International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 1, No. 1, 2026

HRS-Swin: A Hierarchical Representation Separation Swin Transformer for Automated Neonatal Auricular Deformity Classification

Daini Li *, Xiaomeng Yang

College of Computer Science and Artificial Intelligence, Southwest Minzu University, Chengdu 610000, China

Abstract: The incidence of auricular deformities in newborns is notably high, and even experienced clinicians may encounter issues such as misdiagnosis and missed diagnosis due to subjective judgment. Although several studies have explored the use of deep learning methods for auxiliary diagnosis, the highly complex and individualized characteristics of auricular morphology pose significant challenges to existing approaches in achieving automated identification and fine-grained subtype classification. To address this issue, we propose HRS-Swin, a progressive representation reconstruction framework built upon a Swin Transformer backbone. The model integrates a Class Token Fusion module to enhance global semantic representation, a Stable Semantic Enhancement and Residual Compression mechanism for compact and discriminative embedding learning, and a Dynamic Margin Enhancer to enlarge inter-class separability in the embedding space. Experiments on the BabyEar4K dataset (1,926 newborns) demonstrate that HRS-Swin outperforms representative CNN and Transformer baselines. The proposed method achieves an accuracy of 0.8009 and a macro F1-score of 0.7024, showing consistent improvements over standard Swin Transformer. These results indicate that the proposed framework provides a robust and effective solution for automated auricular deformity classification and early clinical assistance.

Keywords: Vision Transformer; Progressive representation reconstruction; Attention mechanism; Adaptive margin learning.

1. Introduction

The ear is one of the most important anatomical components contributing to facial morphology [1]. Congenital auricular deformities (CADs) are among the most common craniofacial anomalies in newborns. Epidemiological studies report that the incidence of CADs is approximately 25% in the United States, 55.2% in Japan, and 43.46% in China [2]. Early diagnosis of auricular deformities largely relies on clinicians' experience, particularly the subjective judgment of pediatricians. To date, objective, quantitative, and reproducible evaluation criteria remain lacking, and diagnostic outcomes may vary considerably among different evaluators. Therefore, establishing an objective, automated, and reliable assessment method for auricular deformities is of significant importance for improving early diagnostic accuracy and guiding clinical decision-making [3].

Due to the anatomical complexity of the auricle—characterized by multiple curved surfaces, folds, and intricate boundary structures—traditional geometry-based modeling or handcrafted feature extraction methods struggle to comprehensively characterize its morphological variations [4]. In recent years, deep convolutional neural networks (CNNs)[5] and vision Transformer architectures based on self-attention mechanisms have achieved remarkable success in medical image analysis [6]. In 2019, Rami R. Hallac et al .[3] applied CNNs to automatically identify adult ear deformities from 2D photographs, demonstrating the feasibility and potential of deep learning in this domain. In 2021, they further compared healthcare providers and CNN models in evaluating outcomes of ear molding therapy [7]. Their findings indicated that CNN models have the potential to approximate provider assessments; however, misclassifications still occurred, and existing models were limited to relatively simple classification tasks.

To address these challenges, this study proposes an intelligent diagnostic framework for auricular deformities based on a Swin Transformer backbone. Unlike the conventional "feature extraction + fully connected classifier" paradigm, our method does not directly perform classification on the global feature vector produced by the backbone network. Instead, we construct a multi-level representation enhancement and compression architecture, including a Class Token Fusion module, a Stable Semantic Enhancement and Residual Compression collaborative mechanism, and a Dynamic Margin Enhancer module. Together, these components establish a progressive reconstruction pathway from high-dimensional semantic space to a compact low-dimensional discriminative embedding space.

The main contributions of this study are summarized as follows:

1. We propose a progressive representation reconstruction framework for neonatal auricular deformity diagnosis, establishing a multi-stage enhancement–compression pathway from high-dimensional semantic features to compact discriminative embeddings.

2. We develop a collaborative mechanism combining Class Token Fusion and Stable Attention Compressor to jointly enhance global semantics and suppress redundant information, improving representation of complex auricular morphology.

3. We introduce a Dynamic Margin Enhancer that performs batch-level adaptive margin expansion in the embedding space, increasing inter-class separability and mitigating class imbalance and boundary ambiguity to enhance robustness and generalization.

2. Method

(1) Overall

The proposed framework follows a progressive representation refinement strategy for auricular deformity classification. Given an input ear image, a Swin Transformer backbone first extracts high-dimensional global semantic features. Instead of directly applying a linear classifier, the features are sequentially processed through three dedicated modules: Class Token Fusion (CTF), Stable Semantic Enhancement and Residual Compression (SAC), and Dynamic Margin Enhancer (DME). CTF introduces a learnable global semantic anchor to strengthen class-aware representation at the high-dimensional stage. SAC subsequently performs channel-wise semantic recalibration and residual dimensionality compression, generating compact yet discriminative embeddings. Finally, DME adaptively expands inter-class margins in the low-dimensional embedding space through batch-level dynamic deviation amplification. Through this hierarchical enhancement–compression–separation pipeline, the framework progressively transforms raw semantic features into compact and highly separable representations suitable for robust auricular deformity classification.

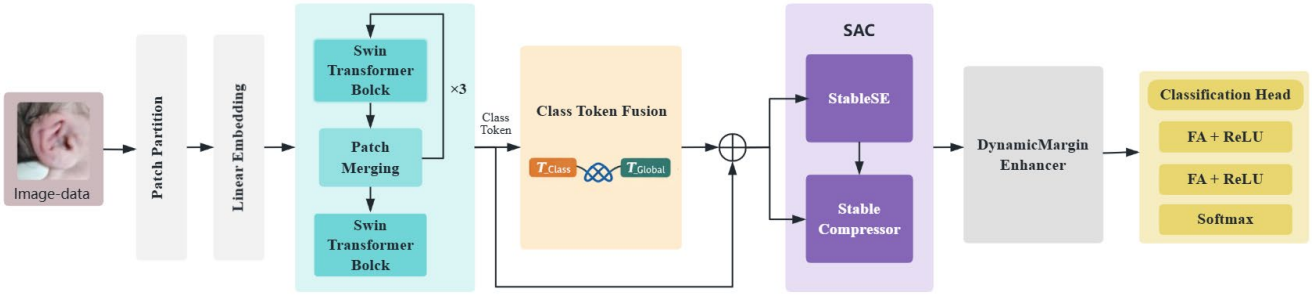

Fig.1 Overall Architecture of HRS-Swin

(2) Class Token Fusion

Traditional Transformer-based classification typically relies on global pooling or a single class token representation. However, in medical imaging tasks where inter-class differences are often subtle, a single feature representation may be insufficient to fully capture discriminative class semantics. To address the lack of an explicit global class representation in pure image feature extraction with Swin Transformer, we design a Class Token Fusion (CTF) module. This module introduces a learnable class token and fuses it with the feature vector produced by the backbone network, thereby enhancing global contextual correlations within the representation.

The core idea of the CTF module is not simple feature concatenation, but the construction of a global semantic anchor. Specifically, the model maintains a learnable class token parameter that gradually converges toward a discriminative direction within the data distribution during training. After concatenation with the sample features and linear transformation, the fused representation is fed back into the original feature space through a residual connection, enabling directional refinement of the original representation. This process is equivalent to introducing a learnable semantic basis vector in the feature space to regulate the projection direction of sample embeddings. Unlike conventional classification heads that perform category mapping only at the final layer, the proposed module introduces category-related constraints at a high-dimensional semantic stage, thereby improving the separability of intermediate representations.

(3) Stable Attention Compressor

This study constructs a two-stage collaborative structure composed of Stable SE and Stable Compressor to address feature redundancy and the curse of dimensionality in auricular deformity classification, achieving channel recalibration and compact representation learning.

The Stable SE module is inspired by channel attention mechanisms. It adaptively learns the importance weight of each feature dimension and performs dynamic recalibration of the original semantic vector. Instead of purely multiplicative scaling, we adopt a residual enhancement strategy, which strengthens informative channels while preserving information continuity. Concretely, the Stable SE module first applies a dimensionality-reduction linear layer $W_1 \in \mathbb{R}^{(d/r)\times d}$ to squeeze the input feature x, followed by a ReLU activation. The representation is then restored through an expansion linear layer $W_2 \in \mathbb{R}^{d\times(d/r)}$, and a Sigmoid function $\sigma(\cdot)$ is used to generate the channel attention weight vector w. as shown in Eq. (1):

$$z = W_2\big(ReLU(W_1 x)\big), w = \sigma(z) \in [0,1]^d \quad (1)$$

The attention weights w are applied via channel-wise scaling and added back to the input in a residual manner, producing the enhanced feature x'.

The Stable Compressor then performs residual dimensionality reduction. The feature is first reduced by $W_1 \in \mathbb{R}^{512\times d}$, followed by BN and GELU to obtain h_1, and further compressed by $W_2 \in \mathbb{R}^{d_{\text{out}}\times 512}$ to produce h_2. A shortcut projection $W_s \in \mathbb{R}^{d_{\text{out}}\times d}$ maps the original feature to the same low-dimensional space. as formulated in Eq. (2):

$$x_{\text{comp}} = h_2 + W_s x \quad (2)$$

This design significantly reduces the feature dimensionality (with a compression rate of approximately 87.5%) while effectively retaining discriminative information. Within the collaborative mechanism, Stable SE first enhances informative features, and Stable Compressor subsequently removes redundancy. This ordered design optimizes information flow. For auricular deformity classification, excessively high-dimensional representations may blur class boundaries, whereas moderate compression facilitates the formation of clearer class clustering structures in the embedding space.

(4) Dynamic Margin Enhancer

Insufficient inter-class margins are a common issue in multi-class medical classification tasks, particularly between morphologically similar categories. To strengthen decision boundaries, the Dynamic Margin Enhancer (DME) introduces a batch-level adaptive margin expansion mechanism in the low-dimensional embedding space. This module addresses class imbalance and boundary ambiguity (e.g., mild auricular deformities versus normal ears) through batch-center computation and learnable deviation scaling for adaptive feature enhancement. Given a batch of features, the batch center μ is computed as shown in Eq. (3):

$$\mu = \frac{1}{B}\sum_{i=1}^{B} x_i \in \mathbb{R}^{d_{\text{out}}} \quad (3)$$

For each sample feature x_i, the deviation from the batch center is scaled element-wise by a learnable vector $s \in \mathbb{R}^{d_{\text{out}}}$(initialized as ones) and controlled by a hyperparameter β(set to 0.6). The enhanced representation is obtained via residual expansion, shown in Eq. (4):

$$x_i' = x_i + \beta \cdot (x_i - \mu) \odot s \quad (4)$$

This dynamic deviation amplification enlarges inter-class margins while promoting intra-class compactness, leading to clearer class separation in the embedding space, particularly for subtly different auricular deformities.

3. Experiment

(1) Dataset

This study utilizes the BabyEar4K dataset, which includes bilateral ear photographs of 1,926 newborns collected from November 2018 to April 2019. Both left and right ear images were acquired for each subject, along with corresponding maternal and neonatal health data. All data were anonymized and publicly released with informed parental consent. The dataset covers several common congenital auricular deformities, including Cup ear, Stahl's ear, Lop ear, and Helical deformities, as well as normal ears. Some categories exhibit subtle morphological differences, posing challenges for fine-grained classification. All images were captured under standardized clinical conditions and were preprocessed (cropping and resizing) before model training. The dataset was split into training, validation, and test sets for performance evaluation.

(2) Metrics

To comprehensively evaluate the proposed model, we adopt widely recognized metrics in medical image classification, including Accuracy, Precision, Recall, and F1-score.

Accuracy measures the overall proportion of correctly classified samples and is defined as:

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} m_{kk}}{\sum_{i=1}^{K}\sum_{j=1}^{K} m_{ij}} = \frac{\text{Tr}(M)}{N} \quad (5)$$

where $M$is the confusion matrix, $\text{Tr}(M)$its trace (correct predictions), and N the total number of samples. Since Accuracy may be biased under class imbalance, additional metrics are considered.

For multi-class evaluation, we adopt macro-averaged Precision, Recall, and F1-score, defined as:

$$\text{Precision}_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} \text{Precision}_k \quad (6)$$

$$\text{Recall}_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} \text{Recall}_k \quad (7)$$

$$\text{F1}_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} \text{F1}_k \quad (8)$$

Precision reflects false positive control, Recall reflects detection capability (false negative rate), and F1-score balances both. In auricular deformity classification with class imbalance and subtle inter-class differences, macro F1-score provides a more comprehensive assessment of model performance.

(3) Setting

All experiments were conducted on a high-performance GPU server with a unified hardware and software configuration. The operating system was Ubuntu 18.04. The server was equipped with eight NVIDIA GeForce GTX 1080Ti GPUs (11 GB memory per GPU), which were used to accelerate parallel training on large-scale image data. The implementation was based on Python 3.8 and TensorFlow 2.13.1. The dataset was divided into training, validation, and test sets with a ratio of 6:2:2. The model was trained for 60 epochs with a batch size of 32. For optimization, Focal Loss was employed to address class imbalance. The AdamW optimizer was used with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 5×10^{-2}. The initial learning rate was set to $1 \times 10^{-5}$and gradually decayed to 1×10^{-7}using a cosine annealing schedule.

(4) Results

The quantitative comparison results on the BabyEar4K dataset are summarized in Table 1. The proposed HRS-Swin model achieves the best performance across all evaluation metrics compared with representative CNN- and Transformer-based baselines, including GoogLeNet, a standard Transformer, YOLOv8, and Swin Transformer.

Table 1. Overall performance comparison between HRS-Swin and existing methods

Model	ACC	F1	Recall	Precision
[8]	0.7281	0.5256	0.5233	0.5189
[9]	0.7241	0.5472	0.5332	0.5568
[10]	0.7689	0.6031	0.6029	0.6600
[11]	0.7728	0.6140	0.6621	0.7130
Ours	**0.8009**	**0.7024**	**0.6776**	**0.7655**

Specifically, HRS-Swin attains an Accuracy of 0.8009, a macro F1-score of 0.7024, a macro-Recall of 0.6776, and a macro-Precision of 0.7655. Compared with the baseline Swin Transformer, HRS-Swin improves Accuracy from 0.7728 to 0.8009 (+2.81%), F1-score from 0.6140 to 0.7024 (+8.84%), Recall from 0.6621 to 0.6776, and Precision from 0.7130 to 0.7655. Notably, the improvement in macro F1-score demonstrates enhanced balance between false positives and false negatives, which is particularly important in medical classification tasks with class imbalance and subtle inter-class variations.

Compared with CNN-based methods such as GoogLeNet and YOLOv8, HRS-Swin shows more substantial gains, indicating that hierarchical representation separation and margin enhancement strategies effectively improve discriminative feature learning beyond conventional convolutional architectures.

Overall, the results confirm that the proposed progressive enhancement–compression–separation framework significantly strengthens feature separability and robustness, leading to superior performance in neonatal auricular deformity classification.

4. Conclusion

This study proposes a hierarchical residual-enhanced multimodal framework, HRS-Swin, for automated classification of neonatal auricular deformities. By leveraging the hierarchical representation capability of the Swin Transformer backbone and integrating structured clinical features through a Transformer-based tabular encoder, the

proposed model effectively captures both fine-grained morphological variations and complementary clinical semantics. The residual cross-modal fusion (RCF) module further strengthens feature interaction at the representation level, improving discriminative consistency across subtypes.

Experimental results demonstrate that HRS-Swin achieves superior overall accuracy and macro-F1 compared with conventional convolutional architectures such as GoogLeNet and generic Transformer-based baselines. The model shows particular robustness in handling inter-class similarity and intra-class variability, which are common challenges in auricular deformity diagnosis. Moreover, the incorporation of automated ear region localization based on YOLOv8 enhances preprocessing standardization and reduces background interference, contributing to stable performance across different image conditions.

In summary, HRS-Swin provides a reliable and scalable auxiliary diagnostic tool for neonatal auricular deformity screening. Its multimodal design not only improves classification performance but also offers a practical framework for integrating medical imaging and structured clinical data in other pediatric diagnostic scenarios. Future work will focus on expanding the dataset scale, exploring self-supervised pretraining strategies, and improving interpretability through attention visualization and clinical correlation analysis.

References

[1] Byrd H S, Langevin C J, Ghidoni L A. Ear molding in newborn infants with auricular deformities [J]. Plastic and reconstructive surgery, 2010, 126(4): 1191-1200.

[2] Wu H, Niu Z, Li G, Li Y, Wang B, Qian J, Wang Y, Jiang H, Chen Y, Han Y. Nonsurgical Treatment for Congenital Auricular Deformities: A Systematic Review and Meta-analysis. Aesthetic Plast Surg. 2022 Feb;46(1):173-182. doi: 10.1007/s00266-021-02427-9. Epub 2021 Jul 6. PMID: 34231021.

[3] Hallac R R, Lee J, Pressler M, et al. Identifying ear abnormality from 2D photographs using convolutional neural networks[J]. Scientific reports, 2019, 9(1): 18198.

[4] Galdámez P L, Raveane W, Arrieta A G. A brief review of the ear recognition process using deep neural networks[J]. Journal of Applied Logic, 2017, 24: 62-70.

[5] Tajbakhsh N, Shin J Y, Gurudu S R, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning?[J]. IEEE transactions on medical imaging, 2016, 35(5): 1299-1312.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[7] Hallac R R, Jackson S A, Grant J, et al. Assessing outcomes of ear molding therapy by health care providers and convolutional neural network[J]. Scientific reports, 2021, 11(1): 17875.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, et al., Going Deeper with Convolutions, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al., Attention Is All You Need, in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[10] Z. Liu, H. Hu, Y. Lin, Z. Yao, et al., Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, arXiv preprint, 2021.

[11] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, 2023. GitHub repository: https://github.com/ultralytics/ultralytics (include as software citation since no formal published paper exists).