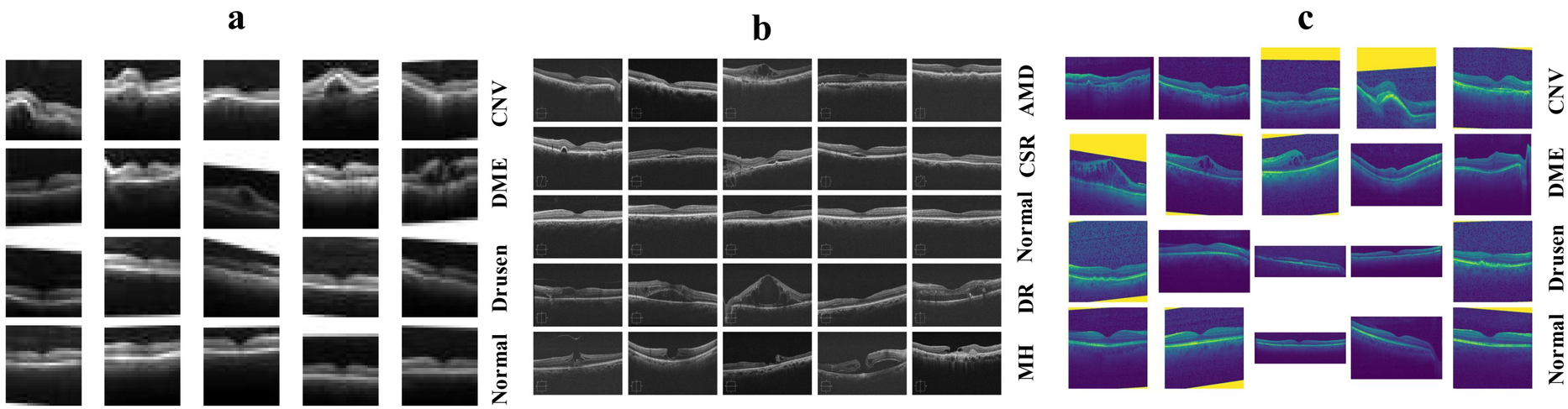
ViT-2SPN: Vision Transformer-based Dual-Stream Self-Supervised Pretraining Networks for Retinal OCT Classification

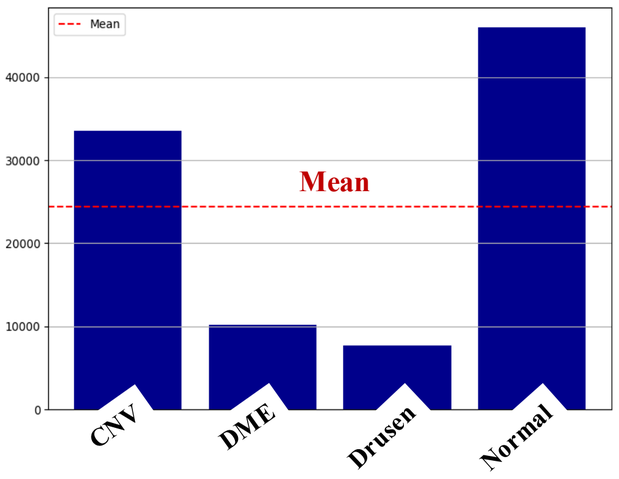
Vision Systems and Intelligence (VSI) Lab.
04/28/2025
Reza

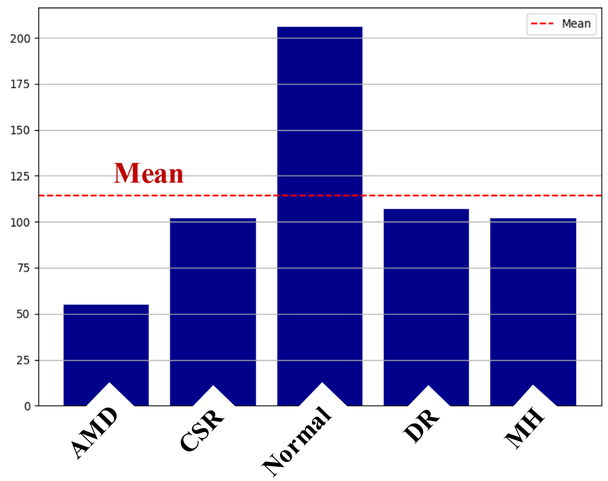
THE UNIVERSITY OF ARIZONA

Datasets

Caption:
Sample images from the datasets used for retinal disease classification. (a), (c) OCTMNIST and UCSD-OCT (including Normal, Diabetic Macular Edema (DME), Choroidal Neovascularization (CNV), and Drusen). (b) OCTID (including Normal, Macular Hole (MH), Age-related Macular Degeneration (AMD), Central Serous Retinopathy (CSR), and Diabetic Retinopathy (DR)).







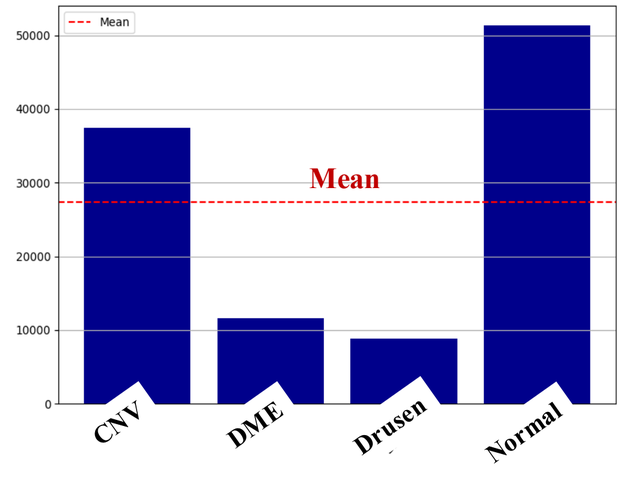
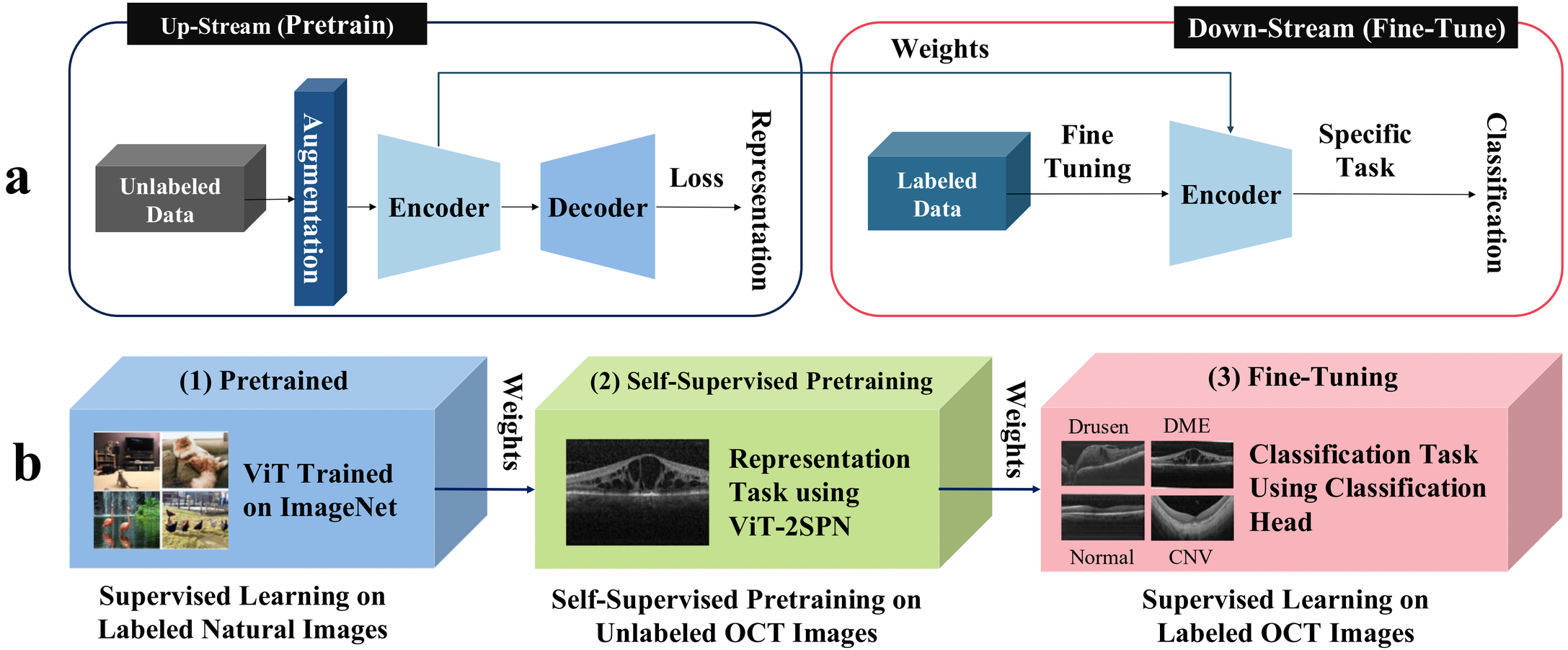


TABLE I
COMPARISON OF THE OCTMNIST, OCTID, AND UCSD OCT DATASETS ACROSS VARIOUS ATTRIBUTES.

Attribute	OCTMNIST (2021)	OCTID (2020)	UCSD OCT (2018)
Number of Images	97,477	572	109,309
Number of Classes	4 (CNV, DME, Drusen, Normal)	5 (Normal, DR, MH, AMD, CSR)	4 (CNV, DME, Drusen, Normal)
Image Dimensions	28×28 pixels	771×514 pixels	Varies (384×1536 to 277×512 pixels)
Image Quality	Low resolution (28×28), less detail	Mid-resolution (771×514), some quality variation	Mid-resolution, better detail retention
Class Distribution	Imbalanced	Imbalanced	Imbalanced

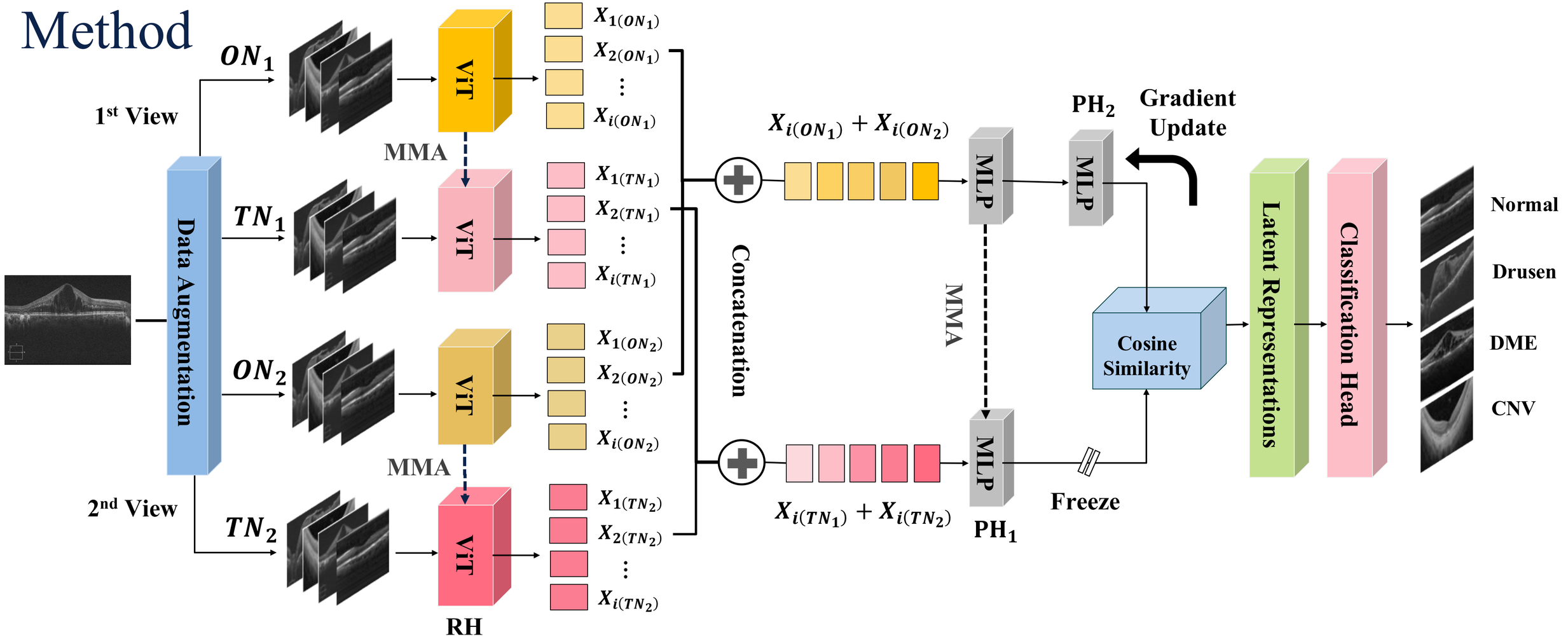
THE UNIVERSITY OF ARIZONA

Approach



(a) Self-supervised learning (SSL) pipeline for the classification task; (b) Self-supervised pertaining (SSP) approach for the classification task. ViT-2SPN stands for Vision Transformer-based Dual-Stream Self-Supervised Pretraining Networks, DME refers to Diabetic Macular Edema, and CNV refers to Choroidal Neovascularization.

Method



Overview of the Vision Transformer-based Dual-Stream Self-Supervised Pretraining Network (ViT-2SPN) architecture. It consists of parallel online ((ON1), (ON2)) and momentum-updated target ((TN1), (TN2)) encoders, all using a ViT backbone pretrained on ImageNet. Each branch processes a different augmented OCT view. Features from both streams are concatenated and passed through projection (PH1) and prediction (PH2) heads, enabling contrastive learning in a shared latent space.

THE UNIVERSITY OF ARIZONA

Method

Loss Function: $$\mathcal{L}_{\text{total}} = \frac{-1}{N \cdot S} \sum_{s=1}^{S} \sum_{i=1}^{N} \frac{\mathbf{p}_{\text{o}}^{(i,s)} \cdot \mathbf{z}_{\text{t}}^{(i,s)}}{\|\mathbf{p}_{\text{o}}^{(i,s)}\| \|\mathbf{z}_{\text{t}}^{(i,s)}\|}$$

- $\mathbf{L_{total}}$ is the total loss that will be used for updating the network weights during backpropagation.
- **N** is the number of samples in each batch — the number of data points processed together at one time.
- **S** is the number of accumulation steps. This is used when the full batch doesn't fit into GPU memory, so we split it and accumulate gradients across multiple mini-batches.
- $\sum_{s=1}^{S}$ represents summing over all accumulation steps, that is, adding the loss contributions from each smaller batch.
- $\sum_{i=1}^{N}$ represents summing over all samples in a batch or accumulation step.
- $\mathbf{P_o^{(i,s)}}$ is the predicted feature from the online network (after the projection head) for the i-th sample in the s-th step.
- $\mathbf{Z_t^{(i,s)}}$ is the target feature from the target network (no predictor head) for the i-th sample in the s-th step.
- $\mathbf{P_o^{(i,s)} \cdot Z_t^{(i,s)}}$ is the dot product between the two feature vectors, measuring how aligned they are without considering their magnitudes.
- $\|\mathbf{P_o^{(i,s)}}\| \cdot \|\mathbf{Z_t^{(i,s)}}\|$ are the norms (magnitudes) of the prediction and target vectors, respectively.
- The fraction calculates the cosine similarity between the prediction and the target feature vectors, resulting in a value between -1 and 1.
- **The negative sign -1** indicates that we are maximizing cosine similarity (because optimization typically minimizes the loss, so we take the negative to flip the objective).

Result

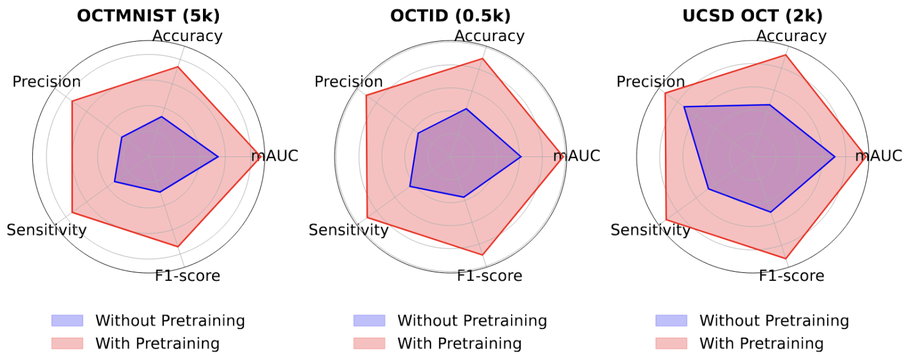


Fig. 5. Comparison of ViT-2SPN performance with and without pretraining on the imbalanced OCTMNIST (5k), UCSD OCT (2k), and OCTID (0.5k) datasets.

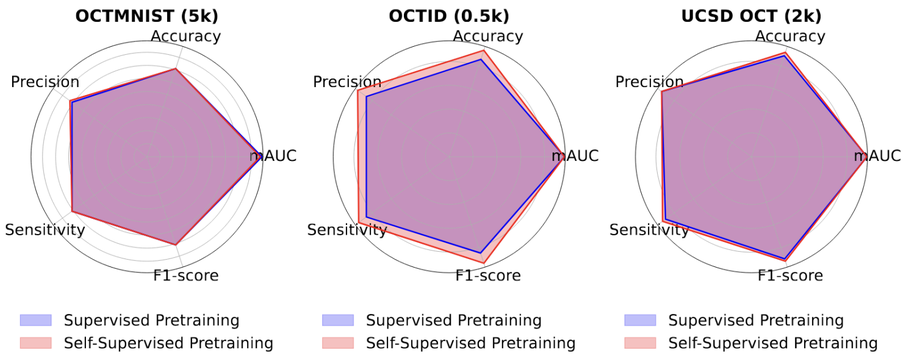


Fig. 4. Performance comparison of supervised and self-supervised pretraining in VIT-2SPN across three imbalanced datasets: OCTMNIST (5k), UCSD OCT (2k), and OCTID (0.5k).

TABLE II

PERFORMANCE COMPARISON OF CNN-BASED AND VIT-BASED BACKBONES IN VIT-2SPN ON THE 5K IMBALANCED OCTMNIST DATASET.

1st Backbone	2nd Backbone	Parameter	FLOPs	Batch Size	mAUC	Accuracy	Precision	Sensitivity	F1-score
ResNet-50	ResNet-50	51.37 M	4.14 G	32	0.837	0.67	0.67	0.67	0.65
ResNet-50	ResNet-50	51.37 M	4.14 G	64	0.854	0.70	0.71	0.70	0.70
ResNet-50	ResNet-50	51.37 M	4.14 G	128	0.839	0.74	0.73	0.74	0.73
ViT-Tiny	ResNet-50	50.60 M	1.07 G	32	0.853	0.71	0.70	0.71	0.70
ViT-Tiny	ResNet-50	50.60 M	1.07 G	64	0.857	0.67	0.68	0.67	0.65
ViT-Tiny	ResNet-50	50.60 M	1.07 G	128	**0.886**	0.73	0.73	0.73	0.73
ViT-Tiny	**ViT-Tiny**	**11.68 M**	1.07 G	32	**0.859**	**0.72**	**0.72**	**0.72**	**0.71**
ViT-Tiny	**ViT-Tiny**	**11.68 M**	1.07 G	64	**0.868**	**0.72**	**0.72**	**0.72**	**0.71**
ViT-Tiny	**ViT-Tiny**	**11.68 M**	1.07 G	128	0.878	**0.75**	**0.76**	**0.75**	**0.75**

TABLE III

PERFORMANCE COMPARISON BETWEEN DUAL-STREAM AND SINGLE-STREAM ARCHITECTURES IN VIT-2SPN ACROSS THREE IMBALANCED OCT DATASETS—OCTMNIST (5K), UCSD OCT (2K), AND OCTID (0.5K).

Dataset	Network	Parameter	FLOPs	mAUC	Accuracy	Precision	Sensitivity	F1-score
OCTMNIST	Single-Stream Network	**5.92 M**	1.07 G	0.868	0.67	0.68	0.67	0.65
	Dual-Stream Network	11.68 M	1.07 G	**0.884**	**0.71**	**0.73**	**0.71**	**0.70**
OCTID	Single-Stream Network	**5.92 M**	1.07 G	0.931	0.78	0.83	0.78	0.78
	Dual-Stream Network	11.68 M	1.07 G	**0.968**	**0.86**	**0.91**	**0.86**	**0.86**
UCSD OCT	Single-Stream Network	**5.92 M**	1.07 G	0.961	0.83	0.86	0.83	0.84
	Dual-Stream Network	11.68 M	1.07 G	**0.964**	**0.84**	**0.87**	**0.84**	**0.85**

THE UNIVERSITY OF ARIZONA

Result

TABLE IV

EVALUATION OF VIT-2SPN PERFORMANCE AGAINST BASELINE SELF-SUPERVISED MODELS ON THE 5K IMBALANCED OCTMNIST DATASET.

Baseline Model	Parameters	FLOPs	mAUC ± SD	Accuracy	Precision	Sensitivity	F1-score
BYOL	34.15 M	4.06 G	0.767 ± 0.0254	0.64	0.64	0.64	0.61
MoCo	25.73 M	4.06 G	0.805 ± 0.0299	0.67	0.66	0.67	0.66
SimCLR	25.73 M	8.13 G	0.810 ± 0.0393	0.69	0.71	0.69	0.69
SwAV	55.65 M	4.06 G	0.801 ± 0.0286	0.66	0.65	0.66	0.65
SimSiam	27.96 M	4.06 G	0.622 ± 0.0957	0.54	0.54	0.54	0.50
SimCLRv2	25.73 M	8.13 G	0.812 ± 0.0193	0.64	0.66	0.64	0.63
MoCov2	25.73 M	4.06 G	0.795 ± 0.0250	0.59	0.61	0.59	0.57
ViT-2SPN (Ours)	**11.68 M**	**1.07 G**	**0.884 ± 0.0070**	**0.71**	**0.73**	**0.71**	**0.71**

TABLE V

EVALUATION OF VIT-2SPN PERFORMANCE AGAINST BASELINE SELF-SUPERVISED MODELS ON THE 0.5K IMBALANCED OCTID DATASET.

Baseline Model	Parameters	FLOPs	mAUC ± SD	Accuracy	Precision	Sensitivity	F1-score
BYOL	34.15 M	4.06 G	0.690 ± 0.0913	0.56	0.44	0.56	0.47
MoCo	25.73 M	4.06 G	0.525 ± 0.0330	0.36	0.24	0.36	0.29
SimCLR	25.73 M	8.13 G	0.558 ± 0.0639	0.42	0.32	0.42	0.29
SwAV	55.65 M	4.06 G	0.548 ± 0.0514	0.38	0.27	0.38	0.32
SimSiam	27.96 M	4.06 G	0.554 ± 0.0671	0.40	0.31	0.40	0.32
SimCLRv2	25.73 M	8.13 G	0.571 ± 0.1009	0.54	0.47	0.54	0.44
MoCov2	25.73 M	4.06 G	0.505 ± 0.0726	0.34	0.34	0.34	0.23
ViT-2SPN (Ours)	**11.68 M**	**1.07 G**	**0.941 ± 0.0881**	**0.84**	**0.85**	**0.84**	**0.84**

TABLE VI

EVALUATION OF VIT-2SPN PERFORMANCE AGAINST BASELINE SELF-SUPERVISED MODELS ON THE 2K IMBALANCED UCSD OCT DATASET.

Baseline Model	Parameters	FLOPs	mAUC ± SD	Accuracy	Precision	Sensitivity	F1-score
BYOL	34.15 M	4.06 G	0.765 ± 0.0672	0.74	0.74	0.74	0.74
MoCo	25.73 M	4.06 G	0.858 ± 0.0585	0.80	0.84	0.80	0.82
SimCLR	25.73 M	8.13 G	0.860 ± 0.0533	0.82	0.83	0.82	0.82
SwAV	55.65 M	4.06 G	0.868 ± 0.0399	0.81	0.86	0.81	0.81
SimSiam	27.96 M	4.06 G	0.616 ± 0.1162	0.72	0.74	0.72	0.73
SimCLRv2	25.73 M	8.13 G	0.574 ± 0.1073	0.59	0.67	0.59	0.61
MoCov2	25.73 M	4.06 G	0.607 ± 0.1194	0.71	0.76	0.71	0.72
ViT-2SPN (Ours)	**11.68 M**	**1.07 G**	**0.959 ± 0.0135**	**0.86**	**0.90**	**0.86**	**0.87**

Result

TABLE VII

EVALUATION OF THE VIT-2SPN MODEL'S PERFORMANCE ACROSS 5 RETRAINING RUNS ON THE IMBALANCED OCTMNIST (5K), UCSD OCT (2K), AND OCTID (0.5K) DATASETS.

Dataset	mAUC ± SD	Top-1 Accuracy ± SD	Confidence Score ± SD	Sensitivity ± SD	Specificity ± SD
OCTMNIST	0.8734 ± 0.0196	0.6626 ± 0.0391	0.6995 ± 0.0269	0.6639 ± 0.0413	0.8885 ± 0.0134
	0.8848 ± 0.0172	0.6926 ± 0.0228	0.7103 ± 0.0125	0.6861 ± 0.0231	0.8973 ± 0.0077
	0.8787 ± 0.0203	0.6888 ± 0.0225	0.7005 ± 0.0263	0.6848 ± 0.0183	0.8963 ± 0.0068
	0.8896 ± 0.0089	0.7074 ± 0.0256	0.7143 ± 0.0209	0.7031 ± 0.0244	0.9025 ± 0.0082
	0.8843 ± 0.0122	0.6948 ± 0.0323	0.7048 ± 0.0207	0.6889 ± 0.0309	0.8981 ± 0.0104
Mean ± SD	0.882 ± 0.0056	0.69 ± 0.0154	0.71 ± 0.0058	0.69 ± 0.0130	0.90 ± 0.0047
OCTID	0.9308 ± 0.1038	0.8280 ± 0.1518	0.7096 ± 0.1522	0.8060 ± 0.1748	0.9578 ± 0.0381
	0.9688 ± 0.0286	0.8200 ± 0.1327	0.7551 ± 0.0805	0.7789 ± 0.1610	0.9537 ± 0.0377
	0.9450 ± 0.0625	0.8500 ± 0.1978	0.7513 ± 0.1594	0.8159 ± 0.2140	0.9618 ± 0.0516
	0.9394 ± 0.0568	0.8040 ± 0.1433	0.7182 ± 0.1353	0.7536 ± 0.1800	0.9487 ± 0.0440
	0.9731 ± 0.0202	0.8440 ± 0.0550	0.7933 ± 0.0365	0.8144 ± 0.0715	0.9606 ± 0.0132
Mean ± SD	0.951 ± 0.0167	0.83 ± 0.0175	0.74 ± 0.0318	0.80 ± 0.0243	0.96 ± 0.0050
UCSD OCT	0.9660 ± 0.0097	0.8434 ± 0.0482	0.8048 ± 0.0388	0.8284 ± 0.0363	0.9518 ± 0.0125
	0.9616 ± 0.0133	0.8379 ± 0.0313	0.7868 ± 0.0380	0.8266 ± 0.0351	0.9501 ± 0.0090
	0.9660 ± 0.0088	0.8571 ± 0.0517	0.8062 ± 0.0370	0.8463 ± 0.0310	0.9563 ± 0.0133
	0.9592 ± 0.0135	0.8490 ± 0.0241	0.7994 ± 0.0231	0.8417 ± 0.0236	0.9519 ± 0.0064
	0.9611 ± 0.0074	0.8692 ± 0.0300	0.7966 ± 0.0333	0.8428 ± 0.0207	0.9570 ± 0.0083
Mean ± SD	0.963 ± 0.0029	0.85 ± 0.0117	0.80 ± 0.0072	0.84 ± 0.0087	0.95 ± 0.0031

Conclusion

1. Dual-Stream Network > Single-Stream Network (without increasing the FLOPs)
2. ViT Backbones in Dual-Stream Network > CNN only or CNN + ViT Backbones in Dual-Stream Network
3. The Pretraining Strategy > Without Pretraining Strategy
4. Self-Supervised Pretraining Strategy (small size data only) > Supervised Pretraining Strategy
5. ViT-2SPN Model > Baseline Self-Supervised Learning Models
6. The ViT-2SPN model demonstrated promising performance in classifying OCT images, achieving high specificity (over 90%) across all datasets. This emphasizes its capacity to accurately identify healthy individuals and reduce false positives, which is particularly beneficial for screening healthy patients and ensuring the correct identification of non-diseased cases. However, its sensitivity, especially on the OCTMNIST dataset, was relatively lower, suggesting that it was less effective at detecting diseased patients. In clinical settings, high sensitivity is essential to minimize false negatives and ensure that patients with conditions are not overlooked, which is critical for timely intervention and treatment.

Thank you! Any Question?