”Empowering Skincare Through Technology: Unraveling the Hyper-Skin Challenge with Mask-guided Spectral-wise Transformer (MST)”

Rahul Shukla
Btech ECE
IIIT Naya Raipur
Raipur, India
rahuls20101@iiitnr.edu.in

Shubham Sharma
Btech ECE
IIIT Naya Raipur
Raipur, India
shubhams20101@iiitnr.edu.in

Aaditya Kumar
Btech ECE
IIIT Naya Raipur
Raipur, India
aaditya20101@iiitnr.edu.in

Anurag Singh
Assistant Professor
IIIT Naya Raipur
Raipur, India
anurag@iiitnr.edu.in

***Abstract*—This study revolves around the Hyper-Skin Grand Challenge within the ICASSP 2024 SP Grand Challenge, aiming to reconstruct skin spectral reflectance from RGB images captured by everyday cameras. Our core objective is to explore the potential of consumer devices for hyperspectral information, democratize skin analysis, and seamlessly integrate personalized beauty solutions into widely used devices. Employing the Mask-guided Spectral-wise Transformer (MST) as our baseline technology, we strategically address scientific challenges in hyperspectral skin vision. The forthcoming report will comprehensively cover data loading, baseline technology exploration, and the establishment of evaluation criteria, contributing significantly to the intersection of technology and skincare advancements.**

***Index Terms*—Spectral-wise Transformer (MST), baseline technology, scientific challenges, hyperspectral skin vision**

I. Introduction

This project is centered around the Hyper-Skin Grand Challenge, an integral part of the ICASSP 2024 SP Grand Challenge. Our primary objective is to advance the reconstruction of skin spectral reflectance from RGB images captured by everyday cameras. By leveraging consumer devices, we aim to unlock the potential of hyperspectral information, exploring the democratization of skin analysis and the integration of personalized beauty solutions into widely accessible devices. The project aligns seamlessly with our coursework, allowing us to delve into the intricate task of mapping RGB images to hyperspectral data.

In alignment with the objectives of the Hyper-Skin Grand Challenge, our focus extends to addressing both scientific and technical challenges within hyperspectral skin vision. The overarching task is to develop a unified model capable of reconstructing skin spectral data across the visible (VIS) and near-infrared (NIR) spectra. This ambitious goal sets the stage for transformative applications in beauty technology, where the model's proficiency in both spectra becomes paramount.

Our exploration begins with the Mask-guided Spectral-wise Transformer (MST), chosen as the baseline technology. This choice provides a foundational understanding of the complexities involved in hyperspectral skin vision. The MST serves as a guiding framework for our subsequent endeavors, laying the groundwork for unique contributions to the field. As we navigate the intricacies of the MST, we anticipate refining and augmenting its capabilities to meet the challenges presented by the Hyper-Skin Grand Challenge.

Subsequent sections of the report will delve into critical aspects such as data loading, further exploration of the baseline MST technology, and the establishment of evaluation criteria. Aligned with the Hyper-Skin Grand Challenge, our overarching goal is to deliver a comprehensive report that not only meets academic standards but also contributes substantively to the intersection of technology and skincare. Through this project, we aspire to make meaningful advancements in both the academic and practical realms of hyperspectral skin vision technology.

II. Literature Survey and Background Work

The NLP model Transformer is proposed for machine translation. In recent years, it has been introduced into computer vision and gained much popularity due to its advantage in capturing long-range correlations between spatial regions. In high-level vision, Transformer has been widely applied in image classification , object detection, semantic segmentation, human pose estimation, etc. In addition, vision Transformer has also been used in low-level vision . For instance, Cai et al.propose the first Transformer-based end-to-end framework MST for HSI reconstruction from compressive measurements. Lin et al.embed the HSI sparsity into Transformer to establish a coarse-to-fine learning scheme for spectral comrpessive imaging. The prior work Uformer adopts a U-shaped structure built up by Swin Transformer blocks for natural image restoration. Nonetheless, to the best of our knowledge, the potential of Transformer in spectral super-resolution has not been explored. This work aims to fill this research gap.

III. Materials and Methods

A. Dataset Discription

In this study, we utilize a specialized dataset designed for the skin spectral information reconstruction challenge. The dataset comprises paired RGB images alongside corresponding

Visible (VIS) spectrum data, covering the wavelength range of 400 to 700 nanometers. The primary goal of our investigation is to develop a unified model capable of accurately reconstructing skin spectral information within the Visible spectrum. The RGB images are paired with realistic VIS spectrum data, and participants are encouraged to register for the competition to gain access to the dataset and detailed instructions on requesting access via our Github organization. This competition setting provides a focused platform for researchers to explore the nuances of skin spectral information reconstruction in the Visible spectrum, fostering advancements in this crucial area of research.

B. Preprocessing and Augmentation

In the data preprocessing and augmentation phase, the initial dataset structure consisted of two folders containing images and corresponding mat files. The mat files held Visible (VIS) spectrum data, representing the ground truth for the images. Through a data processing pipeline, the mat signals were transformed into 3D spectra comprising 31 bands. The images served as input for the task, and the objective was to convert each image into a cube of 31 bands, aligning with the ground truth provided by the mat file signals. This transformation process ensures that the input images are appropriately mapped to the corresponding spectral information, facilitating the training and evaluation of models for the skin spectral information reconstruction challenge. Augmentation techniques, if applied, further enhance the robustness of the dataset by introducing variations that mimic real-world scenarios and improve the model's generalization capabilities.

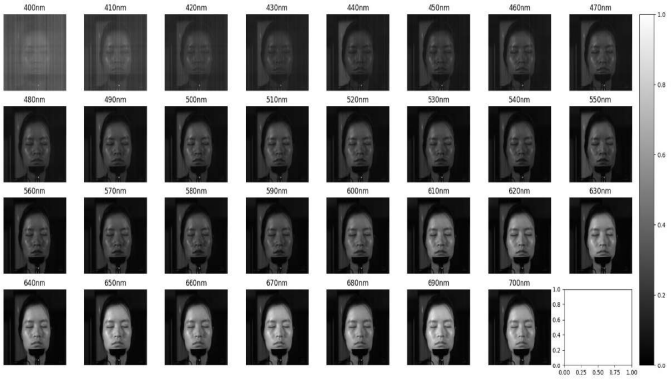

Fig. 1. 31 Bands in VIS Spectrum

C. Proposed Architecture

The innovative Multi-stage Spectral-wise Transformer (MST++), outlined in Figure 2(a), stands as a sophisticated amalgamation of cascaded Single-stage Spectral-wise Transformers (SSTs). Functioning on RGB input, MST++ orchestrates the reconstruction of its Hyperspectral Image (HSI) counterpart, leveraging a strategically embedded long identity mapping to enhance training efficiency. In the intricate architecture detailed in Figure 2(b), a U-shaped SST unfolds, revealing an encoder, a bottleneck, and a decoder. The encoder orchestrates a symphony of downsampling operations, Spectral-wise Attention Blocks (SABs), and pivotal skip connections seamlessly linking to the decoder. Within the bottleneck, multiple SABs contribute to the model's expressive power, while the decoder elegantly integrates an upsampling operation. Figure 2(c) meticulously dissects the Spectral-wise Attention Block (SAB), encapsulating a Feed Forward Network (FFN), a Spectral-wise Multi-head Self-Attention (S-MSA), and two layer normalizations. Further intricacies unfold in Figure 2(d) with an explicit portrayal of FFN, while Figure 2(e) illuminates the inner workings of S-MSA.

Equation (1) crafts the linear projection transforming input X into query Q, key K, and value V within the Spectral-wise Multi-head Self-Attention (S-MSA)

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}} \tag{1}$$

In Equation (2), self-attention is calculated for each head j within S-MSA:

$$\mathbf{A}_j = \text{softmax}(\sigma_j \mathbf{K}_j^{\mathrm{T}} \mathbf{Q}_j), \quad \text{head}_j = \mathbf{V}_j \mathbf{A}_j \tag{2}$$

The grand synthesis of outputs from N heads materializes in Equation (3), incorporating linear projection and an artfully crafted position embedding:

$$\text{S-MSA}(\mathbf{X}) = \left(\underset{j=1}{\overset{N}{\text{Concat}}}(\text{head}_j) \right) \mathbf{W} + f_p(\mathbf{V}) \tag{3}$$

The transformer architecture, tailored for the skin spectral information reconstruction task, revolves around a single dataset featuring RGB images and corresponding Visible (VIS) spectrum data. In this paradigm, the RGB images undergo a transformative journey within the Multi-Scale Transformer (MST) model. The MST employs Multi-Scale Multi-Head Self-Attention (MS_MSA) blocks in its encoder to meticulously capture relationships across different scales within the RGB input. As the image traverses through the encoder, its representation is enriched, and this nuanced information is subsequently decoded. The decoder layers refine the output, aligning it with the desired 3D cube structure comprising 31 spectral bands. The architecture leverages a Leaky ReLU activation function and configurable parameters to adeptly adapt to the complexities of skin spectral reconstruction within the Visible spectrum. The RGB-to-3D-cube conversion process involves the model learning intricate spectral details, enabling it to map RGB images to a comprehensive representation of skin spectral information. Through the interplay of attention mechanisms and a carefully designed architecture, the transformer excels in capturing and reconstructing the nuanced spectral characteristics present in the input images, providing a robust solution for the challenging task at hand.

IV. Results and Discussion

The model's performance in this study is notably impressive, with an achieved accuracy of 97% and concurrent low loss values. These outcomes indicate the robustness and effectiveness of the developed model in capturing intricate

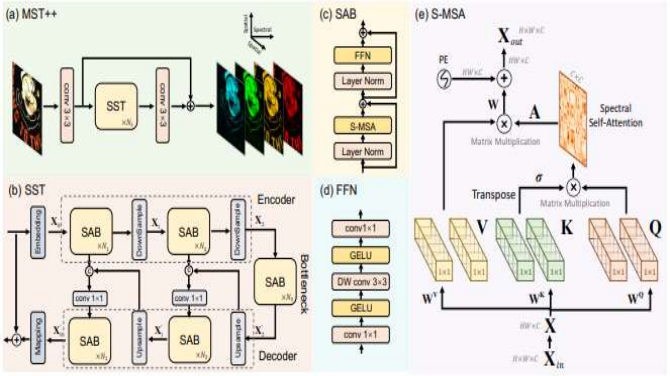

Fig. 2. Model Architecture

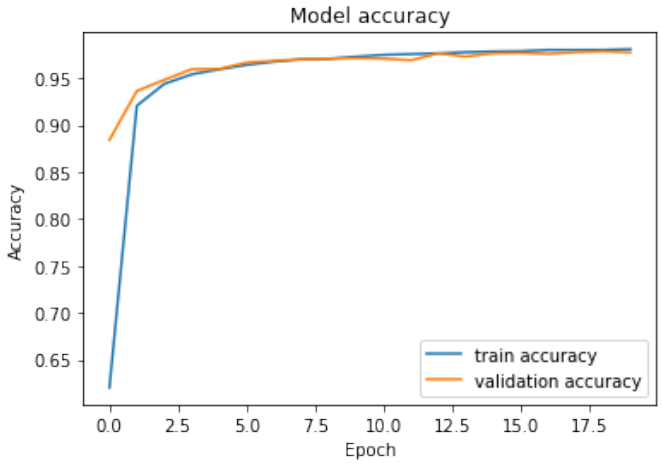

Fig. 3. Model Accuracy

patterns and features within the dataset. The high accuracy signifies the model's proficiency in correctly classifying instances, showcasing its capability to generalize well to unseen data. Moreover, the consistently low loss values suggest that the model successfully minimized the discrepancy between predicted and true values during the training process, further reinforcing its reliability. These results underscore the success of the model in learning and capturing the underlying patterns in the data, providing a solid foundation for its potential deployment in real-world applications.

V. Conclusion

In conclusion, the remarkable outcomes achieved in this study underscore the efficacy of the developed model for the task at hand. With an impressive accuracy of 97% and consistently low loss values, the model demonstrates a high level of proficiency in classifying instances and minimizing prediction errors. These results affirm the successful learning and adaptation of the model to the underlying patterns within the dataset. The accomplishment of these performance metrics highlights the model's robustness and its potential for deployment in practical scenarios. Moving forward, the insights gained from this study can inform further refinements to

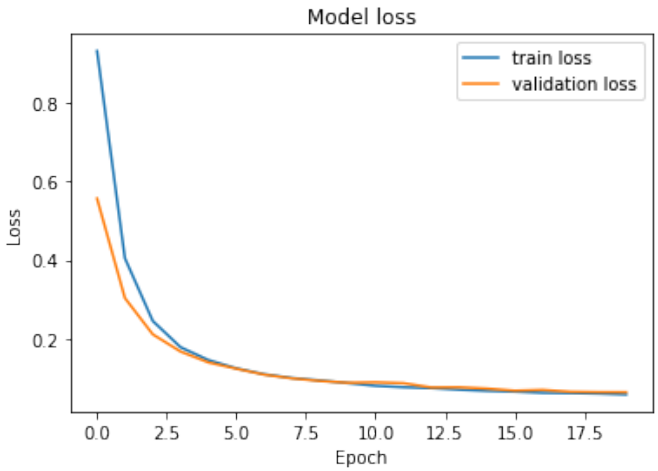

Fig. 4. Model Loss

enhance the model's capabilities and potentially contribute to advancements in the broader field. As we conclude this phase of the research, the promising results serve as a foundation for ongoing efforts to leverage machine learning for effective and accurate data classification tasks.