Multi-View Cross-Fusion Transformer Based on Kinetic Features for Non-Invasive Blood Glucose Measurement Using PPG Signal

Shisen Chen, Fen Qin, Xuesheng Ma, Jie Wei, Yuan-Ting Zhang, *Fellow, IEEE*, Yuan Zhang, *Senior Member, IEEE*, and Emil Jovanov, *Fellow, IEEE*

***Abstract*—Noninvasive blood glucose (BG) measurement could significantly improve the prevention and management of diabetes. In this paper, we present a robust novel paradigm based on analyzing photoplethysmogram (PPG) signals. The method includes signal pre-processing optimization and a multi-view cross-fusion transformer (MvCFT) network for non-invasive BG assessment. Specifically, a multi-size weighted fitting (MSWF) time-domain filtering algorithm is proposed to optimally preserve the most authentic morphological features of the original signals. Meanwhile, the spatial position encoding-based kinetics features are reconstructed and embedded as prior knowledge to discern the implicit physiological patterns. In addition, a cross-view feature fusion (CVFF) module is designed to incorporate pairwise mutual information among different views to adequately capture the potential complementary features in physiological sequences. Finally, the subject-wise 5-fold cross-validation is performed on a clinical dataset of 260 subjects. The root mean square error (RMSE) and mean absolute error (MAE) of BG measurements are 1.129 mmol/L and 0.659 mmol/L, respectively, and the optimal Zone A in the Clark error grid, representing none clinical risk, is 87.89%. The results indicate that the proposed method has great potential for homecare applications.**

***Index Terms*—Blood glucose estimation, deep learning, non-invasive measurement, photoplethysmography.**

NOMENCLATURE

VPG	Velocity plethysmogram.
APG	Acceleration plethysmogram.
FPG	Fasting plasma glucose.
SQA	Signal quality assessment.
PAA	Piecewise aggregation approximation.
KFR	Kinetics feature reconstruction.
SPE	Spatial position encoding.
GAF	Gramian angular field.
RP	Recurrence plot.
MSWF	Multi-size weighted fitting.
CVFF	Cross-view feature fusion.
CST	Cross-scale transformer.
CEG	Clark error grid.
SEG	Surveillance error grid.
CV	Cross-validation.

I. INTRODUCTION

DIABETES is a chronic disease that gravely impairs human health and may trigger serious comorbidities, placing a severe load on the healthcare system [1]. The International Diabetes Association hence emphasizes the vital role of regular at-home blood glucose (BG) monitoring for prevention and early diagnosis of diabetes [2]. However, traditional approaches to BG management primarily rely on invasive and minimally invasive devices, which might bring psychological stress and physiological pain to patients. Their effectiveness is also restricted by factors such as frequency of testing, portability, and cost. Therefore, the research community has been trying for decades to develop a reliable, non-invasive BG monitoring technology, preferably low-cost and portable, that would significantly improve the management of chronic conditions and ease the burden of care.

Photoplethysmography (PPG), an effective solution for detecting multiple physiological parameters at low cost, is successfully applied in commercial wearable devices for measuring

Manuscript received 17 August 2023; revised 5 December 2023; accepted 2 January 2024. Date of publication 9 January 2024; date of current version 5 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62172340, and in part by the Natural Science Foundation of Chongqing under Grants cstc2021jcyj-msxmX0041 and cstc2021jcyj-msxmX0968. *(Shisen Chen and Jie Wei contributed equally to this work.) (Corresponding author: Yuan Zhang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethnic Committee of The Ninth People's Hospital of Chongqing under Application No. 2022-SCI-007.

Shisen Chen, Jie Wei, and Yuan Zhang are with the Chongqing Key Laboratory of Nonlinear Circuits and Intelligent Information Processing, College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China (e-mail: chenshisen@email.swu.edu.cn; hybyyhcx7@swu.edu.cn; yuanzhang@swu.edu.cn).

Fen Qin is with Endocrinology Department, The Ninth People's Hospital, Chongqing 400799, China (e-mail: 285700338@qq.com).

Xuesheng Ma is with the School of Natural Medicine, University of the Western Cape, Bellville 7535, South Africa (e-mail: xma@uwc.ac.za).

Yuan-Ting Zhang is with the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin 999077, China (e-mail: ytzhang@cuhk.edu.HK).

Emil Jovanov is with the Department of Electrical and Computer Engineering, University of Alabama in Huntsville, Huntsville, AL 35899 USA (e-mail: emil.jovanov@uah.edu).

Digital Object Identifier 10.1109/JBHI.2024.3351867

heart rate (HR) and blood oxygen [3]. The technique is based on the Beer-Lambert law and obtains physiological information by measuring changes in light absorption in the blood [4]. In addition, variations in BG levels affect the viscosity of the blood, which in turn affects blood flow and velocity in the microvasculature [5]. This physiological correlation is implicitly reflected in the PPG signal. Therefore, important correlations between hemodynamic characteristics and BG states can be revealed by analyzing PPG signals with the powerful feature representation ability of deep learning models [6].

Notably, acquiring reliable physiological signals is crucial to enhance further the accuracy of non-invasive vital signs monitoring. Unfortunately, many studies have ignored the filtering problem that results in distortion of the beginning and end parts of the signal [7], [8], [9]. Additionally, manual screening of low-quality signals increases labor and time costs and limits the application of real-time monitoring and automated systems [10]. Moreover, partitioning the signal by period requires interpolation to align varying cardiac cycles due to differences in the cardiac cycle duration across subjects. However, this approach may destroy the temporal information in the original physiological sequence while introducing unnecessary computational redundancy [6], [11], [12].

Several advances have been made in recent non-invasive BG measurement studies using conventional machine learning methods [6], [8], [11], [13], [14], [15], [16], [17]. For example, Zhang et al. [6] extracted 28 features in the time-frequency domain of each single-cycle signal using Gaussian fitting. Finally, a Gaussian support vector machine was adopted to categorize the BG into three warning levels, from normal to severe, with an accuracy of 81.49%. Wei et al. [13] used a stacked fusion strategy to extract 33 features related to HR, blood pressure (BP), and the time-frequency domain. Random forest was then used as a regression model, and 86.84% of the test data fell into region A of the CEG. Alonso et al. [14] extracted 13 features related to Mel-Frequency Cepstral Coefficients (MFCC) from the PPG signals and used the AdaBoost multi-model integration method to obtain the best result of MAE = 0.646 mmol/L. These studies focus on extracting a few hand-designed features, such as heart rate variability (HRV), HR, BP, MFCC, and time-frequency domain. Although these features are computationally efficient and, to some extent, interpretable, they often rely on the priori knowledge of domain experts. Thus, providing generic features to adequately express the complex relationship between PPG signal and BG levels is challenging. In addition, the small number of these features limits the model's ability to present individual variability.

Because deep neural networks (DNN) have powerful nonlinear fitting capabilities, researchers have begun to explore applying deep learning to BG measurement to improve performance further. Notably, Li et al. [11] combined the feature extraction capabilities of traditional machine learning and deep learning to manually extract 160 time-frequency domain features from ECG and PPG signals. In comparison, 66,560 spatial morphological features were automatically extracted using DNN. This work demonstrates the potential of deep learning in BG monitoring but requires additional ECG signals, which may cause discomfort to the user. In contrast, PPG signals have the advantage of being cost-effective and easily capturing physiological patterns in vivo, enabling non-invasive BG measurements. Zhang et al. [18] introduced a novel time-frequency graph to enrich the time and frequency information of network learning and finally estimates the BG level of fingertip videos by end-to-end dual-stream DNN. In addition, Lee et al. [19] presented a practical sensor placement approach to obtain higher-quality signals and implemented the measurement of BG levels using a convolutional neural network (CNN). They achieved 84.29% accuracy based on International Organisation for Standardisation (ISO) 15197:2013 standard [20].

Although previously cited research achieves promising preliminary results, their performance can be improved by considering the kinetics and non-stationarity of the physiological time series. Recently, researchers have recognized the limitations of relying only on one-dimensional (1D) PPG signals to estimate BG levels. To surmount this issue, several studies have applied the dynamic trajectory of the PPG signal as a visual indication of BP or BG changes [12], [21]. This method reduces the burden of feature extraction and unveils more important hemodynamic features in the PPG signal. For example, Wang et al. [12] exploited inter-node visibility coding to map PPG waveforms to visual graphs and estimated BP values via a ridge regression algorithm. Ouyang et al. [21] proposed to encode peripheral pulse waveforms as images and classify BG by multi-scale fusion CNN. However, these methods rely solely on the encoded images as input features to a single branch network, omitting the importance of the original physiological sequences and failing to provide sufficient priori knowledge for the model. Furthermore, the adoption of shallow CNN makes it hard to grasp contextual information at long distances, limiting its ability to perceive global information. Moreover, the necessity and effectiveness of signal conversion has not been adequately validated.

Notably, most of the above work adopted the BG meter test value as the ground truth. However, the device's bias makes it impossible to accurately assess the deviation between the measured and actual values. In contrast, utilizing the FPG value (the gold standard) as ground truth shows apparent advantages because it avoids the potential influence of diet, exercise, and other factors on BG measurement [22]. In addition, many studies have not employed CV methods [8], [15], [16], [17], [19] and have divided the dataset in a record-wise manner [13], [14]. In the intra-patient (i.e., record-based) approach, the training and testing sets come from different records of the same patient, which may lead to overfitting and data leakage, resulting in unreliable results. Conversely, our experiments are conducted in a subject-based (i.e., inter-patient) approach. The training and testing sets are from different patients to ensure the model better accommodates inter-individual differences.

In this paper, we propose an elegant paradigm for a PPG based noninvasive BG measurement to address the deficiencies of the existing work. We present a comprehensive methodology, from optimized data pre-processing to the design of a novel multi-view cross-fusion transformer (MvCFT) network. The primary contributions are as follows:

1) We propose an optimized data pre-processing paradigm to enhance the dataset's quality and obtain reliable experimental results. A MSWF algorithm is designed to minimize signal distortion during filtering. The noisy data segments are automatically removed via SQA.
2) We propose a KFR algorithm based on spatial position encoding. The effective fusion of positional information, phase correlation, and periodicity between PPG points facilitates the representation of underlying hemodynamic features in the signal.
3) Our proposed MvCFT network deeply fuses complementary information of heterogeneous features among views and potentially shared knowledge in physiological sequences through the CVFF mechanism. Extensive experiments on clinical datasets have proven that our method outperforms previous state-of-the-art methods and provides a more robust solution for non-invasive BG monitoring.

II. Methodology

A. Problem Definition

This study aimed to regress BG values based on the subjects' PPG signals using a data set. The data set can be set as $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i, i \in [0, n]$ denotes the entire PPG signal of subject i, $Y_i, i \in [0, n]$ represents the FBG value of subject i, and n stands for the number of samples. X_i is segmented into fixed frame length segments using a sliding window, $X_i = \{S_1^t, \ldots, S_l^t\}$, where $S_{ij}^t, j \in [1, l]$ stands for splitting the PPG signal into segments of t seconds (s) and l stands for the number of segments. The first and second-order derivatives of the S_{ij}^t (i.e., S_{ppg}^{ij}), denoted as VPG, i.e., S_{vpg}^{ij} and APG, i.e., S_{apg}^{ij}, respectively, are concatenated as $C_{ij}^t = (S_{ppg}^{ij}, S_{vpg}^{ij}, S_{apg}^{ij}) \in \mathbb{R}^{D \times 1 \times L}$, where C_{ij}^t represents the combined signals obtained by concatenating the jth PPG signal (t(s) length) of subject i and its derivatives. Next, the kinetic features matrix of S_{ij}^t is obtained by the KFR algorithm κ, i.e., $K_{ij}^t = \kappa(S_{ij}^t) \in \mathbb{R}^{D \times S \times S}$, where K_{ij}^t represents the kinetic features obtained from the jth PPG signal (t(s) length) of subject i, D represents the dimension, L represents the length of the subframe signal, and $S \times S$ represents the size of the kinetic features matrix.

The proposed MvCFT network aims to learn the nonlinear mapping relationship $F_{bg}(\cdot)$ between the multi-view data (i.e., C_{ij}^t and K_{ij}^t) and the BG values (Y_i) as shown in the (1).

$$\hat{Y}_i = F_{bg}\left(\left(C_{ij}^t, K_{ij}^t\right); \theta\right) \tag{1}$$

where θ represents the hyperparameters of the model. As shown in (2), a hierarchical smoothing loss is introduced to match the BG measurement standard (ISO 15197:2013),

$$\mathcal{L}_{BG} = \begin{cases} 0.5\left(Y_i - \hat{Y}_i\right)^2/\beta, & \text{if } \left|Y_i - \hat{Y}_i\right| < \beta \\ \left|Y_i - \hat{Y}_i\right| - 0.5 \times \beta & \text{otherwise} \end{cases} \tag{2}$$

where β is set to 0.83, as specified in ISO 15197:2013. This standard dictates that the error should not exceed 0.83 mmol/L for BG values below 5.6 mmol/L. Our ultimate goal is to minimize the loss of label Y_i and the measured value $\hat{Y}_i$.

Algorithm 1: Multi-Size Weighted Fitting (MSWF).

Input: Input signal $\mathbf{X_{ppg}}$, Num_iterations $\mathbf{N}$, Order of the polynomial $\mathbf{k}$, Size of the small sliding window $\mathbf{SW_{size}}$, Size of the large sliding window $\mathbf{LW_{size}}$.
Output: Filtered signal $\mathbf{X_f}$.

// Stage 1. Smoothing filter.
1: For each index i of the X_{ppg} extract the data points in $[\frac{1-SW_{size}}{2}, \frac{SW_{size}-1}{2}]$:
2: Construct the Vandermonde matrix X, where each row represents a data point with the highest order of k;
3: Compute the pseudo-inverse of X: $X_{\text{pinv}} = (X^T X)^{-1} X^T$. $\mathbf{y}$ values in X_{ppg} as column vectors Y;
4: Compute the polynomial coefficients by multiplying X_{pinv} with y: $c = X_{\text{pinv}} y$, X_{ppg} filtered as $Y_f = Xc$.
// Stage 2. Remove baseline drift.
5: For each index i of the X_{ppg} extract the data points in $[\frac{1-LW_{size}}{2}, \frac{LW_{size}-1}{2}]$ and map the interval to $[-1, 1]$:
6: Select the weight function $W(x)$ and calculate the W_i for each point;
7: Use weighted fitting to obtain a locally fitted curve near point x_i: $\hat{Y} = X(X^T w X)^{-1} X^T w Y$;
8: Calculation error $d = |Y - \hat{Y}|$ and the median of the d is noted as d_m;
9: Update the weights as $W_{new}^k = W(\frac{d_k}{d_m})$ and calculate the new Y_b, repeat $\mathbf{N}$ times step 7, 8, 9;
10: Return: $\mathbf{X_f} = \mathbf{Y_f}$ - $\mathbf{Y_b}$;

B. Preprocessing

Despite the efforts to ensure subjects' resting status, their micromovements and breathing during acquisition may still impact the PPG signal, thereby introducing problems such as high-frequency noise and baseline drift [23]. In order to mitigate the problem of signal distortion and avoid misleading the model to make wrong decisions due to signal quality issues, it is crucial to implement data preprocessing. We implemented signal preprocessing as follows:

1) Segmentation: Generally, the longer the signal length is, the more likely it is to accumulate noise and interference, which reduces signal quality. Therefore, the sliding window algorithm is employed to partition the signal into segments of the same frame length.

2) Filtering: When the noise spectrum overlaps with the signal spectrum, the frequency domain filtering algorithm may cause boundary effects and spectral overlap problems, leading to signal distortion. [24]. We propose a filtering algorithm based on MSWF from the time domain perspective as shown in the Algorithm Algorithm 1. The MSWF removes the signal noise and baseline drift while ensuring that the shape and width of the signal are not distorted. Finally, the variability of PPG signals among different individuals is eliminated by normalization of the signal.

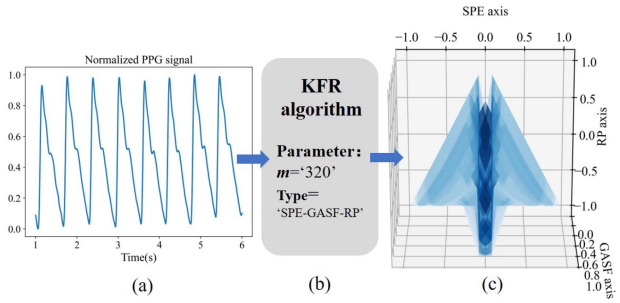

Fig. 1. Pipeline of PPG signal kinetic feature reconstruction, (a) is the PPG signal after normalization, (b) represents the KFR algorithm with parameters set to ***m***='320', ***Type***='SPE-GASF-RP', and (c) 3D visualization of PPG signal kinetic features.

3) Signal Quality Assessment: Automated signal quality monitoring is essential to optimize the quality of model inputs for a more robust BG assessment. 260 subjects participated in this study to produce 4684 PPG segments. Referring to [23], we found that skewness is the best metric for assessing the PPG signal quality, as shown in (3). Therefore, we set the S_{SQI} threshold to 0.3, so that segments with SQA scores greater than 0.3 are accepted and those less than 0.3 are rejected. After this screening, low-quality signals are removed from each subject to retain 3,892 high-quality PPG segments for the following experiments.

$$S_{SQI} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{x_i - \widehat{\mu}_x}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{\mu}_x)^2}}\right]^3 \tag{3}$$

where $\widehat{\mu}_x$, N are the mean value of x_i and the number of samples, respectively.

C. Kinetics Feature Reconstruction

The extreme similarity of PPG signals in the time domain makes it hard for models to distinguish and capture subtle differences. To effectively map the complex nonlinear relationship between PPG signals and BG levels, we propose a KFR algorithm based on spatial position encoding, as shown in Algorithm 2. The algorithm captures the pattern of signal evolution over time by fusing spatial position (SPE), phase correlation (GAF), and periodicity (RP) to ultimately generate the combined nonlinear kinetic features from the original PPG signal, as shown in Fig. 1. The MvCFT network promises to reveal the signal's intrinsic structure and nonlinear kinetics by analyzing the signal and its kinetic features, providing new insights into BG measurement from PPG [25], [26].

Considering the conversion efficiency, the pre-processed PPG can be represented as $X = \{x_1, x_2, \ldots, x_n\}, X \in [0, 1]$ sequence data of length n. The sequence is embedded in m dimensions with PAA, as shown in (4).

$$\overline{X}_{paa}^{i} = \frac{m}{n}\sum_{j=\frac{n}{m}(i-1)+1}^{\frac{n}{m}i} x_j, \quad i \in [0, m] \tag{4}$$

Algorithm 2: KFR Algorithm Flow.

Input: PPG data $X = \{x_i\}_{i=1}^{n}$; the PPG dimensions after dimensionality reduction is *m*; the combination *Type* of kinetic features.
Output: kinetic features F.
// PPG data is dimensionally reduced by PAA to obtain $X_{paa} \in \mathbb{R}^{1\times m}$ by (4).
$X_{paa} \leftarrow \text{PAA}(X, m)$
// The matrix $SPE \in \mathbb{R}^{m\times m}$ is obtained by (5).
$SPE \leftarrow \text{ReconstructSPE}(X_{paa})$
// The matrices $GASF$ and $GADF \in \mathbb{R}^{m\times m}$ is obtained by (7) and (8).
$GASF \leftarrow \text{ReconstructGASF}(X_{paa})$
$GADF \leftarrow \text{ReconstructGADF}(X_{paa})$
// The matrix $RP \in \mathbb{R}^{m\times m}$ is obtained by (9).
$RP \leftarrow \text{ReconstructRP}(X_{paa})$
// According to the parameter *Type*, combining the three kinetic features in $SPE, GASF, GADF$, and RP to obtain $F \in \mathbb{R}^{3\times m\times m}$.
$F \leftarrow \text{CombineFeatures}(SPE, GASF, GADF, RP, Type)$
return kinetic features F

1) Spatial Position Encoding (SPE): Inspired by the relative position embedding, described in [27], SPE incorporates spatial location information with feature representation to augment the distinguishability of the features by the model. The same PPG may have different feature patterns at points in different spatial positions, which reflects the dynamic changes in blood flow. Embedding the SPE matrix can facilitate the model to mine more long-range and local time series dependencies, thus enhancing the generalisability of features. Specifically, the sequence $X_{paa} = \{\vec{x}_i\}_{i=1}^{m}$ after PAA dimensionality reduction. The spatial location information between arbitrary two points in the sequence is computed sequentially by Euclidean norm $\|\vec{x}_i - \vec{x}_j\|, i, j \in [0, m]$. From (5), the SPE matrix $SPE_{ij} \in \mathbb{R}^{m\times m}$ of the sequence can be derived.

$$SPE_{ij} = \|\vec{x}_i - \vec{x}_j\| = \sqrt{(\vec{x}_i - \vec{x}_j)^T(\vec{x}_i - \vec{x}_j)} \tag{5}$$

2) Gramian Angular Field (GAF): GAF maps each time point in the time series to an angular value and a polar radius in the polar coordinate system by calculating the difference or relative angle between different time points. Such representation can help reveal the signal's dynamic evolution pattern at different time points and further explore the implicit cardiovascular kinetics information in the signal. Specifically, $X_{paa} = \{\vec{x}_i\}_{i=1}^{m}$ is transformed from the Cartesian coordinate system to the Polar coordinate system using (6):

$$\phi_i = arccos(\vec{x}_i), r_i = \frac{i}{m}, i \in [0, m] \tag{6}$$

where ϕ_i is the angle vector and r_i corresponds to the radius. Finally, considering the angular sum or difference between different points through (7) and (8) obtains two forms of GAF as

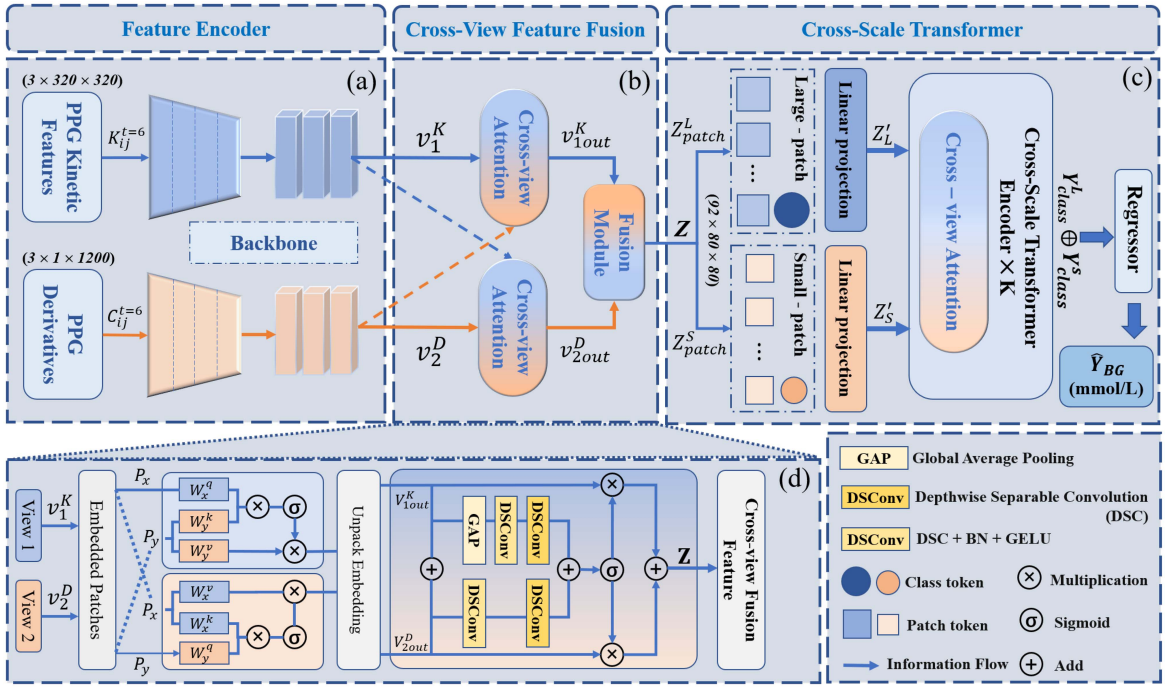

Fig. 2. Pipeline diagram of the proposed MvCFT framework for BG measurement. In the figure, $(\bullet \times \bullet \times \bullet)$ stands for $(C \times H \times W)$, where C, H, and W are the number of channels, height, and width of the features or signals, respectively. $C_{ij}^{t=6}$ and $K_{ij}^{t=6}$ represent the combined signals and combined kinetic features from segment j after segmenting the PPG signal of subject i by 6 s, respectively. The legend is shown in the lower right, where BN stands for batch normalization and GELU (Gaussian Error Linear Units) is the activation function.

$GASF$ and $GADF$ respectively; $GASF, GADF \in \mathbb{R}^{m \times m}$.

$$\begin{aligned} \text{GASF} &= \cos(\phi_i + \phi_j) \\ &= X_{paa}^T X_{paa} - \sqrt{I - X_{paa}^2}^T \sqrt{I - X_{paa}^2} \end{aligned} \tag{7}$$

$$\begin{aligned} \text{GADF} &= \sin(\phi_i - \phi_j) \\ &= \sqrt{I - X_{paa}^2}^T X_{paa} - X_{paa}^T \sqrt{I - X_{paa}^2} \end{aligned} \tag{8}$$

3) Recurrence Plot (RP): RP is a method for analyzing nonlinear kinetics [25]. It reveals a time series' intrinsic structure by analyzing the signal's periodicity, chaos, and non-stationarity. RP enables the capture of the nonlinear kinetic features in PPG signals, which helps to enhance the model's perception of the latent physiological information in PPG signals. Particularly, for $X_{paa} = \{\vec{x}_i\}_{i=1}^m$ the corresponding $RP_{ij} \in \mathbb{R}^{m \times m}$ matrix can be obtained through (9):

$$RP_{ij} = \Theta(\varepsilon - \|\vec{x}_i - \vec{x}_j\|), \quad i, j \in [0, m] \tag{9}$$

where the threshold ε is empirically taken to be 10% of the peak value [28] and $\Theta(\cdot)$ is a step function as shown in (10).

$$\Theta(\cdot) = \begin{cases} 1, & \text{if}(\varepsilon - \|\vec{x}_i - \vec{x}_j\|) \geq 0 \\ 0, & \text{if}(\varepsilon - \|\vec{x}_i - \vec{x}_j\|) < 0 \end{cases} \tag{10}$$

D. Proposed Model

The proposed model MvCFT consists of three significant components: 1) Feature Encoder, 2) Cross-View Feature Fusion, and 3) Cross-Scale Transformer, as shown in Fig. 2. The detailed description is as follows:

TABLE I
STATISTICAL INFORMATION OF THE CLINICAL DATASET

Measure	Maximum	Minimum	Mean	SD
Age (year)	82	16	43	13.8
Height (m)	1.82	1.15	1.62	0.1
Weight (kg)	104.0	63.8	42	11.0
BMI (kg/m^2)	37.8	16.0	24.3	3.5
BG $(mmol/L)$	12.0	4.0	5.7	1.6

SD:Standard deviation, BMI: Body mass index.

1) Multi-View Learning: Multi-view learning aims to distinguish a shared set of high-level semantic features or latent structures from data captured from multiple sources, spaces, and forms via feature encoders [29]. This set of features is consistent and complementary across views, providing a more comprehensive and robust representation of the data for the model. Thus, it is expected to improve the performance and generalization of the model. Specifically, $C_{ij}^t = (S_{ppg}^{ij}, S_{vpg}^{ij}, S_{apg}^{ij}) \in \mathbb{R}^{D \times 1 \times L}$ and $K_{ij}^t = \kappa(S_{ij}^t) \in \mathbb{R}^{D \times S \times S}$ are considered as two input views. The proposed MvCFT learns the potential feature representations of multiple heterogeneous information.

2) Cross-View Feature Fusion (CVFF): Multi-views have complementary information about the identical data, but direct integration of low-level spatial features may impair model performance due to view bias [30]. We propose a CVFF module, as shown in Fig. 2(a), to project the heterogeneous information of multi-views into a common feature space. The underlying feature representations of multi-views are aligned by

incorporating a self-attention mechanism, while preserving intra and inter-view semantic features.

Assume that C_i and K_i are passed through the backbone to obtain two feature maps $v_1^K, v_2^D \in \mathbb{R}^{C\times H\times W}$, C, H, W representing the number of channels, height, and width of the feature maps respectively. Two feature maps are tiled as N patches, using linear projections to obtain the embedded patches $P_x = (p_x^1, \ldots, p_x^N) \in \mathbb{R}^{D_1\times N}$ and $P_y = (p_y^1, \ldots, p_y^N) \in \mathbb{R}^{D_1\times N}$, respectively, where D_1 is the length of each embedded patch. We define three matrices of learnable parameters, W_x^Q, W_y^K, and W_y^V, and finally, the patch embedding is projected to the weight matrix, and V_{1out}^K, V_{2out}^D are calculated by cross-view attention as shown in (11),

$$Q_x = \left(P_x^T\right) W_x^Q, \quad K_y = \left(P_y^T\right) W_y^K, \quad V_y = \left(P_y^T\right) W_y^v,$$
$$V_{1out}^K = U\left(\text{Softmax}\left(\frac{(Q_x K_y^T}{\sqrt{D_2}}\right) V_y\right), \tag{11}$$

where $U(\cdot)$ reverts the embedded patches to the original patches to better aggregate the two view context information. The fused features Z with inter and intra-view information are finally obtained through the local feature extraction block $L(X)$ and the global feature extraction block $G(X)$ in (12).

$$L(X) = DSC_2(GELU(Bn(DSC_1(X))))$$
$$G(X) = DSC_2(GELU(Bn(DSC_1(GAP(X)))))$$
$$Z = (V_{combine}) \otimes (L(V_{combine})) \oplus G(V_{combine}))) \tag{12}$$

where $GELU$ is the activation function, Bn stands for batch normalization, DSC is depthwise separable convolution, GAP is global average pooling, and $V_{combine} = (V_{1out}^K \oplus V_{2out}^D)$.

Fig. 2(c)

3) Cross-Scale Transformer (CST): Traditional single-scale networks are usually designed to handle features at a specific scale [31]. They cannot effectively capture dynamic changes in the data and cross-scale information in the features, leading to the model's incomplete understanding of the overall data. The CST is introduced to facilitate the model to learn generalized features more easily by interacting information at different scales rather than just relying on specific scales. This cross-scale information extraction helps the model focus on both global and local information, which is expected to improve the robustness and generalization ability.

To further capture the dependencies among features, we extract multi-scale information by dividing the fused features into small and large patches, as shown in Fig. 2(c). The small-size patch can capture the details and local information, while the large-size patch can capture the global and integral information. It can boost the model's ability to perceive features at different scales, making it more adaptive and generalizable. Specifically, the feature Z is split into a large size patch Z_{patch}^L and a small size patch Z_{patch}^S with embedded position information (Z_{pos}^S, Z_{pos}^L), as shown in (13):

$$Z_S' = \left[Z_{class}^S; Z_{patch}^S\right] + Z_{pos}^S$$
$$Z_L' = \left[Z_{class}^L; Z_{patch}^L\right] + Z_{pos}^L \tag{13}$$

where Z_{class}^L is the class token for large patches and Z_{class}^S is the class token for small patches. The $(Z_{class}^L, Z_{class}^S)$ act as proxies in charge of exchanging pertinent information about patch tokens of diverse scales. Finally, Z_S' and Z_L' are processed through a stack of CST blocks to derive two class tokens, i.e., Y_{class}^S and Y_{class}^L from the two branches. They are concatenated in the multi-layer perceptron (MLP) and put into the Regressor to obtain the ultimate BG measurement result $\hat{Y}_{BG}$.

III. Experimental Setup

A. Data Acquisition

In this study, we recruited 260 subjects (126 males and 134 females), including 171 healthy ones (BG range: 4.0–6.0 mmol/L) and 89 diabetes-related subjects. Of these, 40 were pre-diabetic subjects (BG range: 6.1–6.9 mmol/L) and 49 were diabetic subjects (BG range: 7.0–12.0 mmol/L). More details are given in Table I. All data were obtained from the Ninth People's Hospital of Chongqing. The hospital's Ethics Committee approved the study (Ethics Approval No. 2022-SCI-007), and participants were asked to sign an informed consent form before the data collection. To improve the reliability of the acquired PPG signals and BG values, a strict acquisition paradigm is utilized to minimize signal interference from diet, respiration, and arterial factors [11]. The experiment was conducted under uniformly restrictive conditions, requiring subjects to avoid glucose-lowering drugs (including insulin) the day before the experiment and to fast for at least 8 hours before collection. It ensures that subjects' PPG signals and BG levels are obtained under relatively consistent physiological conditions, which would help uncover the most authentic relationship between individual PPG signals and BG levels. In contrast, random BG measurements would be influenced by recent meals, resulting in fluctuating BG levels and making it difficult to obtain the actual baseline condition of the patient. In addition, FPG is often used to assess an individual's risk of developing diabetes so that its measurement can be considered an essential reference for individual health management [22].

Before the signal acquisition, subjects were required to rest in the most comfortable sitting position for over 5 minutes to accommodate the acquisition environment and familiarize themselves with the experimental procedure. Data collection phase. Firstly, the nurse collected fasting venous blood from the subject once to obtain FPG values (the gold standard for BG [22]). Then, we continuously collected PPG signals from the subject's fingertip for about 3 minutes using the HKG-07 C sensor (200 Hz sampling rate). Data preprocessing phase. The entire PPG signal of each subject was segmented into sub-signals of fixed frame lengths. Each sub-signal corresponds to the same FPG value for each subject, as the FPG does not change over a short period [32].

B. Experimental Details

Our experiments are based on PyTorch and Python 3.9, trained and validated on two RTX 3090 graphics cards and 64 GB DRAM. This study uses a stratified 5-fold CV method based on

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT FILTERING ALGORITHMS

Filtering Method	RMSE (MEAN ± SD)	MAE (MEAN ± SD)	R^2 (MEAN ± SD)
Butterworth	0.1006 ± 0.0320	0.0797 ± 0.0276	0.8234 ± 0.1559
Wavelet	0.0651 ± 0.0872	0.0517 ± 0.0205	0.9216 ± 0.0872
MSWF(Triweight)	0.0443 ± 0.0261	0.0363 ± 0.0225	0.9540 ± 0.0727
MSWF(Tricube)	0.0421 ± 0.0260	0.0346 ± 0.0222	0.9580 ± 0.0705
MSWF(Gaussian)	0.0391 ± 0.0256	0.0321 ± 0.0219	0.9621 ± 0.0681

$(\cdot)$ represents the different weighting functions adopted in the MSWF algorithm.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT INPUT SIGNAL LENGTHS

Signal length	RMSE (mmol/L)	MAE (mmol/L)	ACC (%)	Memory usage (M)	Inference time with/without GPU (ms)	Params (M)
Single cycle	1.237	0.7403	78.81	594.34	≈ 37/334	11.29
2(s)	1.274	0.7546	80.42	603.61	≈ 35/349	11.29
3(s)	1.232	0.7317	81.78	609.73	≈ 37/369	11.29
4(s)	1.214	0.7135	82.46	616.46	≈ 36/362	11.29
5(s)	1.190	0.7027	81.84	626.19	≈ 36/386	11.29
6(s)	**1.129**	**0.6593**	**83.75**	632.65	≈ 37/386	11.29
7(s)	1.264	0.7412	81.78	642.09	≈ 38/379	11.29
8(s)	1.306	0.7661	80.75	648.90	≈ 37/417	11.29
9(s)	1.354	0.8491	76.44	658.03	≈ 39/417	11.29
10(s)	1.315	0.7768	78.98	664.77	≈ 41/423	11.29

RMSE: Root mean square error, MAE: Mean absolute error, ACC: Accuracy, Params: Number of parameters. Inference time: inference time for per sample with and without the GPU device.

subject numbering, i.e., subject-wise CV, for all experiments. It ensures that the division of the training and testing datasets is based on the subject number, i.e., each subject's data is wholly included in the training or testing set. This division method aims to obtain fairer and more reliable experimental results. All experiments use the same hyperparameter settings with an initial learning rate (0.01), batch size (32), number of training iterations (100), and optimizer (stochastic gradient descent, SGD). The learning rate is periodically adjusted using a cosine annealing strategy, and the model weights are initialized before the start of each fold. Notably, to minimize the impact of data imbalance on the model performance, we combine two strategies, stratified 5-fold CV and weighted random sampling, to ensure that the model pays sufficient attention to samples with different BG ranges during the training process.

C. Evaluation Criteria

In this study, we used RMSE, mean absolute error (MAE), and R-square (R^2) metrics to evaluate the effectiveness of the proposed filtering algorithm. Meanwhile, to evaluate the performance of the proposed model for BG measurements, we use the primary evaluation metrics, including RMSE, MAE, and MARD. In addition, compared with other studies [5], [8], [9], [11], [13], [14], [15], [16], [17], we also use the latest ISO 15197:2013 standards to evaluate the accuracy of BG measurements further comprehensively. CEG [33] and SEG [34] are also applied to provide clinical insights for the proposed BG measurement model. These metrics can facilitate the assessment of the degree of deviation between the measured BG and the reference BG to reveal the clinical implications and potential risks associated with BG measurement.

IV. RESULTS AND DISCUSSION

A. Optimization of Filtering Algorithms

We evaluated alternative filtering algorithms to minimize signal distortion and interference and ensure that the filtered signal has the most authentic correlation with the corresponding BG level. We used RMSE, MAE, and R^2 as measures to evaluate the effectiveness of different filtering methods. As shown in Table II, the proposed MSWF algorithm has the lowest RMSE and MAE and the highest R^2 value. The MTWF employs sliding windows of different sizes to locally smooth the time-domain signals, thereby achieving effective noise elimination while retaining the overall signal morphology. However, filtering methods in the frequency domain (e.g., Wavelet transforms and Butterworth filters) may cause signal distortion due to the problem of spectral overlap among the noise and the signal and boundary effect problems arising from incompleteness at the boundaries of the subframe signals. These problems exist in [7], [8], [9].

Secondly, in the baseline removal stage (Stage 2), we found that the Gaussian weighting function outperforms the Triweight and Tricube functions. MSWF(Gaussian) is smoother in weight decay and matches continuous variations in the baseline better, making the fitting process smoother and reducing the effect of noise.

B. Signal Frame Lengths

We segmented the PPG signal (single-cycle frame or multiple cycles of 2 to 10 seconds in frame length) to investigate the effect of various types of frame length on BG measurement. Meanwhile, we also test the memory usage, inference time, and the number of parameters when the signals are put with different frame lengths into the MvCFT network. The KFR algorithm is utilized to obtain the kinetic features of the partitioned signals. The kinetic features and the signal derivatives are input simultaneously into both branches of the MvCFT network, ultimately obtaining the BG measurements. The experimental results are shown in Table III, and the best balance between performance and efficiency is reached at a frame length of 6 s. The poor performance of the single-cycle segmentation-based method may be attributed to the fact that interpolation is required to align the single-cycle signals of different subjects, which may alter the characteristics of the signal. In addition, the cycle-based segmentation signal could not reflect essential information such as HR and HRV, which are features intimately correlated with BG levels [35].

It can be noticed in Table III that as the signal frame length increases, the performance improves because longer signals contain more internal latent features and thus better indicate BG levels. However, when the signal length exceeded 6 s, the performance of BG measurement decreased significantly. Performance decrease might be caused by the loss of fine-grained features in the longer signal frame when the PPG signal is embedded in low dimensions using PAA. Additionally, longer signals not only

TABLE IV
THE SUBJECT-WISE 5-FOLD CROSS VALIDATION PERFORMANCE OF THE PROPOSED METHOD

Fold	RMSE (mmol/L)	MAE (mmol/L)	ACC (%)	Zone A (%)	Zone B (%)	Zone A+B (%)
1	1.305	0.782	76.36	82.61	17.39	100
2	1.251	0.645	82.16	87.11	12.89	100
3	1.146	0.657	85.29	87.76	12.24	100
4	0.751	0.545	88.67	92.97	7.03	100
5	1.193	0.668	86.28	88.99	11.01	100
Average	1.129	0.659	83.75	87.89	12.11	100

Zone A: No-risk clinical risk, Zone B: Slight clinical risk

TABLE V
RESULTS OF PERCENTAGE OF DIFFERENT ISO RANGES ON THE 4TH FOLD

ISO range	Number of points	Percentage (%)
$\leq$ 5% or 0.3 mmol/L	408	53.13
$>$ 5-10% or 0.3-0.6 mmol/L	204	26.56
$>$ 10-15% or 0.6-0.8 mmol/L	69	8.98
$>$ 15-20% or 0.8-1.1 mmol/L	40	5.21
$>$ 20% or 1.1 mmol/L	47	6.12

ISO range = difference between BGM and REF as percent of REF for REF $>$ 5.55 mmol/L and in mmol/L for REF $\leq$ 5.55 mmol/L.

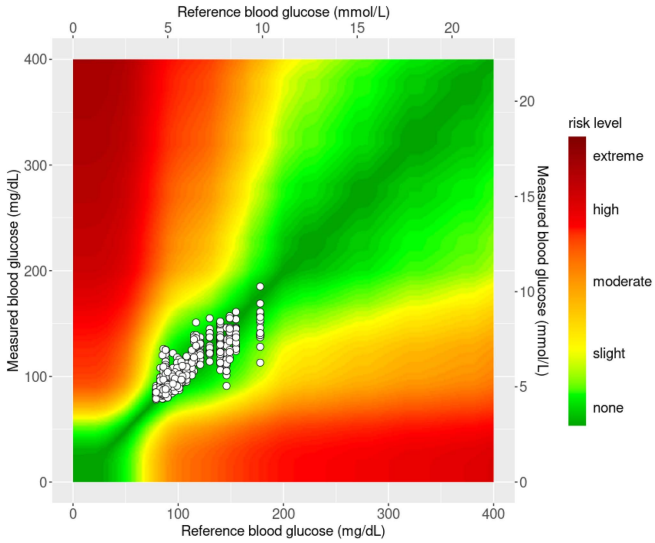

Fig. 3. Assessment of the clinical risk level; the color-coded SEG; the white circles indicate the measured BG and the corresponding reference BG; risk levels are categorized into five categories: none, slight, moderate, high, and extreme.

increase the computational complexity of the model but may also introduce more noise and interference, causing performance degradation. As shown in Table III, memory usage, inference time, and number of parameters vary little at different frame lengths. It only takes about 0.4(s) to complete a test without a GPU, which shows the potential of our approach for home monitoring. Finally, to balance performance and efficiency, we segment the PPG signals into 6(s) frames for processing in all subsequent experiments.

C. Overall Performance Evaluation

Record-wise CV could obscure inter-subject differences and lead to overly optimistic results. We consider the individual discrepancies and variations in physiological features among subjects. We tested a subjects-wise 5-fold CV to more comprehensively assess the adaptability and robustness of the algorithm to different individuals. Table IV shows the detailed results, including the performance metrics for each fold. The critical finding is that all measurements at each fold fall within the zone of none or slight clinical risk in the CEG (Zone A + B = 100%). The best performance is obtained at the fourth fold with RMSE, MAE, and ACC (ISO 15197:2013), obtaining 0.751 mmol/L, 0.545 mmol/L, and 88.67%, respectively (The test set in the fourth fold contained 52 subjects with a total of 768 PPG segments). Notably, 92.97% of the measurements fall into Zone A in the CEG, suggesting that most BG measurements are within the clinically acceptable range.

In addition, the percentage of estimates that fall in different error ranges were tallied based on ISO criteria, and 88.67% of the estimates satisfy this criterion, as shown in Table V. In order to visualize the measurement and assess its clinical risk level, a modern error grid SEG was used to visualize the correspondence between the reference and measured values. In Fig. 3, it can be found that the measurements are all in the clinical no-risk and slight-risk regions, which strongly proves the potential of the proposed method for home care applications.

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT INPUTS

Inputs	Methods	RMSE (mmol/L)	MAE (mmol/L)	ACC (%)
PPG	single-channel	1.280	0.770	76.901
(PPG, VPG, APG)	multi-channel	1.255	0.742	80.422
(GASF, GADF, RP)	single-view	1.346	0.799	77.597
(SPE, GASF, GADF)	single-view	1.344	0.795	78.834
(SPE, GADF, RP)	single-view	1.273	0.747	81.936
(SPE, GASF, RP)	single-view	1.264	0.746	81.992
PVA and SGR	multi-view	**1.129**	**0.659**	**83.751**

PVA and SGR represents (**P**PG, **V**PG, **A**PG) = $C_{ij}^{t=6}$ and (**S**PE, **G**ASF, **R**P) = $K_{ij}^{t=6}$ as the two input views in the MvCFT network, respectively.

D. Ablation Study on MvCFT Network

To verify the necessity of different components in the proposed network. Some ablation experiments are performed to evaluate the effect of each component on the overall performance.

1) Performance Comparison of Different Inputs: Table VI demonstrates the effect of different inputs on the performance of BG measurements. Firstly, the performance of single-channel PPG signals is relatively poor due to the high similarity of the PPG signals in the time domain, which makes it challenging to capture discriminative features. The performance is improved after combining PPG, VPG, and APG into a multi-channel signal as input. Since changes in the cardiovascular system may be related to BG levels [36], VPG and APG are introduced to provide information on changes in the velocity and acceleration of the heartbeat. It enhances the model's understanding of physiological states and improves the accuracy of BG level estimation.

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT BACKBONES

Backbone	RMSE (mmol/L)	MAE (mmol/L)	ACC (%)
Inceptime + EfficientNetB0	1.399	0.832	75.339
Inceptime + CMT	1.310	0.763	79.228
ResNet-1D + EfficientNetB0	1.278	0.748	81.378
ResNet-1D + CMT	1.129	0.659	83.751

TABLE VIII
PERFORMANCE COMPARISON OF DIFFERENT COMPONENTS

CVFF	CST			Architecture	RMSE (mmol/L)	MAE (mmol/L)	ACC (%)
	S	L	K				
✗	✗	✗	0	single-scale	1.354	0.810	78.765
✓	✗	✗	0	single-scale	1.367	0.783	80.429
✓	4×4	20×20	3	cross-scale	1.288	0.758	79.341
✓	5×5	16×16	3	cross-scale	1.351	0.802	79.371
✓	8×8	10×10	1	cross-scale	1.287	0.773	81.143
✓	8×8	10×10	2	cross-scale	1.194	0.701	82.861
✓	8×8	10×10	3	cross-scale	1.129	0.659	83.751

S stands for small patch size, L stands for large patch size, and K is the number of layers of stacked CST blocks.

TABLE IX
PERFORMANCE COMPARISON OF DIFFERENT FUSION METHODS

Fusion module	RMSE (mmol/L)	MAE (mmol/L)	ACC (%)
CONCAT [30]	1.280	0.770	78.933
MS-CAM [40]	1.324	0.788	79.217
AFF [40]	1.318	0.782	79.721
CVFF	**1.129**	**0.659**	**83.751**

CONCAT: Concatenate.

In addition, four different combinations of kinetic features were obtained by adjusting the Type parameter in the KFR algorithm. The poor performance when using (GASF, GADF, RP) as a single-view input may be due to the redundancy of the features produced by GASF and GADF, which increases the difficulty of the model in finding distinguishing features. However, better performance is realized when both SPE and RP are used because SPE provides information about the relative position of the signal in time and space. At the same time, GASF and RP focus more on the signal's nonlinear kinetic features, periodicity, and repetitiveness [21]. Their combination provides additional positional cues for the model to extract kinetic features. It helps better capture the nonlinear relationship between the kinetic features of the PPG signal and BG levels. Rows 3–6 in Table VI, using only kinetic features as single-view inputs, under-explore the potential features of the PPG signal and ignore the importance of the original time-domain signal. Thus, the performance improvement is also limited. Ultimately, the best performance was obtained by simultaneously inputting multi-channel signals and combined kinetic features into the multi-view network MvCFT. This further validates that the performance of BG measurements can be effectively improved by jointing the potential morphological and time-frequency features in the multi-channel signals and the kinetic features of the PPG signal.

2) Performance Comparison of Different Backbones: To validate the adaptability of the proposed network, we performed an experimental comparison with different backbones as feature encoders, as shown in Table VII. One of the branches in the multi-view network will input multi-channel signals. Thus, ResNet-1D [37] and Inceptiontime [38] are considered alternative feature encoders. Likewise, the kinetic features are input in the other branch, and CMT [31] and EfficientNet [39] are considered alternative feature encoders. After four different combinations, we can see that ResNet-1D outperforms Inceptiontime for the time series feature encoder, and CMT exceeds EfficientNet for the 3D feature matrix feature encoder. Ultimately, the feature encoder of the combination of ResNet-1D and CMT is chosen as the feature extractor for MvCFT.

3) Performance Comparison of Different Components: To verify the practicality of different components in MvCFT, a series of experiments were designed to prove the dedication of each component to the overall performance, as shown in Table VIII. The first row of Table VIII shows the performance of the baseline model. It is observed that adding CVFF to the baseline model improves the performance of BG measurement. CVFF effectively enhances the learning capability of the model by fusing high-dimensional representations of $C_{ij}^{t=6} \in R^{3\times1\times1200}$ (PPG and its derivatives) and $K_{ij}^{t=6} \in R^{3\times320\times320}$ (PPG kinetic features), enabling it to derive complementary information from multi-views. Notably, after adding CVFF, the model remains a single-scale architecture without interactive learning of cross-scale features. After adding CST, the model performance is further improved compared to the baseline model. This is attributed to the fact that CST introduces scale transformations and feature interactions, which are more flexible in modeling complex relationships between cross-scale features.

Interestingly, there is a discrepancy in the performance improvement using different patch size scales. This suggests the importance of choosing the right patch size in CST for mapping the complex nonlinear relationship with the BG values. Consequently, we evaluate the impact of different patch sizes. The patches $(S = 4 \times 4, L = 20 \times 20)$ and $(S = 5 \times 5, L = 16 \times 16)$ subjectively deserve better performance as finer-grained features are provided. However, $(S = 8 \times 8, L = 10 \times 10)$ obtained a better performance, for which we consider that due to the excessive disparity in feature granularity between different branches, it is hard to learn features properly. Besides, experiments with stacking different numbers of transformer blocks were also conducted. We set $K = 3$ as the optimal variable considering the model processing speed.

4) Results of Different Fusion Modules: To appraise the validity of the CVFF module, we compared CVFF with the currently prevalent feature fusion methods [30], [40]. As shown in Table IX, the most straightforward feature concatenation approach performs the worst, producing feature redundancy and disregarding feature interactions between different views. MS-CAM and AFF, feature fusion based on the attention mechanism, improved the feature representation somewhat and slightly increased the measurement accuracy. However, learning natural pairwise relationships among different views is still challenging.

TABLE X
COMPARISON RESULTS OF DIFFERENT METHODS ON CLINICAL DATASETS

References	Feature Extraction	Regressor	Model architecture	RMSE (mmol/L)	MAE (mmol/L)	ACC (%)	Zone A (%)	Inference time (ms)	Params (M)
Hina *et al.* [15]	Hand-crafted (6)	SVR	single-view	1.506	0.936	69.79	76.43	≈ 1	——
Nie *et al.* [8]	Hand-crafted (12)	RFR	single-view	1.426	0.887	71.74	76.30	≈ 1	——
Zhang *et al.* [6]	Hand-crafted (28)	GSVR	single-view	1.413	0.803	73.64	79.76	≈ 1	——
Lee *et al.* [19]	End-to-end	FC	single-view	1.374	0.784	75.90	82.68	≈ 7	0.83
Li *et al.* [36]	End-to-end	MLP	single-view	1.346	0.770	76.69	82.94	≈ 28	1.77
Zhang *et al.* [18]	End-to-end	MLP	single-view	1.331	0.761	78.52	82.81	≈ 677	36.17
Ours	End-to-end	MLP	**multi-view**	**1.129**	**0.659**	**83.75**	**87.89**	≈ 386	11.29

Hand-crafted (6) represents manually extracted 6 features, SVR: support vector regression, RFR: random forest regression, GSVR: Gaussian support vector regression, FC: fully connected layer, MLP: multi-layer perception. Inference time is the measurement time per sample without the GPU. Params: Number of parameters.

The CVFF module exploits the complementarity and consistency among views in a shared feature space through feature mapping and all-around fusion across views. Compared with the other three methods, the CVFF module performs best. The results show that the CVFF module outperforms the other three methods in multi-view feature fusion.

E. Comparison With Other BG Measurement Methods

To fairly compare the performance of different solutions on the same dataset, we re-implement the traditional machine learning methods [6], [8], [15] and the end-to-end deep learning methods [18], [19], [36]. As indicated in Table X, our method achieved more promising results. These studies [6], [8], [15] extracted a small number of hand-crafted features (e.g., morphological features, time-frequency features) to achieve BG measurements through different regressors. However, the performance of these methods is not satisfactory. We observe that relying on only a few morphological and time-frequency features makes capturing more distinguishing features on more subject datasets challenging. In addition, these methods depend highly on professional expertise for the number and category of manual features, making it difficult to generalize them to other domain applications.

In end-to-end deep learning studies, we observe that [19], [36] employ shallow CNN networks with fewer parameters and shorter inference time. However, the performance metrics show that shallow CNN networks are limited in mining deep discriminative features, making them less effective. Our previous work [18] achieved promising results by converting the signal into time-frequency maps for glucose measurement using a DNN model. However, the model architecture is single-view, i.e., only the time-frequency map serves as input, which may lose critical features of the time-domain PPG signal. Notably, the inference times of the above studies are at the millisecond level and do not cause significant delays in real-world scenarios.

V. CONCLUSION

Unobtrusive BG measurement can significantly improve the quality of healthcare and patient outcomes. This paper presents a new paradigm for non-invasive BG measurement using PPG signals. The reliability of the input signal is ensured by employing the MSWF algorithm, which avoids boundary effects and spectral overlap problems that may be introduced during filtering. The proposed MvCFT network deeply fuses complementary information among views through the CVFF module. It is designed to capture reconstructed kinetic features and derivatives for potential correlation with BG levels. The results show that our method performs better than the existing works and provides new insights into non-invasive BG measurements. In addition, our proposed KFR algorithm and MvCFT framework can also be applied to other clinical research fields involving physiological signals.

This study performed the training and inference processes on the server. The system's applicability under different conditions (e.g., skin color, environment, etc.) has yet to be considered. Since we focus on fasting BG, subjects are required to perform measurements at rest, which may limit the ability to observe changes in dynamic BG under non-standard conditions. Therefore, the current approach does not apply to ambulatory BG measurement. In future work, we will continue to expand the testing with hyperglycemic samples and conduct external validation studies using datasets from different hospitals. Meanwhile, we will explore ambulatory BG monitoring protocols to build a robust and accurate BG measurement model for diverse real-life BG monitoring scenarios.

REFERENCES

[1] J. M. O'Connell and S. M. Manson, "Understanding the economic costs of diabetes and prediabetes and what we may learn about reducing the health and economic burden of these conditions," *Diabetes Care*, vol. 42, no. 9, pp. 1609–1611, 2019.
[2] M. Franciosi et al., "The impact of blood glucose self-monitoring on metabolic control and quality of life in type 2 diabetic patients: An urgent need for better educational strategies," *Diabetes Care*, vol. 24, no. 11, pp. 1870–1877, 2001.
[3] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, pp. R1–R39, 2007.
[4] W. B. Baker, A. B. Parthasarathy, D. R. Busch, R. C. Mesquita, J. H. Greenberg, and A. Yodh, "Modified beer-Lambert law for blood flow," *Biomed. Opt. Exp.*, vol. 5, no. 11, pp. 4053–4075, 2014.
[5] E. Monte-Moreno, "Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques," *Artif. Intell. Med.*, vol. 53, no. 2, pp. 127–138, 2011.
[6] G. Zhang et al., "A noninvasive blood glucose monitoring system based on smartphone PPG signal processing and machine learning," *IEEE Trans. Ind. Informat.*, vol. 16, no. 11, pp. 7209–7218, Nov. 2020.
[7] B.-J. Wu, B.-F. Wu, and C.-P. Hsu, "Camera-based blood pressure estimation via windkessel model and waveform features," *IEEE Trans. Instrum. Meas.*, vol. 72, 2022, Art. no. 5004113.

[8] Z. Nie, M. Rong, and K. Li, "Blood glucose prediction based on imaging-photoplethysmography in combination with machine learning," *Biomed. Signal Process. Control*, vol. 79, 2023, Art. no. 104179.

[9] W. Liu, G. Wang, A. Huang, and P. Wang, "NIV-NGM-a novel non-invasive blood glucose monitoring method based on near-infrared videos," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 953–957.

[10] F. Mohagheghian et al., "Optimized signal quality assessment for photoplethysmogram signals using feature selection," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 9, pp. 2982–2993, Sep. 2022.

[11] J. Li et al., "Noninvasive blood glucose monitoring using spatiotemporal ECG and PPG feature fusion and weight-based choquet integral multi-model approach," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2023.

[12] W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh, "Cuff-less blood pressure estimation from photoplethysmography via visibility graph and transfer learning," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2075–2085, May 2022.

[13] Y. Wei, J. Liu, L. Hu, B. W.-K. Ling, and Q. Liu, "Clark error grid based stacking method for fusing various time frequency averaged features for reducing measurement errors for non-invasive blood glucose estimation," *IEEE Trans. Consum. Electron.*, vol. 69, no. 3, pp. 510–521, Aug. 2023.

[14] G. A. Alonso-Silverio, V. Francisco-García, I. P. Guzmán-Guzmán, E. Ventura-Molina, and A. Alarcón-Paredes, "Toward non-invasive estimation of blood glucose concentration: A comparative performance," *Mathematics*, vol. 9, no. 20, 2021, Art. no. 2529.

[15] A. Hina and W. Saadeh, "A noninvasive glucose monitoring SoC based on single wavelength photoplethysmography," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 3, pp. 504–515, Jun. 2020.

[16] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Intelligent estimation of blood glucose level using wristband PPG signal and physiological parameters," *Biomed. Signal Process. Control*, vol. 78, 2022, Art. no. 103876.

[17] G. Reddy, K. K. Bhat, U. Lunia, and N. Krupa, "A novel deep learning approach for non-invasive blood glucose measurement from photoplethysmography signals," in *Proc. Adv. Commun. Devices Netw.*, 2022, pp. 377–386.

[18] C. Zhang et al., "Video based cocktail causal container for blood pressure classification and blood glucose prediction," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 2, pp. 1118–1128, Feb. 2023.

[19] E. Lee and C.-Y. Lee, "PPG-based smart wearable device with energy-efficient computing for mobile health-care applications," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13564–13573, Jun. 2021.

[20] N. Jendrike, A. Baumstark, U. Kamecke, C. Haug, and G. Freckmann, "ISO 15197: 2013 evaluation of a blood glucose monitoring system's measurement accuracy," *J. Diabetes Sci. Technol.*, vol. 11, no. 6, pp. 1275–1276, 2017.

[21] C. Ouyang, Z. Gan, J. Zhen, Y. Guan, X. Zhu, and P. Zhou, "Inter-patient classification with encoded peripheral pulse series and multi-task fusion CNN: Application in type 2 diabetes," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 3130–3140, Aug. 2021.

[22] B. Olabi and R. Bhopal, "Diagnosis of diabetes using the oral glucose tolerance test," *BMJ*, vol. 339, p. b4354, 2009.

[23] M. Elgendi, "Optimal signal quality index for photoplethysmogram signals," *Bioengineering*, vol. 3, no. 4, 2016, Art. no. 21.

[24] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, Feb. 2014.

[25] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3939–3945.

[26] A. Goshvarpour and A. Goshvarpour, "Poincaré's section analysis for PPG-based automatic emotion recognition," *Chaos Solitons Fractals*, vol. 114, pp. 400–407, 2018.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[28] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Phys. Rep.*, vol. 438, no. 5/6, pp. 237–329, 2007.

[29] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021.

[30] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[31] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12175–12185.

[32] M. S. Boyne, D. M. Silver, J. Kaplan, and C. D. Saudek, "Timing of changes in interstitial and venous blood glucose measured with a continuous subcutaneous glucose sensor," *Diabetes*, vol. 52, no. 11, pp. 2790–2794, 2003.

[33] W. L. Clarke, D. Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl, "Evaluating clinical accuracy of systems for self-monitoring of blood glucose," *Diabetes Care*, vol. 10, no. 5, pp. 622–628, 1987.

[34] D. C. Klonoff et al., "The surveillance error grid," *J. Diabetes Sci. Technol.*, vol. 8, no. 4, pp. 658–672, 2014.

[35] M. O. Bekkink, M. Koeneman, B. E. d. Galan, and S. J. Bredie, "Early detection of hypoglycemia in type 1 diabetes using heart rate variability measured by a wearable device," *Diabetes Care*, vol. 42, no. 4, pp. 689–692, 2019.

[36] J. Li, I. Tobore, Y. Liu, A. Kandwal, L. Wang, and Z. Nie, "Non-invasive monitoring of three glucose ranges based on ECG by using DBSCAN-CNN," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3340–3350, Sep. 2021.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[38] H. Ismail Fawaz et al., "InceptionTime: Finding alexnet for time series classification," *Data Mining Knowl. Discov.*, vol. 34, no. 6, pp. 1936–1962, 2020.

[39] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[40] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3560–3569.