TYrPPG: Uncomplicated and Enhanced Learning Capability rPPG for Remote Heart Rate Estimation

Taixi Chen[1,2] Yiu-ming Cheung[2†]
[1]School of Computing, Binghamton University, Binghamton, NY, USA
[2]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
tchen51@binghamton.edu; ymc@comp.hkbu.edu.hk

Abstract—**Remote photoplethysmography (rPPG) can remotely extract physiological signals from RGB video, which has many advantages in detecting heart rate, such as low cost and no invasion to patients. The existing rPPG model is usually based on the transformer module, which has low computation efficiency. Recently, the Mamba model has garnered increasing attention due to its efficient performance in natural language processing tasks, demonstrating potential as a substitute for transformer-based algorithms. However, the Mambaout model and its variants prove that the SSM module, which is the core component of the Mamba model, is unnecessary for the vision task. Therefore, we hope to prove the feasibility of using the Mambaout-based module to remotely learn the heart rate. Specifically, we propose a novel rPPG algorithm called uncomplicated and enhanced learning capability rPPG (TYrPPG). This paper introduces an innovative gated video understanding block (GVB) designed for efficient analysis of RGB videos. Based on the Mambaout structure, this block integrates 2D-CNN and 3D-CNN to enhance video understanding for analysis. In addition, we propose a comprehensive supervised loss function (CSL) to improve the model's learning capability, along with its weakly supervised variants. The experiments show that our TYrPPG can achieve state-of-the-art performance in commonly used datasets, indicating its prospects and superiority in remote heart rate estimation. The source code is available at https://github.com/Taixi-CHEN/TYrPPG.**

I. Introduction

Remote heart rate detection has been extensively studied due to its ability to eliminate concerns related to physical contact between doctors and patients [1], [2]. This aspect has become increasingly important in the post-COVID-19 era. Besides, it can be leveraged in affective computing and deepfake detection [3], showing its wide usage domain. In particular, remote photoplethysmography (rPPG) is a typical method for remotely estimating the heart rate. It extracts the Blood Volume Pulse (BVP) to isolate the heart rate by capturing facial color changes resulting from periodic blood circulation. Traditional rPPG algorithms have made valuable contributions to the healthcare aspect. However, most conventional research and tests are conducted in optimal environments, leading to low accuracy and a lack of robust extraction methods under complex conditions. To address these limitations, both unsupervised and supervised learning approaches have been proposed to mitigate the adverse effects of the environment and head motion.

In unsupervised approaches, the absence of well-rounded pre-assumptions significantly restricts their performance. [4]. Moreover, some unsupervised domain adaptation methods [5], [6] aim to increase model generalization ability but may create distorted images without the target domain label. The supervised approaches are commonly based on deep learning, which needs a large amount of video data for training and myriad times for model convergence as well. Thus, they suffer expensive costs in video analysis and denoise [7]. These complicated structures restrict the possibility of deploying the model in portable devices for widespread use. DeepPhys [8] was proposed to leverage the normalized light difference to alleviate the negative impacts of various illumination cases. Besides, numerous experiments proved that the attention mechanism can address head motion issues [9].

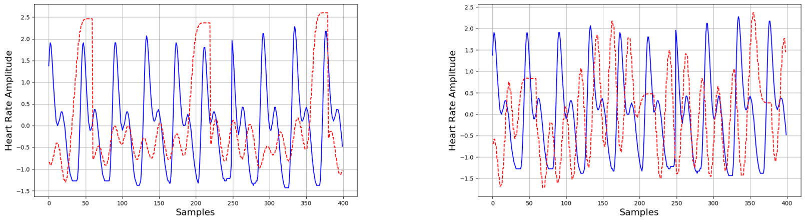
Therefore, existing methods have at least two drawbacks as follows: 1) Using a complicated module in this medical task, which makes them difficult to deploy in a portable device, and 2) Optimizing their model solely based on the unsupervised loss, such as contrastive learning, neglecting to leverage comprehensive information from ground truth. Recently, the Mamba model [10] has demonstrated its better capabilities in natural language tasks and has also proven effective in vision tasks. However, Mambaout [11] has shown that its core component, the SSM module, can be removed for vision tasks, reducing computational and resource costs.

Thus, our paper aims to propose an effective method called: Uncomplica**T**ed and enhanced learning capabilit**Y rPPG** algorithm (**TYrPPG**) that solves the remote heart rate estimation task, based on Mambaout structure. Due to the rPPG signal being periodic, rough physiological signals are not difficult to extract in most cases. Thus, the motivation of TYrPPG is to prove whether the model can estimate the heart rate with a simpler network structure and discover a path to enhance the model's learning ability. Specifically, due to the success of the TSM [12] and Mambaout structure [11], we create a novel gated video understanding block (GVB) based on them. With the GVB, our model can more effectively analyze facial videos and estimate heart rate using a simpler neural network. To the best of our knowledge, this is the first rPPG algorithm utilizing the Mambaout-based structure, distinguishing it from the recently popular transformer-based and Mamba-based models. Figure 2 shows the model structure.

Moreover, the TYrPPG framework proposes a novel Com-

[†]Corresponding author

This work is supported by RGC Senior Research Fellow Scheme under grant: SRFS2324-2S02.



(a) TYrPPG-KL

(b) TYrPPG-video MMD

Fig. 1. Visualization of the heart rate signals estimated by TYrPPG based on KL divergence and proposed video MMD. Straight and dotted lines mark the signals as ground truth and model estimation, respectively. TYrPPG can learn the ground truth better optimized by video MMD, as its estimation signal peaks are more consistent with the ground truth. This shows the effectiveness of the proposed video MMD.

prehensive Supervised Loss (CSL). We hope to improve the ability to learn ground truth distributions by using the proposed CSL. Unlike existing algorithms that utilize KL divergence, we are the first to introduce Maximum Mean Discrepancy (MMD) [13] as a component of the proposed Loss function. To avoid over-fitting and remove temporal redundancy, we propose a Video-MMD, which makes the model better learn the discrepancy between their estimation signals and ground truth from video, inspired by [14]. Due to the complexity of the rPPG signals, models often predict many incorrect values. When using KL divergence, these discrepancies can be excessively exaggerated, leading to infinite results, which is not desirable. In contrast, MMD can mitigate this issue effectively. Figure 1 also proves that its performance surpasses the commonly used KL divergence. In addition, we construct a Weak Supervised Loss (WSL) to utilize only limited distribution information for optimization. This logic is shown in Figure 3. Therefore, this paper makes contributions as below:

- Proposing an efficient gated video understanding block (GVB) to robustly understand the rPPG video and estimate the rPPG signal, which comprises the TSM and modified Mambaout module. It is a combination of 2D-CNN and 3D-CNN without complex structures.
- Based on the proposed video-MMD, designing a novel Comprehensive Supervised Loss (CSL) and its variant Weak Supervised Loss (WSL) to utilize the important information from ground truth distribution, leading to fast and accurate convergence.
- Experiments show TYrPPG can successfully remotely estimate heart rate and obtain a good generalization ability, achieving state-of-the-art performance in the PURE and MMPD datasets, which proves the superiority and promising future of our model.

II. Related Works

The conventional rPPG algorithms have proved the feasibility of remotely detecting heart rate using a web camera through RGB channels [1], [2]. It is also known that the green channel should be more suitable for estimating the rPPG signal [2].

A. Supervised rPPG Algorithm

Conversely, the deep learning-based methods show their power in complicated conditions and domain generalization ability. As the previous section mentioned, the unsupervised domain adaptation method has a stronger generalization ability but may create a distorted image [5]. The deep learning-based algorithm [8] can rely on the ground truth to learn more information. They also utilize the normalized light difference as an input, which is defined as:

$$X_t = \frac{x_{t+1} - x_t}{x_{t+1} + x_t}. \tag{1}$$

Moreover, to extract the rPPG signal, some researchers proposed the transformer-based [14] and mamba-based models [3] to analyze the human facial model, but they ignore the fact that the rPPG signal is periodic and the rough physiological signals are not hard to capture in most cases. Thus, their model costs a bit more and faces over-fitting issues in limited training data conditions and an unseen environment.

B. Mamba Model

Mamba [10] is a new method designed for NLP originally. The core component of Mamba is the State Space Model (SSM), which defines a sequence-to-sequence transformation as:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t), \end{aligned} \tag{2}$$

where $A, B, C \in \mathbb{R}^{n \times n}$ and they are learnable parameters. Further, Mamba [10] uses the gated structure to capture the long-term relationship:

$$\begin{aligned} g_t &= \sigma(Linear(x_t)), \\ h_t &= (1 - g_t)h_{t-1} + g_t x_t. \end{aligned} \tag{3}$$

Besides, the biggest difference between Mambaout [11] and Mamba [10] is that Mambaout does not use SSM. The other details are illustrated in Section III. Moreover, the proposed method belongs to the supervised approach, and we will focus on this approach.

III. Proposed Methodology

Reviewing the existing methods and their drawbacks, we propose a novel supervised TYrPPG model.

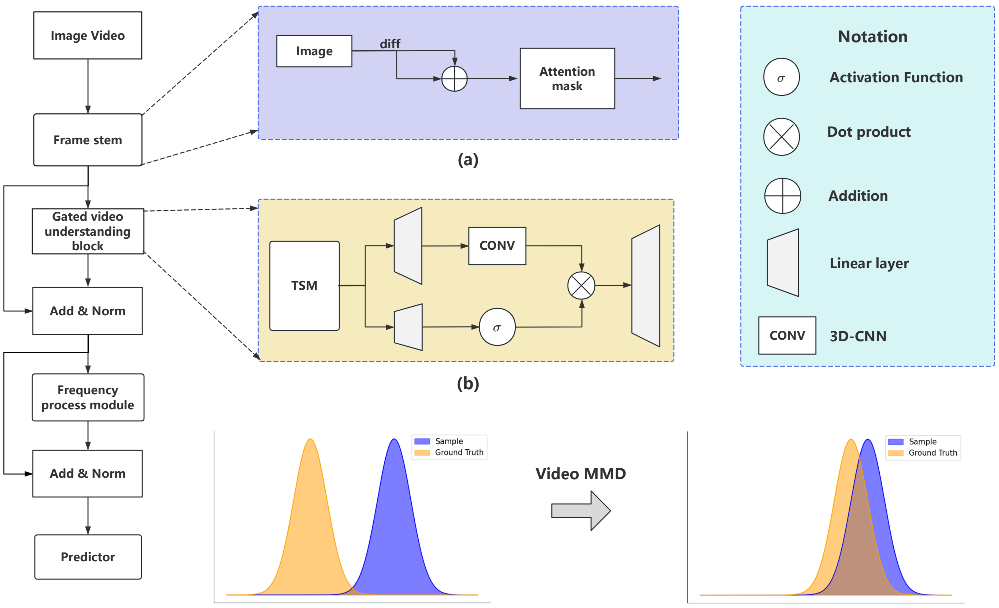


Fig. 2. TYrPPG is a gated 3D-CNN-based model structure. (a) shows the frame stem that is a data augmentation block to help TYrPPG understand the video better. (b) is the GVB, containing a TSM module and a gated 3D-CNN, which is designed to efficiently analyze video. Lastly, the distribution dissimilarity learning process is based on our proposed video MMD. Thus, the whole model structure is simpler than the Mamba-based and transformer-based models.

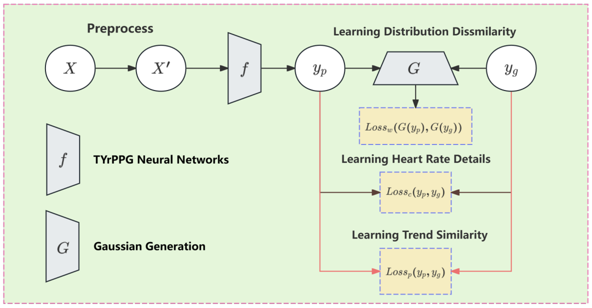


Fig. 3. TYrPPG uses the Comprehensive Supervised Loss (CSL) to optimize the model. CSL contains three loss terms for different purposes, including 1) learning distribution dissimilarity, 2) learning heart rate details, and 3) learning trend similarity. Besides, its variant, Weak Supervised Loss (WSL), comprises only the $\mathbf{Loss_w(.)}$ and $\mathbf{Loss_p(.)}$, exploring the feasibility of obtaining a good generalization ability without learning the signal details.

A. Model strusture

The proposed TYrPPG model utilizes the normalized light difference as an input to mitigate the illumination impact by using Eq.(1). Inspired by [8], we also utilize the attention mask to capture the face, avoiding the head motion impact, which is computed by:

$$Mask = \frac{(H/8)(W/8) \cdot \sigma(X)}{2||\sigma(X)||_1}, \tag{4}$$

where the H is the height and W is the width of the video frame. $\sigma(.)$ is the activation function. Then, the Gated Video Understanding Blocks (GVB) analyze the video frames to understand the video context for heart rate estimation. It comprises two parts: a TSM module [12] based on 2D-CNN and a gated 3D-CNN based on the Mambaout structure [11]. The TSM can first process the video by shifting the channel features in the temporal domain. Assuming that there are T frames, we can get the channel feature vectors of those frames as $X = [X_1, X_2, ..., X_T]$. And the weights of convolution are $W = (w1, w2, w3)$. The motivation for using TSM is that shifting channel features along the time dimension can improve the understanding of video context and make the TYrPPG more robust. The operator is defined as:

$$Y = w_1 X_{t-1} + w_2 X_t + w_3 X_{t+1}, \tag{5}$$

where the X_{t-1} means the previous frame of the X_t, and X_{t+1} is the latter frame. Due to the success of the Mambaout model [11] in vision tasks, we propose a gated 3D-CNN that can be regarded as a variant of it, shown in the model structure in Figure 2 (b). Instead of using the 2D-CNN like the original Mambaout structure, we use a 3D-CNN to replace it, which considers the information from both spatial and temporal domains. Thus, it is more powerful and suitable for analyzing video data compared with the original Mambaout model. The definition of gated 3D-CNN is shown in Eq.(10) and Eq.(11). Based on the Mambaout [11], TYrPPG normalizes input:

$$x' = Norm(x), \tag{6}$$

where $Norm(.)$ is the layer normalization. Moreover, TYrPPG will split the x_t into three parts: x_c, x_i and x_g, which is defined as:

$$x_g = Linear(x'), \quad x_i = Linear(x'), \quad x_c = Linear(x'), \tag{7}$$

TABLE I
INTRA-DATASET RESULTS ON PURE AND MMPD DATASETS

Methods	PURE			MMPD		
	MAE↓	RMSE↓	$\rho\uparrow$	MAE↓	RMSE↓	$\rho\uparrow$
PhysNet [15]	8.21	11.35	0.19	14.46	19.45	0.19
TS-CAN [16]	10.47	14.68	0.23	17.65	22.36	0.07
DeepPhys [8]	7.88	13.32	0.37	16.10	20.71	0.15
PhysFormer [14]	5.29	6.95	0.91	15.15	19.47	0.12
RhythmFormer [17]	0.97	1.10	**0.99**	12.91	16.58	0.31
Ours	**0.86**	**1.02**	**0.99**	**12.68**	**15.71**	**0.34**

TABLE II
INTRA-DATASET RESULTS ON THE PURE DATASET WITH A TRAIN-TEST SPLIT RATIO OF 6:4.

Methods	MAE↓	RMSE↓	$\rho\uparrow$
PhysNet [15]	10.31	14.11	0.21
TS-CAN [16]	11.48	15.28	0.23
DeepPhys [8]	10.02	13.32	0.37
PhysFormer [14]	9.67	12.68	0.21
RhythmFormer [17]	**8.38**	13.26	0.35
Ours	9.25	**11.51**	**0.41**

where $x_g \in D^H$, $x_i \in D^{H-C}$, and $x_c = D^C$. This means that the dimension of the x_g equals the hidden dimension. x_i's dimension equals hidden dimensions subtractive channel dimension, and x_c's dimension equals channel dimension, which will be processed by the 3D-CNN. Then we can constructs x_t' as:

$$x_c' = Conv(x_c), \tag{8}$$

where the $Conv(.)$ is the 3D-CNN. Further, we concatenate the x_i and x_c:

$$x_o = Concatenate(x_i, x_c). \tag{9}$$

Then, the output processed value is computed by:

$$x_{o1} = W_{o1}^T(\sigma(x_g) * x_o), \tag{10}$$

Lastly, we utilize the commonly used residual structure to help the model understand the video information better:

$$x_f = x_{o1} + x, \tag{11}$$

where the $\sigma(.)$ is the activation function and x_f is the final output of the GVB.

Remark 1: The proposed model structure can easily address the head motion issue and illumination impacts without a complex network structure and assumptions.

B. Loss Function

Another critical part of the proposed model is our designed loss function. Specifically, the loss function comprises three different types that respectively focus on the signal's details and distributions. The whole training logic based on different loss functions is shown in Figure 3. Specifically, the training method of our TYrPPG model is called Comprehensive Supervised Loss (CSL). We utilize the ground truth in the frequency domain to optimize our model. The first component is computed by:

$$Loss_c = CE(y_{pred}, y_{pos}), \tag{12}$$

where the $Loss_c$ is a cross-entropy loss to help TYrPPG learn the details of the heart rate signals. Moreover, we also use the negative Pearson correlation to optimize our model to maximize the trend similarity and minimize peak location errors, which is defined as:

$$\begin{aligned} P &= \frac{T\sum_1^T xy - \sum_1^T x\sum_1^T y}{\sqrt{(T\sum_1^T x^2 - (\sum_1^T x)^2)(T\sum_1^T y^2 - (\sum_1^T y)^2)}}, \\ Loss_p &= 1 - P, \end{aligned} \tag{13}$$

where p is the Pearson correlation. Moreover, we design a weakly supervised loss, video-MMD loss function. The original MMD is computed by:

$$\begin{aligned} MMD[f, p, q] &= sup(E_p[f(x)] - E_q[f(x)]) \\ &= sup < \mu_p - \mu_q, f >_H \\ &\leq ||\mu_p - \mu_q||_H. \end{aligned} \tag{14}$$

However, the maximum value of the Power Spectral Density (PSD) is more meaningful for estimating the heart rate. Inspired by [14], to better generalize the distribution of the ground truth, we decide to generate a series of Gaussian distributions based on the maximum values of the PSD, then utilize MMD to optimize our model, which is defined as:

$$\begin{aligned} Loss_w = MMD(&maxG(PSD(PPG_{gt})), \\ &maxG(PSD(PPG_{pred}))), \end{aligned} \tag{15}$$

where $maxG(.)$ regards the maximum Power Spectral Density (PSD) value as the mean value to generate the estimated Gaussian distributions and target Gaussian distributions [14]. Then, we use the MMD between two generated Gaussian distributions as a weak supervision. In addition, we also construct the WSL that will not consider the $Loss_c$. The motivation is to prove that less supervision may also work in heart rate estimation.

Remark 2: It is noteworthy that the rough physiological signals are not hard to extract in most cases, but the model will inevitably predict some incorrect values that do not exist in the ground truth. In this case, however, KL divergence can excessively exaggerate the difference between the predicted signals and ground truth due to those events with zero

TABLE III
CROSS-DATASET RESULTS ON PURE AND MMPD DATASETS

Methods	TrainSet	PURE			MMPD		
		MAE↓	RMSE↓	$\rho\uparrow$	MAE↓	RMSE↓	$\rho\uparrow$
PhysNet [15]	PURE	-	-	-	16.43	20.87	0.01
	MMPD	11.27	16.67	0.39	-	-	-
TS-CAN [16]	PURE	-	-	-	15.65	20.86	0.13
	MMPD	11.31	15.41	0.13	-	-	-
DeepPhys [8]	PURE	-	-	-	15.64	20.39	0.17
	MMPD	12.21	16.11	0.11	-	-	-
PhysFormer [8]	PURE	-	-	-	15.27	19.05	0.06
	MMPD	11.64	**13.56**	0.25	-	-	-
RhythmFormer [17]	PURE	-	-	-	15.42	18.79	0.24
	MMPD	**10.93**	15.31	0.25	-	-	-
Ours	PURE	-	-	-	**13.42**	**16.28**	**0.32**
	MMPD	11.31	15.26	**0.31**	-	-	-

probability, causing incorrect optimization direction. Thus, MMD can provide a more robust evaluation for distribution dissimilarity, better learning the peaks and trends.

In this case, the proposed CSL and WSL are defined as:

$$Loss_{CSL} = \alpha Loss_c + \beta Loss_p + \gamma Loss_w, \qquad (16)$$

and

$$Loss_{WSL} = \beta Loss_p + \gamma Loss_w, \qquad (17)$$

where the α, β, and γ are three hyper-parameters. In our experiments, they are set as 1.0, 1.0, and 2.0.

IV. EXPERIMENTS

In the experiment section, we first introduce the experiment settings and then show the experiment results.

A. Experiment Setting

We use two public datasets to evaluate the performance of TYrPPG: PURE [18] and MMPD [19].

PURE has 60 videos, including 10 different participants and 6 different test conditions (speaking, moving, etc.), with a video resolution of 640x480 and a frame rate of 30Hz.

MMPD including MMPD (370G, 320 x 240 resolution) and Mini-MMPD (48G, 80 x 60 resolution). We chose Mini-MMPD to train our model, which contains 33 subjects. Moreover, this dataset considers more conditions, including skin tone, activities, and lighting.

Besides, we utilize three commonly used indicators to evaluate model performance: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson correlation coefficient (ρ). We set the number of epochs to 30 and the learning rate to 1×10^{-4}. We choose the best results of the models to report. Further, the bold means the best, and the underline means the second best.

B. Comparative results

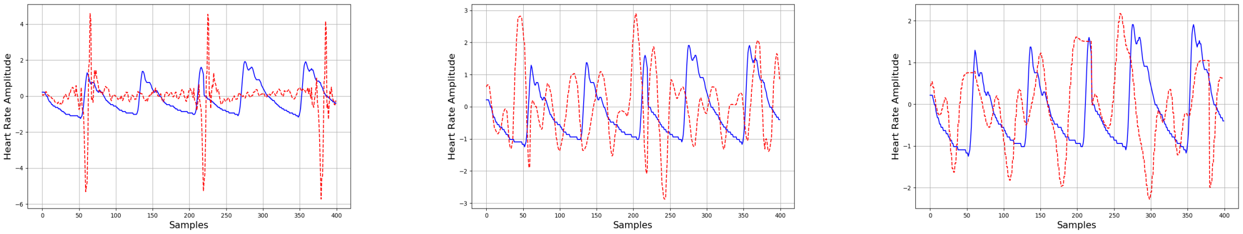
We conduct two main experiments to evaluate TYrPPG, including the intra-dataset and cross-dataset experiments. In particular, we choose some baseline models to make the comparison: PhysNet [15], DeepPhys [8], physFormer [14], TS-CAN [16], and RhythmFormer [17].

TABLE IV
ABLATION STUDY RESULTS ON PURE, WHOSE TRAIN-TEST SPLIT RATIO IS 6:4. THE RESULTS REFLECT THE EFFECTIVENESS OF THE PROPOSED CSL AND PROVE THE PROSPECT OF THE WSL.

Loss terms	MAE ↓	RMSE ↓	$\rho\uparrow$
$Loss_c$	13.41	18.58	0.16
$Loss_t$	11.60	16.54	0.07
$Loss_h$	18.69	22.03	0.04
$Loss_h + Loss_c$	13.40	18.58	0.19
$Loss_c + Loss_t$	12.16	15.84	0.16
$Loss_h + Loss_t$ (WSL)	11.60	15.57	0.04
$Loss_c + Loss_t + Loss_h$ (CSL)	**9.25**	**11.51**	**0.41**

In the first intra-dataset experiment, we split the PURE dataset and MMPD dataset for training and testing. In particular, the first 80% of the PURE dataset is regarded as the training data, and the last 20% of the PURE dataset was utilized for testing. Besides, the ratio of splitting the MMPD dataset is 1:1. Based on Table I, it is clear that our model reaches state-of-the-art performance in the PURE and MMPD datasets. TYrPPG outperforms all supervised methods. It especially outperforms the transformer-based SOTA models, such as Physformer [14] and RhythmFormer [17], which proves that the Mambaout-based GVB structure is more powerful compared with the transformer-based model. Moreover, in the second intra-dataset experiment, we also conduct another experiment on the PURE datasets and set its train-test split ratio to 6:4 according to [14] to further evaluate the performance of the TYrPPG. Based on Table II, we discover that the advantage of TYrPPG is also obvious, as it ranks first in RMSE and Pearson indicator, and ranks second in MAE indicator.

In the cross-dataset experiment, there are two group experiments to evaluate the robustness and generalization abilities of the proposed model. According to Table III, it is obvious that TYrPPG performs best in the MMPD dataset after training on the PURE dataset compared with other baselines, which indicates its significant generalization ability.



(a) Physformer

(b) RythmFormer

(c) TYrPPG

Fig. 4. Visualization of the heart rate signals estimated by Physformer, RythmFormer, and TYrPPG on the PURE dataset. TYrPPG learns significantly better than the other two models. Straight and dotted lines mark the signals as ground truth and model estimation, respectively.

C. *Visualization Analysis*

Based on Figure 4, we can discover that TYrPPG can significantly better learn the signal compared with the other two models. In particular, it can better capture the heart rate peak values from the video, showing its significant estimation ability. Since the peak heart rate has a higher diagnostic significance, TYrPPG provides more meaningful medical diagnosis results, showing its promising future.

D. *Ablation Study*

We have conducted an ablation study in the PURE dataset and set its train-test split ratio to 6:4 to evaluate the effectiveness of the proposed CSL and WSL. Based on the results of Table IV, the CSL has the best learning abilities compared with others, which indicates the necessity to combine the three loss items as the CSL. The results of WSL show the prospect that the model can learn the heart rate without knowing the details of the ground truth.

V. Concluding Remarks

We have proposed a novel rPPG for remotely estimating the heart rate. Compared with the existing method, our model structure is uncomplicated but powerful based on the proposed GVB, which combines the TSM and our modified Mambaout structure. Moreover, we have introduced a novel video MMD loss to enhance the learning of the ground truth distribution. The proposed CSL can provide both the details of the ground truth and the distribution similarity for the training model. TYrPPG achieves SOTA performance in two commonly used datasets, indicating its promising future. Further research may explore more about GVB block.

References

[1] G. Balakrishnan, F. Durand, and J. Guttag, “Detecting pulse from head motions in video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3430–3437, 2013.

[2] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, “Remote plethysmographic imaging using ambient light.,” *Opt. Express*, vol. 16, pp. 21434–21445, Dec 2008.

[3] B. Zou, Z. Guo, X. Hu, and H. Ma, “Rhythmmamba: Fast remote physiological measurement with arbitrary length videos,” *arXiv preprint arXiv:2404.06483*, 2024.

[4] M. van Gastel, S. Stuijk, and G. de Haan, “Motion robust remote-ppg in infrared,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 5, pp. 1425–1433, 2015.

[5] J. Du, S.-Q. Liu, B. Zhang, and P. C. Yuen, “Dual-bridging with adversarial noise generation for domain adaptive rppg estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10355–10364, 2023.

[6] G. Wei, C. Lan, W. Zeng, Z. Zhang, and Z. Chen, “Toalign: Task-oriented alignment for unsupervised domain adaptation,” in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 13834–13846, 2021.

[7] B. Lokendra and G. Puneet, “And-rppg: A novel denoising-rppg network for improving remote heart rate estimation,” *Computers in biology and medicine*, vol. 141, p. 105146, 2022.

[8] W. Chen and D. McDuff, “Deepphys: Video-based physiological measurement using convolutional attention networks,” in *Proceedings of the European Conference on Computer Vision*, pp. 349–365, 2018.

[9] M. Hu, D. Guo, M. Jiang, F. Qian, X. Wang, and F. Ren, “rppg-based heart rate estimation using spatial-temporal attention network,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1630–1641, 2022.

[10] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.

[11] W. Yu and X. Wang, “Mambaout: Do we really need mamba for vision?,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4484–4496, 2025.

[12] J. Lin, C. Gan, and S. Han, “Tsm: Temporal shift module for efficient video understanding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[14] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, “Physformer: Facial video-based physiological measurement with temporal difference transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4186–4196, 2022.

[15] Z. Yu, X. Li, and G. Zhao, “Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks,” *arXiv preprint arXiv:1905.02419*, 2019.

[16] X. Liu, J. Fromm, S. Patel, and D. McDuff, “Multi-task temporal shift attention networks for on-device contactless vitals measurement,” in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 19400–19411, 2020.

[17] B. Zou, Z. Guo, J. Chen, and H. Ma, “Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer,” *arXiv e-prints*, pp. arXiv–2402, 2024.

[18] R. Stricker, S. Müller, and H.-M. Gross, “Non-contact video-based pulse rate measurement on a mobile service robot,” in *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062, IEEE, 2014.

[19] J. Tang, K. Chen, Y. Wang, Y. Shi, S. Patel, D. McDuff, and X. Liu, “Mmpd: Multi-domain mobile video physiology dataset,” in *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pp. 1–5, IEEE, 2023.