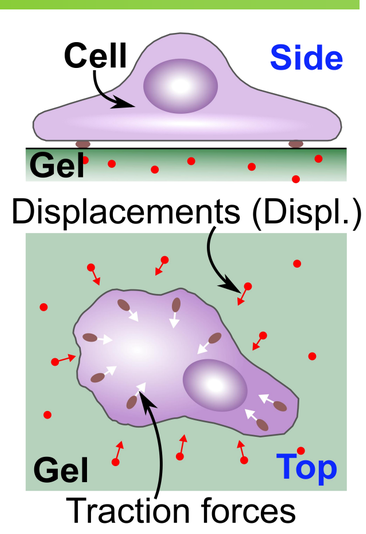
Combined Information for Traction Force Reconstruction Using Metadata with a Vision Transformer–Convolutional Network

Yunfei Huang[1], Elena Van der Vorst[2,3], Benedikt Sabass[1,2,3*]

1, Fakultät Physik, Technische Universität Dortmund, Dortmund, 44227, Germany
2, Faculty of Physics and Center for Nano Science, Ludwig-Maximilians-Universität München, Munich, 80752, Germany
3, Department of Veterinary Sciences, Ludwig-Maximilians-Universität München, Munich, 80752, Germany
*, benedikt.sabass@tu-dortmund.de

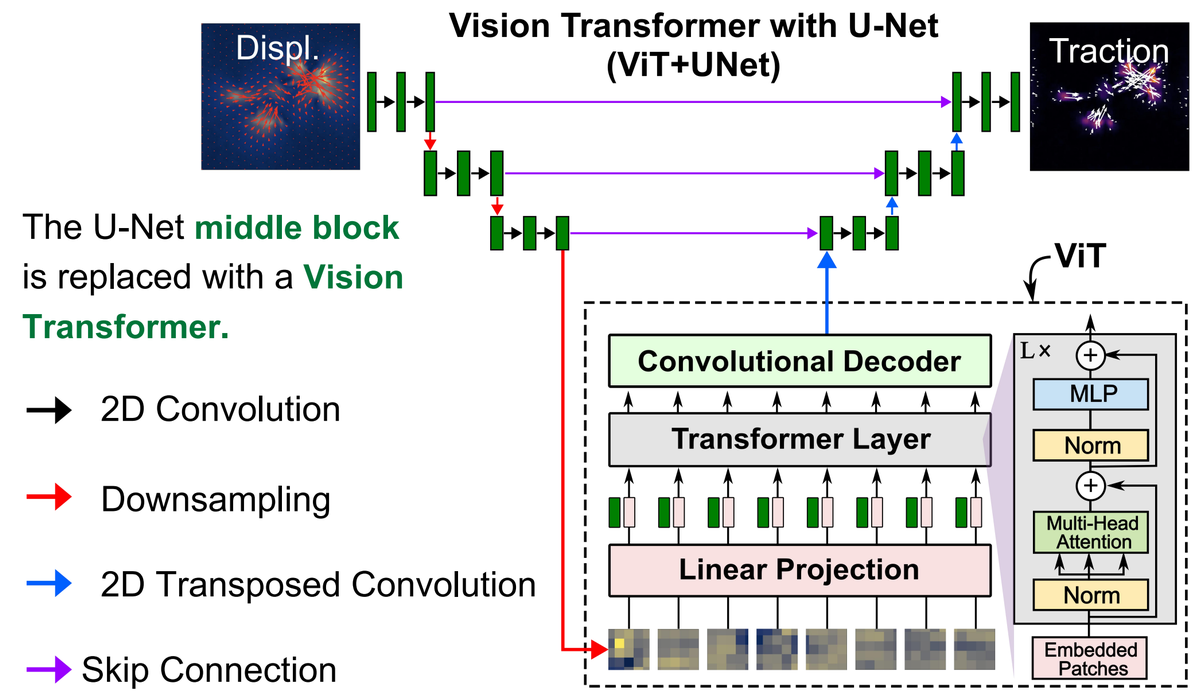
Introduction

- **Traction force microscopy (TFM)** infers cellular forces on a substrate from measured displacements.
- **TFM via deep learning** (DL) has become a powerful approach.
- We aim to develop robust DL models for **accurate** traction force inference and generalization, incorporating **metadata** input.



Hybrid Architecture: ViT+UNet

U-Net captures local patterns. **ViT** learns global context.



The U-Net **middle block** is replaced with a **Vision Transformer.**

→ 2D Convolution

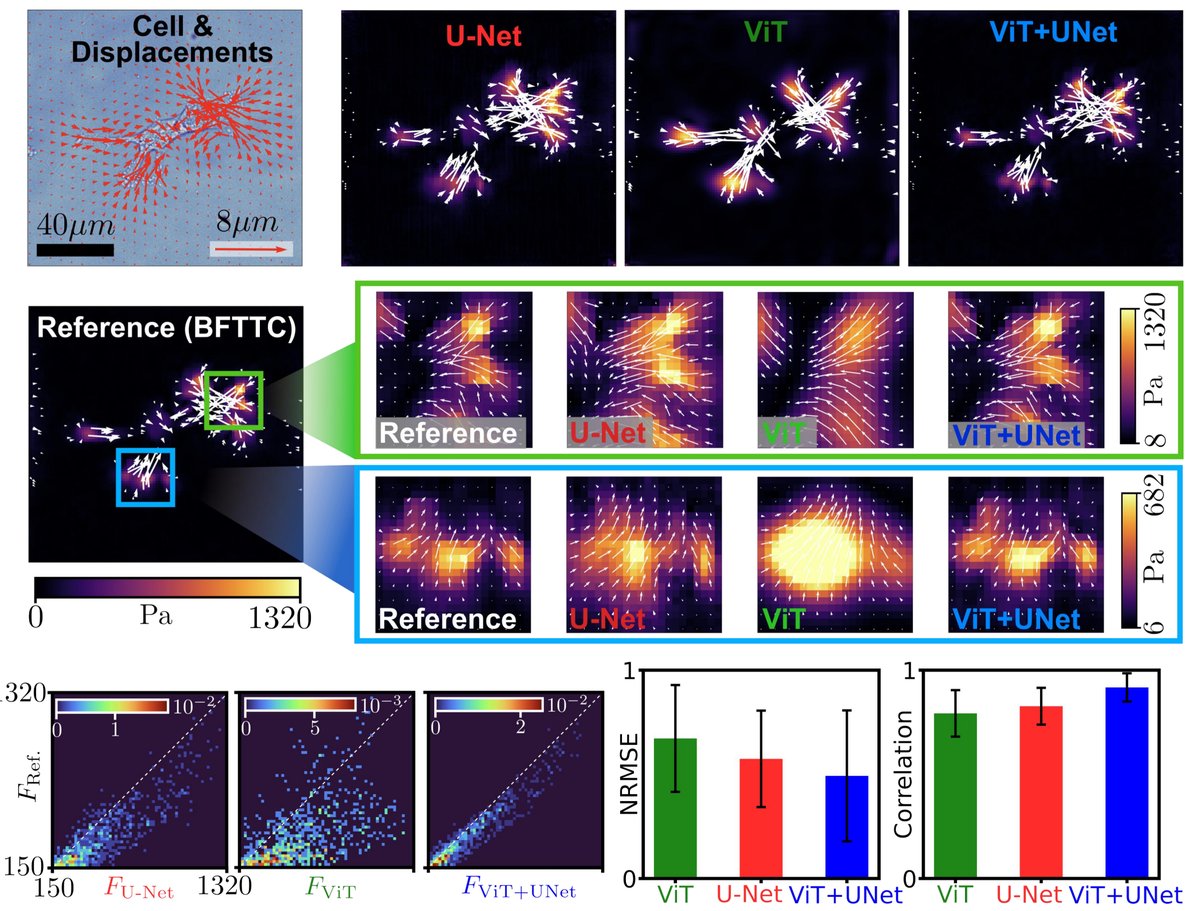
→ Downsampling

→ 2D Transposed Convolution

→ Skip Connection

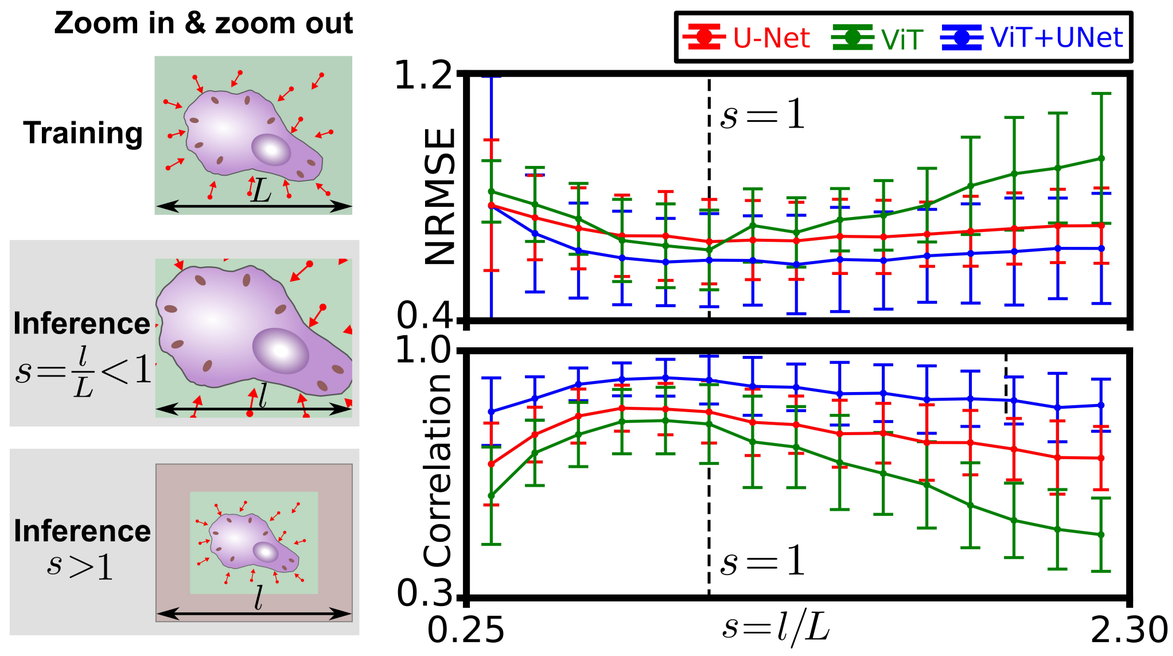
Results

1. ViT+UNet Outperforms U-Net and ViT Models.



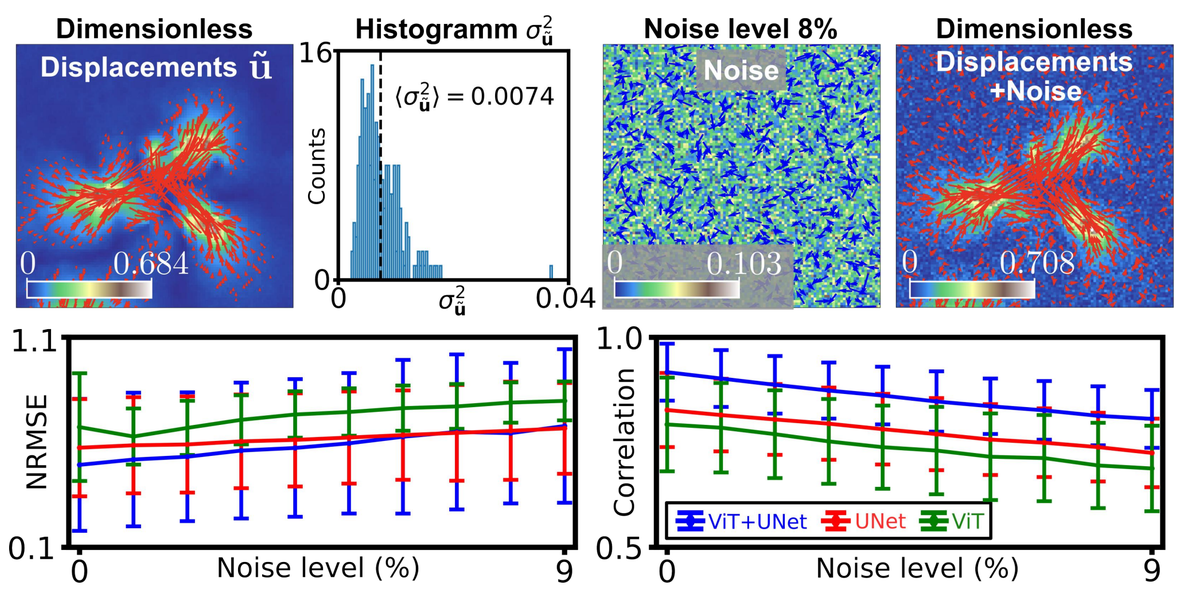
- **ViT+UNet performs better than U-Net and ViT in individual cases.**
- **ViT+UNet also surpasses both models in statistical metrics.**

2. ViT+UNet Shows Superior Generalization Across Spatial Scales.



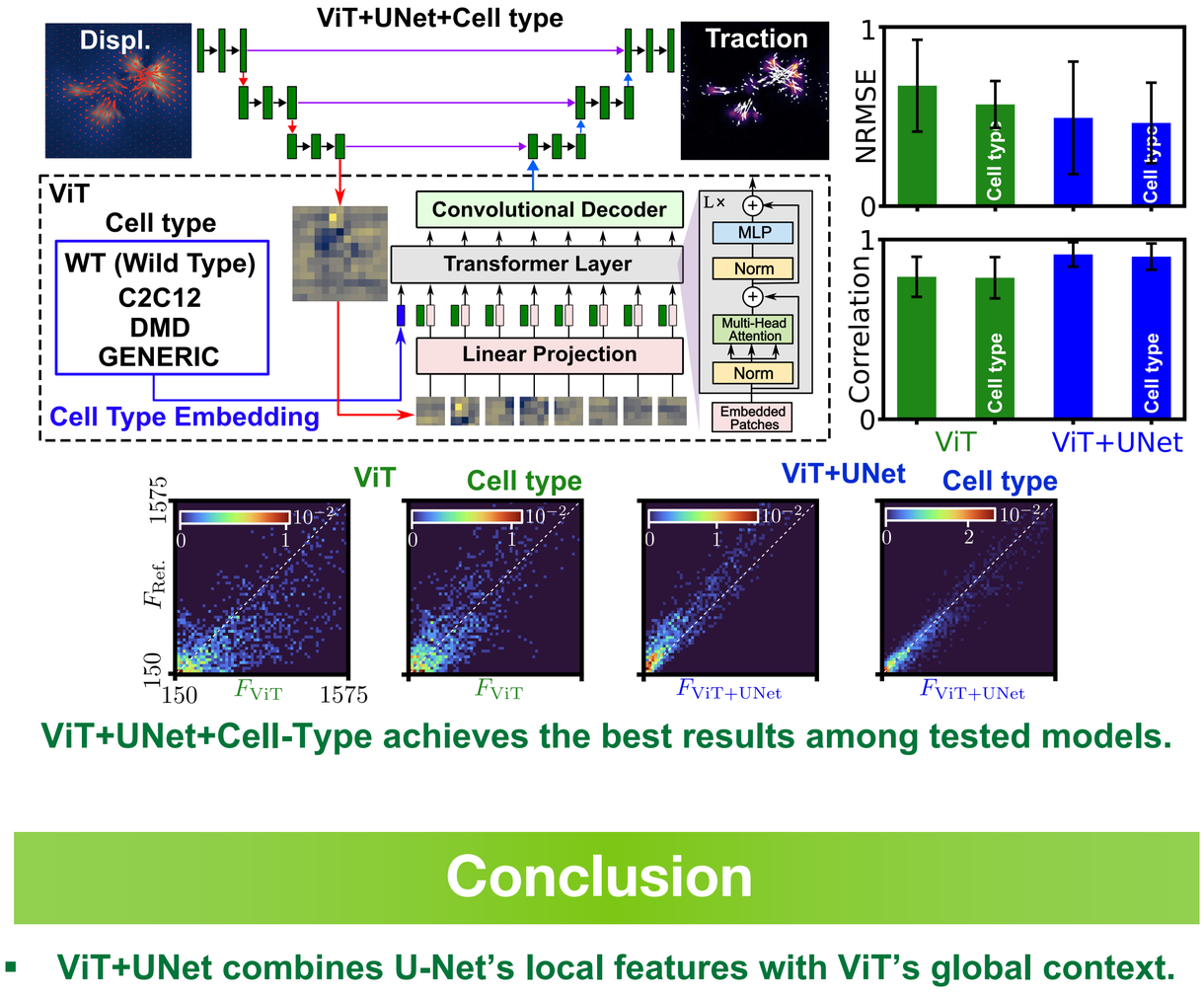
ViT+UNet surpasses U-Net and ViT across spatial scales 0.25–2.3.

3. ViT+UNet Demonstrates Robust Prediction Across Noise Levels.



ViT+UNet outperforms U-Net and ViT across noise levels 0–9%.

4. Cell-Type Input Enhances Prediction Specificity and Accuracy.



ViT+UNet+Cell-Type achieves the best results among tested models.

Conclusion

- **ViT+UNet combines U-Net's local features with ViT's global context.**
- **ViT+UNet achieves robust, accurate inference across spatial scales and noise levels.**
- **Cell-type input significantly improves accuracy and specificity.**