2024 International Conference on Bioinformatics and Biomedicine (BIBM)

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822551

VGG-ST: A VGG-Swin Transformer-Based Model for ROP Disease Diagnosis

1st Xinwei Luo
Computer Science and Engineering
Lehigh University
PA, USA
xil620@lehigh.edu

2nd Songlin Zhao
Computer Science and Engineering
Lehigh University
PA, USA
soz223@lehigh.edu

3rd Yong Chen
Biostatistics, Epidemiology and Informatics
University of Pennsylvania
PA, USA
ychen123@pennmedicine.upenn.edu

4th Gui-shuang Ying
Ophthalmology, Scheie Eye Institute
University of Pennsylvania
PA, USA
gsying@pennmedicine.upenn.edu

5th Lifang He
Computer Science and Engineering
Lehigh University
PA, USA
lih319@lehigh.edu

***Abstract*—The detection of Referral-Warranted Retinopathy of Prematurity (RW-ROP) is crucial for preventing severe visual impairment in premature infants. Recent studies have demonstrated that deep learning models, particularly CNNs, are effective in classifying ROP. However, the comprehensive extraction and integration of ROP-relevant features from retinal images for accurate classification remains challenging. In this paper, we propose a hybrid model based on Very Deep Convolutional Networks (VGG) and Swin Transformer (VGG-ST) for identifying RW-ROP using retinal images. The VGG-ST model first employs the VGG19 architecture to extract detailed local image features and the Swin Transformer V2 architecture to capture comprehensive global contextual information. It then introduces a novel feature enhancement module that combines these local and global features through an adaptive integration strategy, optimizing the feature representation for more accurate RW-ROP detection. Finally, the integrated features are processed through a classification module to predict the probability of RW-ROP, distinguishing between normal and RW-ROP cases. We also present customized data preprocessing techniques to address class imbalance and retinal image blur issues inherent in the e-ROP dataset. Experimental results demonstrate that the VGG-ST model offers improved classification sensitivity compared to existing methods, making it a promising tool for automated ROP screening in clinical settings. The source code is available at https://github.com/hawk-sudo/VGG-ST.**

***Index Terms*—Retinopathy of prematurity, retinal image, deep learning, CNN, transformer**

I. Introduction

Retinopathy of Prematurity (ROP) [1] is a serious vaso-proliferative disease that affects the retinas of premature infants, with the potential to cause visual impairment or blindness. Timely diagnosis is critical, as severe ROP often requires retinal ablative surgery within a narrow diagnostic window for effective treatment. To identify premature infants with ROP who need further intervention, Ells et al. introduced the concept of the Referral-Warranted ROP (RW-ROP) [2]. RW-ROP is defined by the presence of high-risk characteristics in the eyes, such as plus disease, ROP in zone I, or stage 3 ROP or greater. Eyes classified as RW-ROP require thorough evaluation by an ophthalmologist, and a significant proportion of these cases necessitate treatment. The standard method for detecting RW-ROP involves a series of costly diagnostic examinations performed by ophthalmologists on at-risk infants. In the United States, ROP screening guidelines mandate that all infants with a birth weight of 1500 grams or less, or a gestational age of 30 weeks or less, undergo these examinations [3]. However, this approach leads to numerous unnecessary examinations, as fewer than 10% of screened infants ultimately require treatment for ROP [4].

Recent advancements in imaging technology have enabled the use of retinal images for RW-ROP detection, optimizing resource allocation and reducing unnecessary clinical interventions [5]. Wide-angle digital retinal imaging systems, such as RetCam [6], are widely used to examine premature infants for RW-ROP. Despite these technological advances, the interpretation of retinal images remains a manual process that is time-consuming and prone to variability. The shortage of pediatric ophthalmologists and retinal specialists further exacerbates the inefficiency of RW-ROP screenings [7]. As a result, many infants in need of timely treatment may experience delays, leading to potential vision loss. Therefore, there is an urgent need for an automated tool capable of identifying RW-ROP with retinal images, which would streamline the screening process and improve early detection outcomes.

In recent years, deep learning, a subset of Artificial Intelligence (AI), has demonstrated exceptional performance in medical imaging tasks, particularly in diagnosing ROP using retinal images. Convolutional Neural Networks (CNNs) [8] have been widely employed in this domain. For instance, Tan et al. [9] utilized Inception-V3 to classify ROP plus disease, while Huang et al. [10] and Chen et al. [11] employed VGG19 and ResNet152 to determine the stage and severity of ROP, respectively. For an extensive review of the literature preceding 2022, please refer to [12]. More recently, Ebrahimi et al.

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

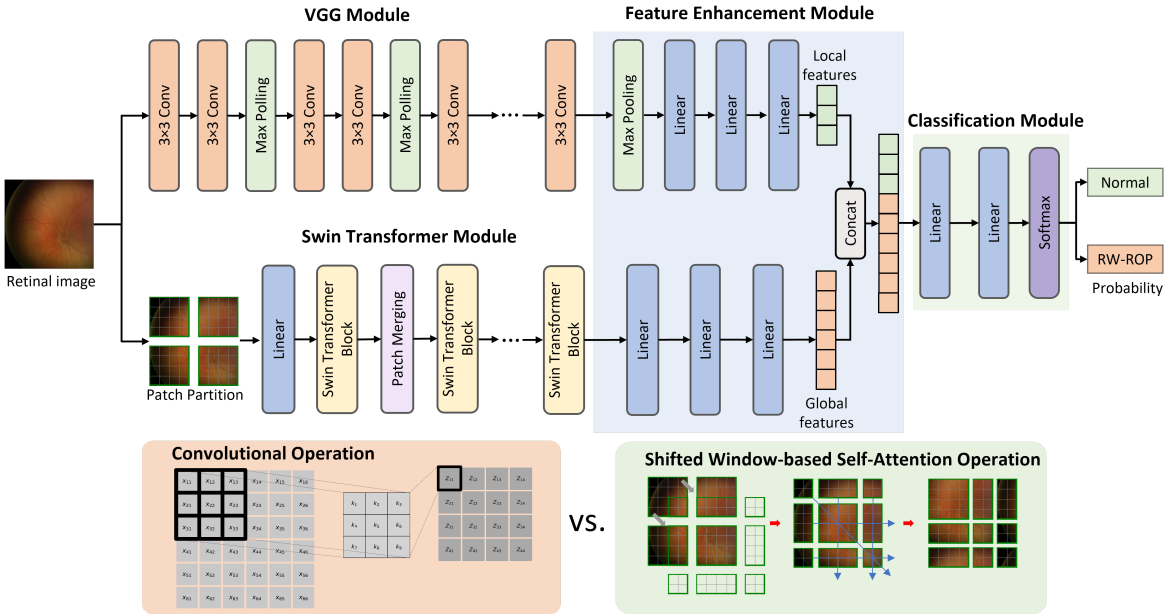

Fig. 1. The architecture of the proposed VGG-ST model. The VGG module (top) performs convolutional operations to extract local features from the input retinal images, while the Swin Transformer module (bottom) partitions the image into patches and applies shifted window-based self-attention mechanisms to capture global features. These local and global features are then fused in the feature enhancement module, where they are integrated into a comprehensive feature vector. The classification module uses this vector to predict the probability of the eye being normal or affected by RW-ROP. Additionally, the lower part of the figure provides a comparison between the convolutional operation and the shifted window-based self-attention operation for clarity.

[13] leveraged EfficientNet-V2 to identify ROP stages, and Wagner et al. [14] employed DenseNet201 for the diagnosis of ROP plus disease. Despite their success, CNNs are inherently limited by their constrained receptive fields, which restrict their ability to capture global contextual information and long-range dependencies within images [15]. This limitation poses significant challenges in diagnosing ROP disease, as retinal images of premature infants often exhibit blurriness and lack distinct local features. To address these challenges, recent studies have increasingly incorporated deep learning models based on transformer architectures to better capture global features within images. Zhao et al. [16] developed a dual-branch model that combines ResNet50 with MaxViT for ROP stage classification. Similarly, Sankari et al. [17] proposed a comprehensive evaluation system that integrates multiple CNNs with a Swin Transformer for ROP diagnosis. However, these approaches primarily focus on combining features or classification results from different models, without fully exploring how to effectively integrate these outputs to enhance model performance for ROP diagnosis.

In this paper, we propose a novel hybrid deep learning model called VGG-Swin Transformer (VGG-ST) for enhancing the detection of RW-ROP using retinal images of premature infants. The VGG-ST model synergistically combines the strengths of VGG19 and Swin Transformer V2 architectures to capture both detailed local features and comprehensive global contextual information. A novel feature enhancement module is introduced, employing an adaptive integration strategy to merge these local and global features into a unified representation. This optimized feature set is then fed into a classification module that accurately differentiates between normal and RW-ROP cases. Furthermore, we developed tailored data preprocessing techniques to address challenges such as class imbalance and image blur in the e-ROP dataset. The VGG-ST model not only exploits diverse feature extraction methods for a thorough analysis but also ensures effective integration for precise RW-ROP identification. The key contributions of this study are as follows:

Novel Hybrid Deep Learning Model. We introduce the VGG-ST, a dual-branch model specifically designed for RW-ROP detection using retinal images. The VGG-ST model synergistically combines convolutional operations with a self-attention mechanism to capture both detailed local features and comprehensive global contextual information, providing a robust foundation for RW-ROP classification.

Adaptive Feature Integration Strategy. We propose an adaptive feature integration strategy that effectively merges local and global features, treating them as complementary sources of critical information. This approach significantly enhances model performance, and our comprehensive hyperparameter analysis explores the impact of varying the local-to-global feature ratio on RW-ROP classification accuracy.

Customized Data Preprocessing Techniques. We develop a set of customized data preprocessing techniques tailored to the specific challenges of the e-ROP dataset, such as image blur and class imbalance. Our ablation studies demonstrate that these preprocessing methods significantly improve the accuracy of RW-ROP classification, underscoring their critical role in model training.

II. Methods

Fig. 1 illustrates the proposed VGG-ST architecture for detecting RW-ROP using retinal images. The model is composed of four modules. First, the VGG module extracts local features through convolutional operations, while the Swin Transformer module captures global features using self-attention mechanisms. These local and global features are then merged into a unified feature vector by the feature enhancement module, following an adaptive integration strategy. Finally, the classification module uses this feature vector to estimate the probability of RW-ROP. In the following, we provide a detailed description of each module and explain how the VGG-ST model effectively combines and leverages both local and global features of retinal images to identify RW-ROP.

A. VGG Module

The VGG module utilizes the VGG19 architecture to extract the local features from the retinal image. The VGG19 comprises a series of 3×3 convolutional blocks and max pooling layers. Each convolutional block includes convolutional layers with 3×3 kernels, batch normalization layers, and Rectified Linear Unit (ReLU) activation layers. Mathematically, the output of a 3×3 convolutional block can be expressed as:

$$\mathbf{h}_{\text{conv}} = \text{ReLU}\left(\gamma \cdot \frac{\mathbf{W} * \mathbf{x} + \mathbf{b} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta\right), \tag{1}$$

where $\mathbf{x}$ represents the input feature map, and $\mathbf{W}$ and $\mathbf{b}$ denote the learnable weights and biases of the convolutional layer, respectively. The parameters γ, μ, σ, and β are the learnable parameters within the batch normalization layers, and ϵ is a small constant added for numerical stability.

The primary advantage of the convolution operation lies in its ability to efficiently capture local image features, such as edges and textures, by convolving the kernel with the feature map. Additionally, the batch normalization layer normalizes each mini-batch feature map to have zero mean and unit variance, which helps stabilize training. The ReLU activation layer introduces non-linearity, enabling the model to learn complex representations. Finally, the Max pooling layer reduces the spatial dimensions of feature maps by selecting the maximum value within each pooling window, preserving the most prominent features while discarding less significant ones.

B. Swin Transformer Module

The Swin Transformer module leverages the Swin Transformer V2 model to learn the global features from retinal image patches. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C denote the height, width, and the channel size of the image, respectively, the module first partitions the image $\mathbf{I}$ into patches $\mathbf{P} = \{\mathbf{p}_i \in \mathbb{R}^{P \times P \times C}, i = 1, 2, \ldots, \frac{H \times W}{P \times P}\}$, with each patch having a size of $P \times P$ ($P = 4$ in our study). Here, $\mathbf{p}_i$ represents the i-th patch. Next, the patches $\mathbf{P}$ are projected into dense vectors (*i.e.*, embeddings) $\mathbf{E} \in \mathbb{R}^{\frac{H \times W}{P \times P} \times D}$ using a linear layer, where D denotes the embedding dimension. Finally, these patch embeddings $\mathbf{E}$ are processed through a series of Swin Transformer blocks and Patch Merging blocks to extract the global features of the image $\mathbf{I}$.

In the Swin Transformer block, the input patch embeddings are divided into non-overlapping windows, each consisting of $M \times M$ patches. Let the j-th window be denoted as W_j, and the patch embeddings within this window are represented as $\mathbf{E}_{W_j} = \{\mathbf{e}_{ji}, i = 1, 2, \ldots, M \times M\}$. To capture the contextual information of $\mathbf{E}_{W_j}$, the Swin Transformer block employs a scaled cosine attention mechanism. The scaled cosine attention for $\mathbf{E}_{W_j}$ is formulated as:

$$\text{Att}(\mathbf{E}_{W_j}) = \text{Softmax}\left(\frac{\cos\left(\mathbf{Q}_{W_j}, \mathbf{K}_{W_j}^{\top}\right)}{\tau} + \mathbf{B}_{W_j}\right)\mathbf{V}_{W_j}, \tag{2}$$

where τ is a learnable scalar, $\mathbf{B}_{W_j}$ is the relative position deviation generated by a Multilayer Perceptron (MLP) component. $\mathbf{Q}_{W_j}$, $\mathbf{K}_{W_j}$, and $\mathbf{V}_{W_j}$ matrices are computed as:

$$\mathbf{Q}_{W_j} = \mathbf{E}_{W_j}\mathbf{W}_{W_j}^{q}, \mathbf{K}_{W_j} = \mathbf{E}_{W_j}\mathbf{W}_{W_j}^{k}, \mathbf{V}_{W_j} = \mathbf{E}_{W_j}\mathbf{W}_{j}^{v}, \tag{3}$$

where $\mathbf{W}_{W_j}^{q}$, $\mathbf{W}_{W_j}^{k}$, and $\mathbf{W}_{W_j}^{v}$ are learnable linear transformation matrices. Specifically, the Swin Transformer block contains two key sub-modules: window multi-head self-attention (W-MSA), followed by shifted window multi-head self-attention (SW-MSA) [18]. The SW-MSA operation performs a window shift before learning the representations of each window. This shift ensures that patches at the borders of the previous windows are repositioned to the center of the new windows, allowing for interactions between patch embeddings from different regions of the image. This process enhances the module's ability to capture global features. To further improve global feature representation, the Swin Transformer module introduces a Patch Merging block, which merges patch embeddings to expand the receptive field of the module.

C. Feature Enhancement Module

The Feature Enhancement (FE) module integrates the features extracted from the VGG and Swin Transformer modules into a unified representation for subsequent RW-ROP classification. To maximize the effectiveness of these fused features, we introduce an adaptive integration strategy that seamlessly combines local and global features. This strategy enables the FE module to emphasize key RW-ROP-related features, thereby enhancing the complementarity between local and global information for more accurate RW-ROP classification.

Let $\mathbf{h}_{\text{vgg}}$ and $\mathbf{h}_{\text{st}}$ represent the feature embeddings output by the VGG module and the Swin Transformer module, respectively. The enhanced feature embeddings, denoted as $\mathbf{h}_{\text{fe}}^{\text{vgg}} \in \mathbb{R}^{l_{\text{vgg}}}$ and $\mathbf{h}_{\text{fe}}^{\text{st}} \in \mathbb{R}^{l_{\text{st}}}$, are derived as follows:

$$\begin{aligned} \mathbf{h}_{\text{fe}}^{\text{vgg}} &= \text{MLP}_{\text{vgg}}\left(\text{MaxPool}\left(\mathbf{h}_{\text{vgg}}\right)\right), \\ \mathbf{h}_{\text{fe}}^{\text{st}} &= \text{MLP}_{\text{st}}\left(\mathbf{h}_{\text{st}}\right). \end{aligned} \tag{4}$$

It is important to note that the MLP further reduces the size of the feature vector to a predetermined dimension while retaining the features relevant to the RW-ROP classification task. The output of the FE module can be expressed as:

$$\mathbf{h}_{\text{fe}} = \text{Concat}\left(\mathbf{h}_{\text{fe}}^{\text{vgg}}, \mathbf{h}_{\text{fe}}^{\text{st}}\right), \tag{5}$$

where $\mathrm{Concat}(\cdot,\cdot)$ represents the embedding concatenation operation. The adaptive feature integration strategy sets the ratio $R = l_{\text{st}} \colon l_{\text{vgg}}$, where l_{st} and l_{vgg} denote the dimensions of $\mathbf{h}_{\text{fe}}^{\text{st}}$ and $\mathbf{h}_{\text{fe}}^{\text{vgg}}$, respectively. Here, R is a hyperparameter that controls the balance between global and local features, ensuring they complement each other effectively in the subsequent RW-ROP classification.

D. Classification Module

Finally, a classification module is employed to identify RW-ROP. The probability of the input image being classified as i-th class label is calculated by the $Softmax$ function as follows:

$$p(\hat{y} = i|\mathbf{h}_{\text{fe}}) = \frac{\exp(\mathbf{W}_i \mathbf{h}_{\text{fe}})}{\sum_{k=1}^{K} \exp(\mathbf{W}_k \mathbf{h}_{\text{fe}})}, \quad (6)$$

where $\mathbf{W}_k$ is the weight of the k-th class. It should be noted that in our study, $K = 2$ because we only have RW-ROP and normal cases.

E. Overall Loss Function

To train our proposed binary classification network, we employ a weighted cross-entropy (WCE) loss function to address the class imbalance between positive and negative samples [19]. The overall loss function is defined as:

$$\mathcal{L}_{\text{final}} = \omega_{\text{vgg}} \mathcal{L}_{\text{vgg}} + \omega_{\text{st}} \mathcal{L}_{\text{st}} + \omega_{\text{cls}} \mathcal{L}_{\text{cls}}, \quad (7)$$

where $\mathcal{L}_{\text{vgg}}$, $\mathcal{L}_{\text{st}}$ and $\mathcal{L}_{\text{cls}}$ correspond to the loss functions of the VGG module, the Swin Transformer (ST) module, and the classification module, respectively. The coefficients ω_{vgg}, ω_{st} and ω_{cls} are hyperparameters that balance the contributions of each individual loss function. Each loss function is computed using a weighted cross-entropy loss, which is defined as:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[\alpha y_i \log(p_i) + (1 - \alpha)(1 - y_i) \log(1 - p_i) \right]. \quad (8)$$

Here, N is the total number of input images, y_i denotes the ground truth label for the i-th retinal image, and p_i represents the predicted probability that the i-th image is classified as RW-ROP. The parameter α is a weight assigned to the RW-ROP class to handle class imbalance. During training, the combined loss from the VGG module, ST module, and classification module serves as the overall supervision signal, guiding the network to optimize both local and global feature extraction for RW-ROP classification.

III. Experiments and Results

A. Experimental Setting

Dataset. The dataset used in this study was sourced from the Telemedicine Methods for Evaluating Acute Retinopathy of Prematurity (e-ROP) study [20]. It comprises retinal images collected from 1,257 infants with birth weights under 1,251 grams, who were admitted to neonatal intensive care units at 13 North American centers. The retinal images were captured using a wide-angle fundus camera during scheduled diagnostic examinations conducted by ophthalmologists. Each retinal image was labeled as either RW-ROP or normal by an experienced ophthalmologist. In total, we utilized 7,905 center-view retinal images, including 814 RW-ROP images and 7,091 normal images.

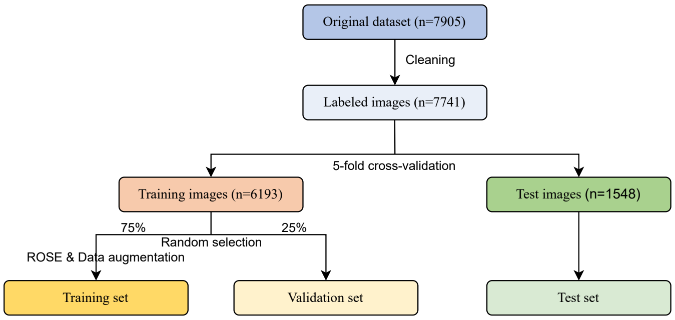

Fig. 2. The data preprocessing workflow for training and evaluation of the RW-ROP classification model.

Data Preprocessing. The overall data preprocessing workflow is illustrated in Fig. 2. The process began with meticulous dataset cleaning, which included removing damaged retinal images and verifying each image's label against its corresponding clinical examination. Subsequently, all retinal images were resized to 224×224 pixels. To enhance image details, Contrast Limited Adaptive Histogram Equalization (CLAHE) [21] was applied to the green channel of each retinal image. For model evaluation, we performed 5-fold cross-validation [22], dividing the labeled images into a training set (80% of the data) and a testing set (20% of the data). The training set was further split into training and validation subsets in a $3\colon 1$ ratio, while preserving the original class distribution. To address the significant class imbalance in the training set, we employed the Random Over Sampling Examples (ROSE) [23] technique to equalize the number of images in each class. Additionally, data augmentation techniques, such as image flipping, cropping, and scaling were applied to the training set to improve the model's generalization capability.

Baselines. We carefully selected 15 deep learning models as baseline comparisons, including Densenet201 [14], Inception-BN [24], Inception-V3 [9], Inception-V4 [25], Xception [26], ResNet50 [27], ResNet101 [17], ResNet152 [11], EfficientNet-V2 [13], VGG16 [28], VGG19 [10], ViT [29], MaxViT [30], Swin Transformer (Swin-T) [17], and ResNet50-MaxViT [16]. These models encompass both CNN-based architectures, such as Inception, ResNet, EfficientNet-V2, and VGG, as well as Transformer-based models like ViT, MaxViT, and Swin-T. All of these models have been employed in ROP-related research, demonstrating strong performance in detecting the disease. By including a diverse set of models with different architectural foundations, we aim to provide a comprehensive evaluation of the effectiveness of our proposed method against state-of-the-art approaches in the field.

Implementation Details. We implemented and evaluated the VGG-ST model, along with other baseline models, using the PyTorch framework. All models were pre-trained on the ImageNet dataset [31], and trained and tested on an NVIDIA

TABLE I
COMPARISON OF RW-ROP DISEASE CLASSIFICATION PERFORMANCE (MEAN ± STD). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Models	AUC	Sensitivity	Specificity
Densenet201 [14]	87.9±1.9	68.7±3.9	87.3±2.3
Inception-BN [24]	81.6±1.1	71.8±3.3	79.6±2.4
Inception-V3 [9]	88.6±2.1	51.7±5.3	**95.6±0.7**
Inception-V4 [25]	88.1±5.8	69.7±5.2	90.2±2.8
Xception [26]	87.4±1.2	68.2±4.2	89.8±2.4
ResNet50 [27]	87.8±1.3	76.7±2.1	84.8±1.4
ResNet101 [17]	88.2±1.5	75.5±2.3	84.7±3.8
ResNet152 [11]	83.2±2.3	68.9±2.7	83.3±1.8
EfficientNet-V2 [13]	87.2±2.7	61.3±4.9	94.5±0.8
VGG16 [28]	85.8±1.6	77.7±1.6	78.2±5.4
VGG19 [10]	86.3±2.1	78.9±2.3	78.5±6.6
ViT [29]	83.1±1.9	65.2±5.2	82.4±2.2
MaxViT [30]	84.8±1.1	67.8±1.7	85.6±2.4
Swin-T [17]	84.9±0.7	70.1±4.8	85.7±4.4
ResNet50-MaxViT [16]	87.5±1.2	79.6±2.7	82.8±2.9
VGG-ST (ours)	**90.3±1.4**	**84.7±3.1**	83.7±1.6

RTX A5000 GPU. The model's learnable parameters were optimized using the Adam optimizer [32]. To fine-tune the VGG-ST model, we conducted a grid search to optimize several hyperparameters, including the learning rate, batch size, number of training epochs, feature integration ratio R, the coefficients ω_{vgg}, ω_{st} and ω_{cls} in the final loss function, and the class weight α. For all models, the learning rate was explored within the range of 0.00001 to 0.01, with batch sizes of 4, 8, 16, 32, and 64 tested, and a maximum of 100 epochs evaluated. Training was stopped early if there was no improvement in validation set performance within 10 epochs. Specifically for the VGG-ST model, the feature integration ratio R was varied between 0.5 and 3 in increments of 0.5, while the coefficients ω_{vgg}, ω_{st} and ω_{cls} were varied from 0 to 1 with a step size of 0.5. The class weight α was adjusted from 0 to 1 in increments of 0.1 to achieve optimal performance.

Evaluation Metrics. We evaluate the model's performance using three widely adopted metrics in medical imaging analysis: Area Under the Curve (AUC), sensitivity, and specificity. Sensitivity, or true positive rate, measures the model's ability to correctly identify patients with the disease, ensuring that affected individuals are detected. Specificity, or true negative rate, assesses the model's accuracy in correctly excluding patients without the disease, minimizing false positives. The AUC score integrates both sensitivity and specificity into a single metric, reflecting the overall diagnostic performance of the model. A higher AUC score indicates superior discrimination between patients with and without the disease. Together, these metrics provide a comprehensive evaluation of the model's effectiveness in diagnosing RW-ROP.

B. Disease Classification Performance

Table I presents the experimental results comparing the RW-ROP disease classification performance of various deep learning models. Among the models tested, our proposed VGG-ST model outperforms all baselines across key metrics. Specifically, VGG-ST achieves the highest AUC of 90.3±1.4, indicating its superior overall performance in distinguishing between disease and normal classes. Additionally, it exhibits the best sensitivity (84.7±3.1), demonstrating its ability to correctly identify true positives, which is critical in medical diagnosis to ensure that cases of RW-ROP are not missed. Although the specificity (83.7±1.6) of VGG-ST is slightly lower than that of some other models, such as Inception-V3 (95.6±0.7) and EfficientNet-V2 (94.5±0.8), it remains highly competitive. Importantly, while these models achieve higher specificity, their sensitivity is much lower, with Inception-V3 at 51.7±5.3 and EfficientNet-V2 at 61.3±4.9, which could lead to missed diagnoses. These findings underscore the effectiveness of our model in capturing both local and global features through its hybrid architecture, resulting in a robust and reliable tool for RW-ROP classification. VGG-ST maintains a crucial balance between minimizing false positives and preserving high sensitivity, a vital consideration in medical diagnostics where both false positives and false negatives can have significant consequences. The superior performance of VGG-ST across these critical metrics highlights its potential to improve clinical outcomes in the diagnosis of RW-ROP.

C. Model Interpretability and Feature Visualization

Figs. 3 and 4 demonstrate the effectiveness of feature extraction and attention mechanisms in the VGG-ST model compared to the other two baseline models with higher sensitivity, VGG19 and ResNet50-MaxViT. In Fig. 3, the t-distributed Stochastic Neighbor Embedding (t-SNE) [33] visualizations reveal that while the original retinal images show minimal differences between normal and RW-ROP cases, the feature vectors extracted by VGG-ST display much clearer and more distinct clustering of RW-ROP samples. This enhanced feature separation strongly suggests that VGG-ST is particularly effective at capturing the discriminative features necessary for accurate classification.

Fig. 4 presents the attention maps generated from the feature maps of VGG19, ResNet50-MaxViT, and VGG-ST for both normal and RW-ROP retinal images. These attention maps, created using Gradient-weighted Class Activation Mapping (Grad-CAM) [34], highlight the regions that each model prioritizes during classification. The results show that VGG19 tends to focus on a small, localized area, potentially overlooking other critical regions, while ResNet50-MaxViT exhibits a more scattered focus across the image. In contrast, VGG-ST more precisely targets relevant regions, such as blood vessel branches, which are crucial for distinguishing between normal and RW-ROP eyes. These findings highlight VGG-ST's ability not only to extract meaningful features but also to focus attention on the most diagnostically significant areas, resulting in improved RW-ROP classification performance.

D. Ablation Studies

We conducted two sets of ablation studies to evaluate the effectiveness of key components in our VGG-ST model and the impact of our data preprocessing methods. The first

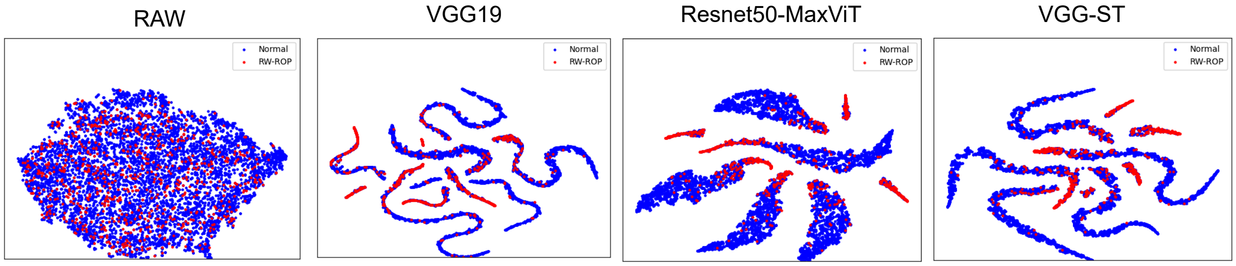

Fig. 3. The t-SNE visualization of raw retinal images and feature embeddings extracted by VGG19, ResNet50-MaxViT, and VGG-ST.

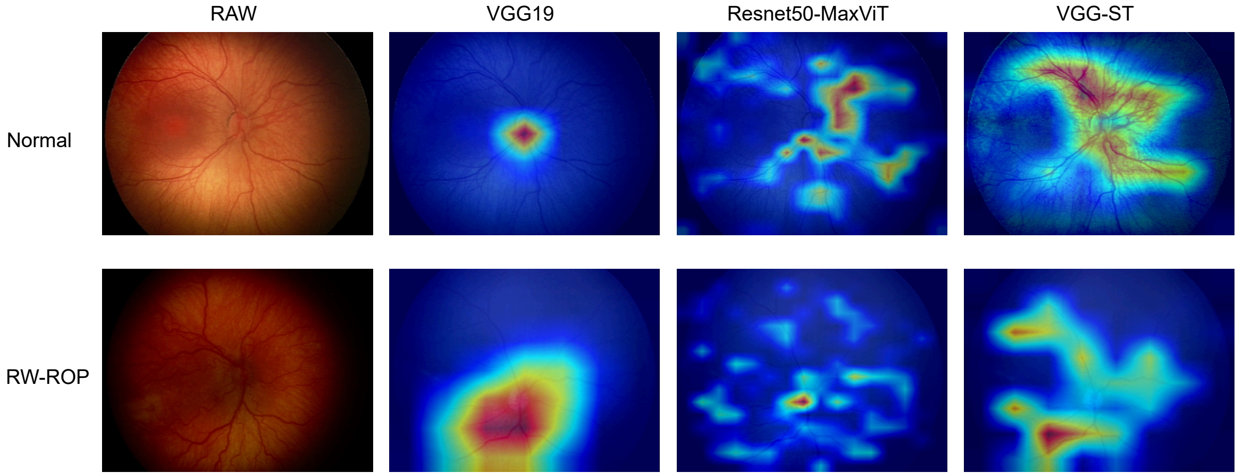

Fig. 4. The attention maps of VGG19, ResNet50-MaxViT, and VGG-ST for normal and RW-ROP retinal images.

TABLE II
ABLATION STUDY RESULTS FOR KEY COMPONENTS (MEAN ± STD). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model Configurations	AUC	Sensitivity	Specificity
w/o VGG	84.9±0.7	70.1±4.8	85.7±4.4
w/o ST	86.3±2.1	78.9±2.3	78.5±6.6
w/o FE	88.2±1.2	79.9±4.0	**85.9±2.1**
Full VGG-ST	**90.3±1.4**	**84.7±3.1**	83.7±1.6

TABLE III
ABLATION STUDY RESULTS FOR DATA PREPROCESSING METHODS (MEAN ± STD). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Data Preprocessing	AUC	Sensitivity	Specificity
Raw	48.6±3.5	0.0±0.0	**1.0±0.0**
Aug	88.7±1.9	75.4±2.6	86.6±2.1
ROSE	88.6±1.5	73.9±6.5	88.7±2.8
Aug & ROSE	89.6±1.3	80.8±4.0	85.0±4.6
CLAHE & Aug	89.0±1.2	79.0±2.3	85.2±2.9
CLAHE & ROSE	87.7±0.9	72.7±3.7	87.7±2.3
CLAHE & Aug & ROSE	**90.3±1.4**	**84.7±3.1**	83.7±1.6

ablation study assesses the contribution of each major component in the VGG-ST model, including the VGG module, Swin Transformer (ST) module, and Feature Enhancement (FE) module. Table II presents the results of different model configurations: “w/o VGG” denotes the model without the VGG module, “w/o ST” denotes the model without the ST module, and “w/o FE” denotes the model without the FE module. The results show that each component plays a crucial role in the overall performance of the model. Specifically, removing the VGG module led to a significant drop in capturing local features, while excluding the ST module weakened the model’s ability to gather global contextual information. The absence of the FE module, which integrates these features, resulted in poorer performance, demonstrating the necessity of effective feature fusion. The full VGG-ST model, which includes all components, shows the best results, underscoring the importance of integrating both local and global feature extraction as well as the feature enhancement step. This study highlights that each module’s contribution is essential for achieving optimal RW-ROP classification performance.

The second ablation study evaluates the impact of our data preprocessing methods on the VGG-ST model’s performance. Table III presents the effects of various image preprocessing techniques on VGG-ST, where “Aug” denotes the data augmentation techniques used to increase the sample size of each class while achieving class balance. Training the model on the original dataset yielded sub-optimal results due to the dataset’s limited size and significant class imbalance. In contrast, the model trained with all preprocessing techniques exhibited superior performance, as indicated in the bottom row of Table III. Among these techniques, ROSE and data augmentation effectively addressed the challenges of class imbalance and the limited number of images in the e-ROP dataset, while CLAHE enhanced the RW-ROP-related features in the retinal images. The ablation study demonstrates the

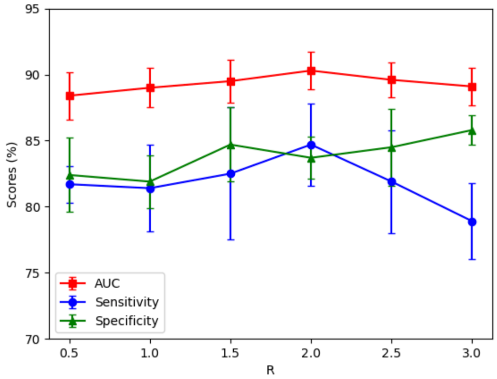

Fig. 5. Visualization of hyperparameter sensitivity analysis for R in the adaptive integration strategy.

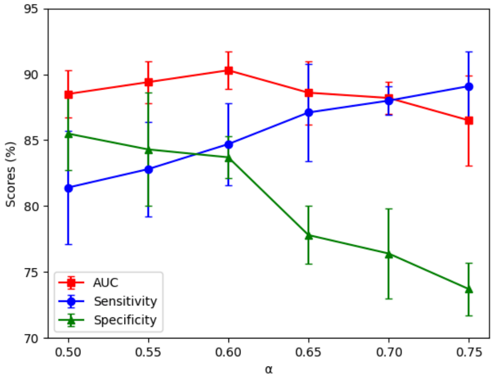

Fig. 6. Visualization of hyperparameter sensitivity analysis for the class weight α in the loss function $\mathcal{L}_{\text{WCE}}$.

TABLE IV
HYPERPARAMTER ANALYSIS OF THE COEFFICIENTS IN THE FINAL LOSS (MEAN ± STD). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

ω_{vgg} ω_{st} ω_{cls}	AUC	Sensitivity	Specificity
0 0 1	89.8±1.3	82.3±4.3	83.1±4.6
0.5 0 1	88.1±1.2	82.7±2.1	81.7±2.7
1 0 1	86.8±1.8	**85.9±1.3**	77.4±2.2
0 0.5 1	89.1±1.4	81.0±2.3	83.3±2.8
0 1 1	89.3±1.5	80.4±3.2	85.3±3.4
1 1 1	89.5±1.6	78.6±2.6	**85.4±2.4**
1 1 0.5	88.9±1.7	80.9±6.1	82.4±6.1
1 0.5 1	88.6±1.2	80.5±5.3	81.9±3.6
0.5 1 1	**90.3±1.4**	84.7±3.1	83.7±1.6

necessity of these preprocessing steps. These findings suggest that training the VGG-ST on a larger and more balanced retinal image dataset would likely maximize its potential and further improve diagnostic performance.

E. Hyperparameter Analysis

In this section, we analyze the impact of key hyperparameters on the VGG-ST model performance. Fig. 5 presents the results of our investigation into the adaptive feature integration strategy, focusing on how the ratio of global features to local features, denoted as R, influences model performance. As R varies from 0.5 to 3, the model's AUC, sensitivity, and specificity show distinct patterns. Notably, the model achieves its highest sensitivity (84.7±3.1%) when R is set to 2, indicating that a higher proportion of global features is beneficial for RW-ROP classification. This result suggests that in the context of retinal images from premature infants, which often suffer from blurriness and lack distinct local features, emphasizing global features helps the model better capture the overall context necessary for accurate classification. These findings highlight the importance of carefully tuning the feature integration ratio to optimize performance for specific medical tasks.

In addition, we examined the effect of the class weight α in the weighted cross-entropy loss function, which modulates the penalty for misclassifying different classes. Fig. 6 shows how varying α from 0.5 to 0.75 impacts the model's AUC sensitivity, and specificity. As α increases, the model's specificity improves, reflecting a higher accuracy in correctly identifying normal retinal images. However, this improvement comes with a decrease in sensitivity, indicating a reduced ability to detect RW-ROP cases. The analysis reveals that setting α to 0.6 achieves the best balance, resulting in the highest AUC and a relatively balanced trade-off between sensitivity and specificity. This balance is crucial in medical diagnostics, where both false positives and false negatives can have significant consequences. By adjusting α, users can fine-tune the model's performance to meet the specific needs of the clinical application, ensuring the most appropriate trade-off for the given medical context.

Furthermore, we investigated the contributions of each component in the overall loss function by controlling the hyperparameters ω_{vgg}, ω_{st} and ω_{cls}. Table IV presents the results of this hyperparameter analysis, where different combinations of these coefficients were evaluated for their impact on the model's AUC, sensitivity, and specificity. The results show that the combination $\omega_{\text{vgg}} = 0.5$, $\omega_{\text{st}} = 1$, and $\omega_{\text{cls}} = 1$ yields the highest AUC (90.3±1.4), suggesting that balancing the contributions of the VGG and Swin Transformer components with a strong emphasis on the classification component optimizes the model's overall performance. Interestingly, when $\omega_{\text{vgg}} = 1$, $\omega_{\text{st}} = 0$, and $\omega_{\text{cls}} = 1$, the model achieves the highest sensitivity (85.9±1.3), indicating that a stronger focus on the VGG component improves the model's ability to detect RW-ROP cases. However, this comes at the cost of specificity, which decreases, as seen with a specificity of 77.4±2.2. The results highlight the importance of carefully tuning these coefficients to balance sensitivity and specificity according to the specific demands of the clinical application.

IV. CONCLUSION

In this study, we introduced the VGG-ST model, a hybrid deep learning architecture that combines the strengths of VGG and the Swin Transformer to automatically detect RW-ROP from center-view retinal images. The VGG-ST model excels at capturing both local and global features, achieving high AUC and sensitivity while maintaining competitive specificity, underscoring its effectiveness as a reliable ROP screening tool. Given the growing burden of ROP on healthcare systems, integrating the VGG-ST model could significantly enhance decision-making in patient management and contribute to the development of primary care-based ROP screening programs for broader populations. This work adds to the ongoing efforts toward the clinical application of AI-driven ROP detection, offering the potential to alleviate the strain on specialized healthcare resources and improve early intervention outcomes. In future research, we plan to incorporate additional data modalities, such as demographic information and longitudinal data, into the model to further enhance its diagnostic performance and adaptability across diverse patient populations.

Acknowledgment

This study was in part supported by the National Institutes of Health (R21EY025686 and R21EY034179), National Science Foundation (MRI-2215789, IIS-2319451), and Lehigh University (Accelerator-S00010293 and CORE-001250).

References

[1] C. Gilbert, "Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control," *Early human development*, vol. 84, no. 2, pp. 77–82, 2008.

[2] A. L. Ells, J. M. Holmes, W. F. Astle, G. Williams, D. A. Leske, M. Fielden, B. Uphill, P. Jennett, and M. Hebert, "Telemedicine approach to screening for severe retinopathy of prematurity: a pilot study," *Ophthalmology*, vol. 110, no. 11, pp. 2113–2117, 2003.

[3] W. M. Fierson, M. F. Chiang, W. Good, D. Phelps, J. Reynolds, S. L. Robbins, D. J. Karr, G. E. Bradford, K. Nischal, J. Roarty *et al.*, "Screening examination of premature infants for retinopathy of prematurity," *Pediatrics*, vol. 142, no. 6, 2018.

[4] R. Hardy, W. Good, V. Dobson, E. Palmer, B. Tung, and D. Phelps, "Early treatment for retinopathy of prematurity cooperative grouprevised indications for the treatment of retinopathy of prematurity. results of the early treatment for retinopathy of prematurity randomized trial," *Arch Ophthalmol*, vol. 121, pp. 1684–94, 2003.

[5] G. E. Q. on behalf of the e ROP Cooperative Group, "Telemedicine approaches to evaluating acute-phase retinopathy of prematurity: study design," *Ophthalmic epidemiology*, vol. 21, no. 4, pp. 256–267, 2014.

[6] C. Wu, R. A. Petersen, and D. K. VanderVeen, "Retcam imaging for retinopathy of prematurity screening," *Journal of American Association for Pediatric Ophthalmology and Strabismus*, vol. 10, no. 2, pp. 107–111, 2006.

[7] K. Altersitz and M. Piechocki, "Survey: Physicians being driven away from rop treatment," *Ocular Surgery News*, 2006.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[9] Z. Tan, S. Simkin, C. Lai, and S. Dai, "Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease," *Translational vision science & technology*, vol. 8, no. 6, pp. 1–11, 2019.

[10] Y.-P. Huang, S. Vadloori, H.-C. Chu, E. Y.-C. Kang, W.-C. Wu, S. Kusaka, and Y. Fukushima, "Deep learning models for automated diagnosis of retinopathy of prematurity in preterm infants," *Electronics*, vol. 9, no. 9, p. 1444, 2020.

[11] J. S. Chen, A. S. Coyner, S. Ostmo, K. Sonmez, S. Bajimaya, E. Pradhan, N. Valikodath, E. D. Cole, T. Al-Khaled, R. P. Chan *et al.*, "Deep learning for the diagnosis of stage in retinopathy of prematurity: accuracy and generalizability across populations and cameras," *Ophthalmology Retina*, vol. 5, no. 10, pp. 1027–1035, 2021.

[12] A. Bai, C. Carty, and S. Dai, "Performance of deep-learning artificial intelligence algorithms in detecting retinopathy of prematurity: A systematic review," *Saudi Journal of Ophthalmology*, vol. 36, no. 3, pp. 296–307, 2022.

[13] B. Ebrahimi, D. Le, M. Abtahi, A. K. Dadzie, A. Rossi, M. Rahimi, T. Son, S. Ostmo, J. P. Campbell, R. Paul Chan *et al.*, "Assessing spectral effectiveness in color fundus photography for deep learning classification of retinopathy of prematurity," *Journal of Biomedical Optics*, vol. 29, no. 7, pp. 076 001–076 001, 2024.

[14] S. K. Wagner, B. Liefers, M. Radia, G. Zhang, R. Struyven, L. Faes, J. Than, S. Balal, C. Hennings, C. Kilduff *et al.*, "Development and international validation of custom-engineered and code-free deep-learning models for detection of plus disease in retinopathy of prematurity: a retrospective study," *The Lancet Digital Health*, vol. 5, no. 6, pp. e340–e349, 2023.

[15] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *CVPR*. IEEE, 2015, pp. 5353–5360.

[16] J. Zhao, H. Lei, H. Xie, P. Li, Y. Liu, G. Zhang, and B. Lei, "Dual-branch attention network and swin spatial pyramid pooling for retinopathy of prematurity classification," in *ISBI*, 2023, pp. 1–4.

[17] V. R. Sankari, U. Snekhalatha, S. Alasmari, and S. M. Aslam, "Automated detection of retinopathy of prematurity using quantum machine learning and deep learning techniques," *IEEE Access*, 2023.

[18] A. G. Alharthi and S. M. Alzahrani, "Do it the transformer way: a comprehensive review of brain and vision transformers for autism spectrum disorder diagnosis and classification," *Computers in Biology and Medicine*, p. 107667, 2023.

[19] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE access*, vol. 8, pp. 4806–4813, 2019.

[20] G. E. Quinn, G.-s. Ying, E. Daniel, P. L. Hildebrand, A. Ells, A. Baumritter, A. R. Kemper, E. B. Schron, K. Wade, e ROP Cooperative Group *et al.*, "Validity of a telemedicine system for the evaluation of acute-phase retinopathy of prematurity," *JAMA ophthalmology*, vol. 132, no. 10, pp. 1178–1184, 2014.

[21] S. M. Pizer, "Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group," in *Proceedings of the first conference on visualization in biomedical computing*, vol. 337, 1990, p. 2.

[22] J. M. Gorriz, F. Segovia, J. Ramirez, A. Ortiz, and J. Suckling, "Is k-fold cross validation the best model selection method for machine learning?" *arXiv preprint arXiv:2401.16407*, 2024.

[23] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and synthetic over-sampling approach to resolve data imbalance in classification," *International Journal of Artificial Intelligence Research*, vol. 4, no. 2, pp. 86–94, 2020.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*. IEEE, 2015, pp. 1–9.

[25] G. Chen, J. Zhao, R. Zhang, T. Wang, G. Zhang, and B. Lei, "Automated stage analysis of retinopathy of prematurity using joint segmentation and multi-instance learning," in *MICCAI-OMIA*. Springer, 2019, pp. 173–181.

[26] O. Attallah, "Diarop: automated deep learning-based diagnostic tool for retinopathy of prematurity," *Diagnostics*, vol. 11, no. 11, p. 2034, 2021.

[27] E. Ndunge Mutua, B. Shibwabo Kasamani, and C. Reich, "Retinopathy of prematurity disease diagnosis using deep learning," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 1097–1110, 2024.

[28] Y. Zhang, L. Wang, Z. Wu, J. Zeng, Y. Chen, R. Tian, J. Zhao, and G. Zhang, "Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images," *IEEE access*, vol. 7, pp. 10 232–10 241, 2018.

[29] H. Lei, J. Zhao, H. Xie, Y. Liu, G. Zhang, and B. Lei, "Dual-branch feature interaction network with structure information learning for retinopathy of prematurity classification," in *BIBM*, 2023, pp. 1230–1235.

[30] Y. Liu, H. Xie, X. Zhao, J. Tang, Z. Yu, Z. Wu, R. Tian, Y. Chen, M. Chen, D. P. Ntentakis *et al.*, "Automated detection of nine infantile fundus diseases and conditions in retinal images using a deep learning system," *EPMA Journal*, vol. 15, no. 1, pp. 39–51, 2024.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] G. Hinton and L. Van Der Maaten, "Visualizing data using t-sne journal of machine learning research," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.