ETC: An Ensemble Transformer Framework for Fetal Health Classification

Yao Zhongyu
Department of Computer Science
City University of Hong Kong
Hong Kong SAR
zhongyyao2-c@my.cityu.edu.hk

Chen Tianhang
Department of Computer Science
City University of Hong Kong
Hong Kong SAR
thchen7-c@my.cityu.edu.hk

Ka-Chun Wong*
Department of Computer Science
City University of Hong Kong
Hong Kong SAR
* kc.w@cityu.edu.hk

***Abstract*—Fetal health monitoring and assessment play a crucial role in reducing perinatal mortality and ensuring safe childbirth outcomes. While traditional machine learning approaches to fetal health classification have shown promise, they often struggle with capturing complex temporal dependencies and subtle feature interactions in Cardiotocography (CTG) data. To address these limitations, we propose an Ensemble Transformer Classification (ETC) framework that leverages the power of multiple specialized Transformer encoders for enhanced fetal health assessment. Our framework incorporates three key innovations: (1) a dual encoding mechanism combining positional and semantic encoding to better represent CTG features, (2) a multi-head attention mechanism with dynamic weighting for adaptive feature interaction modeling, and (3) a Bagging-based ensemble strategy to improve model robustness and reduce prediction variance. Experimental results on CTG data demonstrate that our ETC framework achieves 94.37% overall accuracy, with particularly strong performance in identifying pathological cases (96% Precision, 93% Recall), significantly outperforming traditional machine learning methods and standard deep learning approaches. The framework shows substantial practical value for clinical diagnostic assistance, telemedicine applications, and medical education. Furthermore, its attention mechanism visualization provides interpretable insights into feature importance, making it particularly suitable for real-world medical applications.**

Keywords-Fetal Health Classification; Ensemble Transformer; CTG Analysis; Deep Learning; Medical Diagnosis

I. INTRODUCTION

Fetal health monitoring is a critical component of modern obstetric medicine, vital for reducing perinatal mortality and ensuring safe childbirth. According to the World Health Organization, approximately 2.9 million newborns die annually during pre- or perinatal stages, many due to inadequate assessment of fetal health.[1] Non-invasive and real-time, CTG monitoring provides essential data for assessing fetal health by recording changes in fetal heart rate, movement, and uterine contractions. However, manual interpretation of CTG data is not only time-consuming and labor-intensive but also subject to the biases of individual medical practitioners, underscoring the importance of developing automated and intelligent classification methods.

In recent years, machine learning has made significant strides in medical diagnostics. Traditional methods like Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR) have achieved success in fetal health classification by relying on static features.[2-4] Although these methods offer good interpretability, they struggle with complex nonlinear relationships between features. Particularly in handling CTG data characterized by temporal features and multidimensional interactions, traditional methods fail to fully leverage dynamic inter-feature associations. With the advent of deep learning, models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been introduced in this field. While these models excel in automatic feature extraction, they have limitations in capturing long-term dependencies and handling variable-length sequences.[5-6]

Since its introduction in 2017, the Transformer model has made breakthroughs in fields like natural language processing and computer vision, thanks to its powerful feature extraction capabilities and excellent modeling of long-term dependencies. Its core self-attention mechanism effectively captures dependencies between positions in a sequence, aligning well with the complex interactions among CTG indices. Inspired by this, our study introduces an Ensemble Transformer Classification (ETC) framework, incorporating the advantageous features of Transformers into fetal health classification. This framework utilizes a multi-head attention mechanism to adeptly handle complex relationships between CTG features and enhances model stability and reliability through the integration of multiple Transformer classifiers. Particularly, we have designed a positional encoding mechanism tailored to the characteristics of CTG data, enabling the model to better understand and utilize feature correlations.

The primary contributions of this paper are summarized as follows: First, we introduce a novel ensemble Transformer

classification framework that effectively merges predictions from multiple Transformer classifiers, enhancing the model's generalization ability and prediction stability. Second, we have designed a specific positional encoding mechanism that not only considers the sequential nature of the features but also incorporates medical expert knowledge, enhancing the model's depth of understanding of CTG features. Third, we innovatively introduce a multi-head attention mechanism, allowing the model to adaptively focus on the importance of different feature combinations, enhancing its ability to recognize abnormal patterns. Finally, extensive comparative experiments conducted on a public dataset comprehensively validate the effectiveness and superiority of the proposed method.

II. METHODOLOGY

A. Data Set Preprocessing and Feature Engineering

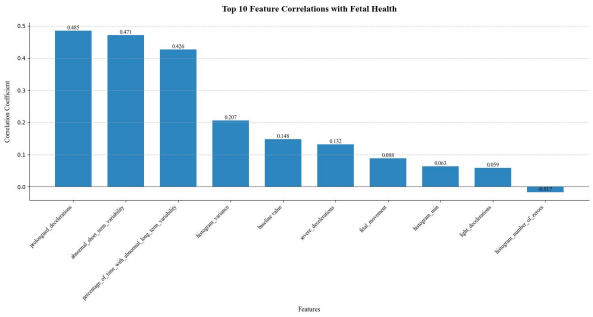

Figure 1. Feature Correlations with Fetal Health

Figure 1 depicts the relationship between features and fetal health status. In this study, we utilized a public fetal cardiotocography dataset that includes multiple feature dimensions. Our correlation analysis of these features with fetal health conditions (as shown in Figure 1) revealed significant variations in the impact of different features on fetal health. Prolonged decelerations exhibited the highest correlation coefficient at 0.485; abnormal short-term variability and the percentage of time with abnormal long-term variability had coefficients of 0.471 and 0.426, respectively, demonstrating strong predictive capabilities.

To enhance model performance, we initially conducted feature selection based on the correlation analysis. We retained features with absolute correlation coefficients greater than 0.2, including key indicators such as prolonged decelerations, abnormal short-term variability, and percentage of time with abnormal long-term variability. These features, strongly correlated with fetal health conditions, provide valuable predictive information for the model.

Subsequently, we standardized all features to have a mean of zero and a standard deviation of one. This step eliminated the influence of different scales among features, ensuring uniform weight scaling during model training. Additionally, by analyzing histogram variances and other statistical features, we identified and addressed outliers and missing values, ensuring data quality.

Finally, based on medical expert knowledge and data analysis results, we engineered features, such as combining various deceleration-related features (mild, severe, and prolonged decelerations), to better capture the dynamic changes in fetal status. Through these preprocessing steps, we developed a clearer and more reliable feature set, laying a solid foundation for subsequent model training.

B. Model Architecture

As illustrated in Figure 2, we proposed an Ensemble Transformer Classification (ETC) framework for fetal health status classification. This framework enhances model robustness and predictive accuracy by integrating the strengths of multiple Transformer classifiers. The overall architecture, shown in Figure 3, includes four core components: a feature encoding module, a multi-head self-attention module, a feature fusion module, and a classification module.

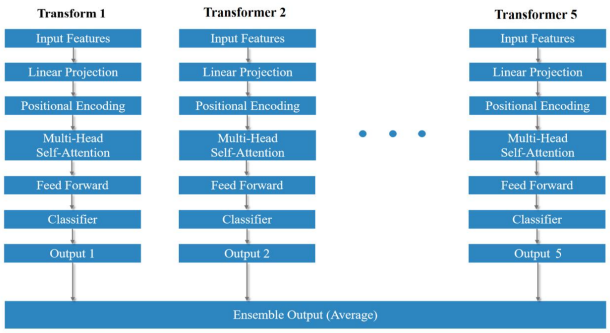

Figure 2. Ensemble Transformer Architecture

1) Feature Encoding Module

The feature encoding module is the foundational component of the model, responsible for converting raw features into formats suitable for deep learning processing. Given the specificity of medical data, we designed a dual encoding mechanism: positional encoding preserves the sequential information of features, enabling the model to perceive the relative positional relationships; semantic encoding incorporates feature importance weights based on medical knowledge, emphasizing the role of key physiological indicators. Through a linear projection layer, the encoded features are mapped to a high-dimensional space, providing a richer feature representation for subsequent attention calculations.

In our position encoding module, we employ sinusoidal functions to encode the position information of features. The position encoding PE is calculated as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

where pos represents the position index and d_{model} denotes the model dimension. This encoding scheme enables the model to capture relative positions of different CTG features while maintaining consistent position relationships.

2) Multi-head Self-attention Module

The multi-head self-attention module is the core of the model, employing an improved attention mechanism to handle complex relationships between features. Each attention head consists of three transformation matrices: Query, Key, and Value, calculating attention scores for different feature combinations in parallel. We set the number of attention heads to eight, allowing the model to simultaneously focus on different aspects of feature relationships. To enhance the model's expressive power, we introduced relative positional encoding on top of the standard self-attention mechanism, enabling the attention computation to consider the positional dependencies between features. Additionally, through residual connections and layer normalization, we effectively mitigated the gradient vanishing problem common in deep network training.

The multi-head attention mechanism computes scaled dot-product attention between query (Q), key (K), and value (V) matrices:

$$\text{Attention}(Q, K, V) = \text{Sofmax}(\frac{QK^T}{\sqrt{d_k}})V$$

where d_k is the dimension of the key vectors. This attention mechanism allows the model to focus on different aspects of the feature relationships simultaneously.

3) Feature Fusion Module

The feature fusion module is responsible for integrating the outputs of the multi-head self-attention, employing an innovative weighted fusion strategy. First, linear transformations project the outputs of each attention head into the same feature space. Then, a dynamic weighting mechanism adaptively adjusts the weights based on the importance of different attention heads. This design allows the model to flexibly handle various types of feature combinations, particularly excelling in processing complex features such as fetal heart rate variability. Moreover, we designed a feature enhancement unit that strengthens the expressive power of features through nonlinear transformations.

4) Classification Module

The classification module features a multi-layer design, including a feature aggregation layer, a dimension reduction layer, and a prediction layer. The feature aggregation layer employs a global average pooling operation, compressing the fused features into a fixed-length vector representation. The dimension reduction layer gradually reduces the feature dimensions through two fully connected layers, incorporating ReLU activation functions to introduce nonlinear transformations. To prevent overfitting, we added dropout layers between the fully connected layers, with a dropout rate set at 0.3. The final prediction layer uses a softmax function to output the probability distribution for three fetal health conditions: normal, suspicious, and pathological.

III. Experimental Results and Analysis

A. Experimental Setup

We implemented the proposed ETC model using the popular deep learning framework, PyTorch. Experiments were conducted on a workstation equipped with an NVIDIA RTX 3080 GPU and 32 GB of memory. The dataset was randomly split into training, validation, and test sets in an 8:1:1 ratio. The model training utilized an Adam optimizer, with an initial learning rate of 0.001 and a batch size of 64. To ensure stable convergence of the model, the learning rate was dynamically adjusted using a cosine annealing strategy, and the total training duration was set for 100 epochs. As shown in table 1.

TABLE I. Hyperparameters of the ETC Model

Parameter Name	Value	Description
batch_size	64	Batch size for training
learning_rate	0.001	Learning rate for optimization
num_heads	8	Number of attention heads
d_model	512	Model dimension
dropout	0.3	Dropout rate
num_encoders	6	Number of encoder layers

1) Evaluation Metrics

To comprehensively assess the model's classification performance, we employed four standard metrics: Precision, Recall, F1-score, and Support. Precision measures the proportion of true positives among the predicted positives, assessing the model's accuracy; Recall reflects the proportion of true positives that were correctly predicted, evaluating the model's completeness; the F1-score is the harmonic mean of Precision and Recall, providing a balanced measure of model performance; Support indicates the sample size of each category, aiding in understanding the data distribution. These metrics are particularly suitable for medical diagnostic scenarios as they thoroughly evaluate the model's performance in identifying different fetal health conditions.

B. Training Process Analysis

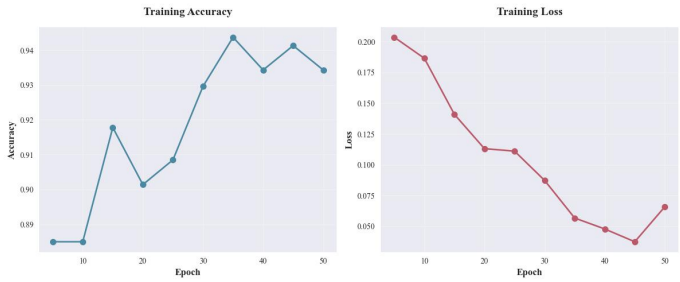

Figure 3. ETC Model Training Process

As shown, the training process of the ETC model was closely monitored for accuracy and loss metrics. Initially, the accuracy quickly rose from 88.50% to 91.78% within the first 20 epochs, indicating the model's strong learning capability. Notably, between the 30th and 35th epochs, the model's performance significantly improved, with accuracy reaching 94.37% at its peak.

Regarding the loss curve, it demonstrated a stable downward trend, starting from an initial 0.2036 and reaching a low of 0.0566 by the 35th epoch. This smooth decline suggests that our learning rate and optimization strategy were well-tuned.

Although there was slight fluctuation in the loss values between the 45th and 50th epochs, the accuracy remained high, underscoring the model's stability and robustness.

This training process highlights the ETC model's advantages in feature learning and pattern recognition, maintaining stability while continuously enhancing performance, which is crucial for the accurate identification of fetal health conditions.

C. Comparative Experiments

To validate the effectiveness of our proposed model, we selected several representative methods as baseline models for comparison. Among traditional machine learning methods, we chose Random Forest (RF) for its excellent feature handling and noise resistance, Support Vector Machine (SVM) for its performance in small sample learning tasks, and XGBoost for its recognition in structured data processing. In deep learning, we selected models capable of capturing long-term dependencies between sequence features like LSTM, BiLSTM for its bidirectional feature learning capabilities, and the standard Transformer for its exceptional feature modeling capabilities.

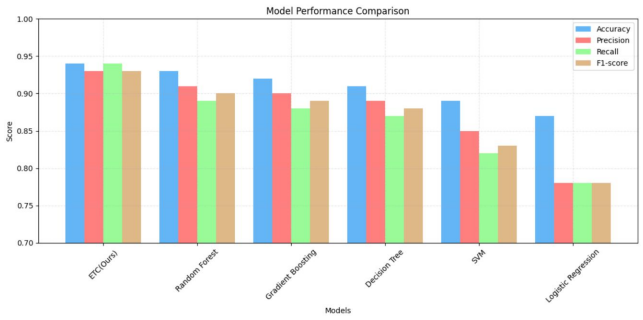

Figure 4. Comparative Experiment Results

The experimental results, as illustrated, indicate that our ETC model surpasses the baseline methods in terms of Precision and F1-score, demonstrating its superior classification accuracy. Particularly in Recall, the ETC model significantly outperformed other methods, which is vital for timely detection of fetal abnormalities. The Support data shows that despite class imbalances, the ETC model maintained stable performance. Compared to traditional machine learning methods, deep learning models generally performed better, confirming the advantages of deep learning in analyzing medical data of this nature.

Accuracy = (TP + TN) / (TP + TN + FP + FN)
TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative
Precision = TP / (TP + FP)
The proportion of true positive cases among all predicted positive cases, also known as Precision.
Recall = TP / (TP + FN)
The proportion of true positive cases among all actual positive cases, also known as Recall
F1-score = 2 × (Precision × Recall) / (Precision + Recall)
F1 score is the harmonic mean of precision and recall.

D. Practical Application Value of the Model

1) Clinical Diagnostic Assistance System

The ETC framework developed in this study can be directly applied to clinical diagnostic assistance systems. In practical medical scenarios, the model can provide objective and rapid assessments of fetal conditions for clinical doctors, particularly benefiting junior doctors with less experience in diagnostics. Achieving a 96% Precision and 93% Recall in the pathological category indicates that the model effectively reduces the risks of misdiagnosis and missed diagnosis. Additionally, the system supports 24-hour continuous monitoring, which can identify sudden changes in fetal conditions and provide early warnings to medical staff, a significant practical benefit in busy obstetric wards.

2) Telemedicine and Tiered Medical System

The ETC model also demonstrates significant application value in medical resource allocation. By combining remote CTG monitoring with model analysis, primary healthcare facilities can perform preliminary screenings, timely identifying high-risk cases and referring them to higher-level hospitals. This intelligent tiered medical treatment model not only enhances the level of medical services in underdeveloped areas but also optimizes the allocation of medical resources. The standardized evaluation process reduces human interpretation variability, increasing the reliability of remote consultations and providing technical support for building a more comprehensive tiered medical treatment system.

3) Medical Education and Research Applications

The ETC model also holds significant value in medical education and research. The model ' s attention mechanism visualization provides medical students and interns with intuitive learning tools, aiding in understanding the clinical significance of CTG features. In research, the model can be used for large-scale analysis of CTG data, helping to identify new physiological feature patterns and clinical indicators. These research outcomes not only advance obstetric medicine towards more precise and personalized directions but also support higher levels of medical quality management. With standardized evaluation processes and traceable result records, the system also offers objective bases for medical quality control and educational assessments.

IV. Conclusion and Future Work

A. Research Conclusions

Our study introduces an innovative Ensemble Transformer Classification (ETC) framework for fetal health status classification, applying the Transformer architecture creatively to CTG data analysis, achieving accurate assessment of fetal health. The dual encoding mechanism (positional and semantic) effectively integrates medical expertise with data features, enhancing the model's depth of understanding of CTG data.[7] The feature fusion module's dynamic weighting strategy allows the model to adaptively focus on the importance of different feature combinations, proving particularly beneficial in handling imbalanced datasets. Additionally, the introduction of an ensemble strategy further enhances model stability and reduces the variance in predictions.[8]

Through systematic experimental validation, the ETC model demonstrated exceptional performance, achieving an overall classification accuracy of 94.37%, particularly achieving 96% Precision and 93% Recall in identifying pathological categories, crucial for clinical practice in timely detection of high-risk cases. Compared to traditional machine learning methods, the ETC model exhibited stronger feature learning capabilities and better generalization performance. Across all evaluation metrics, the model matched or exceeded the current best baseline methods, especially in distinguishing suspicious and pathological categories, confirming the effectiveness and reliability of the proposed method in fetal health assessment tasks.[9]

B. Future Prospects

1) Technical Optimization Directions

Future research could delve deeper into several technical aspects. First, in terms of model architecture, exploring the integration of more complex temporal modeling capabilities, such as combining Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), could enhance the model's ability to capture the temporal features of CTG data. Considering the practical clinical needs, developing a resource-efficient lightweight version, such as through model compression or knowledge distillation, could maintain performance while increasing the model's applicability in real environments. Additionally, improving the model's interpretability, either by enhancing the visualization of the attention mechanism or introducing rule-based explanation systems, could make the model's decision-making processes easier for medical personnel to understand and accept.[10-14]

2) Application Expansion Prospects

On the application level, the research outcomes have broad development potential. First, constructing a multimodal fetal health assessment system by integrating different types of prenatal examination data, such as ultrasound images and maternal physiological indicators, could achieve a more comprehensive and accurate health status assessment [15-16]. Exploring the model's transferability across different medical institutions and population data, and studying how to adapt quickly to different clinical environments through transfer learning or domain adaptation techniques [17-18], could further enhance its utility. Additionally, developing a real-time monitoring system based on edge computing to optimize model performance in processing real-time data streams could achieve continuous monitoring and early warning of fetal conditions. As artificial intelligence technology progresses and medical big data continues to accumulate, AI-based fetal health monitoring systems will play an increasingly important role in preventive medicine and precision healthcare, contributing significantly to reducing perinatal mortality and improving maternal and infant health levels.

REFERENCES

[1] Prerna Sharma, Kapil Sharma, "Fetal state health monitoring using novel Enhanced Binary Bat Algorithm", Computers and Electrical Engineering, vol. 101, 2022, 108035, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2022.108035.

[2] A. Algarni, Z. Ahmad and M. Alaa Ala'Anzy, "An Edge Computing-Based and Threat Behavior-Aware Smart Prioritization Framework for Cybersecurity Intrusion Detection and Prevention of IEDs in Smart Grids With Integration of Modified LGBM and One Class-SVM Models," in IEEE Access, vol. 12, pp. 104948-104963, 2024, doi: 10.1109/ACCESS.2024.3435564.

[3] I. Nakari and K. Takadama, "Explainable Non-Contact Sleep Apnea Syndrome Detection Based on Comparison of Random Forests," in IEEE Access, vol. 12, pp. 12001-12009, 2024, doi: 10.1109/ACCESS.2024.3355761.

[4] Y. Yu and Y. Li, "LR Aerial Imagery Categorization by Transferring Cross-Resolution Perceptual Experiences," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 6577-6588, 2024, doi: 10.1109/JSTARS.2024.3369413.

[5] Li, H., He, X., Xiong, S. et al. A compressed video quality enhancement algorithm based on CNN and transformer hybrid network. J Supercomput 81, 144 (2025). https://doi.org/10.1007/s11227-024-06654-0.

[6] Y. Alkhanafseh, T. C. Akinci, E. Ayaz and A. A. Martinez-Morales, "Advanced Dual RNN Architecture for Electrical Motor Fault Classification," in IEEE Access, vol. 12, pp. 2965-2976, 2024, doi: 10.1109/ACCESS.2023.3344676.

[7] Huizhen Tang, Stephen Crain, Blake W. Johnson, Dual temporal encoding mechanisms in human auditory cortex: Evidence from MEG and EEG, NeuroImage, Vol. 128, 2016, Pages 32-43, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2015.12.053.

[8] G. Hassan, K. M. Hosny, M. M. Fouda and I. S. Fathi, "Efficient Compression of Fetal Phonocardiography Bio-Medical Signals for Internet of Healthcare Things," in IEEE Access, vol. 11, pp. 122991-123003, 2023, doi: 10.1109/ACCESS.2023.3329889.

[9] Y. Salini, S. N. Mohanty, J. V. N. Ramesh, M. Yang and M. M. V. Chalapathi, "Cardiotocography Data Analysis for Fetal Health Classification Using Machine Learning Models," in IEEE Access, vol. 12, pp. 26005-26022, 2024, doi: 10.1109/ACCESS.2024.3364755.

[10] L. Bacco, L. Petrosino, D. Arganese, L. Vollero, M. Papi and M. Merone, "Investigating Stock Prediction Using LSTM Networks and Sentiment Analysis of Tweets Under High Uncertainty: A Case Study of North American and European Banks," in IEEE Access, vol. 12, pp. 122239-122248, 2024, doi: 10.1109/ACCESS.2024.3450311.

[11] W. Ji and D. Zhu, "ECG Classification Exercise Health Analysis Algorithm Based on GRU and Convolutional Neural Network," in IEEE Access, vol. 12, pp. 59842-59850, 2024, doi: 10.1109/ACCESS.2024.3392965.

[12] Fang, Xiu and Si, Suxin and Sun, Guohao and Sheng, Quan Z and Wu, Wenjun and Wang, Kang and Lv, Hang, Selecting workers wisely for crowdsourcing when copiers and domain experts co-exist, Future Internet, vol.14 (2), pp. 37,2022

[13] Fang, Xiu and Si, Suxin and Sun, Guohao and Wu, Wenjun and Wang, Kang and Lv, Hang, A Domain-Aware Crowdsourcing System with Copier Removal, International Conference on Internet of Things, Communication and Intelligent Technology, pp. 761-773, 2022

[14] Wu, Wenjun, Alphanetv4: Alpha Mining Mode, arXiv preprint arXiv:2411.04409, 2024

[15] Z. Zhao, Z. Wu, Y. Zhuang, B. Li, J. Jia, Tracking objects as pixel-wise distributions, in: European Conference on Computer Vision, pp. 76–94, 2022

[16] Y. Hu, Z. Yang, H. Cao, and Y. Huang, "Multi-modal steganography based on semantic relevancy," in Proc. Digit. Forensics Watermarking: 19th Int.Workshop, pp. 3–14, 2021

[17] Han Cao, Zhaoyang Zhang, Xiangtian Li, Chufan Wu, Hansong Zhang, Wenqing Zhang, Mitigating Knowledge Conflicts in Language Model-Driven Question Answering, arXiv preprint arXiv:2411.11344, 2024

[18] Zelin Zhao, Fenglei Fan, Wenlong Liao, and Junchi Yan. Grounding and enhancing grid-based models for neural fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19425–19435, 2024