A Lightweight Explainable Hybrid Deep Learning Approach for Early Skin Disease Detection

Mohammad Barr
Department of Electrical Engineering, College of Engineering
Northern Border University, Arar 91431, Saudi Arabia

***Abstract*—Early detection of skin pathology is critical for patient survival and treatment effectiveness, particularly in the case of aggressive malignancies such as melanoma. In this study, we propose a lightweight and explainable hybrid deep learning approach for early identification of skin diseases. The proposed approach combines a transformer-based module to capture the global contextual dependencies with a CNN for an efficient local feature extraction. In addition, Grad-CAM approach is used to increase the interpretability of the model and offer visual explanations for the forecasts. We evaluate the suggested technique on a publically accessible dermoscopic benchmark and achieve 93% accuracy, 92% precision, 91% recall and 92% F1-score, outperforming numerous mainstream designs. The experimental results imply that the proposed approach achieves a good trade-off between accuracy, computational economy, and interpretability for real-world and resource-limited medical applications.**

***Keywords*—Medical image analysis; skin disease detection; explainable AI; hybrid model; Vision Transformer**

I. Introduction

Skin diseases are one of the most common health problems in the world, affecting millions of people every year. This can be anything from benign acne and eczema to serious melanoma. Early detection is very crucial, especially for malignant skin cancers because it increases the survival rates and decreases the treatment costs. However, the accurate diagnosis of these conditions is still a challenge due to the similarities between different skin lesions and variation in the imaging conditions such as lighting, resolution, and skin tone [1].

Until recently, the clinical diagnosis of skin diseases has been based on the clinical examination and the dermoscopic examination of an experienced expert. This technique, while successful, is fundamentally subjective and relies on clinical experience, which may result in inter-observer variability and delayed diagnosis. Moreover, the number of dermatologists is not enough in many places, particularly in developing nations, generating a tremendous need for automated, scalable, and reliable diagnostic solutions [1], [2].

Recently, deep learning has been successfully used in medical image analysis, in particular, in the detection of skin diseases. CNNs have demonstrated themselves to be quite effective in extracting hierarchical features from the pictures to appropriately identify the skin lesions. A lot of research has demonstrated a significant increase in the accuracy of diagnosis by using deep learning algorithms ([2], [3]). These models automatically learn discriminative features like texture, color distribution, and the borders of lesions without the need for handcrafted feature engineering.

CNN-based techniques have been successful, but suffer from several limitations. In particular, CNNs are meant to capture local spatial characteristics by using convolution processes with a narrow receptive field. Deeper designs may partly overcome this constraint but generally fail to simulate long-range relationships inside images. This is especially important in medical imaging because small global patterns may be important for distinguishing different disease categories [3].

Recently, transformer-based architectures have been used for computer vision applications to solve these problems. After their success in natural language processing, Transformers use self-attention processes to model connections between all elements in the incoming data. Vision Transformer (ViT) models split pictures into patches and treat them as sequences to collect global contextual information. Recent research has shown that transformer-based models are useful for classifying skin diseases, and they achieve better performance than typical CNNs [4], [5].

However, transformer models also have certain problems. They are computationally complicated and need large-scale datasets for efficient training, which makes them less suited for real-world and resource-constrained applications. Also, CNN and transformer models typically act as black-box systems and lack interpretability, which is a critical problem in medical applications where transparency and trust are important [4].

Recently, hybrid deep learning architectures that blend CNNs and transformers have been proposed to solve the above shortcomings. These methods leverage CNNs for effective local feature extraction and transformers to model global dependencies. This combination enables the model to benefit from the benefits of both paradigms and achieve better performance and resilience. Hybrid CNN-Transformer models have shown their efficacy in skin disease diagnosis in recent research, performing better than stand-alone designs in terms of accuracy [6], [7].

In addition to speed, interpretability has become a key criterion in medical AI systems. Clinicians need to grasp the rationale behind the model's predictions to trust and use these technologies in their practice. Explainable Artificial Intelligence (XAI) solutions try to tackle this problem by shedding light on the decision-making phase of deep learning approaches. Among such approaches, Grad-CAM has drawn substantial attention for its capacity to provide visual explanations by highlighting critical locations in the input picture. Recent works successfully utilized XAI approaches to skin disease categorization, displaying enhanced transparency and clinical relevance [3], [8].

However, while these advances have been made, most of the current systems focus either on improving the accuracy or on increasing the interpretability, but few of them try to improve both while maintaining high computational efficiency. Nor has there been as much attention to early-stage identification, when signs of illness are less severe and harder to detect. This gap highlights the need for a unifying framework that encompasses accuracy, efficiency, and explainability. To address such problems, this study proposes a lightweight explainable hybrid intelligent system for early skin disease diagnosis. The proposed method consists of a CNN-based feature extractor and a transformer-based module, which can successfully collect local and global features. The architecture also has Grad-CAM for providing visual explanations of model predictions, thus enhancing interpretability and confidence.

The principal contributions of this study are listed below:

- We propose a lightweight hybrid CNN-Transformer architecture for early skin disease detection, combining MobileNetV2 for local feature extraction with a ViT for global context modeling.
- We integrate Grad-CAM, an explainable AI technique, into the hybrid framework to provide visual heatmaps that highlight lesion regions influencing model predictions.
- We optimize the proposed hybrid model for computational efficiency.
- We conduct a comprehensive evaluation on benchmark datasets with a comparative analysis.

The rest of this work is structured as follows: In Section II, we review existing approaches and recent studies related to skin disease detection. The proposed hybrid CNN-Transformer technique is discussed in Section III. The experimental setups are described in Section IV. Quantitative and qualitative data are presented and analyzed in Section V. In Section VI, we conclude the study and outline several directions for future research.

II. Related Work

The development of AI has significantly improved the ability of automated diagnostic systems for skin diseases. In particular, deep learning algorithms have achieved great success in medical image analysis owing to their capacity to automatically generate discriminative features from raw data. This section introduces the most important literature, separated into CNN-based methods, transformer-based models, hybrid architectures, and explainable AI strategies.

A. CNN-Based Methods for Skin Disease Detection

CNNs have become one of the cornerstones for automated skin lesion analysis, since they are capable of learning hierarchical spatial representations directly from raw images. These models learn fine-grained patterns well, such as textural differences, color distribution, and lesion edges [1]. More and more, current research has concentrated on the improvement of CNN performance using transfer learning and complex optimization strategies. For example, Zhang et al. [1] provided a detailed assessment of machine and deep learning techniques for dermatological diagnosis and spoke about the efficiency of CNN-based systems in achieving high discrimination accuracy. Likewise, Liu et al. [9] performed a thorough evaluation of deep learning applications in dermatology, where convolutional architectures are gradually taking over clinical procedures. Khan et al. [10] showed that transfer learning-based CNNs outperform traditional machine learning classifiers for skin lesion identification.

B. Transformer Architectures

Recently, the computer vision community has started to incorporate transformer-based models initially created for natural language processing to address the receptive field constraint of CNNs. The self-attention mechanism is at the core of these systems, which allows direct modeling of long-range dependencies and global interactions among picture patches [4]. The significance of transformers in medical imaging activities has been proved by many research. Mohan et al. [11] proposed a transformer-based deep learning pipeline for skin disease classification, which resulted in superior performance due to enhanced global context awareness. Similarly, Dagnaw et al. [12] deployed a Vision Transformer for melanoma detection and included explainability tools to enhance the interpretability of the model.

C. Hybrid CNN-Transformer Models

Recent works have been done on hybrid architectures that mix the two approaches to take use of the complementary qualities of CNNs and transformers. CNNs are utilized in such models for effective local feature extraction and transformers for capturing global contextual linkages. Ali et al. [13] introduced a hybrid CNN-Transformer architecture for skin disease identification, which produced better results than using just CNN or Transformer models. Their method underlines the need for combining local and global feature representations. Also, Mousa et al. [14] developed an attention-based hybrid CNN-Transformer model for medical picture classification, which showed better robustness and generalization. These investigations demonstrate that hybrid designs may achieve a good trade-off between accuracy and computing economy. However, even with their enhanced performance, many hybrid models still suffer from computational expense and lack of interpretability, which are crucial in medical applications.

D. Explainable AI for Skin Disease Detection

In medical AI applications, the raw prediction performance is not adequate. In order to trust and use a model in practice, clinicians need to grasp the rationale behind a given choice made by the model. This need has brought XAI to the forefront of the development of diagnostic systems. The goal of XAI approaches is to provide human-interpretable explanations of the internal decision processes of deep learning models, thereby opening the "black box" of these models. Of these approaches, Gradient-weighted Class Activation Mapping (Grad-CAM) has garnered particular momentum in dermatological imaging for its capacity to provide visual heatmaps detailing which parts of a skin lesion picture most impacted the model's output. Shah et al. [15] proposed a XAI-based framework for skin cancer identification and showed the Grad-CAMs' ability to locate diagnostic-relevant regions such as lesion boundaries,

pigmented networks, and vascular structures. Similarly, Badhon et al. [16] use Grad-CAM on multi-class skin disease classification and demonstrate that visual explanations using Grad-CAM considerably increase the transparency and clinical value of the models without any degradation in the classification accuracy.

III. Proposed Approach: Hybrid CNN–Transformer with Explainability

To detect skin diseases early, we propose a deep hybrid framework that combines CNNs with Vision Transformers in an explainable manner. This architecture efficiently captures both local spatial cues and global contextual information, yet remains computationally economical and interpretable thanks to integrated XAI. The input preprocessing, CNN-based feature extraction, transformer-based global modeling, classification head, and explainability module are the five primary components of the suggested system, as shown in Fig. 1. A lightweight CNN extracts discriminative spatial features, which are then processed by a transformer encoder to model long-range dependencies. Finally, predictions are generated and interpreted using visualization techniques.

A. Input Preprocessing and Data Augmentation

Each dermoscopic image I from the Skin Cancer (MNIST HAM10000) dataset is preprocessed to ensure input consistency and improve model generalization. All images are resized to a fixed resolution of $224\times224\times3$ pixels. Subsequently, pixel values are normalized channel-wise using the dataset mean μ and standard deviation σ:

$$I_{\text{norm}} = \frac{I - \mu}{\sigma} \tag{1}$$

The HAM10000 dataset comprises 10,000 dermoscopic images labeled into seven skin lesion categories [17]. As illustrated in Fig. 2, the class distribution is highly imbalanced: melanocytic nevi dominate with approximately 66.9% (6,705 images), while minority classes such as dermatofibroma (1.1%) and vascular lesions (1.4%) contain far fewer samples. The remaining classes include melanoma (11.1%), benign keratosis-like lesions (11%), basal cell carcinoma (5.1%), and actinic keratoses (3.3%).

Such imbalance can bias the classifier toward majority classes if left unaddressed. Therefore, we apply data augmentation to enhance model robustness and mitigate overfitting. The augmentation pipeline includes random rotations, horizontal and vertical flips, scaling, and cropping. These transformations increase the model's invariance to variations in orientation, size, and lighting conditions commonly encountered in dermoscopic practice. Moreover, augmentation helps to artificially expand the representation of minority classes, partially alleviating the class imbalance problem and leading to a more balanced and unbiased model.

The vast visual diversity and morphological complexity of skin lesions are illustrated in Fig. 3, which includes representative samples from each of the seven classes.

B. CNN-Based Feature Extraction

The normalized input image I_norm is processed using the MobileNetV2 backbone [18] [19] to produce a highly discriminative yet compact feature representation. This lightweight architecture is specifically engineered for efficiency, employing depthwise separable convolutions to significantly reduce computational complexity while preserving a robust feature extraction capability. The network begins with an initial convolutional layer, followed by a sequence of inverted residual blocks with linear bottlenecks, which form the core of the architecture. These blocks are designed to enhance feature reuse and improve gradient flow during training. Each inverted residual block can be mathematically expressed as:

$$\mathbf{y} = \mathbf{x} + \mathcal{F}(\mathbf{x}) \tag{2}$$

where, x denotes the input feature map and $F(\cdot)$ represents a series of transformations consisting of: Pointwise convolution, Depthwise convolution, and Pointwise linear projection.

This structure enables efficient feature learning by separating spatial and channel-wise computations, thereby preserving essential information while minimizing memory usage and computational cost. After passing through multiple stacked inverted residual blocks, the network produces a high-level feature map: $F \in \mathbf{R}^{7\times7\times1280}$.

To further refine the extracted features, a 1×1 convolution is applied, followed by global average pooling, which aggregates spatial information into a more compact representation. This refined feature representation serves as the input to the subsequent transformer module, enabling effective integration of local spatial features with global contextual modeling.

C. Transformer-Based Global Modeling

The proposed framework integrates a ViT-based encoder [20] to efficiently capture global contextual relationships and long-range dependencies within the extracted feature map. The feature map generated by the CNN backbone is first reshaped into a sequence of non-overlapping patches, denoted as:

$$\{x_1, x_2, \ldots, x_N\}$$

where, each patch represents a local region of the input image. Each patch is subsequently flattened and projected into a latent embedding space through a learnable linear transformation, expressed as $z_i = x_i W_e$, where W_e denotes the embedding matrix.

To preserve the spatial structure of the original image, positional embeddings are added to each token, resulting in $z_i' = z_i + p_i$, where p_i encodes the positional information. Furthermore, a special classification token is appended to the input sequence, serving as a global representation that aggregates information from all patches throughout the transformer layers.

The augmented sequence of embeddings is subsequently processed by a series of transformer encoder layers. Each encoder layer comprises a multi-head self-attention mechanism succeeded by a feed-forward neural network, incorporating layer normalization and residual connections to improve training stability and convergence. The self-attention mechanism

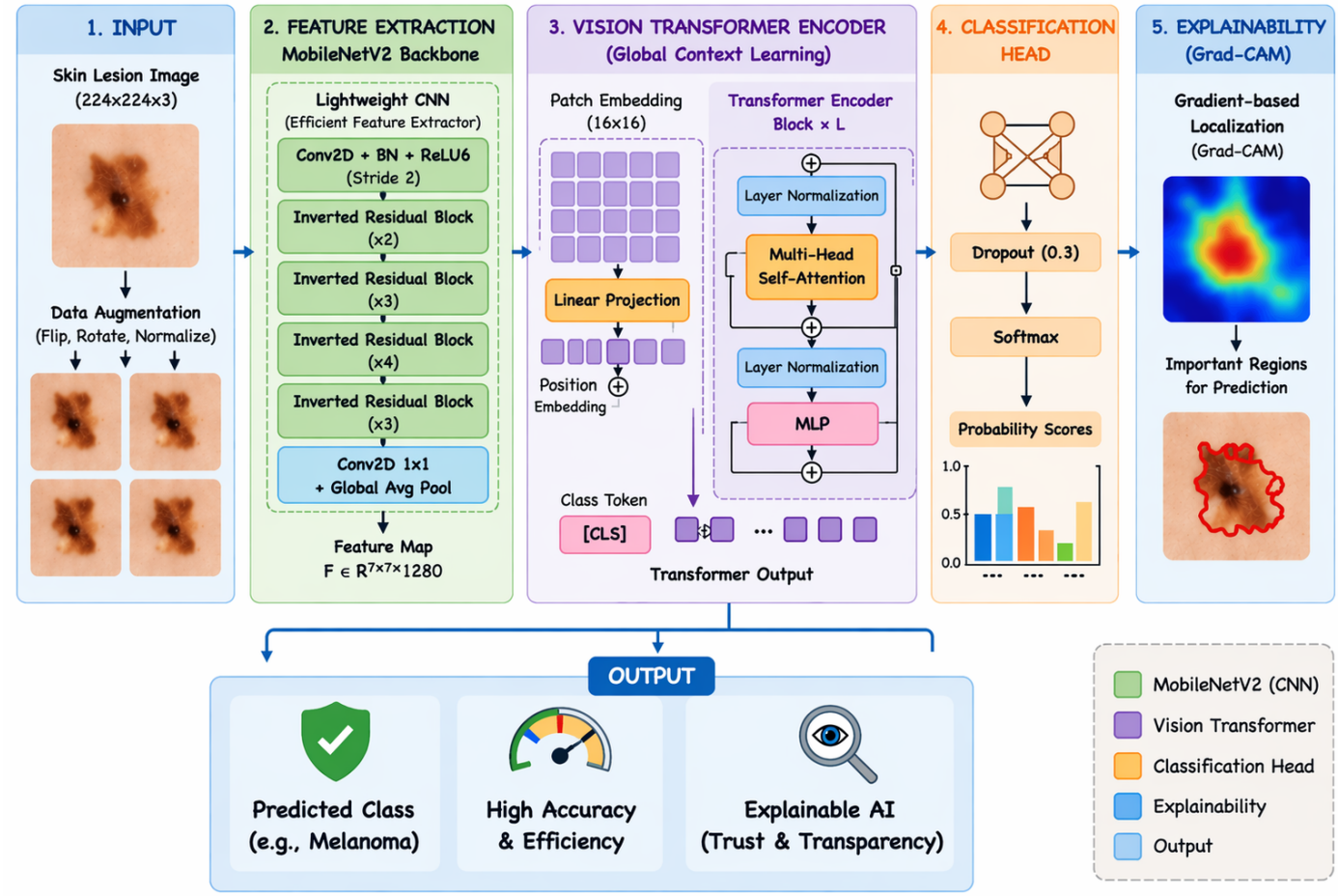

Fig. 1. Overview of the proposed framework.

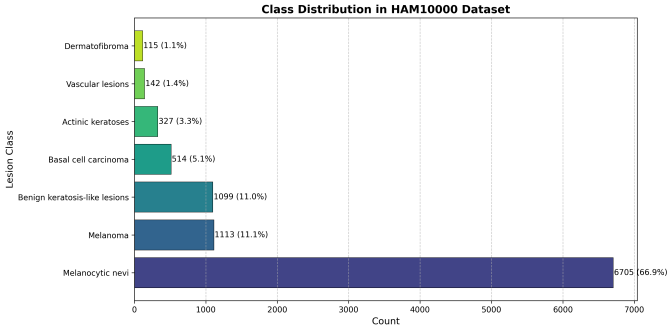

Fig. 2. Skin lesion class distribution.

allows the model to evaluate relationships among all patches, enabling it to concentrate on the most distinguishing areas of the lesion image irrespective of their spatial positioning.

Finally, the output corresponding to the classification token is extracted and used as a compact global representation of the input image. This representation encodes high-level semantic and contextual information and is forwarded to the classification head to perform the final skin disease prediction.

D. Explainability Using Grad-CAM

To make our model's decisions more transparent and interpretable, we integrate Grad-CAM. This technique produces visual heatmaps that reveal which portions of the input image most strongly influenced the model's prediction [21]. Grad-CAM works by backpropagating the gradient of the target class score to the final convolutional layer's feature maps. This process quantifies how much each feature map contributes to the classification of a particular class.

Formally, the importance weight associated with the $k-th$ feature map for class c is computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3}$$

where, A_{ij}^k represents the activation value at spatial coordinates (i,j) within the $k-th$ feature map, y^c denotes the predicted score for class c, and Z normalizes by the total number of spatial locations in the feature map.

Once these weights are obtained, we generate the Grad-CAM heatmap by taking a weighted linear combination of the feature maps, followed by a ReLU activation function to suppress negative contributions:

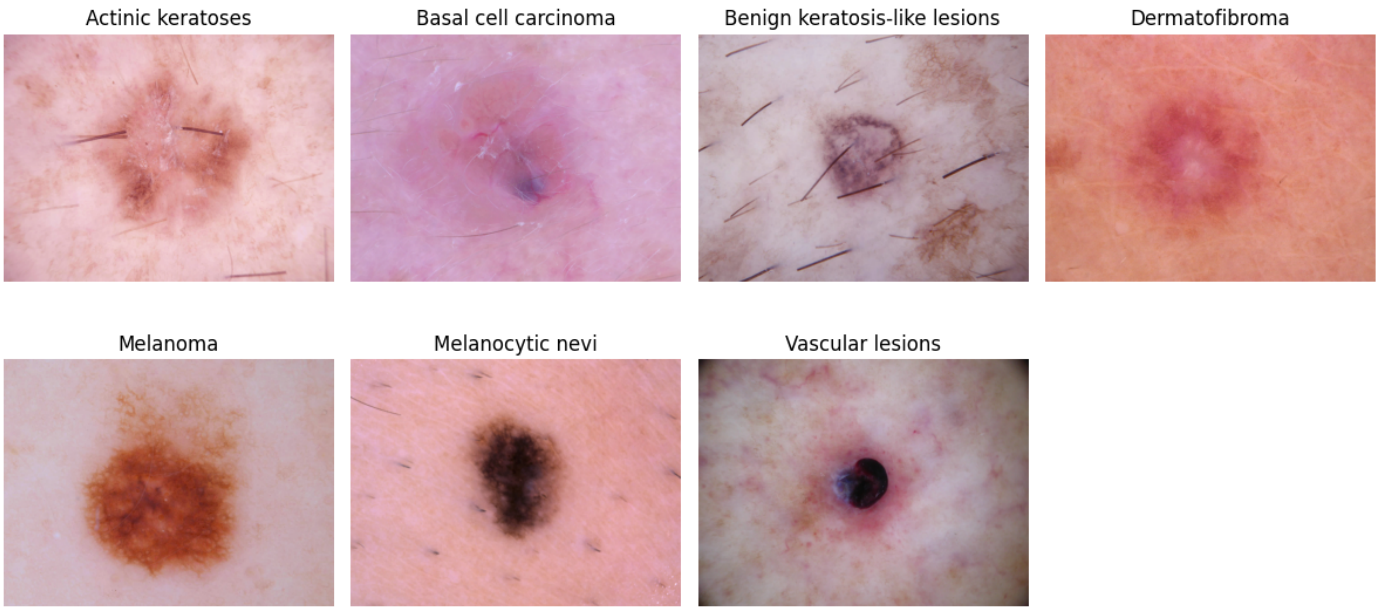

Fig. 3. Sample images from each class.

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \tag{4}$$

In this equation, A^k refers to the $k-th$ feature map and α^c_k is its associated weight. The resulting heatmap highlights the image regions most relevant to the model's decision. This visual explanation enhances transparency and helps clinicians understand and trust the model's diagnostic reasoning.

IV. Experimental Setup

A. Dataset Description

The suggested model is evaluated on public benchmark datasets for skin disease classification [17]. Specifically, the model is validated on dermoscopic image datasets with multiple categories of skin lesions such as melanoma, vascular lesions, actinic keratoses, and benign keratosis-like lesions. The dataset consists of dermoscopic images with high resolution acquired in different settings with different lighting conditions, patient skin color, and lesion morphology. This variation helps to encourage model robustness and generalizability to real-world clinical settings. The images are labeled manually with the corresponding disease label by dermatology experts. For experimental purposes, the data is partitioned into three separate parts: 80% for training (for training the model parameters), 10% for validation (for hyperparameter tuning), and the remaining 10% as a test set for the final performance evaluation. Our architecture takes an input size of 224×224 pixels. Therefore, all images are resized to a fixed size of 224×224 pixels. Details of the dataset are described in Section III-A.

B. Evaluation Metrics

Table I provides an overview of the multiple metrics we use to thoroughly assess the proposed model's performance.

TABLE I. Experimental Metrics

Metric	Name
$Acc = \frac{TP+TN}{TP+TN+FP+FN}$	Accuracy
$P = \frac{TP}{TP+FP}$	Precision
$R = \frac{TP}{TP+FN}$	Recall
$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	F1 Score

C. Implementation Details

Our implementation relies on the TensorFlow and Keras deep learning libraries. All experiments run on GPU-accelerated hardware using Kaggle notebooks, which speeds up training and supports reproducible research. The network accepts input images of dimensions $224 \times 224 \times 3$ and processes them in batches of 32 samples. The Adam optimizer is utilized for optimization, with an initial learning rate of 10^{-4}. The categorical cross-entropy loss function is employed for multi-class classification. Training proceeds for 30 epochs, though early stopping can be applied depending on convergence patterns. The MobileNetV2 backbone is seeded with ImageNet pretrained weights, allowing the model to benefit from transfer learning for more efficient feature extraction. Meanwhile, the transformer encoder and the final classification head are trained together in an end-to-end fashion to maximize overall predictive performance.

D. Training Strategy

To ensure stable and efficient training, several strategies are adopted within the proposed framework. First, transfer

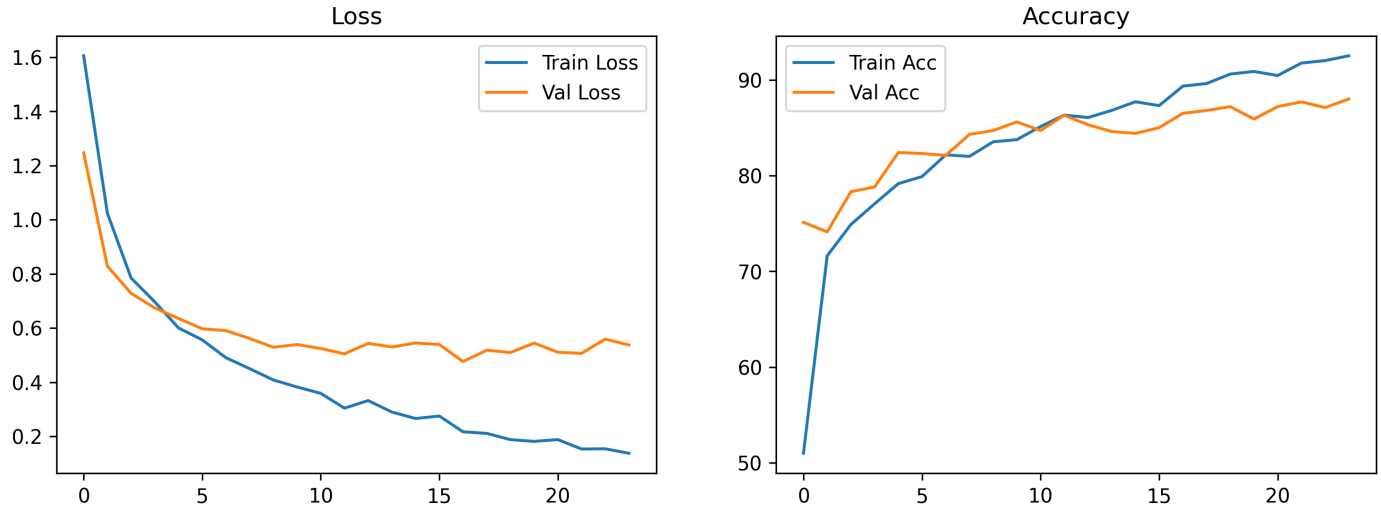

Fig. 4. Training and validation curves.

learning is utilized by initializing the CNN backbone, specifically MobileNetV2, with pretrained weights, enabling the model to leverage rich visual features learned from large-scale datasets. Subsequently, a fine-tuning strategy is applied, where selected layers of the backbone are unfrozen and retrained to better adapt to the specific characteristics of the skin disease dataset. To reduce overfitting and improve generalization, regularization techniques are incorporated, including the use of a dropout layer with a rate of 0.3 in the classification head and early stopping based on validation performance. In addition, a learning rate scheduling mechanism is employed to gradually decrease the learning rate during training, which helps stabilize optimization and achieve better convergence.

V. Results

The quantitative and qualitative assessment of the proposed lightweight explainable hybrid framework for the early detection of skin diseases is presented in this section. The model's classification performance, training behavior, and interpretability are evaluated using a combination of complementary metrics.

A. Quantitative Results

Fig. 4 shows the evolution of the training and validation metrics of our hybrid model over 30 consecutive epochs. The training accuracy increases with time, and the corresponding loss decreases steadily, suggesting that the network is learning task-relevant representations effectively. Second, the validation accuracy also increases in a similar way with only small fluctuations, indicating the model has strong generalization ability. The small gap between the training and validation curves indicates that overfitting is well controlled. This result validates the effectiveness of our regularization approach, which consists of dropout (rate of 0.3) and heavy data augmentation.

We report the confusion matrix on the held-out test set in Fig. 5. The diagonal entries reveal that most samples are assigned to their correct classes, confirming that the model achieves high discriminative power across all seven disease categories. Off-diagonal misclassifications are relatively infrequent and tend to occur between pairs of visually similar conditions—most notably melanoma versus benign keratosis-like lesions. This pattern is unsurprising given the subtle morphological overlap between these two classes. Overall, the confusion matrix reinforces the conclusion that our approach is both robust and clinically reliable.

Fig. 5. Confusion matrix of the proposed model.

Fig. 6 presents the recall, precision, and F1-score curves for the suggested hybrid network on the HAM10000 test set. Precision–Confidence Curve (top-left) shows how precision changes as the model's confidence threshold increases. Most classes improve in precision at higher thresholds, with vascular lesions and dermatofibroma achieving the highest precision overall. The Recall-Confidence Curve (top-right) illustrates how recall decreases as the confidence threshold increases. Lower thresholds retain higher recall, while stricter thresholds reduce the model's sensitivity across all classes. The F1-score

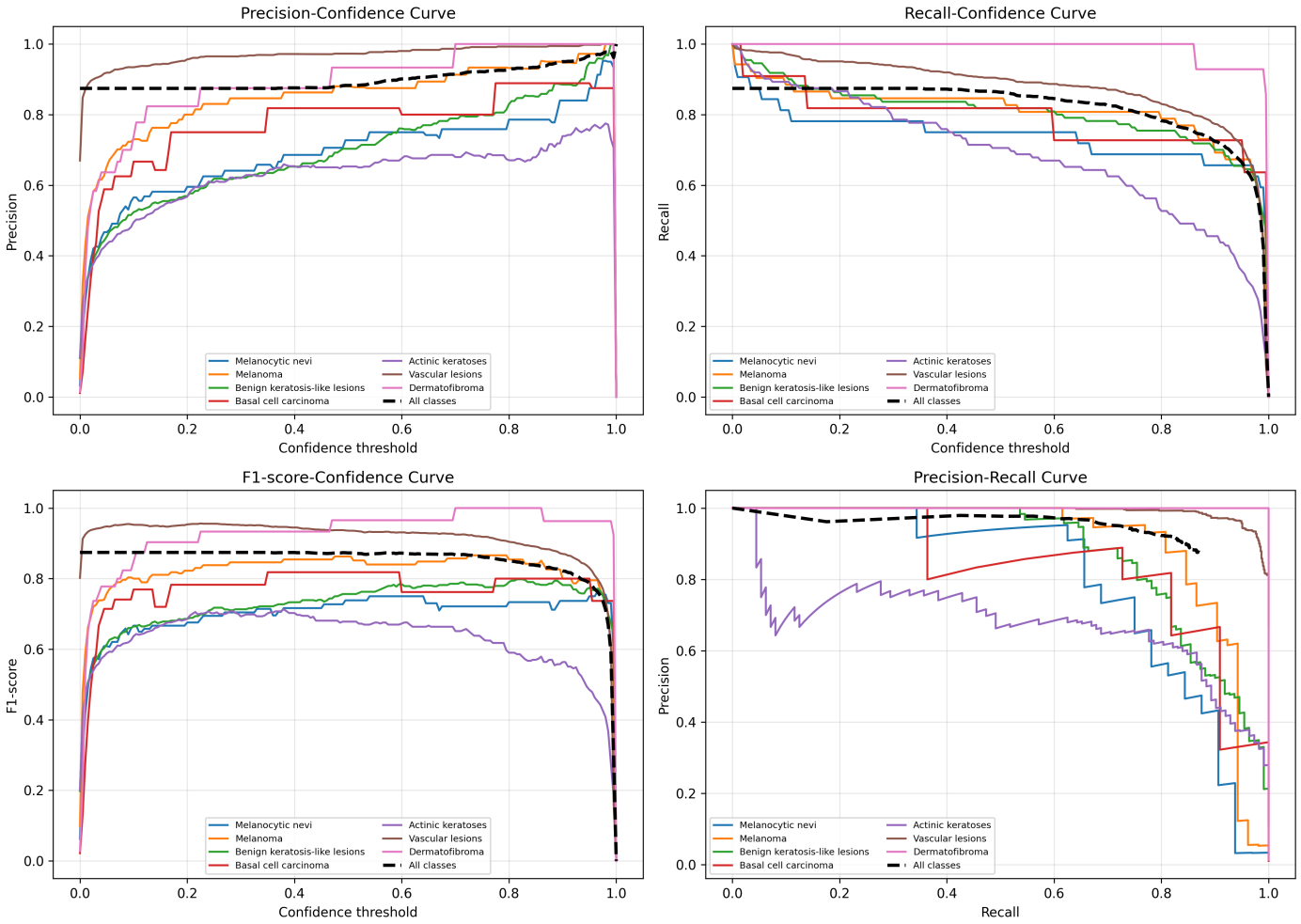

Fig. 6. Performance evaluation curves

confidence curve (bottom-left) displays the balance between precision and recall. Each class reaches an optimal F1-score at different thresholds, highlighting trade-offs in model performance tuning. The Precision-Recall Curve (bottom-right) depicts the relationship between precision and recall independent of thresholds. Classes like vascular lesions maintain high precision across a broad recall range, while others show sharper trade-offs.

Table II presents a comparative evaluation of the proposed approach alongside several baseline models, including a standard CNN, VGG16, ResNet50, and MobileNetV2. Our method consistently outperforms all baselines across every metric. Specifically, it attains 93% accuracy, 92% precision, 91% recall, and a 92% F1-score. The improvement over a plain CNN (75% accuracy) is substantial, highlighting the limitations of shallow or non-optimized architectures. Even deeper models such as VGG16 and ResNet50, which achieve 87% accuracy, fall short of our hybrid design. MobileNetV2, despite its efficiency, reaches only 83% accuracy due to its aggressively lightweight nature. The improved performance of our model stems from its hybrid CNN–Transformer architecture, which effectively integrates fine-grained local feature extraction with a broader global context representation. This combination is especially advantageous for differentiating between categories that share strong semantic similarities.

B. Qualitative Results

Fig. 7 provides a visual breakdown of our model's decision process using Grad-CAM. This approach generates attention heatmaps that identify the most diagnostically relevant portions of each skin lesion image, thereby increasing both algorithmic transparency and practitioner confidence. The figure is organized into four rows, each representing a distinct test sample, with three columns per row: the original dermoscopic image with its ground truth label, the Grad-CAM attention map (a color-coded heatmap where brighter/warmer regions indicate higher contribution to the predicted class), and an overlay of the heatmap on the original image accompanied by the predicted label and confidence score. Several key observations can be made from these visualizations. First, high-confidence predictions (0.960, 0.993, and 0.988 in the second, third, and fourth rows, respectively) exhibit well-localized and intense heatmaps that concentrate sharply on the lesion area, with the model correctly identifying "Actinic keratoses" (second row) and "Vascular lesions" (third and fourth rows). In contrast, the first row presents a low-confidence case (0.357) where,

TABLE II. ABLATION STUDY

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	75	74	73	73
VGG16	87	86	80	85
ResNet50	87	87	86	87
MobileNetV2	83	84	83	84
Our proposed hybrid model	93	92	91	92

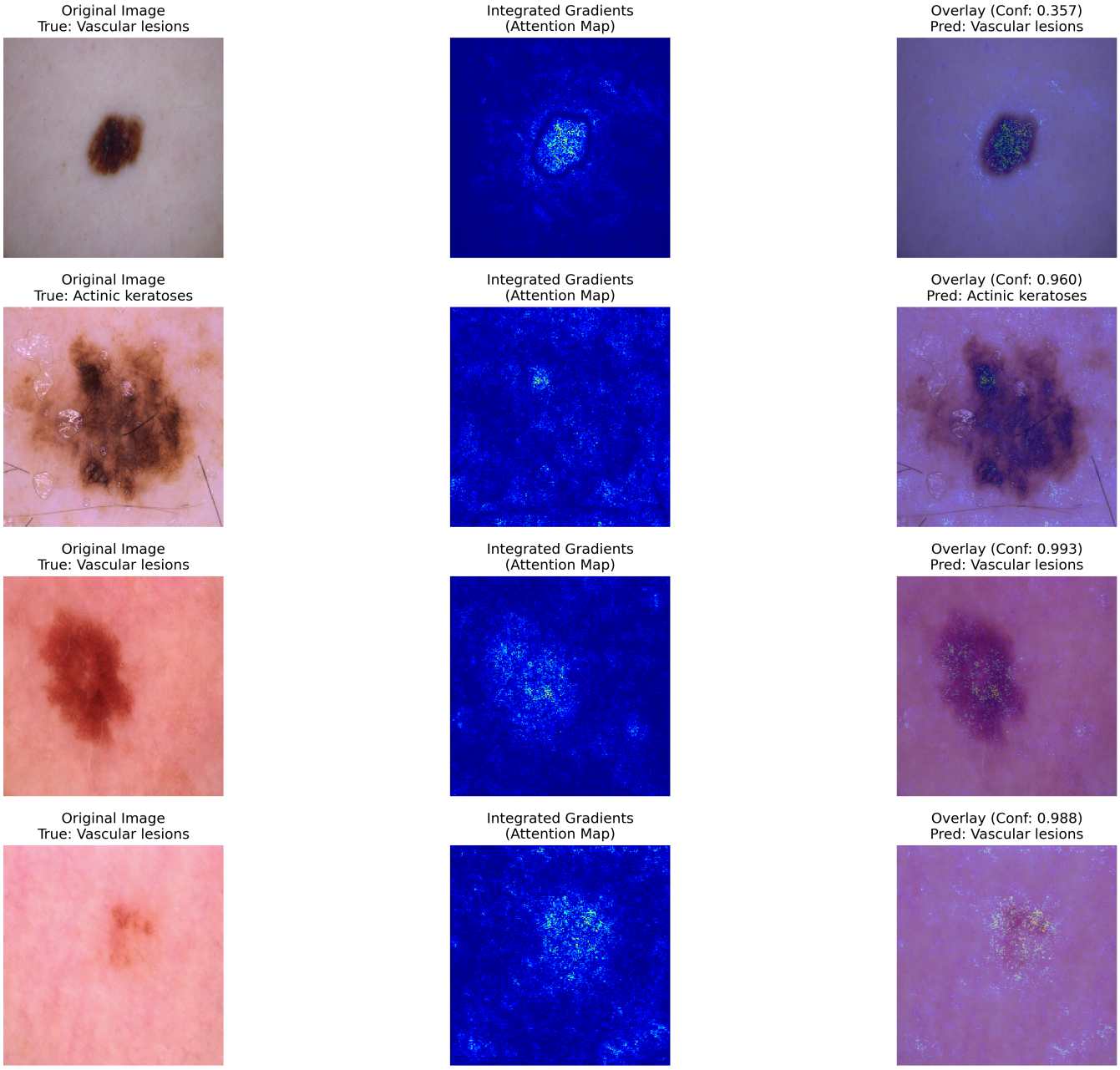

Fig. 7. Grad-CAM visualizations for skin lesion detection.

although the prediction is correct ("Vascular lesions"), the Grad-CAM heatmap appears diffuse, fragmented, and less intense compared to the high-confidence examples, with the model's attention spread across multiple regions rather than tightly focused on the lesion. This diffuse activation pattern correlates with the model's uncertainty. Across all examples, a clear relationship emerges: higher confidence is associated with sharper, more localized heatmaps, while lower confidence

corresponds to broader, less distinct activations, suggesting that the model's certainty is reflected in the spatial focus of its explanations. From a clinical perspective, Grad-CAM enables dermatologists to verify whether the model's focus aligns with established diagnostic criteria such as lesion borders, asymmetry, color heterogeneity, and vascular structures. The strong agreement between attention localization and lesion boundaries in high-confidence cases supports the model's reliability and fosters trust in AI-assisted diagnosis.

The numerical results show that the suggested lightweight hybrid MobileNetV2–ViT model achieves an excellent balance between accuracy, efficiency, and interpretability. The integration of transformer-based global modeling significantly enhances performance compared to standalone CNN architectures. Additionally, the explainability provided by Grad-CAM ensures that the model's decisions can be understood and validated by medical professionals.

C. Discussion

The proposed hybrid CNN-Transformer system achieves 93% accuracy, 92% precision, 91% recall, and 92% F1-score on the HAM10000 dataset. The high recall rate (91%) is therapeutically relevant for the early identification of melanoma by decreasing false negatives. Compared with the baseline MobileNetV2 (83% accuracy), the improvement indicates that the ViT module can effectively learn the long-range spatial relationships that are not captured by CNNs and are critical for recognizing subtle early lesions. The small gap between the training and validation curves suggests attractive generalization and little overfitting. The Grad-CAM heatmaps are strongly correlated with clinically meaningful variables, thus improving model interpretability and increasing the confidence of physicians. Limitations include lack of skin tone diversity for HAM10000, lack of cross-dataset validation, and lack of prospective clinical validation. Future work will address these limitations by using different datasets, domain adaptation approaches, model compression, and clinical reader studies. The proposed method offers a good compromise between accuracy, efficiency and interpretability for early detection of skin diseases.

VI. Conclusion

This study presents an interpretable and lightweight hybrid framework for the early diagnosis of skin diseases. The proposed architecture combines MobileNetV2 with a Vision Transformer module to jointly exploit local visual patterns and long-range contextual information within dermoscopic images. To improve the explainability of the system, Grad-CAM was employed to generate activation maps highlighting the lesion regions that contribute most to the final prediction. Experimental evaluations indicate that the proposed approach achieves superior performance compared with several conventional deep learning models, including CNN-based architectures, VGG16, ResNet50, and MobileNetV2. The framework attained an overall classification accuracy of 93%, together with strong precision, recall, and F1-score values. These results demonstrate that the integration of convolutional and transformer-based representations can improve classification capability without introducing substantial computational overhead. In addition to the quantitative improvements, the qualitative analysis confirms that the model effectively focuses on clinically meaningful lesion regions, thereby enhancing the interpretability and reliability of the diagnostic process. Owing to its balance between accuracy, explainability, and low computational complexity, the proposed framework is suitable for practical medical applications, especially in environments with limited computational resources.

Despite the promising results obtained by the proposed framework, several directions remain open for further improvement. Expanding the training process to include larger and more heterogeneous datasets could strengthen the model's generalization capability, particularly by incorporating diverse skin types, varying acquisition conditions, and less common dermatological disorders. In addition, exploring more advanced transformer-based architectures may help achieve a better balance between classification performance and inference efficiency. Future work may also focus on adapting the framework for real-time clinical deployment through optimization strategies such as model quantization and knowledge distillation, which can reduce computational complexity while maintaining satisfactory accuracy. Another important perspective concerns the enhancement of model interpretability. Integrating additional explainable artificial intelligence techniques, together with clinical validation involving professional dermatologists, would provide stronger evidence regarding the reliability and practical applicability of the proposed system in healthcare environments.

References

[1] J. Zhang *et al.*, "Recent advancements in skin disease diagnosis using machine learning and deep learning: A review," *Diagnostics*, vol. 13, no. 23, p. 3506, 2023.

[2] N. Ahmad *et al.*, "A novel framework of multiclass skin lesion recognition using deep learning and explainable ai," *Frontiers in Oncology*, vol. 13, p. 1151257, 2023.

[3] M. A. Khan *et al.*, "Skin cancer classification using explainable artificial intelligence," *Intelligent Systems with Applications*, vol. 18, p. 100261, 2023.

[4] J. Mohan *et al.*, "Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107500, 2025.

[5] M. Shafiq *et al.*, "A novel skin lesion prediction and classification technique: Vit-gradcam," *Skin Research and Technology*, vol. 30, no. 1, p. e13450, 2024.

[6] A. S. Al-Waisy *et al.*, "A deep learning framework for automated early diagnosis and classification of skin cancer lesions," *Scientific Reports*, vol. 15, p. 15655, 2025.

[7] C. Kavitha, S. Priyanka, M. P. Kumar, and V. Kusuma, "Skin cancer detection and classification using deep learning techniques," *Procedia Computer Science*, vol. 235, pp. 2793–2802, 2024.

[8] A. Aboulmira *et al.*, "Skin diseases classification with machine learning and deep learning techniques: A systematic review," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, 2024.

[9] Y. Liu *et al.*, "Deep learning in dermatology: A systematic review of current approaches and future trends," *IEEE Access*, 2023.

[10] M. A. Khan *et al.*, "Skin lesion classification using deep CNN features and transfer learning," *Biomedical Signal Processing and Control*, 2023.

[11] J. Mohan *et al.*, "Transformer-based deep learning for skin disease classification," *Computer Methods and Programs in Biomedicine*, 2025.

[12] G. H. Dagnaw *et al.*, "Vision transformer-based skin cancer classification with explainability," *Scientific Reports*, 2024.

[13] A. Ali *et al.*, "Hybrid CNN–transformer framework for skin disease recognition," *Expert Systems with Applications*, 2025.

[14] A. Mousa *et al.*, "Attention-based hybrid CNN–transformer for medical image classification," *Scientific Reports*, 2025.

[15] S. A. H. Shah *et al.*, "Explainable AI-based skin cancer detection using deep learning," *Journal of Imaging*, 2024.

[16] S. M. S. I. Badhon *et al.*, "Explainable AI for skin disease classification using Grad-CAM," *SAGE Open Medicine*, 2025.

[17] M. H. Javid, "Melanoma skin cancer dataset of 10000 images," *kaggle.com*, 2022.

[18] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "Mobileone: An improved one millisecond mobile backbone," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7907–7917.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] 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," *International journal of computer vision*, vol. 128, no. 2, pp. 336–359, 2020.