Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.083525

ARTICLE

Decoupling of Multi-View Facial Features for Cushing's Syndrome Diagnosis

Changwei Song[1], Jiaqi Qiang[2], Hongjun Liu[1], Jianqiang Li[1], Hui Pan[2], Qing Zhao[1,*], Jiuzuo Huang[3] and Shi Chen[3]

[1]School of Computer Science, Beijing University of Technology, Beijing, China
[2]Key Laboratory of Endocrinology of National Health Commission, Department of Endocrinology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[3]Department of Plastic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

*Corresponding Author: Qing Zhao. Email: zhaoqing@bjut.edu.cn

Received: 05 April 2026; Accepted: 15 June 2026; Published: 27 July 2026

ABSTRACT: Cushing's syndrome (CS) is a rare endocrine disorder characterized by chronic hypercortisolism, and facial image-based intelligent diagnosis has emerged as a promising non-invasive approach. However, existing diagnostic models suffer from two core bottlenecks: inefficient fusion of deep semantic features and clinical prior features, and insufficient multi-view facial feature disentanglement without CS-specific pathophysiological constraints. To address these limitations, we propose a novel Multi-View Facial Feature Disentanglement Network (MVFFD-Net) for high-precision automatic CS diagnosis. The network takes five standard facial views (frontal, bilateral 45° oblique, and bilateral 90° lateral views) as input, with three key innovations: a bidirectional cross-attention module for synergistic fusion of deep and clinical prior features, a multi-path disentanglement architecture with multi-objective loss for separating view-invariant and view-specific features, and a graph attention fusion framework for adaptive multi-view feature integration. Unlike state-of-the-art models that mainly rely on single-view representations or generic multi-view fusion, MVFFD-Net explicitly integrates clinically guided multimodal fusion, CS-oriented feature disentanglement, and pathology-aware graph-based multi-view aggregation. Experimental results demonstrate that MVFFD-Net achieves an F1-score of 98.78% for CS diagnosis, significantly outperforming representative state-of-the-art methods in the current experimental setting. In this study, the task is defined as subject-level AI-assisted diagnosis to distinguish individuals with CS from matched controls using five standard facial views. Accordingly, the proposed framework is intended to provide non-invasive auxiliary diagnostic support rather than replace standard endocrinological evaluation and biochemical confirmation. More broadly, this clinically informed multi-view learning framework shows promising potential as a non-invasive auxiliary screening tool, though external multi-center validation remains necessary prior to large-scale clinical application.

KEYWORDS: Cushing's syndrome; feature fusion; feature disentanglement; multi-view learning; deep learning

1 Introduction

Cushing's syndrome (CS) is a group of rare endocrine disorders characterized by persistently elevated cortisol levels as the core feature. The complexity of its pathogenesis and the multisystem involvement of its clinical manifestations pose significant challenges to clinical diagnosis and treatment [1]. Due to the long-term excessive secretion of glucocorticoids, the disease induces a series of systemic impairments, including immune dysfunction, metabolic imbalance, and neuropsychiatric symptoms. Typical clinical manifestations include central obesity, supraclavicular fat pad accumulation, skin thinning with striae

formation, proximal muscle weakness, refractory hypertension, abnormal glucose metabolism, acne, and hirsutism. Additionally, impaired cognitive function may occur in some cases. These clinical manifestations exhibit high heterogeneity, which not only increases the difficulty of early disease identification but also often results in a diagnostic delay of 2 to 6 years [2]. This delay significantly elevates the risk of complications and mortality in patients. Therefore, the development of efficient and accurate screening and diagnostic tools has become an urgent need in the current field of clinical management of CS.

Nevertheless, the clinical demand for an accessible, non-invasive, and reliable auxiliary screening tool for CS remains insufficiently addressed. CS is particularly suitable for facial image-based analysis because its characteristic phenotypes are directly manifested in facial morphology and texture; however, these diagnostic cues are distributed across different viewpoints and are difficult to capture adequately by single-view models or generic fusion strategies. Therefore, the core motivation of this study is to develop a clinically informed multi-view diagnostic framework that can more effectively integrate complementary facial information and disease-specific prior knowledge for early and accurate CS identification.

In recent years, advances in artificial intelligence (AI)-driven medical image analysis [3,4] have established novel paradigms for non-invasive CS diagnosis, positioning facial image-based intelligent detection as a burgeoning research hotspot at the intersection of endocrinology and computer vision. Pioneering work by Popp et al. [5] and Kosilek et al. [6] first validated the feasibility of this approach, extracting handcrafted texture and geometric features from frontal and lateral facial images paired with traditional machine learning classifiers for disease identification, and establishing the foundational methodological framework for the field. Subsequently, deep learning emerged as the mainstream methodology: Wei et al. [7] developed a fully automatic convolutional neural network (CNN)-based diagnostic pipeline via facial detection, alignment and deep feature extraction from frontal images, markedly improving diagnostic efficiency and accelerating clinical translation, though CNNs' intrinsic inductive bias toward local feature extraction limits their capture of global facial phenotypes pathognomonic for CS. Liu et al. [8] addressed this limitation using the Vision Transformer (ViT)-based DINOv2 model, whose self-attention mechanism excels at modeling long-range global feature dependencies and achieved superior performance over CNN-based counterparts; critically, both studies relied exclusively on frontal facial images, neglecting diagnostically critical pathological features in lateral and oblique views. To overcome this single-view constraint, Qiang et al. [9] explored multi-view feature fusion, extracting frontal and lateral view features via pre-trained deep models, enhancing feature representation with a hybrid channel-spatial attention module, and optimizing cross-view feature complementarity and consistency via dedicated projection heads to further unlock the diagnostic potential of multi-view facial information.

Despite these advances, two fundamental technical bottlenecks remain unaddressed, which severely restrict the performance ceiling and real-world applicability and generalizability of facial image-based CS diagnostic models. First, the lack of bidirectional synergistic fusion between deep semantic features and clinically interpretable prior features. Existing methodologies are confined to two mutually isolated paradigms: studies relying solely on handcrafted clinical prior features (e.g., Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP) [5,6] capture fine-grained CS-specific pathological alterations with high clinical interpretability, but fail to model global facial pathological semantics; works leveraging exclusively deep learning-derived features [7–9] excel at global phenotypic pattern recognition, but lack principled integration of clinically meaningful prior knowledge. No existing framework has established bidirectional information interaction and collaborative enhancement between these two complementary feature modalities, leading to suboptimal diagnostic performance and limited clinical interpretability. Second, insufficient multi-view feature disentanglement with a critical absence of CS-specific pathophysiological prior constraints. CS-related facial pathological phenotypes exhibit strong view dependency, with distinct

views providing non-redundant diagnostic information [5,6,9]. However, existing multi-view studies fail to accurately disentangle view-invariant core disease features from view-specific pathological phenotypes. The resulting feature entanglement renders models vulnerable to view variations, impeding the retention of cross-view consistent disease signatures and full exploitation of view-specific diagnostic information. While state-of-the-art feature disentanglement techniques [10–12] offer a theoretical solution, their application to CS diagnosis requires integration of a core pathophysiological prior: the pathognomonic moon facies of CS exhibits strict bilateral symmetry, whereas asymmetric facial features are almost exclusively derived from diagnosis-irrelevant imaging or physiological interferences. None of the existing models have embedded this critical clinical prior into the disentanglement workflow, which prevents purification of CS-specific pathological representations, limits detection of subtle early-stage pathological changes, and ultimately makes existing methods unable to meet the high sensitivity and accuracy requirements for real-world early CS screening.

In this work, the target problem is formulated as facial-image-based AI-assisted diagnosis of CS rather than definitive diagnosis in the endocrinological sense. More specifically, the proposed model performs subject-level discrimination between individuals with CS and matched controls from five standard facial views, with the goal of providing non-invasive auxiliary diagnostic support for clinical decision-making and early referral. This positioning is clinically appropriate because CS presents characteristic facial phenotypes, whereas final confirmation still depends on specialist evaluation, hormonal testing, and established diagnostic criteria.

To address these two core limitations, we propose a novel Multi-View Facial Feature Disentanglement Network (MVFFD-Net) for high-precision automatic CS diagnosis, which takes facial images from five standard views (frontal, bilateral 45° oblique, and bilateral 90° lateral views) as input. The key contributions and benefits of this study can be summarized as follows. The network is specifically designed to directly tackle the aforementioned bottlenecks, with three key innovations: First, a bidirectional cross-attention fusion module is introduced to synergistically integrate deep semantic features and clinical prior features. Multi-scale deep features are extracted via the pre-trained DINOv2 vision foundation model, while clinically interpretable handcrafted features (HOG and LBP) are extracted in parallel. This module enables bidirectional information interaction between the two feature modalities, generating single-view fused features that combine global pathological semantics and fine-grained clinical interpretability, thereby improving both diagnostic discriminability. Second, a multi-path feature disentanglement architecture, coupled with a multi-objective optimization loss function, is designed to achieve accurate, lossless separation of view-invariant core disease features and view-specific pathological features. The architecture comprises a shared encoder (for view-invariant features), three view-type-specific encoders (for frontal, oblique, and lateral view-specific features), and a reconstruction encoder (for feature integrity constraints). The multi-objective loss function, centered on feature reconstruction loss and regularized by feature disentanglement constraints (shared feature similarity, feature specificity, and view-specific discriminability with intra-group symmetry consistency), ensures the independence and integrity of the disentangled features, thereby improving robustness to viewpoint variation and preserving clinically relevant disease signatures. Finally, a graph attention fusion framework, aligned with facial anatomical logic and clinical diagnostic rules, is employed to structurally fuse disentangled multi-view features, adaptively weighting the diagnostic contribution of each view while preserving local pathological details and ensuring global cross-view consistency of core disease features. This design enables more effective multi-view aggregation and supports high-sensitivity CS screening. Experimental results show that MVFFD-Net achieves an F1-score of 98.78% for CS diagnosis, significantly outperforming state-of-the-art methods. Collectively, these contributions provide a more accurate, robust, and clinically interpretable solution for non-invasive CS screening, while also offering a

transferable multi-view learning framework for intelligent diagnosis of other diseases with characteristic facial phenotypes.

2 Related Work

CS is a debilitating endocrine and metabolic disorder, whose pathognomonic facial phenotypic alterations serve as critical non-invasive visual biomarkers for AI-assisted diagnosis. In recent years, AI-enabled CS diagnosis has emerged as a rapidly evolving research focus, with existing studies falling into two primary paradigms: clinical data-driven predictive methods and facial image-based phenotypic analysis.

Early foundational work in AI-assisted CS diagnosis centered on structured electronic medical records (EMRs) data and multi-modal clinical datasets. For example, Zhang et al. [13] developed a prediction model integrating structured features and free-text information from EMRs, employing classical machine learning algorithms including Multilayer Perceptron (MLP) [14], Support Vector Machine (SVM) [15], Random Forest (RF) [16], and Logistic Regression (LR) [17]. Furthermore, Isci et al. [18] proposed a multi-modal classification framework, combining clinical features, laboratory indicators, and imaging data using diverse models such as SVM, K-Nearest Neighbors (KNN) [19], LR, Linear Discriminant Analysis (LDA) [20], Decision Trees (DT) based on the Classification and Regression Tree (CART) [21] algorithm, RF, AdaBoost, and Gradient Boosting (GB). However, a fundamental limitation undermines the clinical utility of these methods: their reliance on invasive laboratory tests and specialized medical imaging, which are not routinely available in primary care settings and thus cannot support large-scale, non-invasive early screening of CS.

Given the inherent advantages of non-invasiveness, easy accessibility, and compliance with routine clinical photography workflows, facial imaging has become the core research carrier for AI-enabled CS early screening. Existing facial image-based methods can be categorized into two technical generations: handcrafted feature-driven pipelines and deep learning-based automated feature extraction frameworks.

Pioneering work by Popp et al. [5] and Kosilek et al. [6] first validated the diagnostic value of facial phenotypic features for CS. They extracted handcrafted geometric and texture features from frontal and lateral facial photographs, and achieved disease classification using traditional machine learning (ML) classifiers. However, these methods are heavily dependent on manual annotation and professional image analysis software, which not only fail to fully capture the complex, heterogeneous pathological facial alterations caused by chronic hypercortisolism, but also introduce excessive labor costs, precluding end-to-end automated diagnosis and limiting real-world clinical scalability.

Subsequent work shifted to deep learning frameworks to enable fully automated, scalable diagnosis. Wei et al. [7] proposed the first end-to-end facial image-based CS diagnostic pipeline based on convolutional neural networks (CNNs): they performed face detection and alignment via dlib [22], extracted deep facial features from frontal photographs via transfer learning, and achieved differential diagnosis between CS and acromegaly. This landmark work laid the foundational framework for the clinical translation of facial image-based CS diagnosis. Nevertheless, CNNs have an inherent inductive bias toward local feature extraction, which limits their ability to capture global facial phenotypic features that are pathognomonic for CS. Furthermore, the models developed by Wei et al. [7] and Liu et al. [8] relied exclusively on frontal facial images, neglecting diagnostically critical phenotypic information contained in lateral facial views (e.g., maxillofacial morphological alterations secondary to hypercortisolism), which are routinely evaluated in standard clinical endocrine practice.

To address the inherent limitations of both single-view deep learning models and handcrafted feature-based multi-view pipelines, Qiang et al. [9] developed the first deep learning-integrated multi-view framework for CS diagnosis. The framework extracted deep phenotypic features from frontal and lateral facial images using state-of-the-art pre-trained backbones including ViT, Swin Transformer, ResNet and

DenseNet, implemented feature fusion and enhancement via a hybrid attention module, and introduced complementary and consistency projection heads to exploit cross-view phenotypic complementarity and ensure cross-view feature consistency. While this work first systematically investigated deep learning-enabled multi-view feature fusion for CS diagnosis, it did not resolve the intrinsic coupling of diagnostic and redundant features across different views, nor did it establish an effective fusion mechanism between deep visual features and evidence-based clinical diagnostic priors for CS.

To provide a clearer comparison of representative studies related to AI-based CS diagnosis, Table 1 summarizes their input modality, methodological characteristics, major advantages, and main limitations. As summarized in Table 1, representative studies have advanced AI-based CS diagnosis from both clinical-data-driven and facial-image-based perspectives. Nevertheless, several limitations remain common across existing methods, including reliance on invasive or less accessible data sources, insufficient exploitation of complementary information across facial views, limited integration of clinically meaningful priors, and lack of explicit disentanglement between view-invariant disease signatures and view-specific variations. These observations further motivate the development of a clinically informed multi-view framework for more accurate, robust, and non-invasive CS screening.

Table 1: Comparison of representative facial-image-based methods for Cushing's syndrome diagnosis.

Study	Input/Method	Advantages	Limitations
Kosilek et al. [6]	Frontal and lateral facial photographs; handcrafted features with traditional classifiers	Early proof of feasibility; clinically interpretable handcrafted descriptors	Limited representation capacity; relies on handcrafted features and shallow classifiers; lacks end-to-end automation
Wei et al. [7]	Frontal facial images; CNN-based end-to-end diagnosis	Automated facial analysis; improved diagnostic efficiency	Single-view only; limited cross-view modeling; no explicit clinical prior integration
Liu et al. [8]	Frontal facial images; DINOv2-based facial analysis	Strong global representation capability; better modeling of long-range dependencies	Still restricted to single-view input; no multimodal fusion; no explicit feature disentanglement
Qiang et al. [9]	Multi-view facial images; deep multi-view fusion with hybrid attention and projection heads	Closest prior CS-specific multi-view framework; exploits cross-view complementarity	No explicit disentanglement of common/private features; lacks clinically guided priors and structured graph-based aggregation

Despite these advances, two fundamental, clinically critical bottlenecks remain unresolved in existing facial image-based CS diagnostic models. First, inherent limitation of single-paradigm feature modeling and inefficient integration with clinical diagnostic priors. Existing facial image-based CS diagnostic methods are mostly confined to a single feature modeling paradigm: they either rely exclusively on handcrafted facial

phenotypic features, or only adopt end-to-end extracted deep visual features. Neither strategy effectively embeds well-validated clinical diagnostic priors for CS into the full workflow of feature extraction and fusion. This leads to a disconnection between model feature learning and clinical diagnostic logic, fails to focus the model on CS-specific pathological phenotypic alterations, and ultimately impairs the diagnostic precision and robustness of the model across heterogeneous clinical cohorts. Second, unsolved multi-view feature decoupling with neglected CS facial symmetry prior. Current multi-view CS diagnostic methods cannot disentangle view-specific diagnostic features from redundant ones, underutilizing cross-view complementarity. While general decoupling algorithms work in theory, their CS diagnostic application must incorporate the symmetry prior: CS's hallmark moon facies is inherently bilaterally symmetric, and all asymmetric features are diagnosis-irrelevant interferences. Omitting this prior weakens the model's interference resistance and feature purification, limiting its real-world early screening performance.

3 Methods

To address the critical limitations of inefficient multi-modal feature fusion and unresolved multi-view feature entanglement in existing CS diagnostic models, we propose a novel Multi-View Facial Feature Disentanglement Network (MVFFD-Net) for accurate and non-invasive CS diagnosis. As shown in Fig. 1, the proposed framework is implemented in three cascaded, clinically guided core modules: (1) a multi-modal feature extraction and cross-attention fusion module for single-view feature enhancement; (2) a dual-path multi-view feature decoupling module for separating view-invariant core disease features from view-specific pathological features; and (3) a clinical priority-guided graph attention fusion module for multi-view feature aggregation and final diagnostic classification.

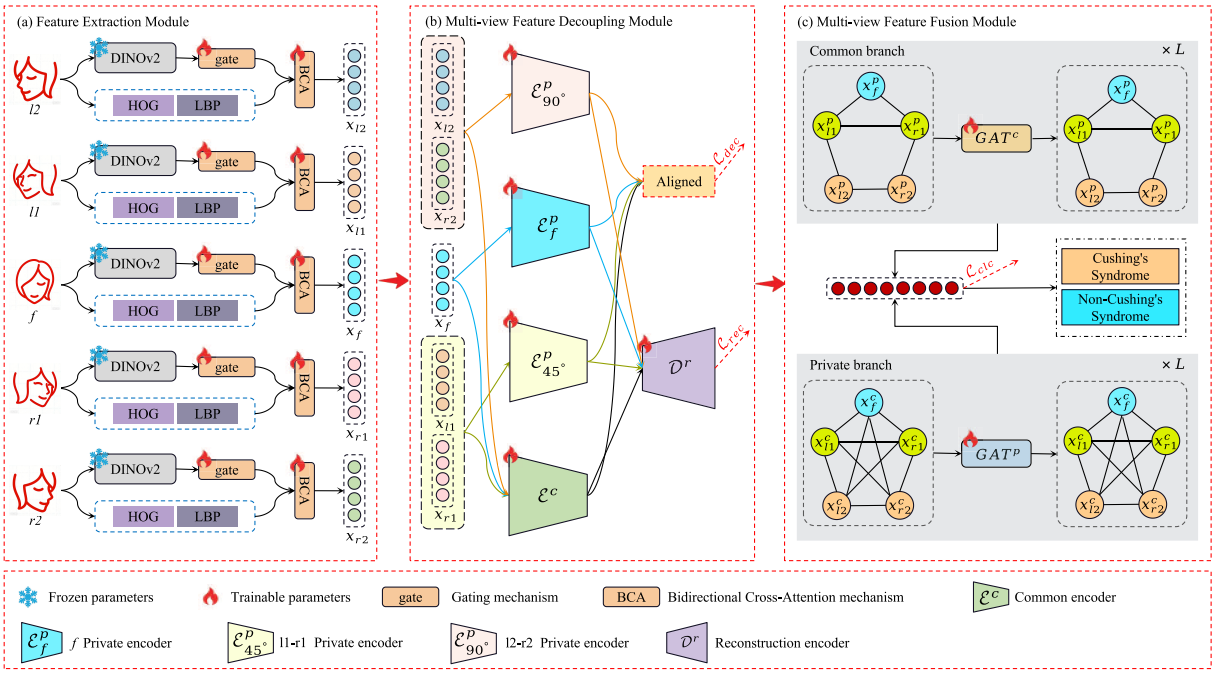

Figure 1: The overall framework of the proposed multi-view facial feature decoupling network (MVFFD-Net). (**a**) illustrates the feature extraction pipeline for five facial images from five standard viewpoints, i.e., frontal, left/right 45° oblique, and left/right 90° lateral views. (**b**) shows the feature decoupling process based on the features extracted from each individual facial viewpoint. (**c**) presents the feature fusion mechanism implemented via the Graph Attention Network.

For clarity, the main mathematical symbols used throughout the proposed framework are summarized in Table 2.

Table 2: Summary of the main mathematical symbols used in the proposed framework.

Symbol	Meaning
v	View index, where $v \in \{f, l_1, r_1, l_2, r_2\}$.
f, l_1, r_1, l_2, r_2	Frontal, left/right 45° oblique, and left/right 90° lateral views.
x_v	Input feature of view v.
x_v^c	View-invariant common feature of view v.
x_v^p	View-specific private feature of view v.
$\hat{x}_v$	Reconstructed feature of view v.
$\mathcal{L}_{rec}$	Reconstruction loss.
$\mathcal{L}_{dec}$	Overall decoupling constraint loss.
$\mathcal{L}_{sim}$	Shared feature similarity loss.
$\mathcal{L}_{spec}$	Feature specificity loss.
$\mathcal{L}_{dist}$	View-specific feature discriminability loss.
$\lambda_{sim}, \lambda_{spec}, \lambda_{dist}, \lambda_{inter}$	Balancing coefficients of the loss terms.
τ	Redundancy threshold in the specificity loss.
$\mathcal{V}$	Node set of the multi-view graph.
$\mathcal{N}(v_i)$	Neighborhood set of node v_i.
α_{ij}^k	Attention coefficient from node v_j to node v_i in attention head k.
h_i	Updated representation of node v_i.
H_{final}	Final fused global representation for classification.
$[\cdot;\cdot]$	Channel-wise concatenation.
$\|$	Concatenation operation.
$\odot$	Element-wise multiplication.
$\cos(\cdot,\cdot)$	Cosine similarity.

Unless otherwise specified, these notations are used consistently throughout the following mathematical formulations.

3.1 Feature Extraction Module

To extract more comprehensive, multi-modal and discriminative facial features for intelligent diagnosis, this study develops a feature extraction workflow implemented in three sequential steps. As shown in Fig. 2, facial images from five standard viewpoints (i.e., frontal, left/right 45° oblique, and left/right 90° lateral views) are taken as the input. For each single-view facial image, independent feature extraction and fusion are performed following the three steps below: First, deep features of each encoder layer are extracted via DINOv2, and handcrafted features including HOG and LBP are obtained synchronously, to fully capture both deep semantic information and underlying texture details of facial images. Second, a gating mechanism is adopted to adaptively fuse the deep features from different encoder layers, to retain validmulti-scale deep

representations. Third, a bidirectional cross-attention mechanism is applied to integrate the fused deep features with the handcrafted features, to achieve complementary fusion of multi-modal feature information.

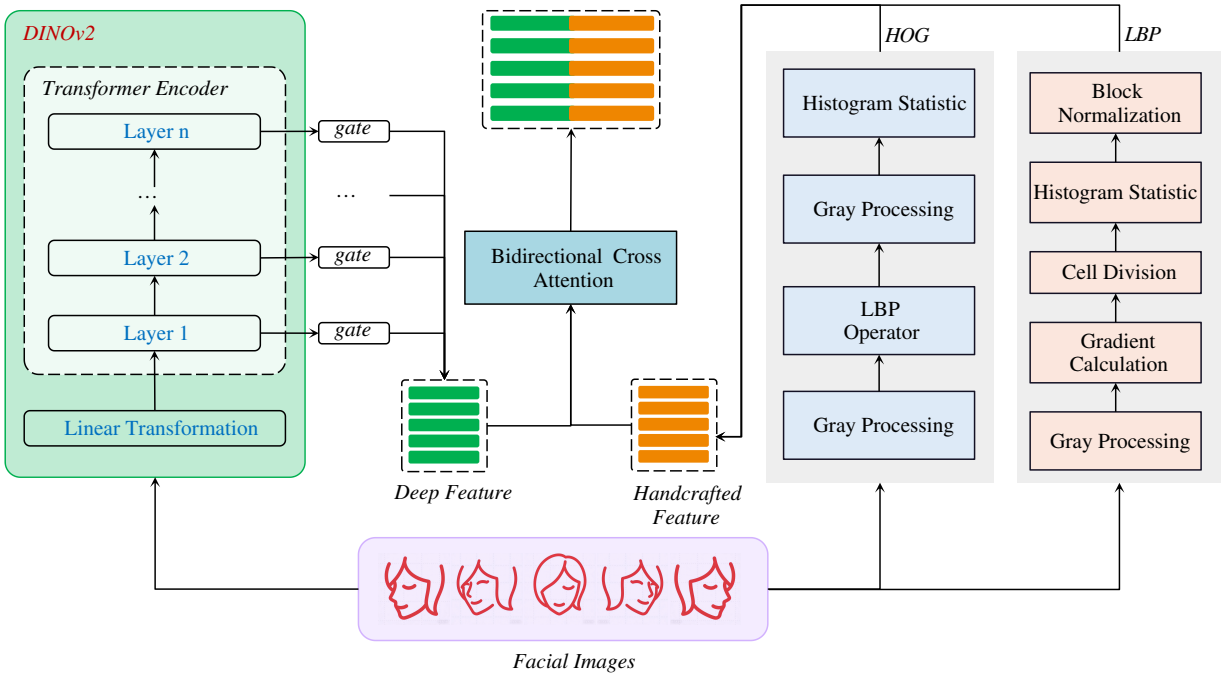

Figure 2: The feature extraction pipeline for five facial images from five standard viewpoints (i.e., frontal, left/right 45° oblique, and left/right 90° lateral views). Initially, for each facial image, deep features from each encoder layer are extracted using DINOv2, while handcrafted features (i.e., HOG and LBP) are simultaneously obtained. Subsequently, a gating mechanism is employed to fuse the deep features across different encoder layers. Finally, a bidirectional cross-attention mechanism is utilized to integrate the fused deep features with the handcrafted features.

3.1.1 Deep and Handcrafted Feature Extraction

In this study, the input facial images are acquired from five standard viewpoints, which are denoted as f (frontal view), l_1 (left 45° oblique view), r_1 (right 45° oblique view), l_2 (left 90° lateral view), and r_2 (right 90° lateral view), respectively. The large-scale self-supervised pre-trained vision Transformer model DINOv2 [23] is adopted as the backbone for deep feature extraction. All trainable parameters of the backbone are fully frozen throughout the entire workflow without any fine-tuning operations. We only leverage the robust general representation capability learned by the model from massive general visual data to extract multi-scale features from facial images. This design effectively mitigates the risk of model overfitting and the degradation of pre-trained generalization ability caused by fine-tuning in medical small-sample scenarios. Meanwhile, its built-in self-attention mechanism can effectively model the characteristic overall facial contour changes of CS (e.g., moon facies, centripetal fat distribution), which is more suitable for the disease phenotype capture requirements of this study than traditional convolutional neural networks (e.g., ResNet, DenseNet).

Features output by encoder blocks at all depths of the Vision Transformer cover a full spectrum of semantic hierarchies, which are highly matched with the multi-scale characteristics of CS facial pathological phenotypes:

1. Shallow encoders focus on low-level visual features such as edges and textures of the image, which can accurately capture microstructural skin changes related to the disease, including acne, hirsutism, and facial plethora;
2. Middle encoders can model the spatial structure and contour features of facial organs, adapting to local pathological changes in fat distribution in areas such as the buccal fat pad and mandibular margin;
3. Deep encoders focus on global semantic representations, which can effectively characterize the overall facial morphological changes such as moon facies, the hallmark phenotype of CS.

Unless otherwise specified, $v \in \{f, l_1, r_1, l_2, r_2\}$ denotes the view index; superscripts c and p denote view-invariant common and view-specific private features, respectively; $[\cdot;\cdot]$ denotes channel-wise concatenation; $\odot$ denotes element-wise multiplication; $\|$ denotes feature concatenation; and $\cos(\cdot,\cdot)$ denotes cosine similarity.

Based on this, we extract the output features from all *Transformer encoder blocks* of the DINOv2 network, to fully retain the pathological representations across low-level, middle-level, and high-level semantic hierarchies. Specifically, for an input facial image, DINOv2 first splits it into a sequence of fixed-size patches. After linear projection and positional encoding, the patch embeddings, along with a special class token, are fed into the Transformer encoder stacks. For each input single-view facial image (i.e., f, l_1, r_1, l_2, r_2), we discard the class token and only retain the patch embeddings output by the l-th encoder layer, resulting in a set of 2D feature tensors F_l ($l \in \{1, 2, \ldots, L\}$, where L is the total number of encoder layers in the adopted DINOv2 model) with dimension $N \times C$. Here, $N = H_p \times W_p$ denotes the total number of patches, where H_p and W_p are the height and width of the patch grid, respectively, and the channel dimension (embedding dimension) $C = 768$. To obtain a fixed-dimensional feature vector, we first reshape the patch embeddings back to a 2D spatial structure of size $H_p \times W_p \times C$, and then perform a Global Average Pooling (GAP) operation on it, as shown in Eq. (1):

$$f_l = \frac{1}{H_p \times W_p} \sum_{i=1}^{H_p} \sum_{j=1}^{W_p} F_l^{reshape}(i, j) \tag{1}$$

where f_l is the one-dimensional feature vector corresponding to the l-th encoder layer, with a fixed dimension of 768. Here, $F_l^{reshape}(i, j) \in \mathbb{R}^C$ denotes the patch embedding at spatial location (i, j) in the reshaped feature map of the l-th encoder layer. This operation averages the embeddings of all spatial patches to obtain a fixed-length global descriptor for the l-th hierarchy, which facilitates subsequent cross-layer fusion.

To compensate for the insufficient explicit representation of clinically key phenotypes by purely data-driven deep features, this study introduces two clinically priority-guided handcrafted features, namely HOG and LBP, to explicitly characterize the fine-grained pathological phenotypes of each single-view facial image. Among them, HOG features can effectively represent the gradient changes of facial contour and fat distribution, adapting to the characteristic facial rounding changes of CS; LBP features can accurately capture the micro-changes of facial skin texture, supplementing the fine-grained pathological information that is easily overlooked by deep features. The extracted HOG and LBP features are first normalized by Z-Score standardization, and then mapped to 768 dimensions through a linear projection layer, finally obtaining a one-dimensional handcrafted feature vector $\boldsymbol{f}_{hand}$ with the same dimension as the hierarchical deep features. This unified dimension design lays a solid foundation for the subsequent hierarchical deep feature fusion via gating mechanism, as well as the cross-modal integration of deep and handcrafted features via bidirectional cross-attention mechanism.

3.1.2 Full-Hierarchy Deep Feature Fusion

Aiming at the limitation that partial hierarchical feature selection and single-depth feature output are difficult to cover the full spectrum of pathological semantic information, and easily lead to the loss of key diagnostic features from shallow to deep encoder layers, this study constructs an adaptive gating fusion mechanism to fuse the full-hierarchy deep features extracted from all encoder blocks of DINOv2. Different from traditional weighted fusion methods that force fixed weight distribution or normalization, the proposed gating mechanism adaptively adjusts the information flux of features from each encoder layer through learnable gating units. It can flexibly retain complementary pathological information across the full semantic hierarchy, while taking the deep-layer global semantic features as the core anchor to ensure the disease discriminability of the fused features. This fusion process is performed independently for the full-hierarchy features of each single-view facial image (i.e., f, l_1, r_1, l_2, r_2).

The core of the gating fusion module is to generate independent adaptive gating weights for features from each encoder layer, to achieve fine-grained control of the hierarchical feature information flow. First, the one-dimensional feature vectors of all L encoder layers are concatenated along the channel dimension and fed into the gating weight learning network. The gating weights $[\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_L]$ corresponding to the full-hierarchy features are generated through two fully connected layers with a Gaussian Error Linear Unit (GELU) nonlinear activation, followed by a Sigmoid activation function, as shown in Eq. (2):

$$[\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_L] = \text{Sigmoid}\left(\text{FC}_2\left(\text{GELU}\left(\text{FC}_1\left([\boldsymbol{f}_1; \boldsymbol{f}_2; ...; \boldsymbol{f}_L]\right)\right)\right)\right) \tag{2}$$

where $[\cdot;\cdot]$ denotes the concatenation operation of feature vectors along the channel dimension; FC_1 is a fully connected layer with an input dimension of $L \cdot C$ and an output dimension of 256, FC_2 is a fully connected layer with an input dimension of 256 and an output dimension of $L \cdot C$, where L is the total number of encoder layers in the adopted DINOv2 model, and the feature channel dimension $C = 768$; the Sigmoid activation function maps each gating weight to the interval $[0, 1]$, realizing element-wise soft gating adjustment of features from each hierarchical encoder layer. Each $\boldsymbol{g}_l \in \mathbb{R}^C$ is an element-wise gate vector assigned to the l-th hierarchical feature, and its entries indicate how much information should be retained from the corresponding feature channels. Therefore, Eq. (2) learns the relative importance of different semantic hierarchies in a data-adaptive manner.

Subsequently, features from each encoder layer are weighted and modulated by their corresponding gating weights. Meanwhile, a residual connection is introduced with the output of the deepest encoder layer (global semantic anchor) as the residual branch, to avoid the loss of core disease-related global semantic information. The final globally fused deep feature f_{deep} after gating fusion is obtained, as shown in Eq. (3):

$$f_{deep} = f_L + \sum_{l=1}^{L} \boldsymbol{g}_l \odot f_l \tag{3}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication), and $\boldsymbol{f}_L$ is the output feature of the deepest encoder layer of DINOv2. This gating fusion strategy can adaptively learn the optimal fusion weights of features at each hierarchy according to their contribution to the CS diagnostic task. It fully exploits the diagnostic value of fine-grained texture features from shallow layers and local structural features from middle layers, while retaining the core global pathological phenotypes encoded in the deepest layer, achieving efficient and complementary fusion of full-spectrum pathological information from all encoder layers. The term $\boldsymbol{g}_l \odot f_l$ reweights the l-th feature vector channel by channel, while the residual term $\boldsymbol{f}_L$ preserves the deepest global disease semantics as a stable anchor. As a result, Eq. (3) integrates shallow texture cues, middle-level structural cues, and deep global cues into a unified diagnostic representation.

3.1.3 Cross-Modal Fusion of Deep and Handcrafted Features

To achieve bidirectional deep interaction and synergistic fusion between data-driven deep features and clinically priority-guided handcrafted features, this module adopts a parallel bidirectional cross-attention mechanism to realize mutual guidance and information augmentation between the two heterogeneous feature modalities. This process is performed independently for each single-view facial image (i.e., f, l_1, r_1, l_2, r_2), and finally outputs a view-specific fused feature that simultaneously inherits the powerful global representation capability of data-driven deep learning and the clinical interpretability of priority-guided handcrafted features. The inputs of this module are two sets of features with a unified dimension of 768 for each single view: the global fused deep feature f_{deep} encoding full-spectrum CS pathological semantics, obtained via the full-hierarchy gating fusion described in Section 3.1.2, and the clinically priority-guided handcrafted feature f_{hand} extracted in Section 3.1.1.

The core of the bidirectional cross-attention mechanism is to construct two symmetric cross-attention mapping branches, and synchronously complete bidirectional information interaction between the two feature modalities in a unified attention embedding space. Specifically, two parallel mapping branches are designed as follows: One branch takes the deep feature f_{deep} with global disease semantics as the Query ($\boldsymbol{Q}$), and the handcrafted feature f_{hand} as both the Key ($\boldsymbol{K}$) and Value ($\boldsymbol{V}$). This branch mines the fine-grained diagnostic-relevant pathological information in handcrafted features that is complementary to deep features, to achieve disease-related information enhancement of clinical prior features. The other symmetric branch takes the handcrafted feature f_{hand} as the Query $\boldsymbol{Q}$, and the deep feature f_{deep} as both the Key $\boldsymbol{K}$ and Value $\boldsymbol{V}$. This branch uses clinical prior knowledge as the anchor to guide deep features to focus on clinically interpretable key pathological phenotypes characterized by handcrafted features (e.g., gradient changes of facial contour, abnormal skin texture), to realize clinical prior constraint enhancement of data-driven deep features. All input features are mapped into the unified attention embedding space via trainable linear projection matrices, and the output of the two bidirectional attention branches is calculated as shown in Eq. (4). The first branch transfers global semantic guidance from deep features to handcrafted priors, whereas the second branch injects clinically interpretable prior guidance into deep features.

$$\begin{cases} \mathrm{Attn}_{deep2hand} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_1\boldsymbol{K}_1^{\top}}{\sqrt{d}}\right)\boldsymbol{V}_1, \quad \boldsymbol{Q}_1 = f_{deep}\boldsymbol{W}_Q^1, \boldsymbol{K}_1 = \boldsymbol{V}_1 = f_{hand}\boldsymbol{W}_{KV}^1 \\ \mathrm{Attn}_{hand2deep} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_2\boldsymbol{K}_2^{\top}}{\sqrt{d}}\right)\boldsymbol{V}_2, \quad \boldsymbol{Q}_2 = f_{hand}\boldsymbol{W}_Q^2, \boldsymbol{K}_2 = \boldsymbol{V}_2 = f_{deep}\boldsymbol{W}_{KV}^2 \end{cases} \tag{4}$$

where $\boldsymbol{W}_Q^1, \boldsymbol{W}_{KV}^1, \boldsymbol{W}_Q^2, \boldsymbol{W}_{KV}^2 \in \mathbb{R}^{C\times C}$ are trainable linear projection matrices for the corresponding mapping branches, with the unified feature dimension $C = 768$; $d = C$ is the dimension of the attention embedding space, and $\sqrt{d}$ is the scaling factor to avoid gradient saturation of the Softmax function caused by excessively large dot-product results; the Softmax function generates normalized attention weights to realize adaptive weighting of CS diagnosis-related feature information in both branches. Accordingly, $\mathrm{Attn}_{deep2hand}$ enhances fine-grained pathological evidence complementary to deep features, whereas $\mathrm{Attn}_{hand2deep}$ steers deep features toward clinically meaningful phenotypes.

Subsequently, the enhanced features output by the two bidirectional mapping branches are concatenated along the channel dimension. The dimension is then aligned to the unified feature dimension $C = 768$ via a fully connected (FC) layer with Gaussian Error Linear Unit (GELU) nonlinear activation, to obtain the fused feature $f_{bi-attn}$ after bidirectional interaction, as shown in Eq. (5):

$$f_{bi-attn} = \mathrm{GELU}\left(\mathrm{FC}\left(\left[\mathrm{Attn}_{deep2hand}; \mathrm{Attn}_{hand2deep}\right]\right)\right) \tag{5}$$

where $[\cdot;\cdot]$ denotes the concatenation operation of feature vectors along the channel dimension, and the FC layer has an input dimension of $2C = 1536$ and an output dimension of $C = 768$. Concatenation preserves the complementary outputs of the two directional branches, and the FC layer followed by GELU projects them back into a unified 768-dimensional representation space for subsequent fusion.

Finally, to alleviate the gradient vanishing problem caused by network depth increase and preserve the core semantic integrity of the original input features, a residual connection and Layer Normalization (LN) operation are introduced to optimize the bidirectional attention fused features, yielding the final view-specific fused feature f_{final}, as shown in Eq. (6):

$$f_{final} = \text{LayerNorm}\left(f_{deep} + f_{hand} + f_{bi-attn}\right) \tag{6}$$

Here, f_{deep} is the hierarchically fused deep feature, f_{hand} is the clinically guided handcrafted feature, and $f_{bi-attn}$ is the bidirectional interaction feature. The residual summation preserves the original modality-specific evidence, whereas LayerNorm stabilizes the feature distribution before the output is fed into the subsequent decoupling module.

This bidirectional cross-attention fusion strategy breaks through the information interaction bottleneck of traditional unidirectional feature fusion methods. Through the symmetric parallel mapping branches, it simultaneously realizes the constraint guidance of clinical prior knowledge on data-driven deep features and the disease information enhancement of handcrafted features by deep semantic representations. It fully injects clinically interpretable prior knowledge while retaining the full-spectrum global pathological semantics of CS, and effectively improves the disease discriminability and clinical interpretability of the final view-specific fused features.

3.2 Multi-View Feature Decoupling Module

Existing multi-view facial diagnosis methods still face two core limitations that restrict their clinical practicality:

1. Most methods directly aggregate multi-view facial features without explicit semantic decoupling, failing to accurately separate the view-invariant core pathological features of CS from view-specific nuisance features related to pose variation and geometric deformation. This defect makes disease-discriminative features vulnerable to viewpoint interference, leading to insufficient generalization performance in real clinical scenarios with non-standardized shooting conditions.
2. Existing studies generally neglect the inherent left-right symmetry of human facial physiological structure, and fail to impose consistency constraints on the decoupled features from paired symmetric viewpoints (i.e., left/right 45° oblique views l_1/r_1 and left/right 90° lateral views l_2/r_2 defined in this study). This makes it difficult to suppress feature noise caused by viewpoint differences, and further impairs the robustness and disease discriminability of the extracted features.

To address the above limitations, this study constructs a dual-path feature decoupling architecture based on a shared feature encoder and a view-specific feature encoder, which achieves accurate separation of view-invariant core disease features and view-specific features from the final fused feature f_{final} of each single viewpoint. Meanwhile, a multi-objective model optimization function is established, consisting of a reconstruction loss, a feature decoupling constraint loss, and a newly introduced facial symmetry consistency loss. This multi-objective optimization strategy jointly guarantees the information integrity and semantic independence of the decoupled features, as well as the cross-view consistency of core disease features under symmetric viewpoints. It thus provides standardized, robust, and highly discriminative feature inputs for the subsequent multi-view feature fusion module.

3.2.1 Feature Decoupling Encoder and Reconstruction Encoder

The overall module is built on a dual-branch decoupling architecture consisting of a shared feature encoder $\mathcal{E}^c$, three view-specific feature encoders ($\mathcal{E}_f^p$, $\mathcal{E}_{45°}^p$, $\mathcal{E}_{90°}^p$), and a reconstruction encoder $\mathcal{D}^r$. Parameters between the shared branch and view-specific branches are completely independent to avoid mutual interference and semantic aliasing in the feature encoding space.

The core optimization objective of the shared feature encoder $\mathcal{E}^c$ is to extract view-invariant core pathological features that are consistent across all viewpoints, which capture the universal CS disease phenotypes observable in all views. This encoder takes the normalized features of all 5 views as input, with weights fully shared across all viewpoints to ensure that cross-view features are mapped into a unified semantic space. The main body of $\mathcal{E}^c$ is composed of 3 cascaded residual fully connected blocks. Each residual block adopts a standard structure of layer normalization (LN), dimensionality mapping fully connected (FC) layer, Gaussian Error Linear Unit (GELU) activation function, Dropout layer, inverse dimensionality mapping FC layer, and residual shortcut connection. The input and output dimensions of a single residual block are both 768, the intermediate hidden layer dimension is set to 512, and the Dropout rate is set to 0.3. For the input feature of each viewpoint, the shared encoder outputs the corresponding view-invariant common feature:

$$x_v^c = \mathcal{E}^c(x_v), \quad v \in \{f, l_1, r_1, l_2, r_2\} \tag{7}$$

where $\mathcal{E}^c(\cdot)$ denotes the nonlinear mapping function of the shared feature encoder, and the superscript c represents the common view-invariant semantic attribute of the features. Eq. (7) maps the input feature of each viewpoint into a shared semantic space so that viewpoint-invariant disease cues can be compared and aggregated across all views.

The core optimization objective of the view-specific feature encoders is to extract viewpoint-unique pathological features that cannot be fully captured by other perspectives, while complying with the physiological prior of human facial left-right symmetry. Specifically: (1) The frontal view-specific encoder $\mathcal{E}_f^p$ exclusively processes the frontal view feature $\boldsymbol{x}_f$, focusing on fine-grained pathological changes such as facial skin texture abnormalities and facial organ spacing changes under frontal observation; (2) The 45° oblique view-specific encoder $\mathcal{E}_{45°}^p$ is shared by the paired symmetric left/right 45° oblique views ($\boldsymbol{x}_{l_1}$, $\boldsymbol{x}_{r_1}$), focusing on buccal fat pad accumulation and zygomatic contour pathological changes; (3) The 90° lateral view-specific encoder $\mathcal{E}_{90°}^p$ is shared by the paired symmetric left/right 90° lateral views ($\boldsymbol{x}_{l_2}$, $\boldsymbol{x}_{r_2}$), focusing on key supplementary diagnostic phenotypes such as mandibular retrusion, temporal lipoatrophy, and mandibular margin fat accumulation.

The weight-sharing strategy for symmetric viewpoints aligns with the inherent symmetry of human facial physiology, which lays the foundation for the subsequent facial symmetry consistency constraint, and helps suppress feature noise caused by viewpoint differences. All view-specific encoders are completely independent of the shared encoder $\mathcal{E}^c$ in terms of parameters, to avoid information confusion between the two semantic spaces. The main body of each view-specific encoder is composed of 2 residual fully connected blocks, with the single-block structure consistent with that of the shared encoder, only the Dropout rate is adjusted to 0.2 to further reduce overfitting risk while ensuring sufficient feature encoding capability. For the input feature of each viewpoint, the corresponding view-specific encoder outputs the view-unique

pathological feature:

$$x_v^p = \begin{cases} \mathcal{E}_f^p(x_v), & v = f \\ \mathcal{E}_{45^\circ}^p(x_v), & v \in \{l_1, r_1\} \\ \mathcal{E}_{90^\circ}^p(x_v), & v \in \{l_2, r_2\} \end{cases} \tag{8}$$

where $\mathcal{E}_\cdot^p(\cdot)$ denotes the nonlinear mapping function of the corresponding view-specific encoder, and the superscript p represents the view-specific semantic attribute of the features. Eq. (8) selects the corresponding view-specific encoder according to the viewpoint group and extracts pathological information that is unique to frontal, oblique, or lateral observations.

To verify the information integrity of the decoupled features and construct the reconstruction optimization objective, we design a reconstruction encoder $\mathcal{D}^r$. Its core function is to reconstruct the original input feature x_v from the concatenated decoupled common feature x_v^c and view-specific feature x_v^p, so as to ensure that no key diagnostic information is lost during the decoupling process. The main body of $\mathcal{D}^r$ consists of 2 cascaded residual fully connected blocks, with input and output dimensions both 768, consistent with the feature dimension throughout the framework. The reconstruction process is formulated as:

$$\hat{x}_v = \mathcal{D}^r\left(\left[x_v^c; x_v^p\right]\right), \quad v \in \{f, l_1, r_1, l_2, r_2\} \tag{9}$$

where $\hat{x}_v$ is the reconstructed feature of the v-th viewpoint, and $[\cdot;\cdot]$ denotes the channel-wise concatenation operation. This reconstruction step explicitly checks whether the combination of common and private features still contains sufficient information to recover the original input representation, thereby preventing diagnostically useful information from being discarded during decoupling.

Finally, the module outputs the view-invariant common disease features and view-specific pathological features of all 5 viewpoints. All output features are processed by independent layer normalization to ensure the stability of feature distribution, which can be seamlessly connected with the subsequent multi-view feature fusion module.

3.2.2 Multi-Objective Loss Function System

This study constructs a multi-objective loss function system, which takes the feature reconstruction loss as the core guarantee for information integrity, and the feature decoupling constraint loss as the core specification for semantic independence. This system achieves accurate decoupling of view-invariant common features and view-specific pathological features while preserving complete diagnostic information, and further embeds the physiological prior of human facial left-right symmetry to enhance the robustness and clinical consistency of the decoupled features.

Feature Reconstruction Loss The feature reconstruction loss is designed to ensure that the decoupled common features x_v^c and view-specific features x_v^p fully retain the effective diagnostic information of the original input features, and avoid the loss of key CS pathological phenotypes during the decoupling process. Consistent with the reconstruction encoder $\mathcal{D}^r$ defined in Section 3.2.1, we reconstruct the original input feature $\boldsymbol{x}_v$ from the channel-wise concatenation of the corresponding decoupled common feature and view-specific feature, and construct the reconstruction constraint with mean square error (MSE). The loss function is defined as:

$$\mathcal{L}_{rec} = \frac{1}{5} \sum_{v \in \{f, l_1, r_1, l_2, r_2\}} \left\| x_v - \hat{x}_v \right\|_2^2 \tag{10}$$

where $\hat{x}_v = \mathcal{D}^r\left(\left[x_v^c; x_v^p\right]\right)$ is the reconstructed feature of the v-th viewpoint output by the reconstruction encoder, and $[\cdot;\cdot]$ denotes the channel-wise concatenation operation consistent with the previous definition. Notably, the parameters of the reconstruction encoder are shared between left-right symmetric viewpoints (i.e., l_1/r_1 and l_2/r_2), which conforms to the physiological prior of human facial symmetry. This loss term constrains the information integrity of the decoupled features, ensuring that no effective pathological information is lost during decoupling, and providing complete feature input for subsequent diagnostic tasks. The coefficient 1/5 averages the reconstruction error over the five viewpoints, ensuring that all views contribute equally to the information-preservation objective.

Feature Decoupling Constraint Loss The feature decoupling constraint loss is designed to achieve accurate semantic decoupling between view-invariant common features and view-specific features, ensure the semantic independence of the two feature spaces, and eliminate information redundancy and aliasing. As shown in Fig. 3, it consists of three complementary components: shared feature similarity loss, feature specificity loss, and view-specific feature discriminability loss, which jointly standardize the decoupling process from three dimensions: cross-view consistency, intra-view independence, and inter-view discriminability. The overall decoupling constraint loss is defined as:

$$\mathcal{L}_{dec} = \lambda_{sim} \cdot \mathcal{L}_{sim} + \lambda_{spec} \cdot \mathcal{L}_{spec} + \lambda_{dist} \cdot \mathcal{L}_{dist} \tag{11}$$

where λ_{sim}, λ_{spec} and λ_{dist} are the balance weight coefficients of each component, set to 0.4, 0.3 and 0.3 respectively, to coordinate the multi-dimensional optimization objectives of decoupling. The three terms respectively enforce cross-view consistency of common features, semantic independence between common and private features, and discriminability of private features across viewpoint groups.

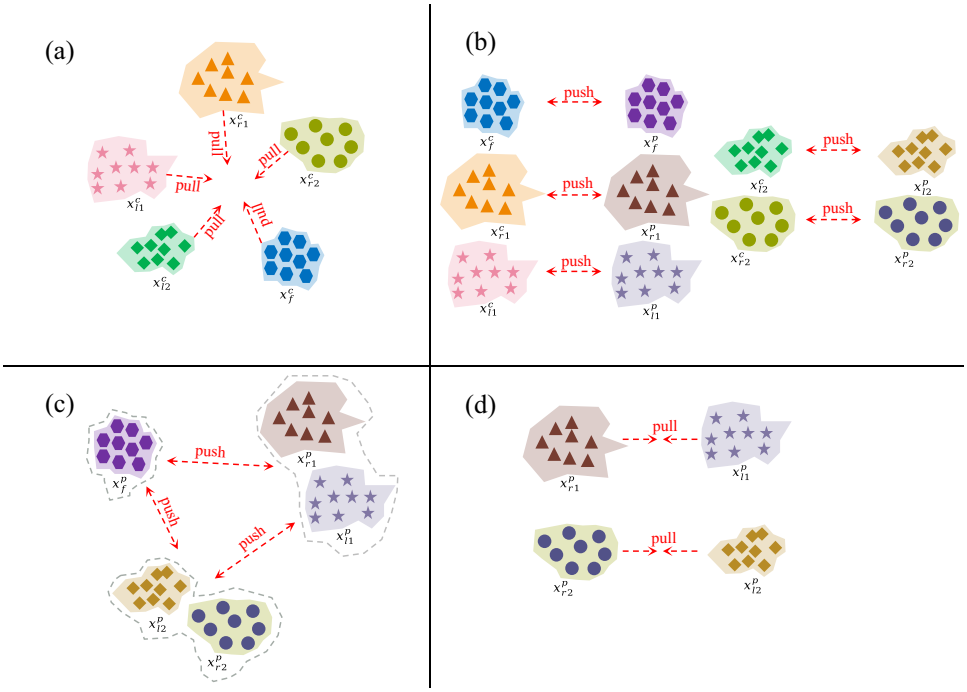

Figure 3: Illustration of the feature decoupling constraint loss, which consists of three complementary components: (**a**) shared feature similarity loss, (**b**) feature specificity loss, and (**c**,**d**) view-specific feature discriminability loss (including (**c**) inter-group feature discriminability constraint and (**d**) intra-group symmetry consistency constraint).

Shared Feature Similarity Loss This loss term constrains the view-invariant common features learned by the shared encoder $\mathcal{E}^c$ to maintain high semantic consistency across different viewpoints of the same subject, so as to eliminate the interference of pose fluctuation and viewpoint change on the core disease features. We take the common feature of the frontal view (the most standardized clinical view) as the semantic anchor, and constrain the cosine similarity between the frontal common feature and the common features of other viewpoints of the same sample. The loss function is defined as:

$$\mathcal{L}_{sim} = 5 - \sum_{v \in \{f, l_1, r_1, l_2, r_2\}} \cos\left(x_v^c, x_f^c\right) \tag{12}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity calculation function, whose theoretical value ranges from 0 to 1. The ideal state is that the common features of all 5 viewpoints are completely consistent, with the sum of cosine similarity reaching the maximum value of 5, and the loss value approaching 0. Minimizing this loss term can effectively enhance the view invariance of the shared common features, ensuring that the core CS pathological phenotypes are stably encoded regardless of viewpoint changes. Because the summation includes the frontal feature itself, $\cos(x_f^c, x_f^c) = 1$, and the theoretical maximum of the similarity sum is 5. Therefore, the loss is formulated as $5 - \cdots$ so that the optimal value approaches zero when all common features are maximally aligned.

Feature Specificity Loss This loss term constrains the semantic independence between the common feature and view-specific feature of the same viewpoint, to eliminate redundant information between the two feature spaces. Different from traditional methods that force complete orthogonality between features, we adopt a hinge loss strategy with a redundancy threshold, which only penalizes excessive information redundancy between the two types of features, while avoiding the loss of key overlapping pathological information caused by forced orthogonality. The loss function is defined as:

$$\mathcal{L}_{spec} = \sum_{v \in \{f, l_1, r_1, l_2, r_2\}} \max\left(0, \cos\left(x_v^c, x_v^p\right) - \tau\right) \tag{13}$$

where $\tau = 0.1$ is the preset redundancy threshold. Only when the cosine similarity between the common feature and the corresponding view-specific feature exceeds the threshold, the penalty term is activated to suppress information redundancy. This design balances the decoupling accuracy and information integrity, ensuring the semantic independence of the two feature spaces without sacrificing effective diagnostic information. This hinge-style formulation penalizes only excessive redundancy beyond the threshold τ, thereby avoiding overly strict orthogonality that could remove diagnostically useful shared information.

View-Specific Feature Discriminability Loss This loss term is designed to constrain the view-specific features to capture the unique pathological information of the corresponding viewpoint, suppress the homogenization of specific features from different perspective groups, and simultaneously embed the physiological prior of human facial left-right symmetry. It consists of two complementary constraints: intra-group symmetry consistency constraint and inter-group feature discriminability constraint, defined as:

$$\mathcal{L}_{dist} = \mathcal{L}_{dist\_intra} + \lambda_{inter} \cdot \mathcal{L}_{dist\_inter} \tag{14}$$

where $\lambda_{inter} = 0.3$ is the balance weight of the inter-group constraint. Eq. (14) combines symmetry preservation within paired views and feature separation across different viewpoint groups, so that the private branch remains both physiologically consistent and semantically discriminative.

The intra-group symmetry consistency constraint $\mathcal{L}_{dist\_intra}$ is designed to enforce the semantic consistency of view-specific features from paired left-right symmetric viewpoints, which strictly conforms to the inherent left-right symmetry of human facial physiological structure, and is consistent with the

weight-sharing strategy of the view-specific encoders for symmetric viewpoints. The loss function is defined as:

$$\mathcal{L}_{dist\_intra} = 2 - \cos\left(x_{l_1}^p, x_{r_1}^p\right) - \cos\left(x_{l_2}^p, x_{r_2}^p\right) \tag{15}$$

Minimizing this loss term can effectively suppress feature noise caused by asymmetric shooting angles, and enhance the robustness of view-specific features. The constant 2 corresponds to the maximum possible cosine-similarity sum of the two symmetric pairs (l_1, r_1) and (l_2, r_2). Thus, the loss approaches zero when both pairs are maximally consistent.

The inter-group feature discriminability constraint $\mathcal{L}_{dist\_inter}$ is used to maximize the feature space distance between specific features from different viewpoint groups (consistent with the grouping of the three view-specific encoders: frontal view, 45° oblique view group, 90° lateral view group), ensuring that the view-specific features of each group can capture differentiated and unique pathological information. The loss function is defined as:

$$\mathcal{L}_{dist\_inter} = -\left(\left\|\bar{x}_f^p - \bar{x}_{45}^p\right\|_2^2 + \left\|\bar{x}_f^p - \bar{x}_{90}^p\right\|_2^2 + \left\|\bar{x}_{45}^p - \bar{x}_{90}^p\right\|_2^2\right) \tag{16}$$

where $\bar{x}_f^p = x_f^p$ is the view-specific feature of the frontal view, $\bar{x}_{45}^p = \frac{1}{2}\left(x_{l_1}^p + x_{r_1}^p\right)$ is the mean feature of the 45° oblique view group, and $\bar{x}_{90}^p = \frac{1}{2}\left(x_{l_2}^p + x_{r_2}^p\right)$ is the mean feature of the 90° lateral view group. By minimizing this loss term, the specific features of different viewpoint groups are forced to be separated in the feature space, which guarantees the differentiated expression ability of view-specific features from each perspective, and avoids feature homogenization. The negative sign is introduced because the overall training objective is minimized; therefore, minimizing Eq. (16) is equivalent to maximizing the squared distances between different viewpoint groups.

3.3 Multi-View Feature Fusion Module

To address the core limitations of existing multi-view fusion methods, this study constructs a clinical priority-guided dual-path Graph Attention Network (GAT) fusion framework based on the view-invariant common disease features and view-specific pathological features obtained from the multi-view feature decoupling module. This framework adopts GAT as the sole core operator for feature aggregation throughout the entire fusion pipeline, and designs two parallel fusion branches specifically tailored to the distinct optimization objectives of the two types of decoupled features. Through graph topological designs complying with human facial anatomical logic and clinical diagnostic rules for CS, combined with an adaptive attention weighting mechanism, the framework achieves efficient and complementary fusion of multi-view pathological features, and finally outputs a robust global fused representation for CS classification and diagnosis.

3.3.1 Clinical Priority-Guided Multi-View Graph Topology Definition

The core of graph neural network-based feature fusion lies in rational graph topology design. Unreasonable graph connections will introduce redundant noise and invalid information interaction, thus destroying the semantic consistency of features. For the two types of features with heterogeneous semantic properties obtained via the decoupling module (i.e., view-invariant common disease features and view-specific private pathological features), this study designs two sets of differentiated graph topological structures corresponding to the dual-path GAT fusion framework, as shown in Fig. 4, namely the fully connected graph for common features and the facial anatomical structure-based graph for private features. All designs strictly

follow the clinical observation rules of CS facial phenotypes and the anatomical symmetry prior of the human face, providing a clear and clinically interpretable structural foundation for subsequent GAT fusion, and fundamentally solving the problems of vague graph structure definition and disconnection from clinical priors in existing methods.

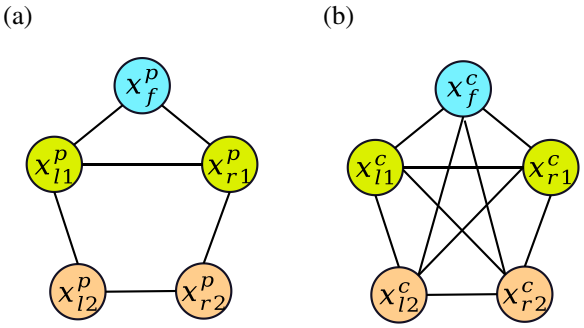

Figure 4: Illustration of the two sets of differentiated graph topological structures for the dual-path GAT fusion framework: (**a**) view-invariant common feature graph topology, (**b**) view-specific private feature graph topology.

The node set $\mathcal{V} = \{v_f, v_{l_1}, v_{r_1}, v_{l_2}, v_{r_2}\}$ is shared by the two sets of graph topologies, which consists of five nodes corresponding to the five standard viewpoints: frontal, left 45° oblique, right 45° oblique, left 90° lateral, and right 90° lateral views, respectively. The initial input features of each node are the view-invariant common feature and view-specific private pathological feature of the corresponding viewpoint obtained via decoupling in Section 3.2, with both feature dimensions uniformly set to 768, consistent with the unified dimension design throughout the framework.

View-Invariant Common Feature Graph Topology The common features encode the view-invariant core pathological phenotypes of CS (e.g., moon facies, global facial fat distribution) that are stably expressed across all five viewpoints. To fully exploit the cross-view consistency of core disease features and enable sufficient global information interaction, we construct a fully connected undirected graph for the common feature fusion branch. For this graph, undirected connections are established between every pair of nodes to realize full information interaction between all viewpoints. Meanwhile, node self-loops are retained to ensure the integrity of the node's own feature information. Correspondingly, a 5×5 all-one binary adjacency matrix $\mathcal{A}^c$ is constructed, where $\mathcal{A}^c_{i,j} = 1$ holds for all $i, j \in \{1, 2, 3, 4, 5\}$, providing a global neighborhood range for the attention calculation of the common feature GAT branch.

View-Specific Private Feature Graph Topology The private features encode the unique pathological information that can only be observed under specific viewpoints, with strong view specificity. Undifferentiated full connection will introduce invalid cross-view information interaction and redundant noise, which will destroy the uniqueness of view-specific features. Therefore, we strictly follow the human facial anatomical symmetry prior and clinical observation continuity prior to construct a sparsely connected undirected graph, and only establish edges between viewpoints with anatomical correlation and clinical logical association. The construction of the edge set $\mathcal{E}^p$ and adjacency matrix $\mathcal{A}^p$ complies with two clinical rules: 1. Based on the facial anatomical symmetry prior, fixed undirected edges are added between bilaterally symmetric viewpoints, i.e., $v_{l_1} \leftrightarrow v_{r_1}$ and $v_{l_2} \leftrightarrow v_{r_2}$. This design ensures the information interaction and consistency verification of pathological features between symmetric views, which is fully aligned with the facial symmetry consistency constraint in the loss function system and the weight-sharing strategy of symmetric view-specific encoders. 2. Based on the clinical observation continuity prior, fixed undirected edges are established between spatially adjacent viewpoints that are sequentially observed in clinical diagnosis, i.e.,

edges between the frontal view and bilateral oblique views $v_f \leftrightarrow v_{l_1}$, $v_f \leftrightarrow v_{r_1}$, and edges between ipsilateral oblique and lateral views $v_{l_1} \leftrightarrow v_{l_2}$, $v_{r_1} \leftrightarrow v_{r_2}$. This design conforms to the standard clinical facial diagnosis routine of "frontal view → oblique view → lateral view", and avoids noise caused by invalid information interaction between non-adjacent and non-symmetric views.

Based on the above rules, a 5 × 5 binary adjacency matrix $\mathcal{A}^p$ is constructed, where $\mathcal{A}^p_{i,j} = 1$ if an effective edge exists between node i and node j, and 0 otherwise. Meanwhile, $\mathcal{A}^p_{i,i} = 1$ is set to preserve node self-loop information, defining a clear and clinically reasonable local neighborhood range for the attention calculation of the private feature GAT branch.

3.3.2 Dual-Path Differentiated Graph Attention Fusion Architecture

To accommodate the distinct semantic attributes and optimization objectives of the decoupled view-invariant common disease features and view-specific private pathological features, this study constructs a dual-path GAT fusion architecture with two structurally consistent but functionally independent branches, corresponding exactly to the two sets of graph topologies defined in Section 3.3.1. The common feature fusion branch is designed to enhance the cross-view global consistency of core disease features, while the private feature fusion branch focuses on complementary aggregation of viewpoint-unique pathological information. This differentiated design avoids semantic aliasing and critical diagnostic information loss caused by mixed fusion of heterogeneous features. Both branches take the corresponding features of the five viewpoints as input, adopt an identical GAT computation pipeline, and adapt to their respective feature types and graph topologies through completely independent trainable parameters. Specifically, an orthogonality constraint is introduced into the feature aggregation process of the private feature fusion branch, which enforces low semantic redundancy between the fused private features of different viewpoint groups. This design is fully aligned with the feature decoupling constraint in the loss function system, effectively avoiding feature homogenization during fusion and ensuring the complete preservation of unique pathological details from each viewpoint.

Both branches of the dual-path architecture implement inter-node information interaction using a single-layer multi-head GAT with 4 parallel attention heads. The output dimension of each attention head is set to 192, and the concatenated multi-head features maintain a unified dimension of 768, consistent with the input feature dimension throughout the entire framework. To align with the clinical diagnostic rules of CS, an initial bias term is added to the attention weight calculation, which assigns higher initial weights to the frontal view node and bilateral lateral view nodes. This design conforms to the clinical routine that the frontal view is the primary observation perspective, while the bilateral lateral views provide critical supplementary diagnostic information for CS hallmark phenotypes such as mandibular retrusion and temporal lipoatrophy. For any node v_i in the graph, the attention weight of its neighborhood node $v_j \in \mathcal{N}(v_i)$ (where $\mathcal{N}(v_i)$ includes the node v_i itself, consistent with the self-loop design in the adjacency matrix) is calculated as shown in Eq. (17). Node feature update and aggregation are then performed based on the computed attention weights, as formulated in Eq. (18). For simplicity, the head index k is omitted in Eq. (17), although the attention calculation is performed independently in each attention head.

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^\top \left[\boldsymbol{W}x_i \parallel \boldsymbol{W}x_j\right]\right)\right)}{\sum_{k\in\mathcal{N}(v_i)} \exp\left(\text{LeakyReLU}\left(\boldsymbol{a}^\top \left[\boldsymbol{W}x_i \parallel \boldsymbol{W}x_k\right]\right)\right)} \tag{17}$$

Here, α_{ij} denotes the normalized importance of neighbor node v_j to target node v_i, and the normalization is performed over all nodes in $\mathcal{N}(v_i)$. If the clinically guided initial bias is used in implementation,

it should be understood as being added before Softmax normalization.

$$\boldsymbol{h}_i = \Big\|_{k=1}^{4} \sigma\left(\sum_{j\in\mathcal{N}(v_i)} \alpha_{ij}^{k} \boldsymbol{W}^{k} x_j\right) \tag{18}$$

where $\boldsymbol{W} \in \mathbb{R}^{768\times192}$ denotes the trainable feature projection matrix shared by all attention heads in each independent branch, $\boldsymbol{W}^k$ is the projection matrix corresponding to the k-th attention head, $\boldsymbol{a} \in \mathbb{R}^{384\times1}$ denotes the trainable attention weight vector for each branch, $\|$ denotes the channel-wise concatenation operation of feature vectors, the negative slope of the LeakyReLU activation is set to 0.2, k is the index of the attention head, and σ denotes the nonlinear activation function. For the common feature fusion branch, x_i and x_j represent the view-invariant common features x_i^c and x_j^c; for the private feature fusion branch, they correspond to the view-specific private pathological features x_i^p and x_j^p. Here, $\boldsymbol{h}_i$ denotes the updated node representation of node v_i. Each attention head first computes a weighted sum of projected neighbor features according to α_{ij}^k, and the outputs of the 4 heads are then concatenated to obtain the final 768-dimensional representation. Since each head outputs 192 dimensions, the concatenation preserves the unified feature dimensionality required by the subsequent fusion and classification stages. This operation enables the model to capture multiple complementary interaction patterns among views while adaptively emphasizing diagnostically informative neighbors.

After updating the features of all five nodes via the respective GAT branches, global average pooling is performed independently on the updated node features of the dual-path branches to eliminate viewpoint dimension differences, yielding a 768-dimensional common feature fusion vector $\boldsymbol{H}_c \in \mathbb{R}^{768}$ and a 768-dimensional private feature fusion vector $\boldsymbol{H}_p \in \mathbb{R}^{768}$, respectively. The two fusion vectors are concatenated along the channel dimension to obtain the final 1536-dimensional global fused feature $\boldsymbol{H}_{\text{final}}$, as calculated in Eq. (19).

$$\boldsymbol{H}_{\text{final}} = \boldsymbol{H}_c \,\|\, \boldsymbol{H}_p \tag{19}$$

where $\|$ denotes the channel-wise concatenation of feature vectors. Here, $\boldsymbol{H}_c$ and $\boldsymbol{H}_p$ are the pooled outputs of the common-feature and private-feature branches, respectively. Their concatenation preserves both view-invariant disease evidence and view-specific pathological details, providing the final 1536-dimensional input to the classification head.

The global fused feature $\boldsymbol{H}_{\text{final}}$ is fed into a binary classification head for CS diagnosis, which consists of two cascaded fully connected (FC) layers. The first FC layer maps the 1536-dimensional global feature to a 512-dimensional hidden representation, followed by a GELU nonlinear activation function. The second FC layer maps the hidden feature to 1 dimension, and the Sigmoid activation function finally outputs the predicted probability of CS. This design realizes end-to-end model inference from multi-view facial images to subject-level CS diagnostic predictions. Here, the prediction target is formulated as subject-level binary diagnosis between the CS group and the matched control group. For each subject, the model outputs a probability score indicating the likelihood of belonging to the CS group based on five-view facial photographs. Therefore, the proposed framework is designed as a non-invasive auxiliary diagnostic model for CS, rather than a replacement for standard biochemical testing or specialist clinical diagnosis.

4 Experiments and Dataset

4.1 Dataset

The dataset in this study comprises 343 subjects, including 49 patients diagnosed with Cushing's syndrome (disease group) and 294 BMI-matched controls. As shown in Fig. 5, for each subject, five real

two-dimensional facial photographs were acquired under a standardized imaging protocol, including one frontal view, left and right 45° oblique views, and left and right 90° lateral views. Therefore, the complete dataset contains 1715 facial images in total. These multi-view images were directly photographed and were not synthetically generated. Since the dataset is featured by a limited scale and a pronounced class imbalance, 5-fold cross-validation was adopted for all experiments. Furthermore, random seeds were fixed, and each set of experiments was implemented with five distinct random seeds, with the final experimental results reported as the mean ± standard deviation over five random seeds. All data collection was approved by the Ethics Committee of Peking Union Medical College Hospital.

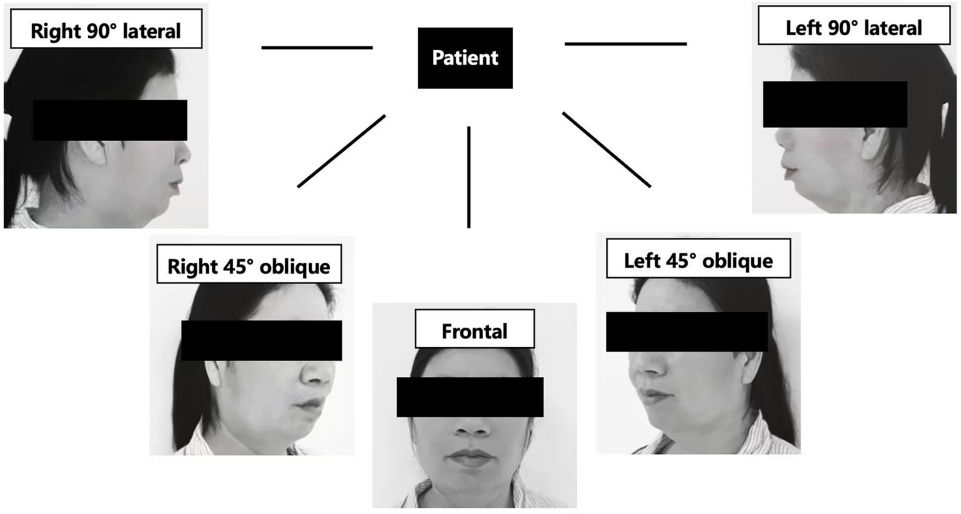

Figure 5: Representative case of a patient with Cushing syndrome, including frontal view, left and right 45° oblique views, and left and right 90° lateral views.

Inclusion Criteria for Disease Group: (1) Age 18–65 years; (2) Diagnosis confirmed by surgical pathology or by an endocrinologist based on clinical presentation, hormone tests, and functional assays; (3) Image acquisition within 10 days pre- or post-surgery; (4) Exclusion of facial trauma/plastic surgery history, systemic diseases affecting facial features (such as scleroderma, lupus), and images of insufficient quality.

Control Group Construction: Control samples (n = 294) were collected from April 2022 to April 2024, with criteria including: (1) $BMI \geq 24$ kg/m^2 and matched to the disease group; (2) Absence of Cushing's syndrome facial features (e.g., moon facies, plethoric appearance); (3) No history of glucocorticoid use; (4) Exclusion criteria consistent with the disease group. Collaboration between orthopedic and endocrinology departments ensured sample diversity.

Data Quality Control: A standardized imaging protocol was employed. For both patients and controls, facial images were captured by trained technicians using a Canon EOS 90D camera (resolution 5472 × 3648) under a consistent and uniform light source and against a fixed-color background. Participants were instructed to sit upright, ensure that the entire face was clearly visible, keep their eyes open and looking straight ahead, close their lips naturally, maintain a neutral and relaxed expression, and remain still during image acquisition. The camera lens was positioned at the same height as the participant's eye level. Five photographs were obtained for each participant, corresponding to the frontal view, left/right 45° oblique views, and left/right 90° lateral views. All images underwent double-blind review, with samples exhibiting motion blur, exposure anomalies, or facial occlusion excluded.

4.2 Implementation Details

All experiments were conducted on a high-performance computing platform equipped with two NVIDIA RTX A6000 48 GB GPUs, an Intel Xeon Gold 6230 CPU, and 512 GB RAM. The software environment was built on Python and implemented using the PyTorch deep learning framework, with CUDA/cuDNN acceleration for GPU-based training and inference. For data preprocessing, Dlib was used for initial face detection and cropping, while the SAM was adopted for refined facial region extraction. Standard scientific computing and image processing libraries were further used for image normalization and handcrafted feature extraction, including Histogram of HOG and LBP descriptors. All comparative experiments and ablation studies were carried out under the same hardware and software settings with fixed random seeds to ensure fairness and reproducibility. To further facilitate reproducibility, the complete source code of this study, together with the environment setup and dependency configuration, has been released in a public GitHub repository: https://github.com/songchangwei/MVFFD-Net. The repository includes the model implementation, training and evaluation scripts, dependency list, and detailed instructions for setting up the experimental environment and reproducing the reported results. The repository includes the full codebase, environment configuration files, package dependencies, and step-by-step instructions for installing the runtime environment and reproducing the experiments.

To comprehensively evaluate the classification performance, we adopted a series of classical metrics including Accuracy, Precision, and Recall. Specifically, the F1-score was introduced as a comprehensive evaluation metric to address potential class imbalance in the dataset, ensuring reliable assessment of overall model performance under imbalanced distributions of positive and negative samples. The implementation was based on the PyTorch deep learning framework [24], which guarantees both reproducibility and development efficiency. The model training followed a standardized protocol: We employed the Stochastic Gradient Descent (SGD) optimizer [25] with an initial learning rate of 5×10^{-4} maintained constant over 200 training epochs. A batch size of 32 was selected to balance training stability with computational resource utilization. The cross-entropy loss function was adopted to effectively measure the discrepancy between predicted probabilities and ground-truth labels, thereby driving the optimization of model parameters.

4.3 Data Preprocessing

Due to variations among photographers and shooting environments, differences in image backgrounds can significantly affect experimental results. To mitigate this interference, we extract facial regions from the images. Initially, we employ the Dlib library, a powerful tool widely used in computer vision for tasks such as face extraction and recognition. After processing the initial photo with Dlib, most non-facial areas are removed, resulting in an image that focuses on the face. However, some unwanted elements, such as hair, clothing, or background, may still remain. Therefore, we utilize the Segment Anything Model (SAM) model for further facial extraction. SAM is a foundational model in computer vision that has demonstrated remarkable performance in segmentation tasks. During the extraction process, we place a label point at the center of the face; SAM generates a mask centered on this point, ultimately yielding a clean facial image devoid of any distracting information.

After obtaining the extracted facial images, we standardize them to meet model input requirements. We first remove the white areas surrounding the face to emphasize the facial features. Next, we uniformly resample the images to a size of 224×224, ensuring consistency in dimensions and resolution throughout the analysis.

For the HOG features, the image is first preprocessed via grayscale conversion and scale normalization. Subsequently, the Sobel operator is employed to compute the image gradient information in the horizontal and vertical directions, from which the gradient magnitude and orientation are further derived. After that,

the image is partitioned into local cells of fixed size. Within each cell, the gradient orientations are quantized into 9 orientation bins, and an orientation histogram is constructed to characterize the local structural features. Finally, blocks are formed by grouping adjacent cells, and intra-block normalization is performed. The features extracted from all blocks are concatenated to form the final HOG descriptor, with a resulting feature dimensionality of 6084.

For LBP features, the extraction is implemented based on a 3×3 pixel neighborhood structure. Specifically, by comparing the grayscale intensity of the central pixel with that of its 8 surrounding neighboring pixels, the local texture pattern is encoded into a binary pattern, which is then converted to the corresponding decimal value. In this way, a local texture descriptor for each pixel is obtained. Furthermore, a 256-dimensional histogram is constructed by counting the distribution of LBP values across the entire image, which serves as the global texture feature of the image, with a fixed feature dimensionality of 256.

4.4 Ethical Statement

This study strictly adheres to academic ethics guidelines, prioritizing participant rights protection throughout the research process. Participant privacy preservation constitutes the fundamental principle governing all research procedures. All personal information underwent immediate de-identification during collection and maintained strict anonymity throughout subsequent analyses. Data storage and transmission processes employed encrypted technologies to prevent unauthorized access and leakage, ensuring comprehensive security throughout the entire data lifecycle. All analytical procedures exclusively serve non-commercial objectives aimed at advancing scientific knowledge and promoting public welfare. The research outcomes contain no commercial applications or financial interest exchanges, preserving the academic integrity and public benefit orientation of this investigation. Prior to study commencement, all participants received explicit notification regarding the exclusive use of their data for anonymized analysis. Informed consent was obtained following full disclosure of research purposes and procedures. The research team assumes full academic responsibility for the study's methodology, content, and conclusions, maintaining rigorous commitment to scholarly objectivity and research reliability.

5 Results

5.1 Contrast Experiment

To comprehensively verify the performance superiority of the proposed Multi-View Facial Feature Disentanglement Network (MVFFD-Net) in the Cushing's syndrome diagnostic task, two types of contrast experiments were designed: performance comparison of different backbone models under single-view setting and performance comparison with state-of-the-art multi-view fusion methods. Four classical evaluation metrics were adopted, including Accuracy, Precision, Recall and F1-Score. Among them, F1-Score was used as the core comprehensive evaluation metric to address the class imbalance characteristics of the experimental dataset (49 Cushing's syndrome patients and 294 BMI-matched controls).

The experimental results of the two types of contrast experiments are integrated into a single table for intuitive comparison, as shown in . The single-view group tests the diagnostic performance of mainstream deep learning backbones on only frontal facial images, while the multi-view group verifies the effectiveness of the proposed MVFFD-Net by comparing with advanced multi-view feature fusion methods on five standard facial views (frontal, bilateral 45° oblique, and bilateral 90° lateral views).

The experimental results reveal two core findings that validate the performance superiority of the proposed MVFFD-Net framework:

Table 3: Contrast experiments of single-view backbones and multi-view fusion methods for Cushing's Syndrome diagnosis.

Setting	Model/Backbone	Accuracy	Precision	Recall	F1 Score
Single-view	Swin [26]	96.56 ± 0.24%	83.22 ± 0.86%	95.10 ± 1.12%	88.76 ± 0.80%
	ResNet [27]	94.98 ± 0.25%	80.46 ± 0.73%	85.71 ± 1.45%	83.00 ± 0.89%
	ViT [28]	94.69 ± 0.24%	77.90 ± 0.68%	87.76 ± 1.45%	82.53 ± 0.87%
	DenseNet [29]	95.04 ± 0.21%	79.20 ± 0.78%	88.57 ± 1.12%	83.62 ± 0.69%
	DINOv2 [23]	97.61 ± 0.13%	90.18 ± 1.20%	93.47 ± 0.91%	91.79 ± 0.40%
Multi-view	GCCA [30]	97.90 ± 0.24%	90.36 ± 1.23%	95.51 ± 0.91%	92.86 ± 0.81%
	DCCA [31]	98.60 ± 0.24%	96.24 ± 0.91%	93.88 ± 1.44%	95.04 ± 0.89%
	Qiang [9]	98.48 ± 0.24%	95.44 ± 0.87%	93.88 ± 1.44%	94.65 ± 0.88%
	TMC [32]	99.13 ± 0.21%	95.29 ± 1.02%	98.78 ± 1.12%	96.99 ± 0.71%
	QMF [33]	99.01 ± 0.16%	94.54 ± 0.82%	98.78 ± 1.12%	96.61 ± 0.55%
	Our Method (MVFFD-Net)	**99.65 ± 0.13%**	**97.99 ± 0.02%**	**99.59 ± 0.91%**	**98.78 ± 0.46%**

Note: Bold values indicate best performance.

For the single-view setting, the self-supervised pre-trained vision foundation model DINOv2 achieves the optimal diagnostic performance among all tested backbones, with an F1-Score of 91.79%, which is 3.03 percentage points higher than the second-best Swin Transformer. This performance advantage stems from two core strengths of DINOv2: (1) its built-in self-attention mechanism effectively models long-range global feature dependencies, which is inherently more suitable for capturing the overall facial morphological changes of CS (e.g., moon facies, centripetal fat distribution) than convolutional neural networks with intrinsic local feature extraction bias; (2) the multi-scale hierarchical features extracted by DINOv2 cover the full semantic spectrum from shallow texture details to deep global phenotypic representations, which is highly matched with the multi-scale characteristics of CS facial pathological phenotypes. These results fully validate the rationality of selecting DINOv2 as the feature extraction backbone, and establish a high-performance baseline for the subsequent multi-view diagnostic framework.

For the multi-view setting, the proposed MVFFD-Net comprehensively outperforms all mainstream state-of-the-art (SOTA) multi-view fusion methods across all four evaluation metrics, achieving the state-of-the-art diagnostic performance. Specifically, MVFFD-Net achieves an F1-Score of 98.78%, which is 1.79 percentage points higher than the second-best TMC method, and 2.17 percentage points higher than the classical deep multi-view representation learning method DCCA. Paired t-tests across five random seeds further confirmed the robustness of the proposed method. Compared with TMC, MVFFD-Net achieved significant improvements in Precision ($t = 6.03$, $p = 0.0039$) and F1-score ($t = 8.95$, $p = 0.0009$). Similarly, compared with QMF, MVFFD-Net also showed significant gains in Precision ($t = 9.49$, $p = 0.0007$) and F1-score ($t = 10.71$, $p = 0.0004$). Furthermore, the ROC curves and corresponding AUC values of six multi-view models are depicted in Fig. 6. It can be observed that the proposed MVFFD-Net achieves the best performance with an AUC of 0.9949 ± 0.0040. This performance improvement suggests that the three core components of the proposed framework jointly contribute to more effective CS diagnosis under the current experimental setting: the bidirectional cross-attention module realizes the synergistic fusion of deep semantic features and clinical prior features, the multi-path disentanglement architecture with multi-objective loss achieves accurate separation of view-invariant and view-specific pathological features, and the clinical priority-guided graph attention fusion framework realizes structured and adaptive aggregation of

multi-view features. The proposed framework effectively breaks through the core technical bottlenecks of existing methods, and achieves high-precision and robust CS diagnosis based on multi-view facial images. This result is also broadly consistent with recent findings in facial-image-based medical AI, where richer view coverage and better integration of complementary features generally lead to improved diagnostic performance. In the context of CS diagnosis, the present results further indicate that incorporating clinically guided priors into a multi-view framework may offer advantages beyond those of generic multi-view fusion models. From the perspective of AI-assisted diagnosis, these results indicate that the proposed framework can effectively distinguish CS patients from matched controls under the current experimental setting. In particular, the high recall and F1-score suggest that the model has strong potential to reduce missed diagnoses while maintaining overall diagnostic reliability, which supports its role as a non-invasive auxiliary diagnostic tool.

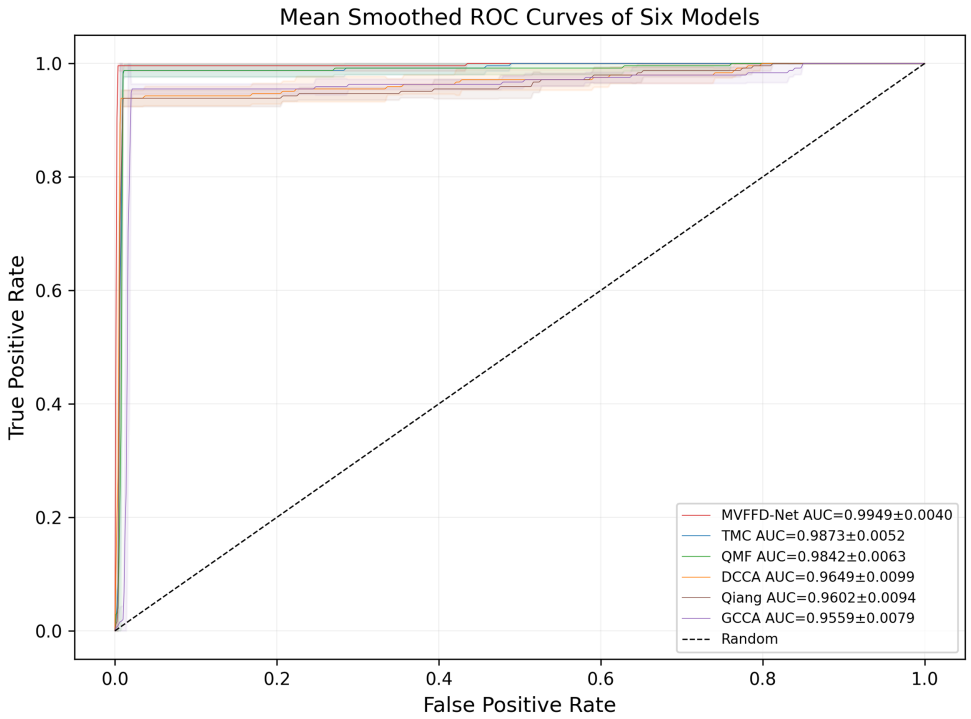

Figure 6: Smoothed mean receiver operating characteristic curves (mean ± standard deviation) for six multi-view models.

5.2 Ablation Experiment

5.2.1 Ablation Experiment of Feature Combinations

To quantitatively validate the contribution of different feature modalities (pure handcrafted features, pure deep features, and deep-handcrafted fused features) to the diagnostic performance of CS, and to verify the effectiveness of the proposed bidirectional cross-attention (BCA) module in multimodal feature fusion, a series of ablation experiments on feature combinations were designed in this study. All experiments were conducted based on the unified MVFFD-Net framework, with only the feature input and fusion methods replaced. The detailed experimental results are presented in Table 4.

The ablation results of feature combinations reveal three core findings that underpin the design rationale of our multimodal feature fusion framework, with direct implications for the clinical translation of CS diagnostic models.

Table 4: Ablation experiment results of feature combinations for Cushing's syndrome diagnosis. Results are reported as mean ± standard deviation over five random seeds.

Feature Combination	Accuracy	Precision	Recall	F1-Score
HOG only	89.04 ± 0.35%	62.07 ± 0.83%	58.77 ± 1.61%	60.37 ± 1.15%
LBP only	87.93 ± 0.14%	58.27 ± 0.67%	54.69 ± 1.71%	56.41 ± 0.92%
DINOv2 only	99.25 ± 0.18%	96.05 ± 1.32%	98.78 ± 1.12%	97.39 ± 0.54%
HOG+LBP	90.84 ± 0.33%	68.80 ± 1.13%	65.72 ± 1.71%	67.22 ± 1.34%
DINOv2+HOG	99.30 ± 0.16%	96.07 ± 1.32%	99.18 ± 1.15%	97.59 ± 0.54%
DINOv2+LBP	99.48 ± 0.13%	97.22 ± 1.04%	99.18 ± 1.12%	98.18 ± 0.45%
DINOv2+HOG+LBP	**99.65 ± 0.13%**	**97.99 ± 0.02%**	**99.59 ± 0.91%**	**98.78 ± 0.46%**

Note: Bold values indicate best performance.

First, handcrafted features alone showed limited diagnostic value. HOG only and LBP only achieved F1-scores of 60.37 ± 1.15% and 56.41 ± 0.92%, respectively, and even their combination reached only 67.22 ± 1.34%, with recall remaining below 66%. This indicates that handcrafted descriptors alone are insufficient to capture the full spectrum of CS-related facial phenotypes. Second, replacing handcrafted features with DINOv2 led to a marked performance improvement, yielding an F1-score of 97.39 ± 0.54% and a recall of 98.78 ± 1.12%. This result confirms the strong discriminative ability of deep semantic representations for CS facial analysis. Third, incorporating handcrafted priors into DINOv2 further improved performance. DINOv2+HOG and DINOv2+LBP increased the F1-score to 97.59 ± 0.54% and 98.18 ± 0.45%, respectively, while the full fusion model (DINOv2+HOG+LBP) achieved the best overall results, with an accuracy of 99.65 ± 0.13%, precision of 97.99 ± 0.02%, recall of 99.59 ± 0.91%, and F1-score of 98.78 ± 0.46%. These results support the effectiveness of the proposed BCA module in exploiting complementary information between deep and handcrafted features.

Collectively, the results of this ablation experiment provide evidence supporting two central assumptions of our study: (1) data-driven deep semantic features and clinically interpretable handcrafted features have strong complementary value for CS diagnosis; (2) the BCA module can effectively achieve bidirectional synergistic fusion of cross-modal features, breaking through the performance ceiling of single-paradigm feature modeling. The full multimodal fusion strategy not only maximizes the diagnostic performance of the model under class-imbalanced clinical datasets, but also injects evidence-based clinical prior knowledge into the deep learning framework, which significantly improves the clinical interpretability of the model and lays a solid foundation for its real-world clinical translation. These findings also align with the broader view that clinically interpretable handcrafted priors and deep semantic features are not mutually exclusive, but can be complementary when integrated in a principled fusion framework.

5.2.2 Ablation Experiment of Facial Symmetry Consistency Loss

To quantitatively validate the necessity of the facial symmetry consistency loss (a core component of the view-specific feature discriminability loss in the multi-objective loss function system) and its contribution to suppressing viewpoint noise, enhancing feature purification, and improving diagnostic robustness, a targeted ablation experiment was designed. The baseline model is the full MVFFD-Net with the complete multi-objective loss function, while the ablation model only removes the facial symmetry consistency loss from the view-specific feature discriminability loss. The detailed experimental results of the baseline model and the ablation model (without facial symmetry consistency loss) are summarized in Table 5.

Table 5: Ablation experiment results of facial symmetry consistency loss for CS diagnosis.

Model Setting	Accuracy	Precision	Recall	F1-Score
Baseline (with facial symmetry consistency loss)	**99.65 ± 0.13%**	**97.99 ± 0.02%**	**99.59 ± 0.91%**	**98.78 ± 0.46%**
Ablation (without facial symmetry consistency loss)	99.36 ± 0.13%	96.82 ± 1.05%	98.78 ± 1.12%	97.78 ± 0.45%

Note: Bold values indicate best performance.

The ablation experiment quantitatively validates the essential role of the proposed facial symmetry consistency loss in optimizing the diagnostic performance, feature purification capability, and clinical robustness of the MVFFD-Net framework. As shown in Table 5, the full model with the facial symmetry consistency loss achieves the best performance across all four evaluation metrics, with an accuracy of 99.65 ± 0.13%, precision of 97.99 ± 0.02%, recall of 99.59 ± 0.91%, and F1-score of 98.78 ± 0.46%. In contrast, removing the facial symmetry consistency loss leads to consistent performance degradation, reducing accuracy to 99.36 ± 0.13%, precision to 96.82 ± 1.05%, recall to 98.78 ± 1.12%, and F1-score to 97.78 ± 0.45%. In particular, the F1-score decreases by 1.00 percentage point, highlighting the contribution of the facial symmetry consistency loss to robust CS diagnosis. This result directly confirms that the facial symmetry consistency loss is an indispensable component of the multi-objective optimization system, and its integration of CS-specific pathophysiological prior knowledge effectively addresses the core challenge of viewpoint noise interference in multi-view facial feature learning. This observation is clinically meaningful because characteristic CS-related facial changes are often manifested in a globally symmetric manner rather than as isolated asymmetric local patterns. Therefore, explicitly modeling symmetry-related constraints may be particularly appropriate for CS-oriented multi-view facial representation learning.

5.2.3 Ablation Experiment of GAT Graph Topology

To quantitatively validate the rationality and superiority of the dual-path differentiated graph topology design for the GAT fusion framework, as well as the clinical value of the clinical priority-guided facial anatomical sparse graph topology (ASG) and fully connected graph topology (FCG) tailored for heterogeneous feature branches, a targeted ablation experiment was conducted. This experiment aims to clarify the contribution of the proposed topology design to reducing invalid information interaction, suppressing redundant noise, preserving feature semantic independence, and enhancing the discriminability of multi-view fused features for CS diagnosis.

To ensure fair comparison, all topology settings maintain the same node set (5 nodes corresponding to the five standard facial views: frontal, left/right 45° oblique, left/right 90° lateral views) and retain self-loop connections for each node to preserve the integrity of the node's own feature information. The detailed topology combination settings for the baseline and ablation groups are as follows:

1. Baseline (FCG + ASG): Original dual-path differentiated design. The common feature branch adopts FCG (10 undirected edges excluding self-loops, full connection between all nodes), and the private feature branch adopts ASG (6 undirected edges excluding self-loops, established only between symmetric and clinically adjacent viewpoints).
2. Ablation 1 (FCG + FCG): The private feature branch is replaced with FCG, while the common feature branch retains the original FCG setting. This group verifies the necessity of ASG for view-specific private feature fusion.

3. Ablation 2 (ASG + ASG): The common feature branch is replaced with ASG, while the private feature branch retains the original ASG setting. This group verifies the necessity of FCG for view-invariant common feature fusion.
4. Ablation 3 (ASG + FCG): Both the common and private feature branches are replaced with mismatched topologies (ASG for the common branch and FCG for the private branch), completely discarding the dual-path differentiated design aligned with feature semantic attributes.

The detailed experimental results of the three graph topology settings are summarized in Table 6. The ablation results quantitatively validate the rationality and superiority of the proposed dual-path differentiated graph topology design. The baseline model (FCG + ASG) achieves optimal performance across all four core evaluation metrics, while all three ablation groups show consistent and significant performance degradation. In summary, the ablation experiment results fully confirm that the proposed dual-path differentiated graph topology design is the optimal solution for multi-view feature fusion in this task. The combination of FCG for the common feature branch and ASG for the private feature branch not only ensures sufficient global interaction of cross-view consistent core disease features, but also preserves the uniqueness of view-specific pathological information and suppresses invalid noise interference. This design strictly aligns with the semantic properties of the decoupled features and clinical diagnostic priors, and is an indispensable component of the MVFFD-Net framework to achieve high-precision CS diagnosis. Unlike generic graph-based multi-view fusion strategies that mainly rely on data-driven connectivity, the proposed topology design explicitly incorporates anatomical and clinical prior knowledge. This may explain why the differentiated topology setting is more suitable for CS diagnosis, where the diagnostic relevance of each facial view is not uniform.

Table 6: Ablation experiment results of Dual-Path GAT graph topology for CS diagnosis.

Topology Combination (Common Branch + Private Branch)	Accuracy	Precision	Recall	F1-Score
Baseline (FCG + ASG)	**99.65 ± 0.13%**	**97.99 ± 0.02%**	**99.59 ± 0.91%**	**98.78 ± 0.46%**
Ablation 1 (FCG + FCG)	99.30 ± 0.16%	96.07 ± 1.32%	99.18 ± 1.12%	97.59 ± 0.54%
Ablation 2 (ASG + ASG)	99.30 ± 0.16%	96.42 ± 0.86%	98.78 ± 1.12%	97.58 ± 0.56%
Ablation 3 (ASG + FCG)	99.13 ± 0.21%	95.66 ± 1.58%	98.37 ± 0.91%	96.99 ± 0.69%

Note: Bold values indicate best performance.

6 Discussion

The present study demonstrates that integrating multi-view facial representations with clinically guided feature fusion and structured graph-based aggregation can improve automatic CS diagnosis under the current experimental setting. Compared with prior single-view approaches, the proposed framework benefits from the complementary diagnostic information provided by frontal, oblique, and lateral views, which is particularly important for CS because several characteristic facial phenotypes are not equally visible from a single viewpoint. In comparison with traditional handcrafted-feature-based methods, the proposed model captures both global facial morphology and fine-grained pathological details, thereby achieving a more balanced representation of clinically relevant phenotypic cues.

Compared with existing multi-view learning methods, the performance gain of MVFFD-Net is not only reflected in the overall F1-score, but also in the high recall rate, which is especially meaningful for screening-oriented clinical tasks where missed diagnosis should be minimized. The results suggest that the combination

of bidirectional cross-attention fusion, multi-view feature disentanglement, and clinically guided graph topology design contributes to more effective utilization of complementary information across views. In particular, the ablation experiments indicate that handcrafted priors still provide useful supplementary information to deep representations, while the symmetry-aware disentanglement strategy helps suppress nuisance variation caused by viewpoint differences.

Importantly, the term "diagnosis" in this study should be understood as AI-assisted auxiliary diagnosis based on facial phenotypic analysis rather than definitive endocrine diagnosis. The proposed framework is intended to support clinicians by providing an additional non-invasive source of diagnostic evidence, especially in situations where characteristic facial manifestations are present. Final confirmation of CS should still rely on standard hormonal evaluation, imaging, and specialist clinical judgment.

From a clinical perspective, the proposed framework provides a promising non-invasive auxiliary screening tool for CS. Its high recall suggests potential value in early identification and referral, especially in scenarios where laboratory testing or advanced imaging is not immediately available. In addition, the explicit incorporation of clinical priors, such as handcrafted texture descriptors and symmetry-related constraints, improves the interpretability of the model design and may facilitate future clinician-oriented analysis of model behavior.

Despite these encouraging findings, several limitations should be acknowledged. First, the dataset size remains limited, and the number of positive cases is relatively small, which may affect the statistical stability of the reported performance. Second, all samples were collected from a single institution, and no cross-institution external validation was performed. Therefore, the generalizability of the model across centers, populations, imaging devices, and clinical workflows remains unconfirmed. Third, although the preprocessing pipeline reduces background interference, the robustness of the framework under real-world clinical conditions, such as illumination variation, occlusion, pose inconsistency, and image quality degradation, has not yet been fully established. Fourth, the current framework relies on five-view facial inputs, which may increase acquisition and standardization requirements in practical deployment.

Future work should therefore focus on several directions. A larger multi-center dataset with more diverse acquisition conditions should be established to validate the robustness and generalizability of the proposed framework. External validation and prospective clinical studies are also needed to better assess real-world applicability. In addition, future research may explore more robust feature learning under noisy conditions, lightweight model design for efficient deployment, and the integration of three-dimensional facial imaging or multimodal clinical information to further improve diagnostic reliability and clinical usability.

7 Conclusion

To address the core bottlenecks of inefficient multi-modal feature fusion and insufficient multi-view feature disentanglement in existing facial image-based CS diagnostic models, we propose a novel Multi-View Facial Feature Disentanglement Network (MVFFD-Net) for high-precision non-invasive CS diagnosis. The framework integrates a bidirectional cross-attention fusion module, a multi-path feature disentanglement architecture with multi-objective optimization, and a clinical priority-guided graph attention fusion framework, taking five standard facial views as input. Overall, MVFFD-Net provides a promising non-invasive framework for AI-assisted auxiliary diagnosis of CS based on multi-view facial images. Under the current experimental setting, MVFFD-Net achieved an F1-score of 98.78% on the collected dataset and outperformed several representative multi-view baselines. These findings suggest that combining multi-view facial information with clinically guided feature fusion and structured graph-based aggregation is a promising direction for CS-oriented intelligent diagnosis. Nevertheless, this study possesses several

limitations, such as limited dataset scale, single-center data origin, and the lack of external validation under diverse real-world acquisition scenarios. Future work will focus on larger multi-center validation, more robust learning under noisy clinical settings, and the integration of richer imaging modalities to further improve generalizability and clinical applicability.

Acknowledgement: The authors gratefully acknowledge the patients and control participants who contributed facial image data to this study. The authors also thank the clinical staff and collaborators from the Department of Endocrinology and the Department of Plastic Surgery at Peking Union Medical College Hospital for their assistance with subject recruitment, data collection, and clinical evaluation.

Funding Statement: This study was supported by the National High Level Hospital Clinical Research Funding [No.: 2025-PUMCH-A-124].

Author Contributions: Conceptualization, Changwei Song, Jiaqi Qiang and Jianqiang Li; methodology, Changwei Song, Hongjun Liu, Qing Zhao and Jiaqi Qiang; software, Changwei Song and Hongjun Liu; validation, Changwei Song and Jiaqi Qiang; formal analysis, Changwei Song; investigation, Changwei Song and Jiaqi Qiang; resources, Jianqiang Li, Hui Pan, Jiuzuo Huang and Shi Chen; data curation, Jiaqi Qiang, Hui Pan, Jiuzuo Huang and Shi Chen; writing—original draft preparation, Changwei Song; writing—review and editing, Changwei Song and Jiaqi Qiang; visualization, Changwei Song; supervision, Qing Zhao; project administration, Qiang Zhao; funding acquisition, Shi Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are not publicly available due to ethical, legal, and privacy restrictions. The dataset consists of five-view facial photographs of patients with Cushing's syndrome and matched controls, together with associated clinical information collected in a hospital setting. Public sharing of these data could compromise participant privacy and is not permitted under the current ethics approval and institutional data-use agreement. Access to de-identified derived data may be considered upon reasonable request and only after approval by the relevant institutional ethics committee and data-governance authority.

Ethics Approval: This study was approved by the Clinical Ethics Committee of Peking Union Medical College Hospital (approval ID: I-22PJ418). Written informed consent was obtained from each participant. Researchers were authorized to use the participants' facial images and clinical data for academic research and publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Arnaldi G, Angeli A, Atkinson AB, Bertagna X, Cavagnini F, Chrousos GP, et al. Diagnosis and complications of Cushing's syndrome: a consensus statement. J Clin Endocrinol Metab. 2003;88(12):5593–602. doi:10.1210/jc.2003-030871.
2. Psaras T, Milian M, Hattermann V, Freiman T, Gallwitz B, Honegger J. Demographic factors and the presence of comorbidities do not promote early detection of cushing's disease and acromegaly. Exp Clin Endocrinol Diabetes. 2010;119(1):21–5. doi:10.1055/s-0030-1263104.
3. Peng H, Li J, Cheng W, Zhao L, Guan Y, Li Z, et al. Learn to learn: a mirror meta-learning method for retinal disease diagnosis on fundus images. IEEE Trans Artif Intell. 2025;6(12):3391–405. doi:10.1109/tai.2025.3566082.
4. Li J, Zhu C, Zhao M, Xu X, Zhao L, Cheng W, et al. OMGMed: advanced system for ocular myasthenia gravis diagnosis via eye image segmentation. Bioengineering. 2024;11(6):595. doi:10.3390/bioengineering11060595.
5. Popp KH, Kosilek RP, Frohner R, Stalla GK, Athanasoulia-Kaspar AP, Berr CM, et al. Computer vision technology in the differential diagnosis of Cushing's syndrome. Exp Clin Endocrinol Diab. 2019;127(10):685–90. doi:10.1055/a-0887-4233.

6. Kosilek R, Schopohl J, Grunke M, Reincke M, Dimopoulou C, Stalla G, et al. Automatic face classification of Cushing's syndrome in women—a novel screening approach. Exp Clin Endocrinol Diabetes. 2013;121(9):561–4. doi:10.1055/s-0033-1349124.
7. Wei R, Jiang C, Gao J, Xu P, Zhang D, Sun Z, et al. Deep-learning approach to automatic identification of facial anomalies in endocrine disorders. Neuroendocrinology. 2019;110(5):328–37. doi:10.1159/000502211.
8. Liu H, Song C, Qiang J, Li J, Pan H, Lu L, et al. Comparative analysis of pre-trained deep learning models and DINOv2 for Cushing's syndrome diagnosis in facial analysis. In: Proceedings of the 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC); 2025 Jul 8–11; Toronto, ON, Canada. New York, NY, USA: IEEE; 2025. p. 324–9. doi:10.1109/compsac65507.2025.00052.
9. Qiang J, Liu H, Guo X, Song C, Lu L, Long X, et al. Deep learning-based multiview facial identification as a screening tool for Cushing syndrome. Endocr Pract. 2025;31(12):1601–7. doi:10.1016/j.eprac.2025.07.016.
10. Li Y, Wang Y, Cui Z. Decoupled multimodal distilling for emotion recognition. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. New York, NY, USA: IEEE; 2023. p. 6631–40. doi:10.1109/cvpr52729.2023.00641.
11. Zhao M, Gao H, Zhao L, Wang Z, Wang F, Zheng W, et al. Decoupled multi-perspective fusion for speech depression detection. IEEE Trans Affect Comput. 2025;16(3):1772–86. doi:10.1109/taffc.2025.3538519.
12. Zhao X, Li X, Jiang R, Tang B. Decoupled cross-attribute correlation network for multimodal sentiment analysis. Inf Fusion. 2025;117(5):102897. doi:10.1016/j.inffus.2024.102897.
13. Zhang W, Li D, Feng M, Hu B, Fan Y, Chen Q, et al. Electronic medical records as input to predict postoperative immediate remission of Cushing's disease: application of word embedding. Front Oncol. 2021;11:754882. doi:10.3389/fonc.2021.754882.
14. Pinkus A. Approximation theory of the MLP model in neural networks. Acta Numer. 1999;8:143–95. doi:10.1017/S0962492900002919.
15. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. doi:10.1023/a:1022627411411.
16. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/a:1010933404324.
17. Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B Stat Methodol. 1958;20(2):215–32. doi:10.1111/j.2517-6161.1958.tb00292.x.
18. Isci S, Kalender DSY, Bayraktar F, Yaman A. Machine learning models for classification of Cushing's syndrome using retrospective data. IEEE J Biomed Health Inform. 2021;25(8):3153–62. doi:10.1109/jbhi.2021.3054592.
19. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85. doi:10.1080/00031305.1992.10475879.
20. Tharwat A. Principal component analysis—a tutorial. Int J Appl Pattern Recognit. 2016;3(3):197. doi:10.1504/ijapr.2016.079733.
21. Loh WY. Classification and regression trees. WIREs Data Min Knowl Discov. 2011;1(1):14–23. doi:10.1002/widm.8.
22. Soni DG. A comprehensive study on Dlib-ML: a machine learning toolkit. Int J Res Appl Sci Eng Technol. 2025;13(11):1478–83. doi:10.22214/ijraset.2025.75328.
23. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: learning robust visual features without supervision. 2024 [cited 2026 Apr 5]. Available from: https://hal.science/hal-04376640.
24. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d Alché-Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems. Vol. 32. Red Hook, NY, USA: Curran Associates, Inc.; 2019 [cited 2026 Apr 5]. Available from: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
25. Bottou L. Large-scale machine learning with stochastic gradient descent. Heidelberg, Germany: Physica-Verlag HD; 2010. p. 177–86. doi:10.1007/978-3-7908-2604-3_16.
26. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. New York, NY, USA: IEEE; 2021. p. 9992–10002. doi:10.1109/iccv48922.2021.00986.

27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. New York, NY, USA: IEEE; 2016. p. 770–8. doi:10.1109/cvpr.2016.90.
28. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the Proceedings of the International Conference on Learning Representations; 2021 May 4; Vienna, Austria. [cited 2026 Apr 5]. Available from: https://openreview.net/forum?id=YicbFdNTTy.
29. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 2261–9. doi:10.1109/CVPR.2017.243.
30. Afshin-Pour B, Hossein-Zadeh GA, Strother SC, Soltanian-Zadeh H. Enhancing reproducibility of fMRI statistical maps using generalized canonical correlation analysis in NPAIRS framework. NeuroImage. 2012;60(4):1970–81. doi:10.1016/j.neuroimage.2012.01.137.
31. Wang W, Arora R, Livescu K, Bilmes J. On deep multi-view representation learning. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. Vol. 37 of Proceedings of Machine Learning Research. Lille, France: PMLR; 2015. p. 1083–92. [cited 2026 Apr 5]. Available from: https://proceedings.mlr.press/v37/wangb15.html.
32. Han Z, Zhang C, Fu H, Zhou JT. Trusted multi-view classification with dynamic evidential fusion. IEEE Trans Pattern Anal Mach Intell. 2023;45(2):2551–66. doi:10.1109/tpami.2022.3171983.
33. Zhang Q, Wu H, Zhang C, Hu Q, Fu H, Zhou JT, et al. Provable dynamic fusion for low-quality multimodal data. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning. Vol. 202 of Proceedings of Machine Learning Research. Lille, France: PMLR; 2023. p. 41753–69. [cited 2026 Apr 5]. Available from: https://proceedings.mlr.press/v202/zhang23ar.html.