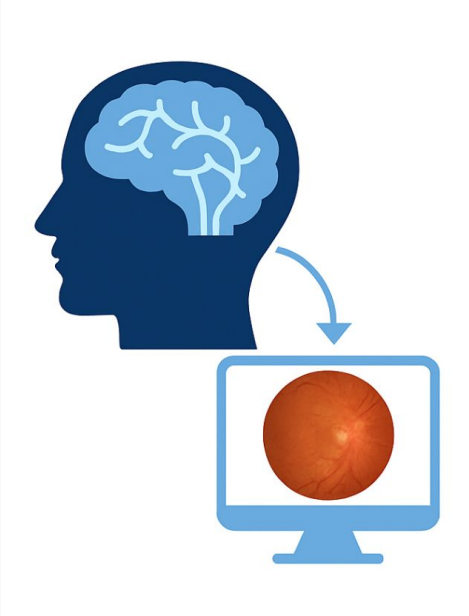
BlendED

Retinal Vessel Segmentation with Masked Auto Encoders

Team 1: Sin Jie Chia, Haruki Kinoshita, Daniel Liang, Yasuaki Matsuda, Minji Ryu, William Wang, Takumi Yamashita

Table of Contents

Research Background

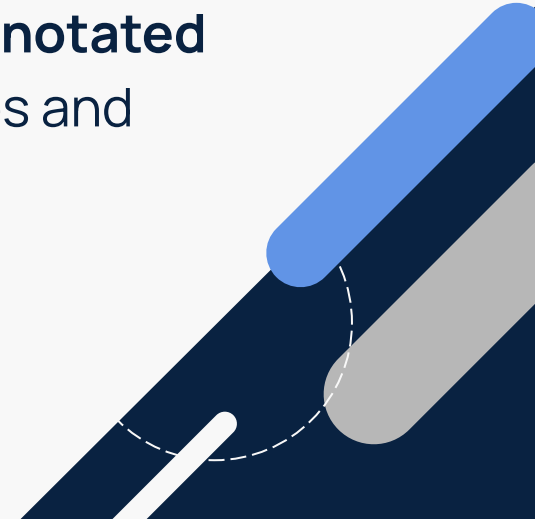


- Medical value
- AI progress in medical imaging
- CNNs excel in segmentation
- Data scarcity challenge
- U-Net strengths and limits
- Pretraining gap
- Unlabeled images

Research Objectives

1. Study and apply **U-Net** for medical image segmentation and **MAE** for self-supervised learning

2. Pretrain the U-Net encoder with MAE on **unannotated** retinal images to capture fine-grained features and improve accuracy



Research Process

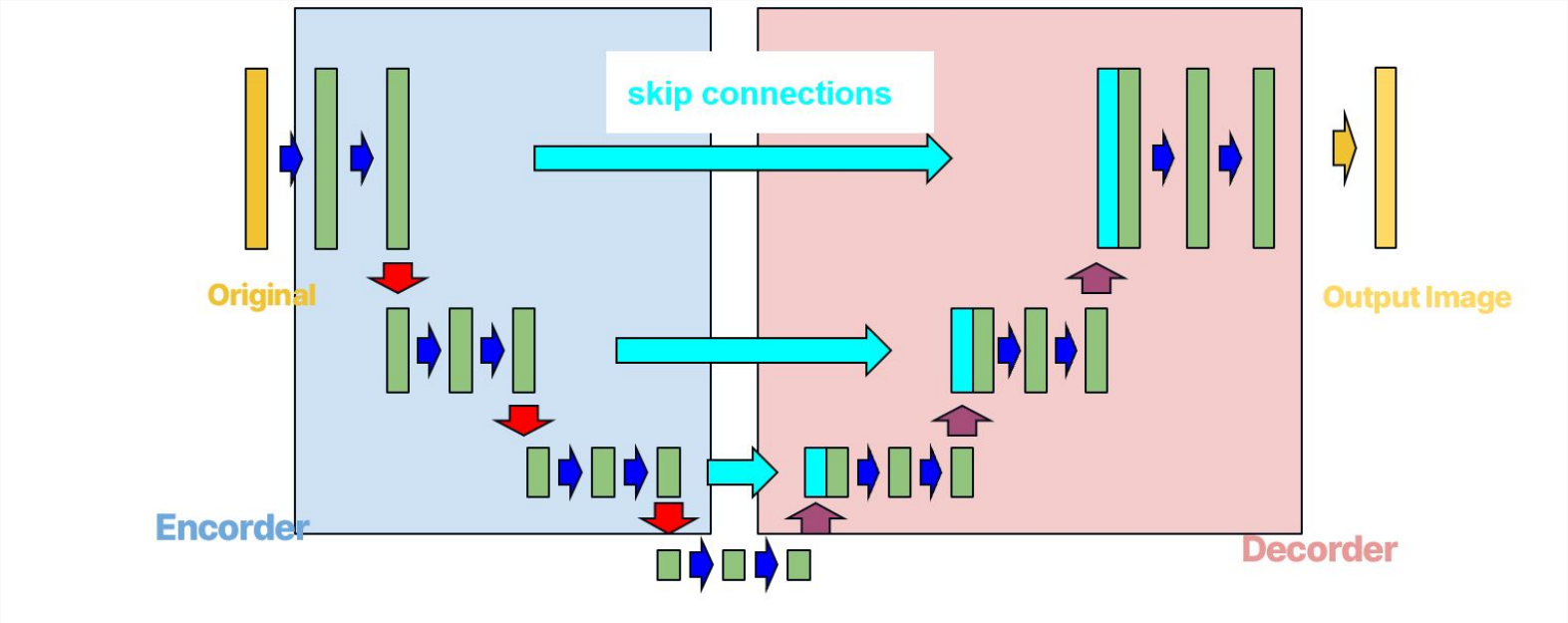
UNet Result

MAE + UNet Result

RETfound + Unet Result

What is UNet?

- Convolutional neural network architecture for medical image segmentation
- Encoder-decoder structure and skip connections.
- Shaped like the letter "U"



Supervised Learning

- *Data Preparation*

- **RGB to Grayscale**: Convert 3-channel color images to single-channel.
- **Dataset Standardization**: Perform Z-score standardization on the entire dataset.
- **CLAHE Enhancement**: Contrast Limited Adaptive Histogram Equalization.
- **Gamma Correction**: Apply γ=1.2 correction to enhance dark details.
- **Normalization**: Scale pixel values to the 0-1 range.

.hdf5

Supervised Learning - *Training*

Patch Extraction

Random Extraction Strategy: Randomly extract 48×48 small patches from 565×565 complete images.
Sample Quantity: 190,000 training samples.
Data Augmentation: Increase data diversity through random position extraction.

Test Data Patching:
Sliding Window Strategy: Extract overlapping 48×48 patches with a step size of 5×5.
Boundary Processing: Automatically expand the image size to ensure it is divisible by the patch size.
Overlapping Prediction: Average the overlapping areas to improve prediction quality.

Network Architecture Construction

Skip Connections: Fuse features of the encoder and decoder through concatenate operations.

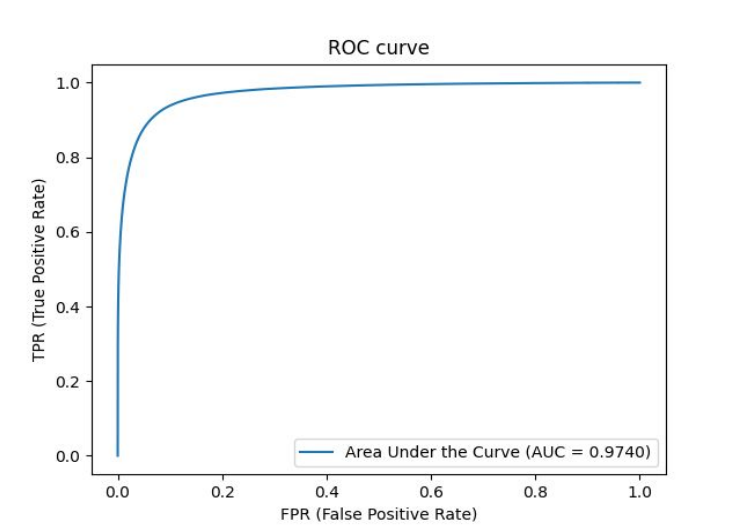
Regularization: Add 20% Dropout after each layer to prevent overfitting.

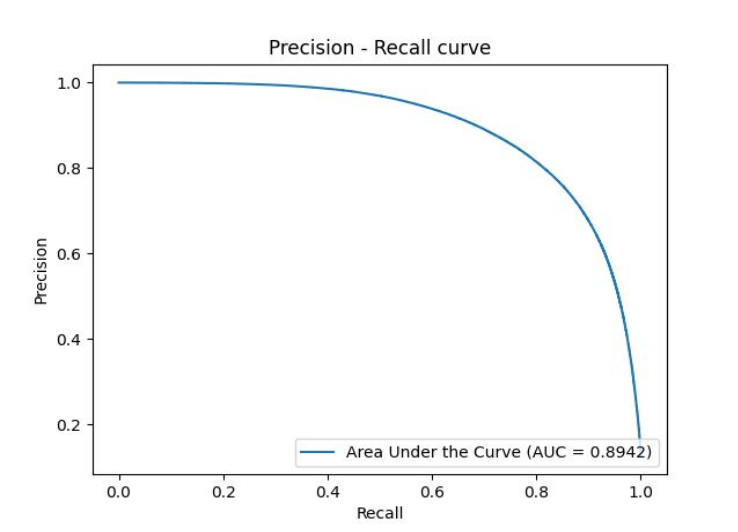
Optimizer Settings

- # of Epochs: 20.
- Batch Size: 32
- Validation Split: 10%.
- Randomly shuffle training data in each epoch.

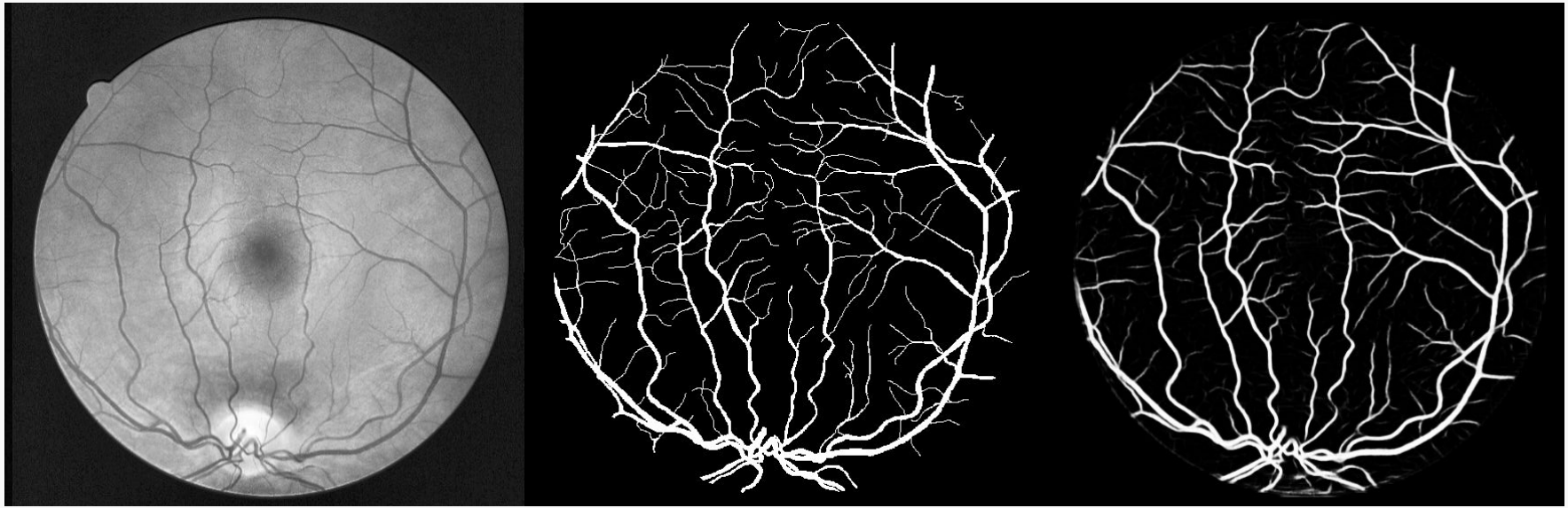
Supervised Learning *- Evaluation*

- Load the trained model for prediction
- Process test images using a sliding window strategy
- Calculate various evaluation metrics (Dice coefficient, ROC curve, precision-recall, etc.)
- Generate visualization images of prediction results





Result of UNet



Original

Ground Truth

Predicted

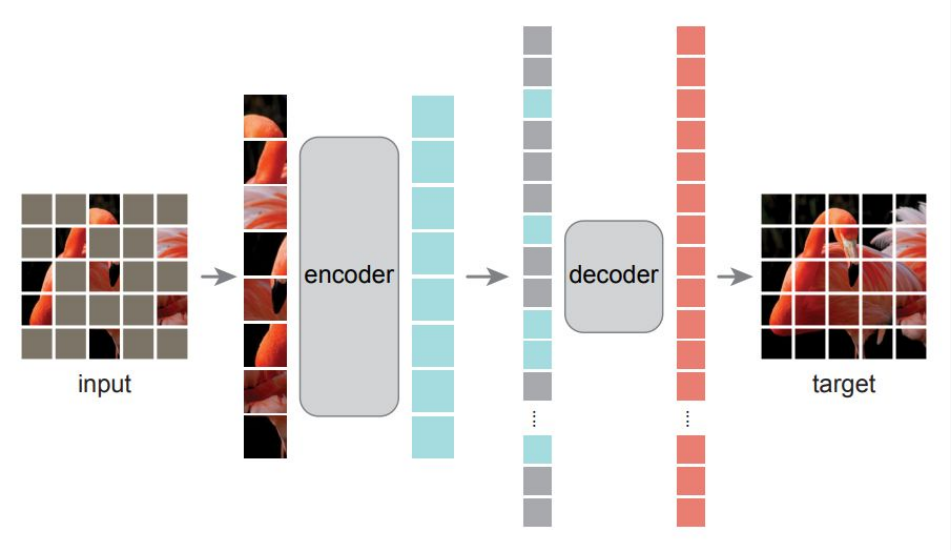
Masked Autoencoders (MAE)

Core idea: Mask 75% of an image → train model to **reconstruct hidden parts** from visible ones

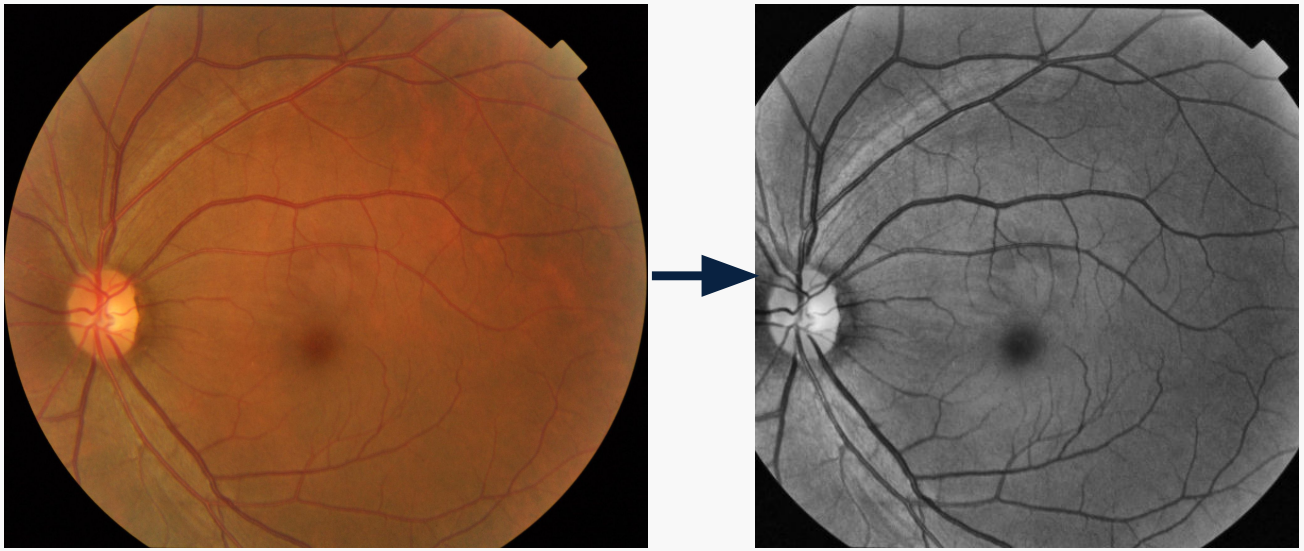
Architecture: Asymmetric autoencoder

- Encoder: Processes visible patches → latent space
- Decoder: Rebuilds full image using latent + mask tokens

Performance: ViT-Huge + MAE → 87.8% top-1 accuracy on ImageNet-1K (self-supervised pre training)



Pre-Processing Stage

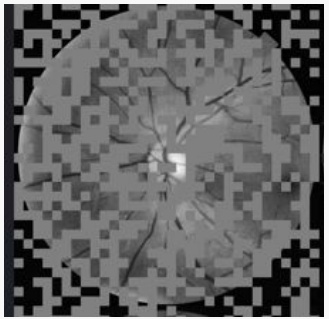


- Resize to 512x512
- Center square crop
- Green channel only
- Contrast Limited Adaptive Histogram Equalization.
- Left-right flip to double data

Our Architecture

Input

512x512 image
4 × 4 patches
50% masked



MAE Encoder

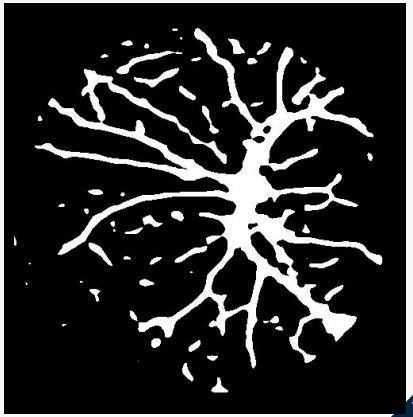
Feature map of 128 × 128 each cell = 192-number vector (embedding)

UNet decoder

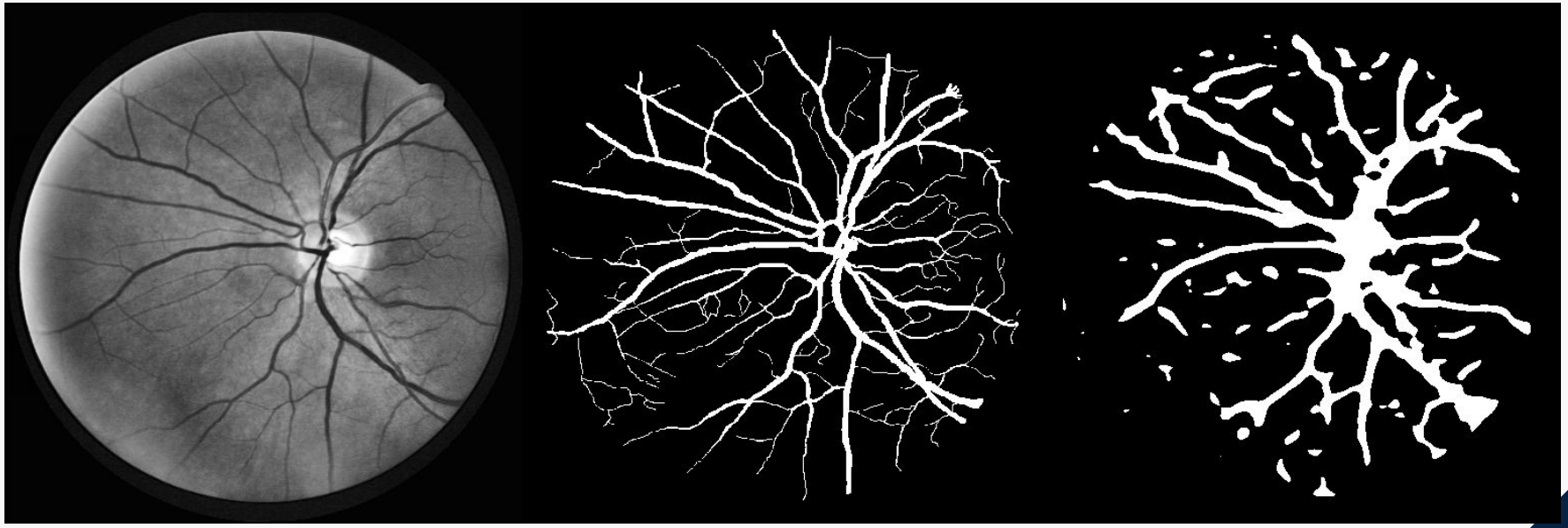
Upsampling

Skip connections

Mask



Result of MAE + UNet

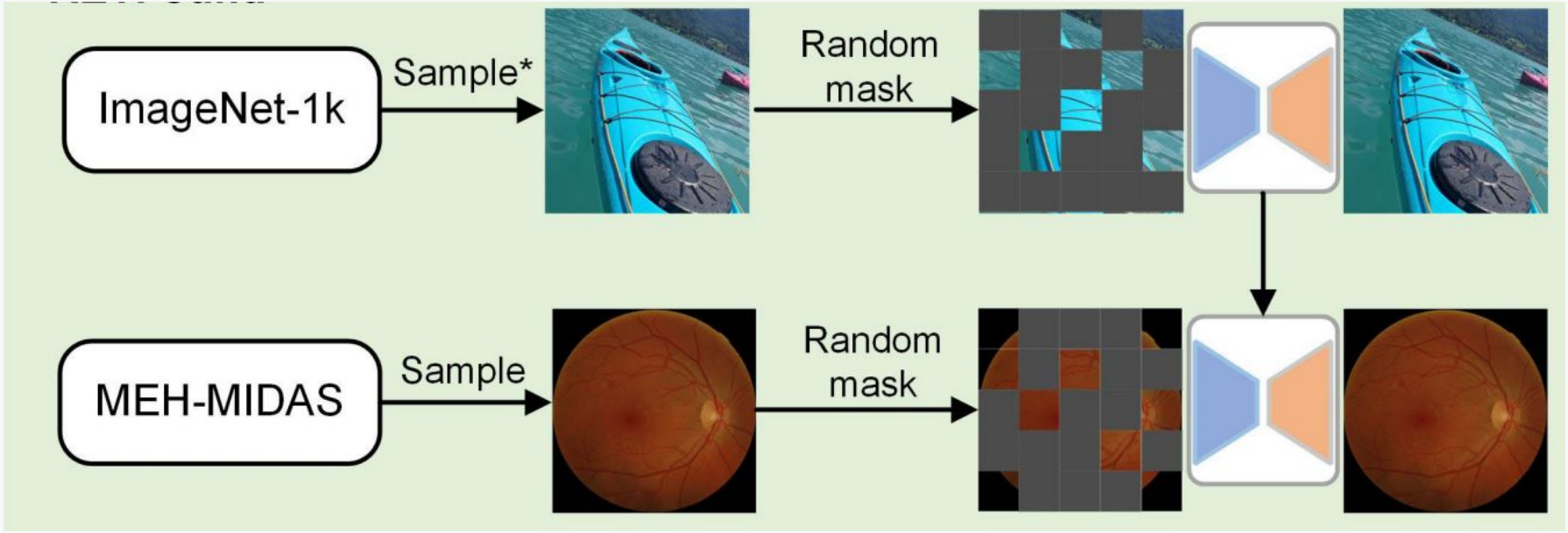


Original

Ground Truth

Predicted

RETFound - *A vision foundation model*



Differences

	Our MAE + Unet	RETFound + Unet
Data	5834	1.28M (ImageNet-1k) + 1.6M (MEH_MIDAS/ CFP)
Masking ratio	50%	75%
Epochs	100 + 100	800 + 30
Image size	512x512	224x224
Encoder size	embedding dim = 192, 12 encoder blocks (ViT tiny), 3 attention heads	emb 1024, 24 encoder blocks (ViT large), 16 attention heads
Colour	Grey scale	RGB

Differences

	Our MAE + Unet	RETFound + Unet
Freeze the encoder	start: All 12 blocks of encoder frozen ›5th epoch: unfreeze all 12 blocks	start: 22 blocks of encoder frozen ›10th epoch: unfreeze all 24, halve learning rate
decoder	1024 → 512 → 64 channels (upsampling stages) Decoder upsamples +skip connections	kernel size 1x1 convolution to reduce channel from 1024 to 256 256→128→64→32→16 channels

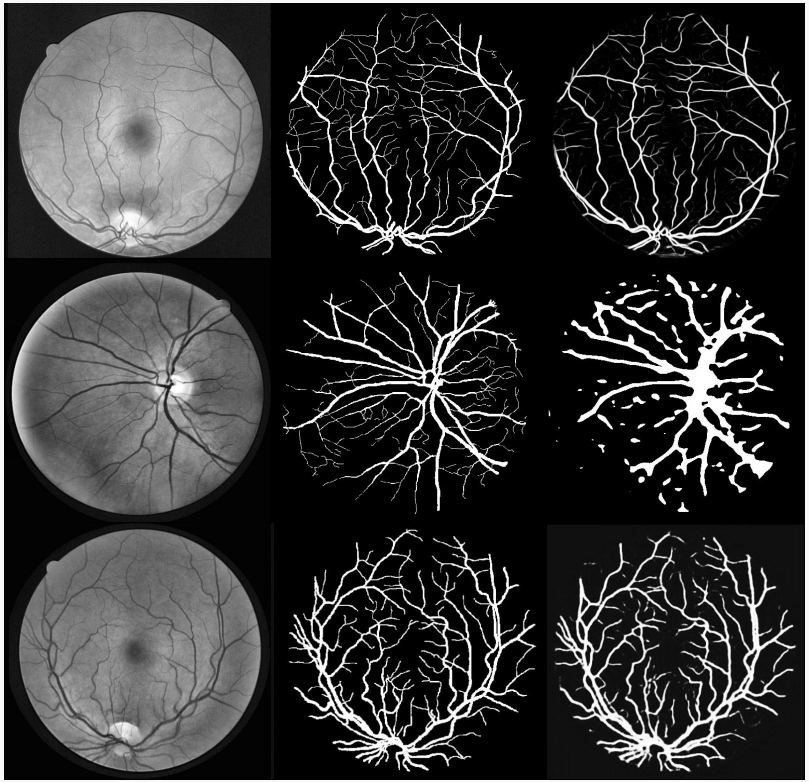
Result of RETFound

Original

Predicted

Ground Truth

Comparison



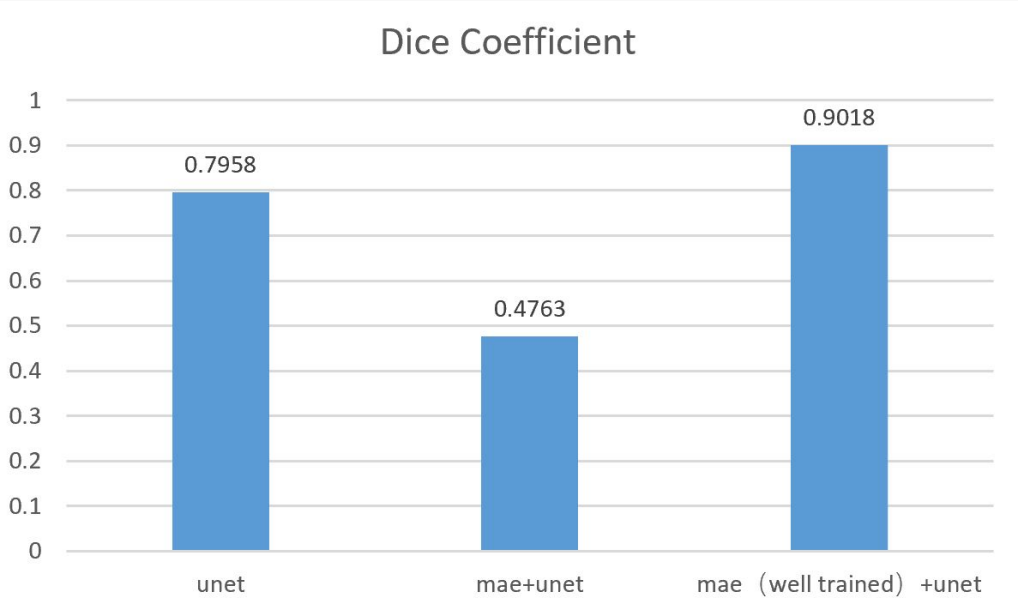
Original Ground Truth Predict

The ordinary U-Net performs well in overall blood vessel segmentation, but it is insufficient in identifying small blood vessels.

Due to insufficient training of the MAE, this model can only barely identify vascular regions and cannot effectively segment blood vessels.

Performance is almost perfect.

Comparison



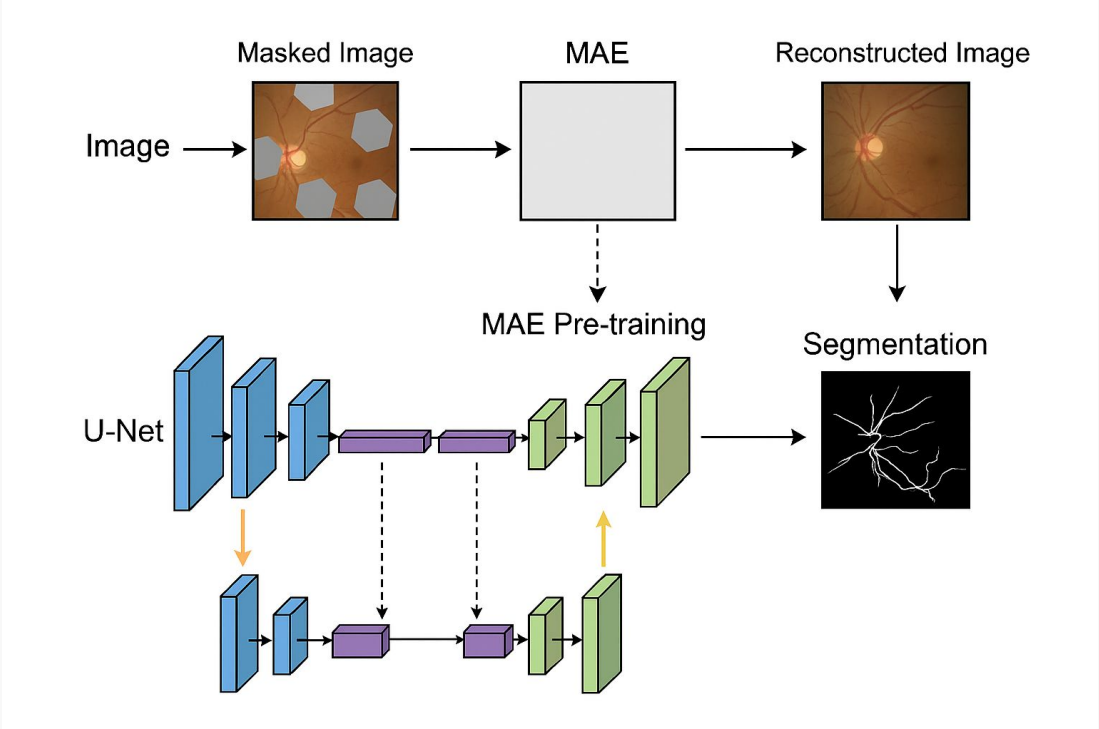
DICE score of the ordinary U-Net model is not low,

Fully trained MAE+UNet **outperforms** the previous two in all metrics

$$DICE(A,B) = 2 * |A \cap B| / |A| + |B|$$

Conclusion

- MAE pre-training can effectively improve the performance of U-Net in retinal vessel segmentation under small-sample conditions
- Shows significant advantages in capturing detail



Future Optimization

- Train on a larger dataset for longer periods
- Experiment with higher masking ratios
- Pretrain on coloured retinal fundus images and compare
- Increase embedding dimension from 192 → 1024 and attention heads from 3 → 16
- Introduce more diverse medical imaging modalities
- Explore alternative vision foundation models

References

- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000-16009).
- Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.
- Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., ... & Keane, P. A. (2023). *A foundation model for generalizable disease detection from retinal images*. Nature, 622(7981), 156-163.

BlendED

Thank You!