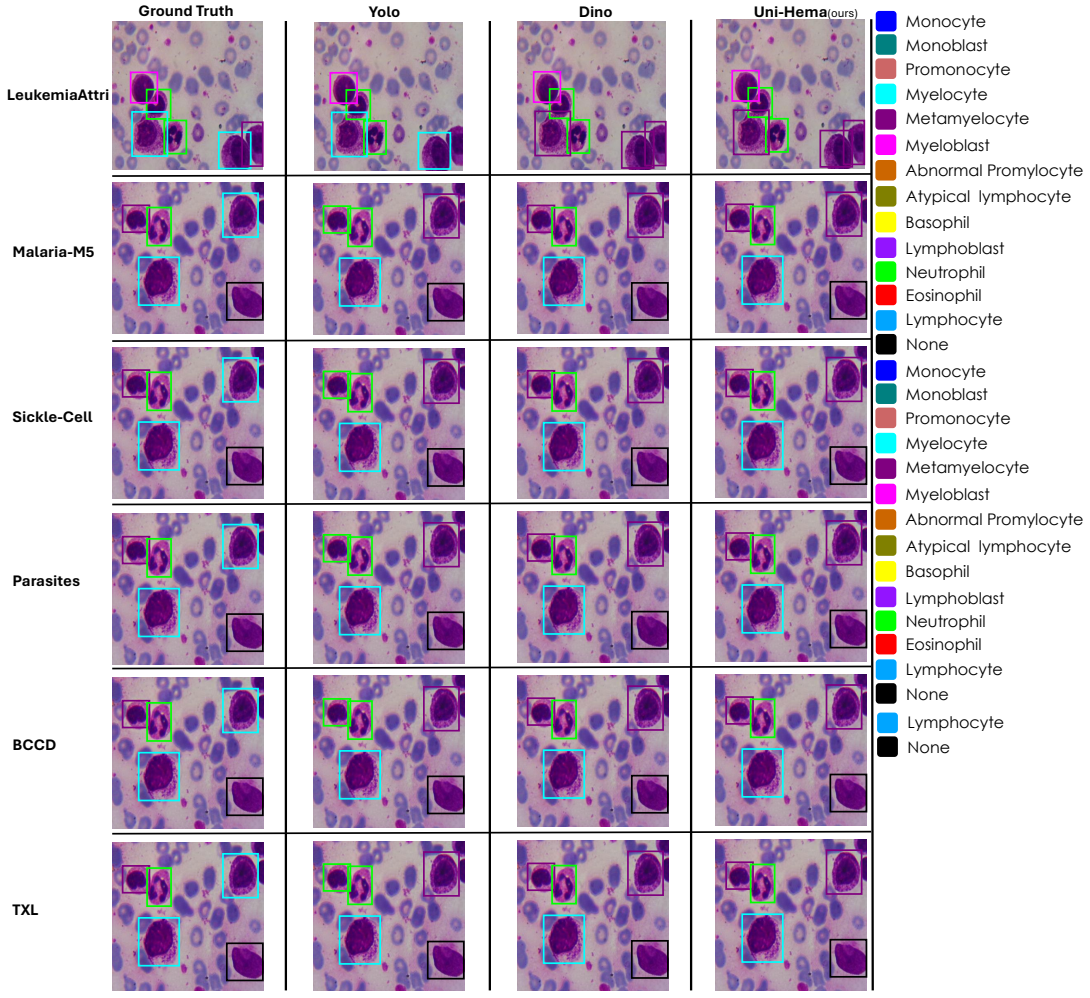
Dataset	Encoder	Backbone	Backbone+Encoder
Single Cell Classification (F1)			
Raabin [24]	98.8	90.8	**98.8**
BMC[28]	86.0	60.3	**86.2**
Unseen			
Acevedo [2]	97.9	93.0	**98.1**
Minisit [49]	98.2	96.9	**98.6**
C-NMC [30]	68.8	72.0	**72.8**
RV-PBS_8 [33]	93.6	88.7	**93.6**

8. Limitation and future work

Our work has a few limitations. First, we did not train or test the model for weakly labeled classification on microscopy field-of-view images. Second, the current datasets have cell-level annotations but lack captions for single cells or microscopy field-of-view images, which limits prompt-based detection and segmentation. In the future, we plan to address these issues by working on multi-cell images with weak labels, prompt-controlled detection and segmentation, and adding cell-level captions to improve multimodal learning. Given the resource constraints limitation, our approach is designed to be efficient, and with access to greater computational resources, we plan to further enhance Uni-Hema to improve multimodal functionality across digital hematopathology tasks.



(a) Qualitative segmentation results of U-Net, NanoNet, AttenNet, TransNetR, and our method on anemia, malaria, WBC, and healthy cell images. True Positives (TP), False Positives (FP), and False Negatives (FN) are highlighted in light yellow, blue, and red, respectively. Our method consistently reduces false detections and provides more accurate localization across all four datasets, demonstrating competitive and robust performance compared to existing state-of-the-art models.



(b) Dotted white boxes denote missed detections, while dotted colored boxes represent incorrect class predictions. Across diverse datasets, our single unified model achieves strong localization and class consistency, outperforming or matching the specialized detectors in challenging cases.

Figure 1. Qualitative results for segmentation (a) and detection (b) tasks across multiple datasets.

References

[1] Syed Saiden Abbas and Tjeerd MH Dijkstra. Detection and stage classification of plasmodium falciparum from images of giemsa stained thin blood films using random forest classifiers. *Diagnostic pathology*, 15(1):130, 2020.

[2] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30:105474, 2020.

[3] Haval Taha Ali, Fattah Alizadeh, and Nawsherwan Sadiq Mohammad. White blood cell microscopic image dataset for segmentation. *Available at SSRN 4617448*.

[4] JR Alipo-on, FI Escobar, JL Novia, MM Atienza, S Manay, MJ Tan, N AlDahoul, and E Yu. Dataset for machine learning-based classification of white blood cells of the juvenile visayan warty pig. 2022.

[5] Md Zahangir Alom, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net). In *NAECON 2018-IEEE National Aerospace and Electronics Conference*, pages 228–233. IEEE, 2018.

[6] Alexandra Bodzas, Pavel Kodytek, and Jan Zidek. A high-resolution large-scale dataset of pathological and normal white blood cells. *Scientific Data*, 10(1):466, 2023.

[7] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

[8] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022.

[9] Deponker Sarker Depto, Shazidur Rahman, Md Mekayel Hosen, Mst Shapna Akter, Tamanna Rahman Reme, Aimon Rahman, Hasib Zunair, M Sohel Rahman, and MRC Mahdy. Automatic segmentation of blood cells from microscopic slides: a comparative analysis. *Tissue and Cell*, 73: 101653, 2021.

[10] Eden. parasite detection dataset. `https://universe.roboflow.com/eden-lcx9y/parasite-detection`, 2024. visited on 2025-10-14.

[11] Ahmed Elsafty, Ahmed Soliman, and Yomna Ahmed. 1 million segmented red blood cells with 240 k classified in 9 shapes and 47 k patches of 25 manual blood smears. *Scientific Data*, 11(1):722, 2024.

[12] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

[13] Lu Gan and Xi Li. Txl-pbc: a freely accessible labeled peripheral blood cell dataset. *arXiv preprint arXiv:2407.13214*, 2024.

[14] Manuel Gonzalez-Hidalgo, FA Guerrero-Pena, Silena Herold-García, Antoni Jaume-i Capó, and Pedro D Marrero-Fernández. Red blood cell cluster separation from digital images for use in sickle cell disease. *IEEE journal of biomedical and health informatics*, 19(4):1514–1525, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Matthias Hehr, Ario Sadafi, Christian Matek, Peter Lienemann, Christian Pohlkamp, Torsten Haferlach, Karsten Spiekermann, and Carsten Marr. A morphological dataset of white blood cells from patients with four different genetic aml entities and non-malignant controls (aml-cytomorphology_mll_helmholtz). *(No Title)*, 2023.

[17] Jane Hung and Anne Carpenter. Applying faster r-cnn for object detection on malaria images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 56–61, 2017.

[18] Debesh Jha, Nikhil Kumar Tomar, Sharib Ali, Michael A Riegler, Håvard D Johansen, Dag Johansen, Thomas de Lange, and Pål Halvorsen. Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 37–43. IEEE, 2021.

[19] Debesh Jha, Nikhil Kumar Tomar, Vanshali Sharma, and Ulas Bagci. Transnetr: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In *Medical Imaging with Deep Learning*, pages 1372–1384. PMLR, 2024.

[20] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - yolov5 sota realtime instance segmentation, 2022.

[21] Yasmin M Kassim, Feng Yang, Hang Yu, Richard J Maude, and Stefan Jaeger. Diagnosing malaria patients with plasmodium falciparum and vivax using deep learning for thick smear images. *Diagnostics*, 11(11):1994, 2021.

[22] Valentin Koch, Sophia J Wagner, Salome Kazeminia, Ece Sancar, Matthias Hehr, Julia A Schnabel, Tingying Peng, and Carsten Marr. Dinobloom: a foundation model for generalizable cell embeddings in hematology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 520–530. Springer, 2024.

[23] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: a deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020.

[24] Zahra Mousavi Kouzehkanan, Sepehr Saghari, Eslam Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmaeil Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, et al. Raabin-wbc: a large free access dataset of white blood cells from normal peripheral blood. *bioRxiv*, pages 2021–05, 2021.

[25] Andrea Loddo, Cecilia Di Ruberto, Michel Kocher, and Guy Prod'Hom. Mp-idb: the malaria parasite image database for image processing and analysis. In *Sipaim–Miccai Biomedical Workshop*, pages 57–65. Springer, 2018.

[26] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023.

[27] Christian Matek, Simone Schwarz, Carsten Marr, and Karsten Spiekermann. A single-cell morphological dataset of leukocytes from aml patients and non-malignant controls (aml-cytomorphology_lmu). *The Cancer Imaging Archive (TCIA)[Internet]*, 2019.

[28] Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology*, 138(20):1917–1927, 2021.

[29] Mostafa Mohamed and Behrouz Far. An enhanced threshold based technique for white blood cells nuclei automatic segmentation. In *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 202–207. IEEE, 2012.

[30] S Mourya, S Kant, P Kumar, A Gupta, and R Gupta. All challenge dataset of isbi 2019 (c-nmc 2019)(version 1)[dataset]. the cancer imaging archive, 2019.

[31] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[33] Jimut Bahan Pal, Aniket Bhattacharyea, Debasis Banerjee, and Br Tamal Maharaj. Advancing instance segmentation and wbc classification in peripheral blood smear through domain adaptation: A study on pbc and the novel rv-pbs datasets. *Expert Systems with Applications*, 249:123660, 2024.

[34] John A Quinn, Alfred Andama, Ian Munabi, and Fred N Kiwanuka. Automated blood smear analysis for mobile malaria diagnosis. *Mobile point-of-care monitors and diagnostic device design*, page 115, 2018.

[35] Abdul Rehman, Talha Meraj, Aiman Mahmood Minhas, Ayisha Imran, Mohsen Ali, and Waqas Sultani. A large-scale multi domain leukemia dataset for the white blood cells detection with morphological attributes for explainability. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer, 2024.

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] Muhammad Shahzad, Syed Hamad Shirazi, Muhammad Yaqoob, Zakir Khan, Assad Rasheed, Israr Ahmed Sheikh, Asad Hayat, and Huiyu Zhou. Anerbc dataset: a benchmark dataset for computer-aided anemia diagnosis using rbc images. *Database*, 2024:baae120, 2024.

[39] Usman Ali Shams, Isma Javed, Muhammad Fizan, Aqib Raza Shah, Ghulam Mustafa, Muhammad Zubair, Yehia Massoud, Muhammad Qasim Mehmood, and Muhammad Asif Naveed. Bio-net dataset: Ai-based diagnostic solutions using peripheral blood smear images. *Blood Cells, Molecules, and Diseases*, 105:102823, 2024.

[40] Eugene Shenderov. Acute promyelocytic leukemia (apl). Kaggle Dataset, 2019. Accessed: Aug. 6, 2025.

[41] shenggan, Nicolas Chen, cosmicad, and akshaylamba. Bccd: Blood cell count and detection, 2018.

[42] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[43] Waqas Sultani, Wajahat Nawaz, Syed Javed, Muhammad Sohail Danish, Asma Saadia, and Mohsen Ali. Towards low-cost and efficient malaria detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20655–20664. IEEE, 2022.

[44] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[45] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[46] Florence Tushabe, Samuel Mwesige, Vicent Kasule, Emily Nsiimire, Sarah C Musani, David Areu, and Emmanuel Othieno. An image-based sickle cell detection method. *Authorea Preprints*, 2024.

[47] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021.

[48] Feng Yang, Mahdieh Poostchi, Hang Yu, Zhou Zhou, Kamolrat Silamut, Jian Yu, Richard J Maude, Stefan Jaeger, and Sameer Antani. Deep learning for smartphone-based malaria parasite detection in thick blood smears. *IEEE journal of biomedical and health informatics*, 24(5):1427–1438, 2019.

[49] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d

biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[50] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[51] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018.