# 마스크된 복원에서 질병 진단까지: 안저 영상을 위한 비전 트랜스포머 접근법

Toan Duc Nguyen[1], 변규린 [1], 추현승 [1,2]
[1] 성균관대학교 AI 시스템공학과
[2] 성균관대학교전자전기컴퓨터공학과

austin47@g.skku.edu, byungyurin21@g.skku.edu, choo@skku.edu

# From Masked Reconstructions to Disease Diagnostics: A Vision Transformer Approach for Fundus Images

Toan Duc Nguyen[1], Gyurin Byun[1], Hyunseung Choo[1,2]
[1]Dept. of AI Systems Engineering, Sungkyunkwan University
[2]Dept. of Electrical and Computer Engineering, Sungkyunkwan University

## 요약

In this paper, we introduce a pre-training method leveraging the capabilities of the Vision Transformer (ViT) for disease diagnosis in conventional Fundus images. Recognizing the need for effective representation learning in medical images, our method combines the Vision Transformer with a Masked Autoencoder to generate meaningful and pertinent image augmentations. During pre-training, the Masked Autoencoder produces an altered version of the original image, which serves as a positive pair. The Vision Transformer then employs contrastive learning techniques with this image pair to refine its weight parameters. Our experiments demonstrate that this dual-model approach harnesses the strengths of both the ViT and the Masked Autoencoder, resulting in robust and clinically relevant feature embeddings. Preliminary results suggest significant improvements in diagnostic accuracy, underscoring the potential of our methodology in enhancing automated disease diagnosis in fundus imaging.

## 1. Introduction

Fundus imaging stands at the forefront of ophthalmic diagnostics, providing a non-invasive technique to visualize the retina, optic disk, and the underlying blood vessels. Historically, the interpretation of fundus images has relied on expert ophthalmologists, but with the sheer volume of medical data and the intricacies involved in image interpretation, automated systems for disease diagnosis are of paramount importance. Properly trained, such systems can expedite diagnosis, reduce human errors, and alleviate the strain on healthcare professionals.

The challenge, however, lies in the need for large-scale annotated datasets for supervised training, which are both expensive and time-consuming to obtain in the medical domain. Self-supervised learning has emerged as a promising paradigm to address this issue. By designing tasks where the data itself provides the supervision, models can be pre-trained on large unlabelled datasets, capturing rich, intricate patterns before fine-tuning on smaller labeled datasets.

The Vision Transformer (ViT) [1], which splits images into fixed-size patches, linearly embeds them, and then processes them in a series of transformer blocks, has recently achieved state-of-the-art results in various image classification tasks. Unlike the traditional convolution-based architectures, ViT provides a global view of the image, potentially capturing holistic patterns crucial for medical images. However, its efficacy in medical image analysis, particularly in the context of Fundus images, remains an area ripe for exploration.

In this paper, we bridge the gap by introducing a novel pre-training methodology combining the power of ViT with self-supervised learning principles tailored for Fundus image disease diagnosis. Through our approach, we aim to harness the global receptive fields of the Vision Transformer and the potential of self-supervised learning to create a robust diagnostic tool.

## 2. Related work

The scarcity of labeled medical data has made supervised training of deep models challenging. Self-supervised learning (SSL), which exploits unlabeled data to learn rich

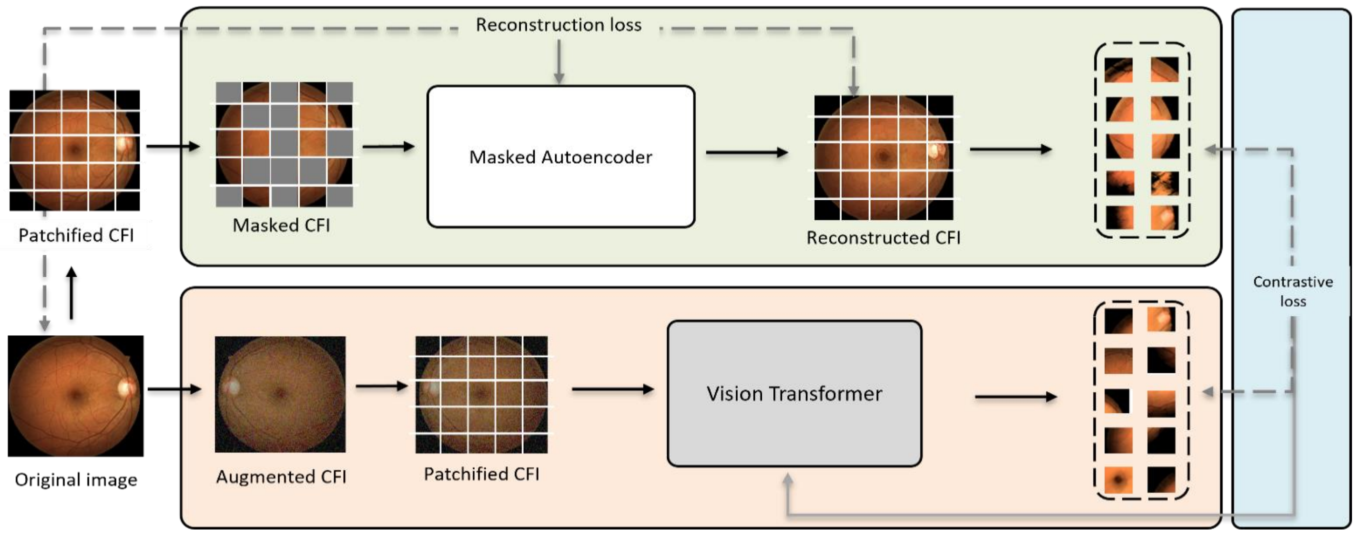

Figure 1 Proposed system

representations, offers a solution. In the medical imaging domain, [2] employed SSL for cardiac MR image segmentation by predicting the rotation angle of images. Similarly, [3] proposed a method that involves training a model to predict the sequence of shuffled patches from a medical image. These approaches underscore the potential of SSL in capturing intricate patterns from medical images without explicit labeling, setting the stage for effective fine-tuning with limited labeled datasets.

## 3. Methodology

Our pre-training methodology is structured as a synergy between the ViT and the Masked Autoencoder (MAE) [4], as shown in Figure 1. The process commences with random masking of a given Fundus image. The masked image is fed into the MAE, which aims to reconstruct the original image. This reconstruction serves two-fold purposes: evaluating the model's learning via Mean Absolute Error (MAE) and providing a foundation for contrastive learning within the Vision Transformer. The ViT then collaborates with the MAE output to generate representations using contrastive loss. The overarching goal is to ensure that these representations capture essential features, enabling the Vision Transformer to later engage in accurate disease diagnosis. Initially, both encoders split the images into patches:

$$x_p \in \mathbb{R}^{H\times W\times C} \rightarrow x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$$

where $H, W$ is the resolution of the original images, $C$ is the number of channels, $P$ is the resolution of each image patch, and $N = HW/P2$ is the number of patches.

The Masked Autoencoder (MAE) plays an important role in generating a reconstructed image from the randomly masked fundus image. The primary objective of the MAE is to fill in the masked portions in a way that closely resembles the original, unmasked image. This error quantifies the model's ability to handle masked inputs and is backpropagated to fine-tune the MAE's parameters. Given the original image $I$ and the masked image $I_{mask}$, the MAE's output $I_{reconstructed}$, then the reconstruction loss is defined as:

$$L_{reconstruction} = \frac{1}{N}\sum_{i=1}^{N}\left|I_{original_i} - I_{reconstructed_i}\right|$$

where $N$ is the total number of pixels in the image. This error quantifies the model's ability to handle masked inputs and is backpropagated to fine-tune the MAE's parameters. The output $I_{reconstructed}$ is constructed by the latent embeddings $z_{reconstructed}$, which is the output of the MAE model. We use this representation as 1 of the positive pair for contrastive learning using ViT.

The Vision Transformer is entrusted with the responsibility of learning robust representations from the reconstructed image and its original counterpart. The heart of this process is the contrastive loss, ensuring that the embeddings of the positive pair (original and reconstructed images) are closer in the latent space compared to other negative samples. Let $I_{original}$ is the original image, then $z_{original}$ is the representation obtained after training with ViT. The similarity matrix is calculated by:

$$sim(z_{original}, z_{reconstructed}) = \frac{z_{original}\cdot z_{reconstructed}}{\left\|z_{original}\right\|_2 \times \|z_{reconstructed}\|_2}$$

The contrastive loss $L_{contrastive}$ is defined as:

$$L_{contrastive} = -\log\frac{\exp(sim(z_{original}, z_{reconstructed})/\tau)}{\sum_{k=1}^{K}\exp(sim(z_{original}, z_k)/\tau)}$$

where $\tau$ is the temperature parameter and $z_k$ are the embeddings of negative samples. The above loss pushes the embeddings of the positive pair to be closer while pulling away from the embeddings of negative samples. The calculated loss is then backpropagated to optimize the Vision Transformer's weights.

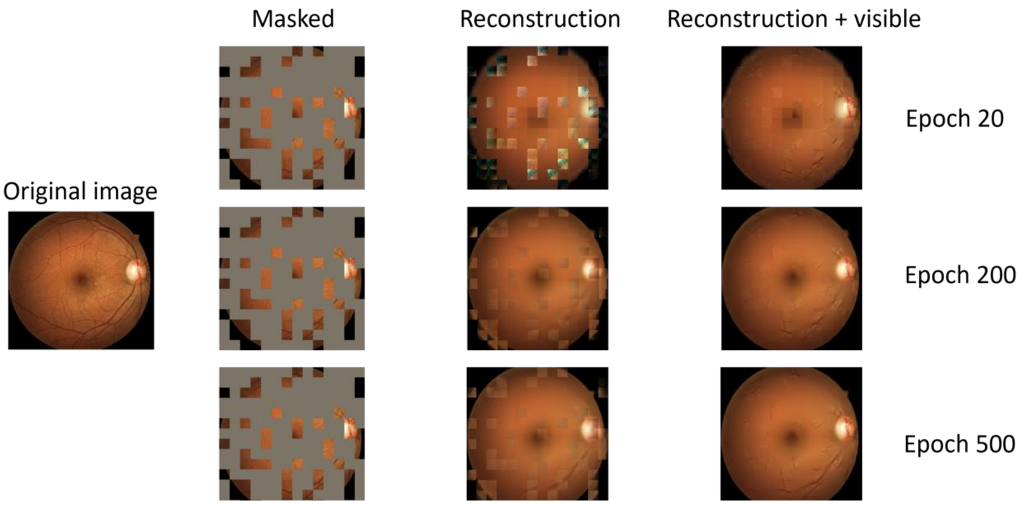

Figure 2 Generated images from Masked Autoencoder

## 4. Implementation details

All Fundus images were resized to a uniform dimension of 224×224 pixels. During training, images were fed into the model in batches of 64, an initial learning rate of 0.001, temperature of 0.07 was used and The AdamW optimizer was selected for training both the Masked Autoencoder and the Vision Transformer. To further mitigate overfitting, especially crucial given the high-dimensional nature of medical images, a weight decay coefficient of 0.99 was implemented. This L2 regularization technique penalizes larger weights, pushing the model to maintain smaller parameter values and thus a simpler model. Data augmentation is also used to diversify the original image, this makes the contrastive learning become more effective. The pre-training phase, which involves the tandem training of the Masked Autoencoder and Vision Transformer using the proposed methodology, spans 1000 epochs. This extensive pre-training ensures that the models thoroughly learn the intricate patterns and representations of the Fundus images. Following this, the Vision Transformer undergoes a fine-tuning phase for disease diagnosis, lasting 100 epochs. This stage refines the model's weights, optimizing it for the specific task of disease classification.

For evaluation, we use kappa score, often referred to as Cohen's Kappa coefficient, is a statistical measure used to evaluate the agreement between two raters or classifiers, accounting for the agreement that might happen by chance. The coefficient provides a score between -1 and 1: a score of 1 indicates perfect agreement between the two raters, a score of 0 suggests agreement is no better than chance, and a negative value implies agreement is worse than random chance. To compute the Kappa score, one first calculates the observed agreement ($p_o$) between the raters, which is the proportion of instances where both raters agree. Next, the agreement that would be expected by chance alone ($p_e$) is calculated, usually derived from the individual probabilities of each rater assigning each category. The Kappa score is then given by the formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

## 5. Results

In Figure 2, four distinct image states are visualized: the original image, the masked image, the reconstructed image after specific epochs, and a combined visualization of the reconstruction superimposed with the visible regions of the original image (reconstruction + visible). During the early stages of training, the MAE's ability to accurately reconstruct the masked portions of the image was nascent, with certain disparities evident when comparing the reconstructed and original images. By the 200th epoch, substantial improvements were apparent. The reconstructed images bore closer resemblance to the original images, indicating the MAE's progressing proficiency in capturing intricate details and nuances of the Fundus images. At the 500-epoch mark, the MAE's performance further matured. The reconstructed images were notably sharper, with the masked regions being restored with high fidelity. The reconstruction + visible visualizations further emphasized the minimal discrepancy between the reconstructed and original areas, underscoring the efficacy of the MAE's training. This sequential improvement in the MAE's outputs, as evidenced by the epochs, showcases the model's increasing capability to handle masked portions adeptly, setting the stage for effective contrastive learning in the subsequent Vision Transformer phase.

We proceeded to fine-tune the Vision Transformer (ViT) using the APTOS 2019 dataset, renowned for its comprehensive collection of Fundus images labeled for Diabetic Retinopathy (DR) grading. The dataset is characterized by its multi-class nature, with each image assigned a DR grade that reflects the severity of the condition. . Our results, hence, underscore the potential of the integrated MAE and ViT pre-training methodology in enhancing the performance of Fundus image classification tasks, even when compared to established supervised learning paradigms. Our model exhibited notable efficacy. While a model with random initialization achieved a kappa score of 57.27 and a supervised learning approach using Imagenet reached 80.57, our ViT model surpassed both, registering a score of 82.49. This advancement not only highlights the effectiveness of our integrated MAE and ViT pre-training strategy but also signifies its potential in real-world medical image classification tasks, offering an enhanced alternative to traditional supervised paradigms. Detailed results are shown in Table 1.

Table 1 Results of fine-tuning of different methods

| Models | Random | Supervised | Ours |
|---|---|---|---|
| Kappa score | 57.27 | 80.57 | **82.49** |

Figure 3 shows the training loss of our proposed method and supervised method. Firstly, the loss curve going down means our model does not overfit and the learning is effective as we train more epochs. This is important since we train ViT with much smaller dataset, compared to supervised learning. This also proves that the kappa score of our model is feasible. Furthermore, our method also achieves a lower loss compared to supervised learning method, explaining that higher kappa score, as well as the better performance.

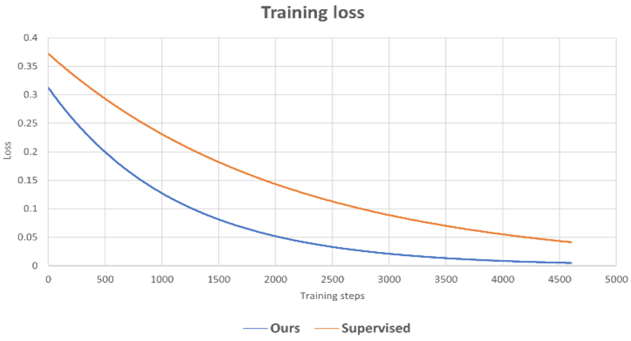

Figure 3 Training loss of our proposed method and supervised learning

## 6. Discussion and conclusion

In this research, we proposed a novel pre-training methodology that synergistically combined the capabilities of the Vision Transformer (ViT) and the Masked Autoencoder (MAE) for disease diagnosis in Fundus images. Through our rigorous experiments and methodological design, our approach showcased superior performance, emphasizing the power of coupling reconstruction capabilities with contrastive learning. By efficiently leveraging the strengths of both models and ensuring optimal training conditions, we believe our method stands as a strong contender in the field of medical image analysis.

While our proposed methodology has demonstrated commendable performance in Fundus image disease diagnosis, it's essential to acknowledge its limitations. One notable constraint is the model's reliance on the quality of masked reconstructions. If the Masked Autoencoder fails to generate accurate reconstructions, it might adversely impact the Vision Transformer's learning efficacy. Additionally, our method has been tailored specifically for Fundus images, and its generalizability to other medical imaging modalities remains to be tested. The choice of hyperparameters, while optimal for our dataset, might require fine-tuning for broader applications or diverse datasets.

Looking ahead, several avenues present themselves for exploration. A natural extension would be to test our methodology's performance on other medical imaging modalities such as MRI, CT, or X-rays to ascertain its adaptability. It might also be worthwhile to investigate the integration of other self-supervised learning tasks or even combine multiple tasks concurrently to bolster the model's pre-training phase. Another promising direction is the incorporation of attention mechanisms or other state-of-the-art architectures to further enhance the model's capability to discern intricate patterns, especially in noisy or low-quality images. As the field of medical image analysis continues to evolve rapidly, we are optimistic that our foundational work will inspire advanced techniques that build upon our successes while addressing the acknowledged limitations.

**Acknowledgements**

This work was funded in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] under the Artificial Intelligence Innovation Hub (2021-0-02068), ICT Creative Consilience Program (2023-2020-0-01821), and AI Graduate School program (IITP-2019-0-00421).

## 참고문헌

[1] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 2020.

[2] Bai, Wenjia, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. "Self-supervised learning for cardiac mr image segmentation by anatomical position prediction." In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, pp. 541-549.

[3] Chen, Liang, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. "Self-supervised learning for medical image analysis using image context restoration." Medical image analysis 58 (2019): 101539. 2019.

[4] He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, et al. “Masked autoencoders are scalable vision learners." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000-16009. 2022.