NEURIPS 2026

Beyond ViT Tokens: Masked-Diffusion Pretrained Convolutional Pathology Foundation Model

for Cell-Level Dense Prediction

Weiming Chen*, Xitong Ling*, Zhenyang Cai, Xidong Wang,
Jiawen Li, Tian Guan, Benyou Wang†, Yonghong He†

Tsinghua Shenzhen International Graduate School, Tsinghua University
The Chinese University of Hong Kong, Shenzhen

* Equal contribution † Corresponding authors

Dense prediction needs cell-level detail

- **Motivation**

  ViT patch tokenization can obscure nuclear boundaries, chromatin texture, and narrow gaps between adjacent cells.

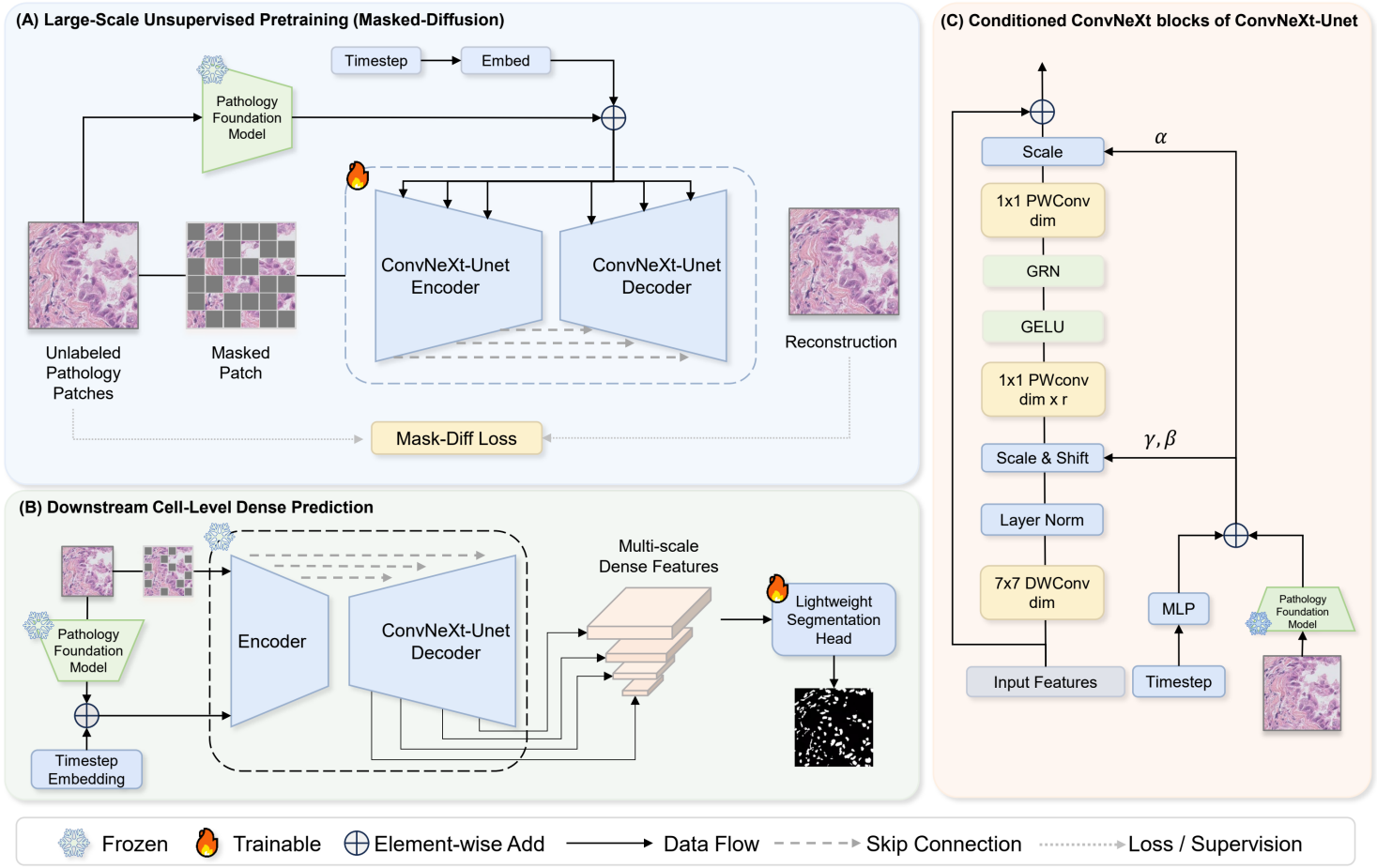
- **Approach**

  Pretrain a convolutional backbone with masked diffusion on unlabeled pathology patches, then freeze it for downstream segmentation.

- **Evaluation**

  Compare nuclei segmentation on CPM-15, CPM-17 and TNBC, including full-data and few-shot settings.

Masked-diffusion pretraining



Pretraining

1,048,349 unlabeled HistAI patches

Mask 8 × 8 tiles and reconstruct 256 × 256 RGB

Timestep and frozen H0-mini features condition the ConvNeXt U-Net

Source: poster Figure 1. The backbone produces multi-scale features for downstream prediction.

Frozen backbone, trainable segmentation heads

CMD-L backbone

602M frozen parameters

Multi-scale decoder features

256 × 256 input

Two head designs

A Top-down fusion

7.15M trainable parameters

Project, upsample and combine scales

B Concatenate + convolution

21.21M trainable parameters

Resize and concatenate features, then decode

Both heads use cross-entropy and Dice losses.

Source: poster Figure 4 and Table 4.

Full-data segmentation

CMD-L with frozen backbone

Mean Dice / boundary F1

Segmentation head	CPM-15	CPM-17	TNBC
Top-down fusion (7.15M)	0.864 / 0.911	0.865 / 0.875	0.854 / 0.874
Concat + conv (21.21M)	0.889 / 0.947	0.880 / 0.866	0.854 / 0.838

Best BF1 in the poster comparison

0.947 CPM-15 0.875 CPM-17 0.874 TNBC

Metrics are means. Head choice varies by dataset; the full paper reports 95% confidence intervals.
Input sizes differ across compared model families. Source: poster Table 4.

Few-shot segmentation

One labeled example

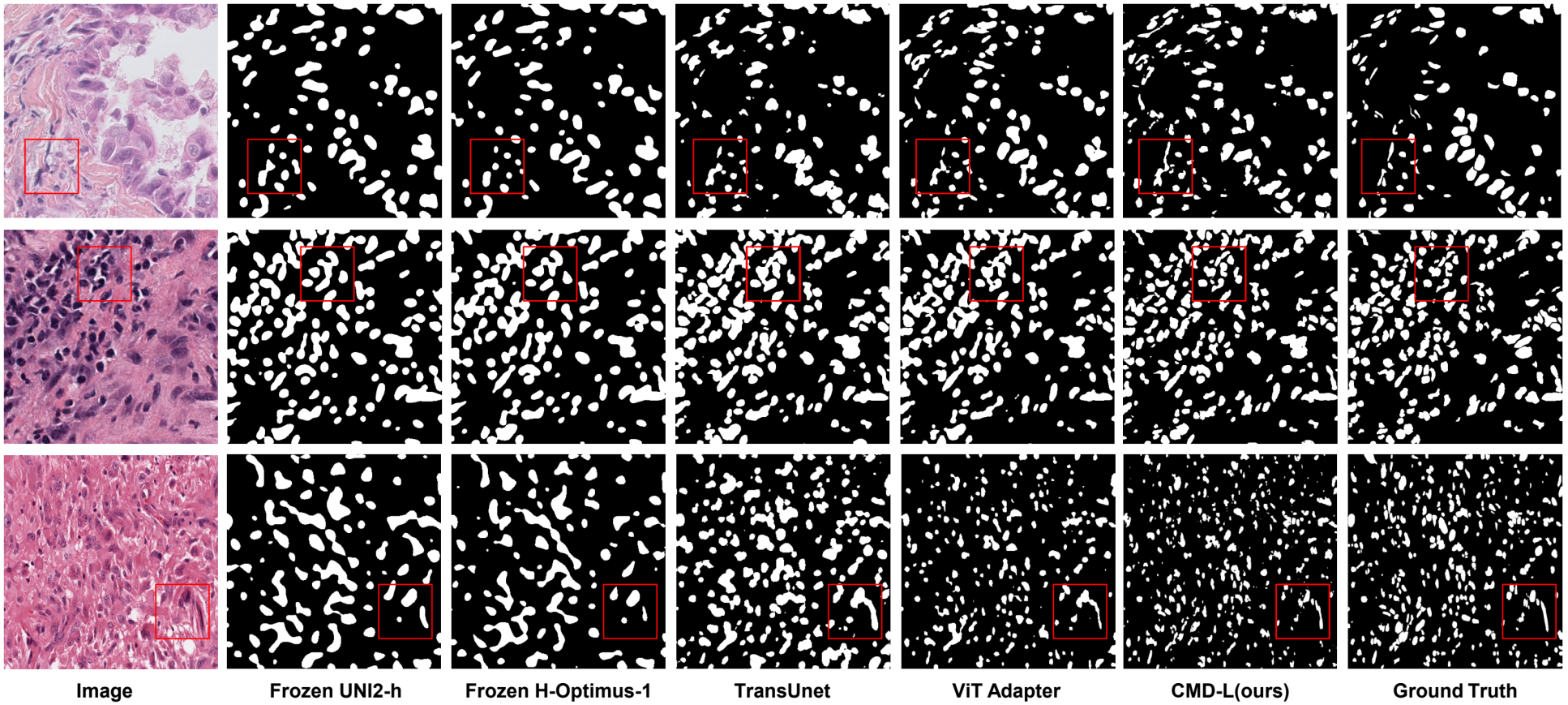
Method	CPM-17 Dice	TNBC Dice	Trainable
CMD-L + fusion head	**0.769**	**0.714**	7M
ViT-Adapter	0.714	0.498	364M

CMD-L uses a frozen 602M-parameter backbone and a trainable head.

At 5 shots, ViT-Adapter has slightly higher Dice on both datasets.

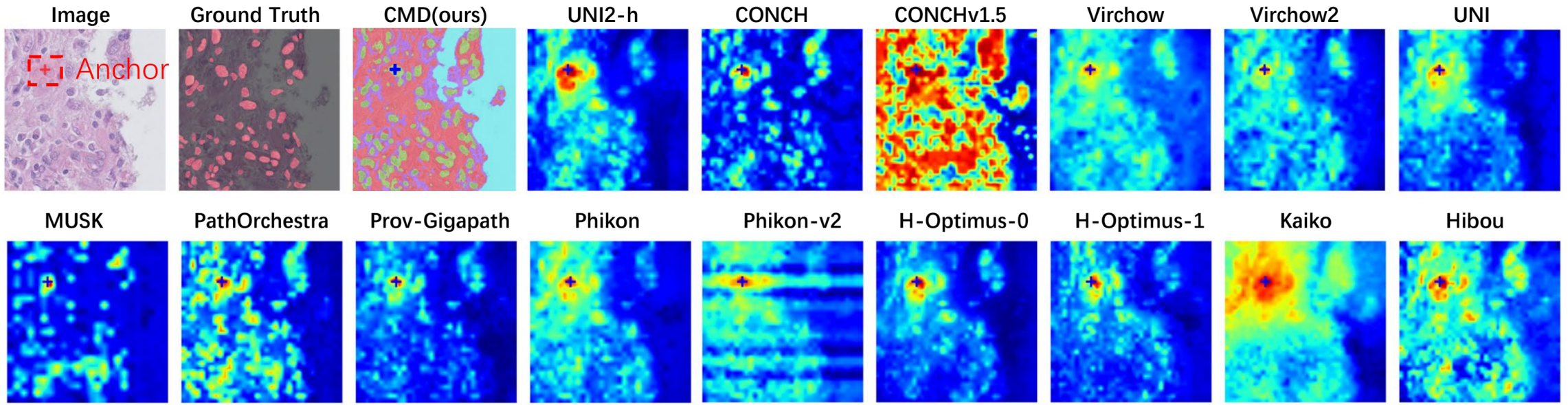
One-shot Dice, 95% CIs: CMD-L CPM-17 [0.696, 0.832], TNBC [0.618, 0.817];
ViT-Adapter CPM-17 [0.683, 0.749], TNBC [0.470, 0.529]. Source: poster Table 5.

Nuclei segmentation in crowded tissue



Rows: TNBC, CPM-17 and CPM-15. Red boxes mark crowded or ambiguous regions.
CMD-L better separates adjacent nuclei and preserves boundaries in these examples.

Cell-aware dense features



CMD: K-means clusters of dense convolutional features (K = 4).

ViT maps: cosine similarity to a nucleus anchor token.

The CMD clusters follow local nuclear morphology more closely in this example. Source: poster Figure 3.

Takeaways

- **Dense convolutional features**

  Masked-diffusion pretraining provides multi-scale features for cell-level segmentation.

- **Boundary quality**

  CMD-L achieves the best reported BF1 on all three full-data benchmarks, using different segmentation heads across datasets.

- **Label efficiency**

  With one labeled example, CMD-L outperforms the listed ViT-Adapter results on CPM-17 and TNBC in Dice.

Code: github.com/M4A1TasteGood/ConvNeXt-Masked-Diffusion