RhythmGuassian: Repurposing Generalizable Gaussian Model For Remote Physiological Measurement

Hao Lu[1,2*], Yuting Zhang[1,2*], Jiaqi Tang[2], Bowen Fu[3], Wenhang Ge[1,2],
Wei Wei[3], Kaishun Wu[1,2], Yingcong Chen[1,2‡]
[1]The Hong Kong University of Science & Technology (Guangzhou),
[2]The Hong Kong University of Science & Technology, [3]Northwestern Polytechnical University,
{hlu585,yzhang430}@connect.hkust-gz.edu.cn, yingcongchen@ust.hk
* Equal contribution, ‡ Corresponding author

Abstract

Remote Photoplethysmography (rPPG) enables non-contact extraction of physiological signals, providing significant advantages in medical monitoring, emotion recognition, and face anti-spoofing. However, the extraction of reliable rPPG signals is hindered by motion variations in real-world environments, leading to entanglement issue. To address the challenge, we employ the Generalizable Gaussian Model (GGM) to disentangle geometry and chroma components with 4D Gaussian representations. Employing the GGM for robust rPPG estimation is non-trivial. Firstly, there are no camera parameters in the dataset, resulting in the inability to render video from 4D Gaussian. The "4D virtual camera" is proposed to construct extra Gaussian parameters to describe view and motion changes, giving the ability to render video with the fixed virtual camera parameters. Further, the chroma component is still not explicitly decoupled in 4D Gaussian representation. Explicit motion modeling (EMM) is designed to decouple the motion variation in an unsupervised manner. Explicit chroma modeling (ECM) is tailored to decouple specular, physiological, and noise signals, respectively. To validate our approach, we expand existing rPPG datasets to include various motion and illumination interference scenarios, demonstrating the effectiveness of our method in real-world settings. Code is available at https://github.com/LuPaoPao/RhythmGuassian.

1. Introduction

Remote Photoplethysmography (rPPG) [11, 17, 49, 54, 55] is a non-contact technique that utilizes camera devices to extract physiological signals, such as heart rate, from facial videos. It has been widely applied in various fields, including medical monitoring, emotion recognition [23], and face anti-spoofing [64]. Compared to traditional contact-based methods for physiological signal acquisition, rPPG offers the advantages of being non-invasive and more convenient. But, the model cannot accurately track the relative 4D geometric changes of face skin in the real scene, resulting in motion noise. The relative position between the light source and the face can also introduce illumination noise. Therefore, addressing the challenges of motion and illumination interference has become a key focus in current research efforts [26, 28, 63, 65, 68].

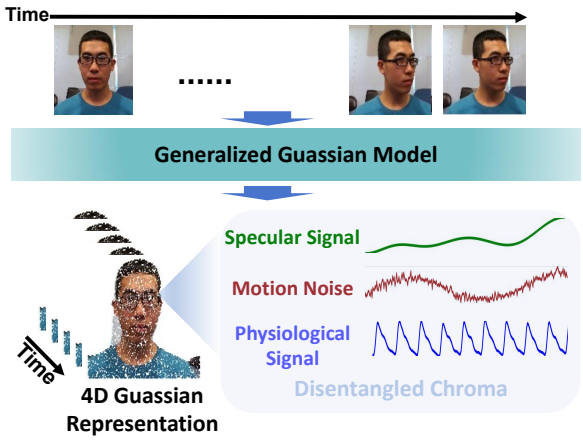

Figure 1. The main idea of this paper. Based on the input video, the Generalizable Gaussian Model (GGM) predicts a 4D Gaussian Spaltting representation. The chroma dimension is decoupled into specular signals, physiological signals, and motion noises.

Traditional methods primarily rely on Blind Source Separation (BSS) [21, 37] and subspace methods [12, 17, 54, 55]. These methods typically operate under specific assumptions, such as the ability to separate noise and signal in certain subspaces or assuming that the interference from motion and lighting is linear. While these assumptions are effective in some cases, their generalization ability is lim-

ited in more complex dynamic scenarios. In recent years, deep learning (DL) based approaches have shown their great potential in rPPG measurement [28, 29, 35]. By learning dedicated rPPG feature representations, these methods achieve promising performance in much more complicated environments [14, 22, 51, 68].

Besides, self-supervised learning approaches have gradually been adopted in the rPPG domain [26, 42, 45, 50]. These methods do not rely on external labels but rather design specific tasks to learn physiological signal features directly from the video. For instance, rPPG-MAE [26] employs a masked autoencoder (MAE) to learn the periodic characteristics within the spatio-temporal map (STMap). Another method, Contras-Phys [45], constructs a contrastive learning framework where different frames from the same video serve as positive samples.

However, motion noise remains the most challenging obstacle to address, as most existing methods lack a 4D geometric understanding. This understanding is crucial for the model to not only capture the changes in surface pixels but also fully utilize 4D geometric information to predict skin reflectance under varying positions accurately. Fortunately, the Generalizable Gaussian Splatting (GGS), a technique initially developed for 3D or 4D reconstruction, can fully understand the geometry of the scene [6, 38]. Applying GGS to the rPPG domain offers a promising approach for achieving 4D geometric understanding, which can, in turn, accurately separate non-physiological noise. Yet, there are two significant challenges: (1) We do not have accurate camera parameters in the rPPG dataset. (2) Unsupervised explicit decoupling of noise and physiological signals has not been explored in 4D space.

To this end, a novel method called RhythmGuassian is proposed, which leverages the power of the Generalizable Gaussian Model (GGM) to explicitly separate the chroma and geometric variation as shown in Fig. 1. Firstly, without using camera parameters, the "4D virtual camera" is proposed to simplify the complex 4D representation of facial video, allowing the model to efficiently learn geometry and chroma variations. This virtual camera can render key parameters like geometric position, various reflections, and motion flowing, enabling a focused representation of the face. Then, we leverage unsupervised task knowledge to ensure that the chroma and geometric variation are accurately tracked and separated. The 4D geometric position is represented as a combination of moving optical flow and 3D position, which an be decoupled by cleverly designed reconstruction constraints. The chroma changes are carefully decomposed into components representing specular reflection, the desired physiological signal, and motion-induced noise. These three components are decoupled under different unsupervised domain knowledge. To validate the effectiveness, we have compiled and expanded the largest existing rPPG dataset, By conducting extensive experiments on this dataset, we can comprehensively evaluate the generalization ability of our proposed method and ensure its applicability in real-world scenarios. In summary, the main contributions of this paper are as follows:

- We introduce the Generalizable Gaussian Model (GGM) into the rPPG domain to disentangle the chroma and geometric changes explicitly.
- We propose the novel concept of the "4D virtual camera" to simplify the complex 4D representation of facial video, allowing the model to efficiently learn geometry and chroma variations without camera parameters.
- We tailored a series of unsupervised task knowledge, which can explicitly decouple different chroma and geometric components.
- We have curated and expanded an rPPG dataset containing diverse interference scenarios to validate the proposed method's effectiveness thoroughly.

2. Related Work

2.1. Remote Physiological Measurement

In the field of rPPG, environmental noise such as lighting and motion has consistently posed significant challenges. Researchers have proposed various methods, from traditional approaches [4, 12, 17, 52, 54, 55, 67] to deep learning [8, 11, 24, 31, 44, 63, 65, 69, 70], to mitigate the impact of these environmental factors. For example, some methods convert video into STMaps [27, 34] to minimize the influence of localized skin fluctuations on the overall signal, while others attempt to decouple relatively clean rPPG signals from noisy data [1, 15, 28, 35]. Recently, self-supervised learning has been explored in rPPG research, reducing dependence on labeled data while enhancing model robustness against noise [26, 42, 45]. However, most of these methods address noise mitigation at the outcome level rather than tackling the root causes of noise. This work presents a novel approach that, for the first time, focuses on the capability of 4D geometric understanding. By employing Generalizable Gaussian Model (GGM) to reconstruct facial videos, this method effectively enables the disentanglement of motion and illumination variations, thereby promoting the reliable extraction of physiological signals.

2.2. Generalized Reconstruction Models

Some works have proposed to greatly speed this up by training neural networks to directly learn the full reconstruction task in a way that generalizes to novel scenes [30, 53, 58, 62]. Recently, LRM [20] was among the first to utilize large-scale multiview datasets including Objaverse [13] to train a transformer-based model for 3D reconstruction. The resulting model exhibits better generalization and higher quality reconstruction of object-centric

3D shapes from sparse posed images in a single model forward pass. Similar works have investigated changing the representation to Gaussian splatting [47, 66], introducing architectural changes to support higher resolution [41, 60], and extending the approach to 3D scenes [5, 10]. Recently, L4GM [38] utilize temporal cross attention to fuses multiple frame information to predict the Gaussian representation of a dynamic object. However, for the rPPG task, Gaussian splatting can effectively decouple motion, light, and physiological signals, which has not yet been explored.

3. Preliminaries

3.1. Remote Physiological Measurement

rPPG is a non-contact technique that captures periodic chrominance changes in facial skin, reflecting blood volume variations due to the cardiac cycle. The observed chrominance changes of a facial pixel can be modeled as [8]:

$$C_k(t) = I \cdot (v_s(t) + v_d(t) + v_n(t)), \tag{1}$$

where $C_k(t)$ represents the RGB values of the k-th pixel over time, while I is the luminance intensity, which varies with the light source, its distance from the skin, and the camera. According to the dichromatic reflection model, the reflected light consists of the following components: the specular reflection $v_s(t)$, which carries no physiological information and presents a DC signal pattern; the diffuse reflection $v_d(t)$, which results from light absorption and scattering within the skin tissues and contains the physiological periodic changes; and the noise accounts for environmental light fluctuations, motion artifacts, and camera quantization errors. Among these factors, motion-induced noise $v_n(t)$, which exhibits a pattern similar to motion flow, remains a significant challenge for accurate signal extraction. The three components have different patterns, which gives the possibility of decoupling the three components. In this paper, we try to explicitly decouple these three components using Gaussian splatting.

3.2. Generalizable Guassian Model

Our method builds on the success of generalizable Gaussian models (GGM) [20, 47]. GGM accepts a set of images and directly outputs a 3D or 4D representation of the video, especially, a set of Gaussian points P. Each Gaussian is represented by 14 parameters, including a center $\mathbf{u} \in \mathbb{R}^3$, a scaling factor $\mathbf{s} \in \mathbb{R}^3$, a quaternion rotation $\mathbf{q} \in \mathbb{R}^4$, an opacity $\alpha \in \mathbb{R}$, and a color feature $\mathbf{c} \in \mathbb{R}^3$. For rPPG tasks, applying these methods (the videos as input) directly to decouple motion and color change is very redundant. Proper adaptation of these methods to rPPG tasks is necessary to help the network better decouple motion and color change.

3.3. Spatial-Temporal Map

Our method select Spatial-Temporal Map (STMap) as the input of neural network [28, 29, 33, 35]. This is because its robustness and real-time abilities are leading, compared with the video inputs [22, 51, 56]. STMap flaten the average of each patch of the face image into a 2-dimensional vector $\mathbf{p} \in \mathbb{R}^{3\times N}$ as each column of the image. These vectors of different times are arranged in corresponding columns to form STMap $\mathbf{M} \in \mathbb{R}^{3\times N\times T}$. N is the number of patches ($N = \frac{W}{P_s} \times \frac{H}{P_s}$) and T is the length of time.

4. Method

RhythmGaussian is proposed to overcome entanglement issue when extracting rPPG signals by explicitly decoupling 4D chromatic and geometric changes using the Generalizable Gaussian Model, as shown in Fig. 2. The STMap, derived from the input video, is processed by an encoder and passed into the physiological decoder to predict the physiological signal. Another path is directed into the Gaussian Adaptor to obtain the 4D Gaussian Map. Then, Explicit Motion Modeling (EMM) learns the motion flow to accurately describe motion changes. Next, the 4D virtual camera is proposed to render videos from the 4D Gaussian Map. Finally, Explicit Chroma Modeling (ECM) is proposed to decouple specular reflection, physiological signals, and diffuse reflection in an unsupervised manner.

4.1. 4D Virtual Camera

Motivation. Employing the GGM for robust rPPG estimation is a non-trivial task. The rPPG dataset lacks intrinsic and extrinsic camera parameters as well as geometric information. This information absence significantly hinders the adaptation of the Gaussian splatting.

Solution. Face video can be converted into STMap and then fed to Encoder, physiological Decoder, and GS Adapter to get physiological signal and 4D Guassian Map. The corresponding 4D Gaussian map is predicted by the 4D Gaussian adapter G. The relationship between the face video and the 4D Gaussian map is shown in Fig. 3. The 4D Gaussian adapter employs up-sample blocks to convert features to the 4D Gaussian map $\mathbf{M}_{gs} \in \mathbb{R}^{22\times N\times T}$, including the depth value $\mathbf{d_r} \in \mathbb{R}^1$, specular reflection $\mathbf{v_s} \in \mathbb{R}^3$, diffuse reflection $\mathbf{v_d} \in \mathbb{R}^3$, motion noises $\mathbf{v_n} \in \mathbb{R}^3$, alpha $\mathbf{a} \in \mathbb{R}^1$, scale $\mathbf{s} \in \mathbb{R}^3$, rotation $\mathbf{r} \in \mathbb{R}^3$, and motion flow $[\Delta h, \Delta w, \Delta s] \in \mathbb{R}^5$. The final color representation of the Gaussian point is:

$$\mathbf{c} = \mathbf{v_s} + \mathbf{v_d} + \mathbf{v_n}, \tag{2}$$

where, $\mathbf{v_s}$, $\mathbf{v_d}$ and $\mathbf{v_n}$ are specular reflections, physiological signal components, and motion noises, as the same definition of Sec. 3.1. The final position of the Gaussian point is:

$$\mathbf{p} = KE(ud, vd, d), \tag{3}$$

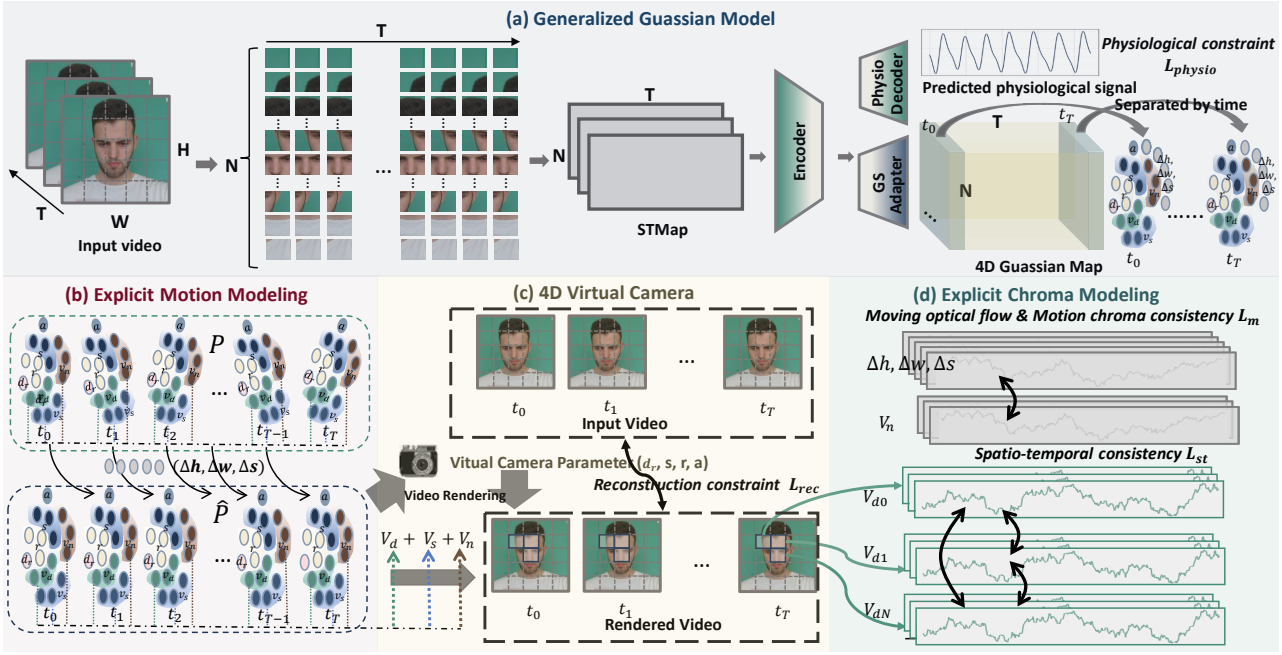

Figure 2. An overview of the proposed method. a) **Generalized Guassian Model:** The STMap, derived from the input video, is processed through an encoder and fed into the physiological decoder to predict the physiological signals. Another path feeds into the Gaussian Adaptor to obtain 4D Gussian Map. **b) Explicit Motion Modeling (EMM)** achieves spatial motion compensation in 4D Gaussian space by applying Δh (horizontal displacement), Δw (width variation), and Δs (scale adjustment) parameters to the Gaussian splatting set P, ultimately generating the motion-modified point set $\hat{P}$. **c) 4D Virtual Camera** use the motion-modified point set to render videos with tri-stream chrominance components (V_s, V_d, and V_n), while optimizing the model through pixel-wise reconstruction loss between rendered and input videos. **d) Explicit Chroma Modeling (ECM):** Spatiotemporal Consistency Loss enforces inter-frame stability of V_d across facial regions to preserve physiological signal consistency. Motion-Aware Chroma Consistency Loss $\mathcal{L}_m$, guided by dynamic optical flow, synchronizes pose parameters $[\Delta h, \Delta w, \Delta s] \in \mathbb{R}^{5\times N\times T}$ with noise-associated chroma component $V_{dn} \in \mathbb{R}^{3\times 1\times T}$.

where u and v are the uv coordinates in 4D Gaussian map, d means the vertical distance between the Gaussian point and the 4D virtual camera. The activation functions for RGB color, alpha, scale, and rotation are consistent with those in [47]. Based on the above representation, the Gaussian points predicted by the network can be rendered into the face video via the 4D virtual camera.

With reasonable simplification, the virtual 4D camera is proposed to render face video from the 4D Gaussian map as shown in Fig. 2 (c). The 4D virtual cameras use fixed virtual camera intrinsic parameters E and extrinsic parameters K to render the face video instead of using real camera parameters. The camera intrinsic parameters E use the face center as the focal point use half the size of the face as the focal length, and use an identity matrix for the extrinsic parameters K. The rendered video and the original video are constrained by L1 $\mathcal{L}_{rec}$.

Further explain. The above steps have the following three key points: 1. Simplification of the 4D position. The different patches of the original face image are flatted into a column in the 4D Gaussian map, which still retains the 4D location information. The shifts parallel to the original camera plane can be represented by the v-coordinate in the 4D Gaussian map. The depth perpendicular to the original camera plane can be reflected in the scale of the Gaussian point s, because the pinhole imaging principle is larger near and smaller. 2. Simplification of the 4D motion. We define the offset of pixels on the face video in adjacent frames as $[\Delta h, \Delta w] \in \mathbb{R}^2$. The depth change is considered to be a change in the size $\Delta s \in \mathbb{R}^3$ of a Gaussian point, which is the pinhole principle. The full motion flow representation is $[\Delta h, \Delta w, \Delta s] \in \mathbb{R}^5$. 3. Chroma decomposition. As explained in Sec. 3.1, the chrominance signal is disentangled into three components $\mathbf{v_s}$, $\mathbf{v_d}$, and $\mathbf{v_n}$. These three components are significantly different and can be explicitly decoupled as introduced in Sec. 4.3.

4.2. Explicit Motion Modeling

Motivation. The 4D virtual camera can be used to ensure 4D reconstruction. However, the rPPG task needs to accurately track each area's movement and then decouple the physiological signals in the same area. Only reconstruction processing cannot meet the requirements of the rPPG task. Explicit motion modeling (EMM) is proposed to force the neural network to predict the motion flow in an unsupervised manner.

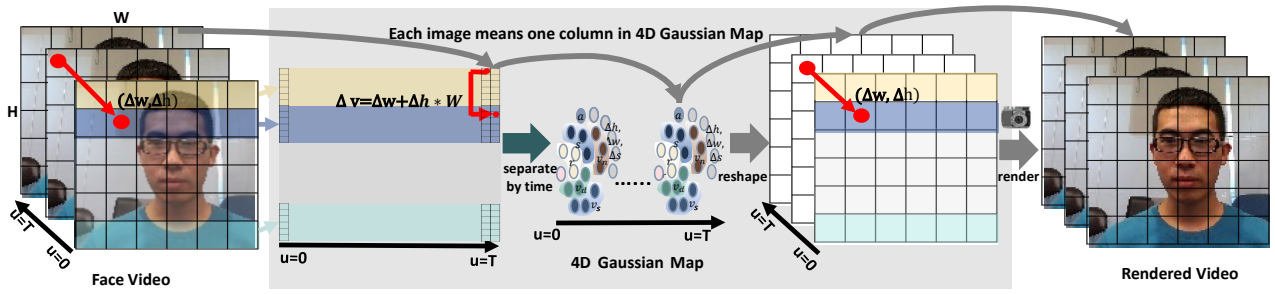

Figure 3. The coordinate relationship between face video and 4D Gaussian Map. In the adjacent time, the position changes of the face image $(\Delta w, \Delta h)$ correspond to the number of rows $(\Delta v = \Delta w + \Delta h * W)$ in 4D Gaussian Map. When moving Δh vertically in the face video, the longitudinal change in the STMap will jump $\Delta h * W$. Different times in the face video correspond to different columns in the 4D Gaussian map, that is, one column of STMap represents one image. In addition, the variation in the distance between the face and the camera is considered to be represented by the size change Δs of the Gaussian point.

Solution. To achieve this, explicit motion modeling (EMM) is proposed to accurately decouple motion changes. 4D Gaussian adapter G is used to predict the motion flow $\Delta h, \Delta w, \Delta s$. Here, $\Delta h, \Delta w$ represent motion flow parallel to the original camera plane (face image) install of the uv coordinates of the 4D Gaussian map. Δs stands for depth changes perpendicular to the original camera plane. The modified Guassian point position is:

$$\hat{\mathbf{p}} = KE((u+1)d, (v + \Delta w + \Delta h * W)d, d), \quad (4)$$

where, Δw represents the change in the v coordinate on the 4D Gaussian map, that is, the change in the motion on the column axis of the original face image. Δh represents the change of motion on the row axis of the original face image, because the face image patch is rearranged, it needs to be multiplied by the column number of face image patch W. This relationship between the face video and 4D gaussian model is illustrated in Fig. 3. Using motion flow can find a correspondence position between the current frame and the next frame. The modified points $\hat{\mathbf{p}}$, physiological signals, and specular reflection signals can be rendered into the original video. We also applied L1 reconstruction $\mathcal{L}_{rec}$ constraints to the rendered video. To simplify the description, the constraints on both the modified point and the original point before the correction are expressed as $\mathcal{L}_{rec}$, and they are balanced with the same weight. It is mentioned that the modified point is considered to be free of chroma motion noises.

4.3. Explicit Chroma Modeling

Motivation. Reconstruction rendering and EMM can accurately model the video's 4D geometry and motion information. The ultimate goal of rPPG is to decouple physiology from all chroma components. Further explicit decoupling of the three chroma components (v_d, v_s, v_n) should be efficiently designed.

Solution. Many papers have successfully used rPPG task knowledge to disentangle the physiological v_d efficiently. Similarly, we apply spatio-temporal consistency loss $\mathcal{L}_{st}$ to the physiological v_d signal like paper [45]. $\mathcal{L}_{st}$ can ensure that physiological signals are periodic and similar in different areas of the face, specific details refer to [45].

Different from existing papers, an elegant loss $\mathcal{L}_m$ is proposed to disentangle the motion noise v_n, which can make a clever connection between chroma motion noises changes and motion flow. Specifically, we calculate the modulus l of all motion representing (including $\Delta h, \Delta w, \Delta s$) at each time, and then calculate the negative Pearson loss $\mathcal{L}_m$ as a consistent constraint on the motion noise and the moving optical flow. The specular reflection light v_s is guaranteed by the reconstruction loss. Anyway, all the constraints of the paper are:

$$\mathcal{L}_{all} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{st}\mathcal{L}_{st} + \lambda_m\mathcal{L}_m + \mathcal{L}_{physio} \quad (5)$$

where $\mathcal{L}_{physio}$ is the supervision of physiological signals as same as [29], which is not used under the unsupervised protocol. The remaining three constraints are unsupervised. λ_{rec}, λ_{st}, and λ_m are balanced parameters.

5. Experiment

5.1. Datasets

We have collected the largest existing rPPG dataset to date, specifically designed with lighting and motion interferences, aiming to more broadly validate the effectiveness of the proposed method. Below is description of the 10 rPPG datasets: **UBFC-rPPG [2]** consists of 42 face videos captured under both sunlight and indoor illumination conditions. **UBFC-Phys [40]** dataset contains videos from 20 subjects, with a total of over 1000 video clips. **VIPL-HR [32]** includes recordings from nine scenarios using three different RGB cameras, under varying illumination conditions, and with different levels of movement. **PURE [43]** contains 60 RGB videos from 10 subjects engaged in six different activities, including sitting still, talking, and four head rotation and movement variations.

Target	Method	LF-(u.n)			HF-(u.n)			LF/HF			HR-(bpm)		
		MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑
UBFC	GREEN [49]	0.2355	0.2841	0.0924	0.2355	0.2841	0.0924	0.6695	0.9512	0.0467	8.0184	9.1776	0.3634
	CHROM [17]	0.2221	0.2817	0.0698	0.2221	0.2817	0.0698	0.6708	1.0542	0.1054	7.2291	8.9224	0.5123
	POS [55]	0.2364	0.2861	0.1359	0.2364	0.2861	0.1359	0.6515	0.9535	0.1345	7.3539	8.0402	0.4923
	NEST [29]	0.0597	0.0782	0.2017	0.0597	0.0782	0.2017	0.2138	0.2824	0.3179	4.7471	6.8876	0.8546
	Baseline	0.0621	0.0813	0.1873	0.0621	0.0813	0.1873	0.1985	0.2667	0.3043	5.1542	7.4672	0.8165
	Ours	**0.0483**	**0.0621**	**0.2304**	**0.0483**	**0.0621**	**0.2304**	**0.2183**	**0.2789**	**0.3278**	**4.2704**	**6.4505**	**0.8952**
PURE	GREEN [49]	0.2539	0.3002	0.0326	0.2539	0.3002	0.0326	0.6525	0.8932	0.0417	10.3247	14.2693	0.4952
	CHROM [17]	0.2096	0.2751	0.1059	0.2096	0.2751	0.0759	0.5404	0.8266	0.1173	9.7914	12.7568	0.3732
	POS [55]	0.1959	0.2571	0.1684	0.1959	0.2571	0.1684	0.5373	0.846	0.1433	9.8273	13.4414	0.3432
	NEST [29]	0.0635	0.0874	0.6422	0.0635	0.0874	0.6422	0.2255	0.3505	0.5734	7.6889	10.4783	0.7255
	Baseline	0.0671	0.0923	0.6046	0.0671	0.0923	0.6046	0.2864	0.4184	0.5526	8.2542	11.1765	0.6832
	Ours	**0.0524**	**0.0725**	**0.6632**	**0.0524**	**0.0725**	**0.6632**	**0.2279**	**0.3482**	**0.5811**	**7.0732**	**9.7975**	**0.7726**
BUAA	GREEN [49]	0.3472	0.3951	0.0871	0.3472	0.3951	0.0871	0.6453	0.8632	0.0921	5.8231	7.9882	0.5624
	CHROM [17]	0.3786	0.3237	0.0682	0.3786	0.3237	0.0682	0.6813	0.8836	0.0715	6.0934	8.2938	0.5165
	POS [55]	0.3198	0.3762	0.0962	0.3198	0.3762	0.0962	0.6275	0.8424	0.1127	5.0407	7.1198	0.6374
	NEST [29]	0.1436	0.1665	0.2955	0.1436	0.1665	0.2955	0.5514	0.6884	0.3004	3.3723	5.8806	0.7647
	Baseline	0.1451	0.1681	0.2891	0.1451	0.1681	0.2891	0.5564	0.6904	0.2914	3.7852	6.3237	0.7492
	Ours	**0.1307**	**0.1583**	**0.3078**	**0.1307**	**0.1583**	**0.3078**	**0.5283**	**0.6725**	**0.3155**	**3.0401**	**5.2022**	**0.7917**

Table 1. HRV and HR estimation results on the MSDG protocol.

Method	VIPL-HR			V4V		
	MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑
GREEN [49]	12.18	18.23	0.25	15.64	21.43	0.06
CHROM [17]	11.44	16.97	0.28	14.92	19.22	0.08
POS [55]	14.59	21.26	0.19	17.65	23.22	0.04
DeepPhys [8]	12.56	19.13	0.14	14.52	19.11	0.14
TS-CAN [25]	12.34	18.94	0.16	14.77	19.96	0.12
Rhythmnet* **[33]**	8.97	12.16	0.49	10.16	14.57	0.34
Dual-GAN* **[28]**	8.88	11.69	0.50	10.04	14.44	0.35
BVPNet* **[11]**	8.45	11.64	0.51	10.01	14.35	0.36
NEST*+ **[29]**	7.86	11.15	0.58	9.27	13.79	0.41
HiBa*+ **[56]**	7.34	10.41	0.61	9.68	12.54	0.42
Ours	7.05	10.36	0.68	9.06	11.07	0.55
Ours Pro	6.84	10.14	0.71	8.62	10.32	0.57

Table 2. HR estimation results on VIPL-HR and V4V datasets. Methods marked with * use the STMap as the input of CNN, and + indicates methods based on the baseline (Rhythmnet without GRU). Ours Pro means that we used the all the remaining 9 datasets for training.

V4V [39]: Designed to capture drastic changes in physiological indicators, this dataset simulates ten tasks such as watching a funny joke, making a 911 emergency call, and experiencing strong odors. **BUAA-MIHR [59]** was created to evaluate algorithm performance under various illumination conditions. **MR-NIRP [36]** includes both NIR and RGB videos of a passenger's face, along with pulse oximeter readings. **UCLA-rPPG [57]** includes videos from 50 subjects, with over 3000 video clips in both controlled and real-world environments. **MMPD [46]** includes 660 one-minute videos recorded at 30 fps using a Samsung Galaxy S22 Ultra, with a resolution of 320x240 pixels. **VV100 [61]** includes 100 subjects recorded with a Basler acA1920-40uc camera at 1920x1200 pixels, 30 fps, and various indoor lighting conditions.

5.2. Implement and Metrics

Implement. The proposed method is implemented using PyTorch and is trained on an A6000 GPU. The optimizer is AdamW, with a learning rate of 0.0001 and a batch size of 256. λ_{rec}, λ_{st}, and λ_m are set as 0.5, 0.4, 0.1 respectively. For the generation of STMap from Video, we use the FAN to detect the 2D landmarks of faces [3], and other steps completely follow [29, 34]. STMap in each dataset is sampled with a time window of 256 with step 10. Following [29, 34], we also design spatial and temporal augmentation for STMap, i.e., shuffle each row of STMap and slide the time window. The teration is set to 20000 as same as [29].

Metrics. The most commonly used performance metrics for evaluation include the standard deviation (STD), mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (r) are reported.

5.3. Multi-Source Domain Generalization

Following NEST [29], we conduct Multi-Source Domain Generalization (MSDG) experiments for both heart rate (HR) and heart rate variability (HRV) estimation as shown in Tab. 2 and 1. The experimental results in Tab. 1 demonstrate our method's superior performance in both HRV anal-

Method	SD↓	MAE↓	RMSE↓	r↑
SAMC [48]	18.0	15.9	21.0	0.11
POS [55]	15.3	11.5	17.2	0.30
CHROM [17]	15.1	11.4	16.9	0.28
I3D [4]	15.9	12.0	15.9	0.07
DeepPhy [8]	13.6	11.0	13.8	0.11
BVPNet [11]	7.75	5.34	7.85	0.70
RhythmNet [33]	8.11	5.30	8.14	0.76
CVD [35]	7.92	5.02	7.97	0.79
Physformer [65]	7.74	4.97	7.79	0.78
Dual-GAN [28]	7.63	4.93	7.68	0.81
NEST [29]	7.49	4.76	7.51	0.84
Baseline	8.04	5.21	8.07	0.77
Ours w/o EMM	7.43	4.72	7.74	0.84
Ours w/o ECM	7.48	4.64	7.81	0.86
Ours	**7.19**	**4.22**	**7.12**	**0.92**

Table 3. HR estimation results by our method and several state-of-the-art methods on the VIPL-HR database. EMM means explicit motion modeling, and ECM means explicit chroma modeling.

Method	Pretraining	linear-probing	HR (bpm)		
			MeanAE↓	RMSE↓	r↑
MoCo [18]	VIPL	VIPL	9.27	13.05	0.04
SIMSIAM [9]	VIPL	VIPL	8.43	11.73	0.14
BOYL [16]	VIPL	VIPL	8.98	12.43	0.08
SIMCLR [7]	VIPL	VIPL	8.57	11.94	0.10
rPPG-MAE [26]	VIPL	VIPL	7.83	11.19	0.48
Ours	VIPL	VIPL	7.24	9.28	0.52
	10-All	VIPL	**6.43**	**8.21**	**0.56**

Table 4. Self-Supervision result on VIPL-HR.

ysis and HR estimation across three cross-dataset evaluations. Notably in HRV analysis, our method shows particular strength in high-frequency (HF) component estimation with 22.4% MAE improvement over traditional methods in PURE. These findings highlight that our method not only sets a strong baseline but also greatly benefits from additional training data, leading to more effective domain generalization.

Then, we evaluate our method on the most challenge VIPL-HR and V4V datasets, comparing it with several state-of-the-art approaches, as summarized in Tab. 2. In these two datasets with relatively large movements, our approach has achieved significant improvements. This shows that our algorithm can effectively cope with more complex scenes by more accurate 4D geometry understanding. To further verify the scale up capability of our algorithm, we introduce 5 additional data sets to join the training to further improve the generalization performance of the model, defined as Our Pro. Stronger experimental results show that our algorithm can learn the stronger ability to decouple geometry and chroma in more data.

5.4. Self-Supervision on VIPL-HR.

All additional supervision introduced in this paper is unsupervised. To verify the pre-training capabilities, we followed the self-supervised methods in the rPPG field to conduct heart rate (HR) estimation experiments using the proposed method on the VIPL-HR dataset [26]. During the pre-training phase, we removed the BVP decoder while retaining the branch containing the GS Adapter. The loss function used in this phase was $L_{\text{pre}} = L_{\text{rec}} + L_{\text{st}} + L_{\text{m}}$. We employed the linear probing technique [19] to evaluate the performance of the proposed self-supervised method (GGM). During this phase, the encoder was frozen, and the BVP decoder was adapted to predict the BVP signal and heart rate. The loss function during this phase was $L_{\text{physio}} = L_{\text{BVP}} + L_{\text{HR}}$. Our method was compared to four popular contrastive learning approaches: MoCo [18], SIMSIAM [9], BYOL [16], and SIMCLR [7], as well as rPPG-MAE [26], a self-supervised approach specifically designed for rPPG, as shown in Table 4.

The results demonstrate that our method achieves exceptional performance, even without label supervision, outperforming the current state-of-the-art approaches. Additionally, we performed linear-probing experiments on the VIPL-HR dataset after pretraining on 10 rPPG datasets. The results showed a significant improvement, with the MAE value decreasing from 7.24 to 6.43, which notably outperforms rPPG-MAE, which achieved an MAE of 7.83 when pretrained solely on the VIPL-HR dataset. These findings suggest that increasing the volume of pretraining data substantially enhances the performance of the generalizable Gaussian model (GGM). Moreover, our self-supervised Gaussian reconstruction task helps the encoder distinguish between different chromatic signals from facial videos, enabling it to converge to useful physiological signals during the linear probe phase.

5.5. Intra-Dataset Evaluation.

The performance of our method for heart rate (HR) estimation on the VIPL-HR dataset was evaluated, and it was compared to several state-of-the-art methods, as shown in Table 3. The methods evaluated include SAMC [48], POS [55], CHROM [17], I3D [4], DeepPhy [8], BVPNet [11], RhythmNet [33], CVD [35], Physformer [65], Dual-GAN [28], and NEST [29]. Our method is tested in three configurations: the baseline, "Ours w/o EMM" (without explicit motion modeling), and "Ours w/o ECM" (without explicit chroma modeling), with the full model denoted as "Ours."

The results indicate that our method outperforms all state-of-the-art methods in all evaluation metrics. Specifically, the full model achieves the best performance with an MAE of 4.22, RMSE of 7.12, and a correlation of 0.92, significantly outperforming the baseline, which had an MAE of 5.21 and a correlation of 0.77. Notably, even without the explicit motion modeling (EMM) or explicit chroma modeling (ECM), our method still shows considerable improve-

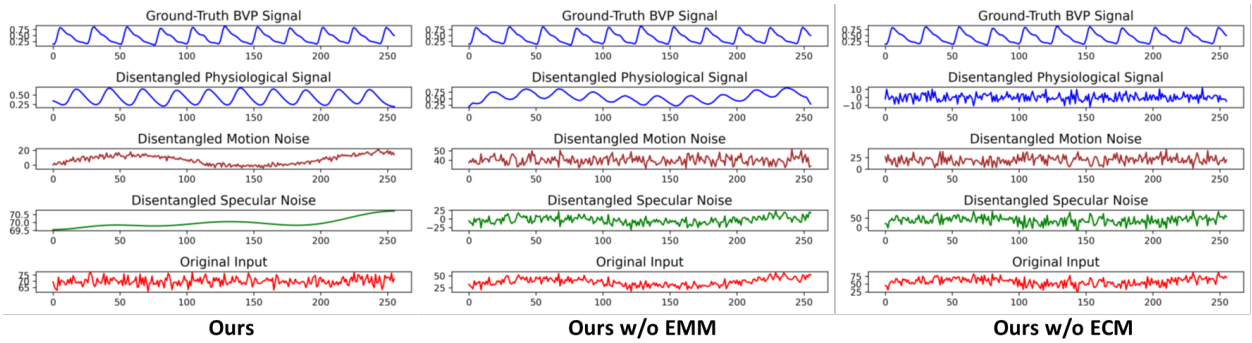

Figure 4. Visualization of disentangled signal components on VIPL-HR. Signal decomposition showing the separation of physiological signals, motion noise, and specular noise from the original input, alongside the ground-truth BVP (blood volume pulse) signal.

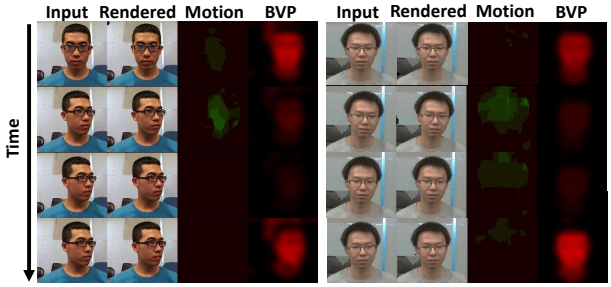

Figure 5. Visualization of rendered results. Here, we ensure more accurate face reconstruction through a more fine-grained STMap. In fact, we can achieve a very good decoupling effect by using a low-resolution STMap (all the other experimental results in this paper were completed with a low resolution).

ments over previous approaches, with "Ours w/o EMM" reaching an MAE of 4.72 and "Ours w/o ECM" achieving an MAE of 4.64. These results suggest that both the EMM and ECM modules contribute to improving HR prediction, with each module complementing the other to enhance the overall performance of the GGM.

6. Visualization

The proposed method effectively disentangles the physiological signal, which closely resembles the ground-truth (GT) BVP signal. Moreover, as shown in Fig. 4, both the motion and specular noise components show a clear reduction in noise, with the signals presenting smoother, more coherent patterns. In the second column, where the EMM module is removed, the motion and specular noise components remain disordered and cannot be separated from the input signal. In the third column, where the ECM module is removed, none of the three components can be disentangled, with the physiological signal showing a significantly higher noise level compared to the GT BVP signal. These results highlight the critical role of the EMM and ECM components in enabling the model to effectively separate the motion and specular noise, while preserving the purity of the physiological signal.

Further, the rendered videos of the motion component, physiological signal component, and the overall signal are compared with the input video, as shown in the Fig. 5. The video rendered from the BVP physiological signal component, however, shows rhythmic variations in brightness, which aligns with the rhythmic nature of the heartbeat. This demonstrates that our method effectively disentangles the subtle physiological signals from facial videos.

7. Conclusion

In conclusion, this paper proposes a novel approach to addressing the motion noise challenges in remote photoplethysmography (rPPG) by introducing the Generalizable Gaussian Model (GGM) and the concept of a "4D virtual camera." Our method provides an innovative solution for disentangling chromatic and geometric variations in facial video, particularly in the presence of motion and illumination interference. By leveraging unsupervised learning techniques, we can explicitly decouple the physiological signal from various noise sources, offering improved performance in complex and dynamic environments. Our extensive experiments on an expanded rPPG dataset demonstrate the robustness and generalization ability of the proposed method, making it highly applicable for real-world rPPG applications.

8. Acknowledge

This research is supported by HKUST-HKUST(GZ) Cross-Campus Collaborative Research Scheme (Project No. C036); Guangdong Provincial Department of Science and Technology's '1+1+1' Joint Funding Program for Guangdong-Hong Kong Universities, and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

References

[1] Boris Bačić, Chengwei Feng, and Weihua Li. Jy61 imu sensor external validity: A framework for advanced pedometer algorithm personalisation. *ISBS Proceedings Archive*, 42(1): 60, 2024. 2

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 5

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proc. ICCV*, 2017. 6

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 7

[5] D. Charatan, S. Li, A. Tagliasacchi, and V. Sitzmann. Pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. *arXiv preprint arXiv:2312.12337*, 2023. 3

[6] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 7

[8] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. *european conference on computer vision*, 2018. 2, 3, 6, 7

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 7

[10] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 3

[11] Abhijit Das, Hao Lu, Hu Han, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 1, 2, 6, 7

[12] Gerard De Haan and Arno Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 1, 2

[13] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2

[14] Jingda Du, Si-Qi Liu, Bochao Zhang, and Pong 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*, pages 10355–10364, 2023. 2

[15] Chengwei Feng, Boris Bačić, and Weihua Li. Sca-lstm: A deep learning approach to golf swing analysis and performance enhancement. In *International Conference on Neural Information Processing*, pages 72–86. Springer, 2025. 2

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 7

[17] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 2013. 1, 2, 6, 7

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 7

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv: Computer Vision and Pattern Recognition*, 2021. 7

[20] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2, 3

[21] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011. 1

[22] Haodong Li, Hao Lu, and Ying-Cong Chen. Bi-tta: Bidirectional test-time adapter for remote physiological measurement. In *European Conference on Computer Vision*, pages 356–374. Springer, 2025. 2, 3

[23] Zhihua Li and Lijun Yin. Multimodal facial action unit detection with physiological signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[24] Qian Liang, Yan Chen, and Yang Hu. Continual learning for remote physiological measurement: Minimize forgetting and simplify inference. In *European conference on computer vision*, pages 126–144. Springer, 2024. 2

[25] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 6

[26] Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements. *IEEE Transactions on Multimedia*, 2024. 1, 2, 7

[27] Hao Lu and Hu Han. Nas-hr: Neural architecture search for heart rate estimation from face videos. *Virtual Reality & Intelligent Hardware*, 3(1):33–42, 2021. 2

[28] Hao Lu, Hu Han, and S. Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. *computer vision and pattern recognition*, 2021. 1, 2, 3, 6, 7

[29] Hao Lu, Zitong Yu, Xuesong Niu, and Ying-Cong Chen. Neuron structure modeling for generalizable remote physiological measurement. In *CVPR*, pages 18589–18599, 2023. 2, 3, 5, 6, 7

[30] Zhengyang Lu, Bingjie Lu, and Feng Wang. Causalsr: Structural causal model-driven super-resolution with counterfactual inference. *Neurocomputing*, page 130375, 2025. 2

[31] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. *International Conference on Pattern Recognition*, 2018. 2

[32] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian conference on computer vision*, pages 562–576. Springer, 2018. 5

[33] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 3, 6, 7

[34] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 2020. 2, 6

[35] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. *european conference on computer vision*, 2020. 2, 3, 7

[36] Ewa M Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE transactions on intelligent transportation systems*, 23(4):3589–3600, 2020. 6

[37] Ming-Zher Poh, Daniel McDuff, and Rosalind W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 2011. 1

[38] J. Ren, K. Xie, A. Mirzaei, H. Liang, X. Zeng, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, et al. L4gm: Large 4d gaussian reconstruction model. *arXiv preprint arXiv:2406.10324*, 2024. 2, 3

[39] Ambareesh Revanur, Zhihua Li, Umur A Ciftci, Lijun Yin, and László A Jeni. The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2760–2767, 2021. 6

[40] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14(1):622–636, 2023. 5

[41] Q. Shen, X. Yi, Z. Wu, P. Zhou, H. Zhang, S. Yan, and X. Wang. Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. *arXiv preprint arXiv:2403.18795*, 2024. 3

[42] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474, 2023. 2

[43] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 5

[44] Weiyu Sun, Xinyu Zhang, Hao Lu, Ying Chen, Yun Ge, Xiaolin Huang, Jie Yuan, and Yingcong Chen. Resolve domain conflicts for generalizable remote physiological measurement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8214–8224, 2023. 2

[45] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 492–510. Springer, 2022. 2, 5

[46] Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: multi-domain mobile video physiology dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–5. IEEE, 2023. 6

[47] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3, 4

[48] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016. 7

[49] Wim Verkruysse, Lars O. Svaasand, and J.S. Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 2008. 1, 6

[50] Hao Wang, Euijoon Ahn, and Jinman Kim. Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2431–2439, 2022. 2

[51] Jiyao Wang, Hao Lu, Hu Han, Yingcong Chen, Dengbo He, and Kaishun Wu. Generalizable remote physiological measurement via semantic-sheltered alignment and plausible style randomization. *IEEE Transactions on Instrumentation and Measurement*, 2024. 2, 3

[52] Jiyao Wang, Hao Lu, Ange Wang, Xiao Yang, Yingcong Chen, Dengbo He, and Kaishun Wu. Physmle: Generalizable and priors-inclusive multi-task remote physiological measurement. *arXiv preprint arXiv:2405.06201*, 2024. 2

[53] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and

T. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2

[54] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 1, 2

[55] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 2017. 1, 2, 6, 7

[56] Yin Wang, Hao Lu, Ying-Cong Chen, Li Kuang, Mengchu Zhou, and Shuiguang Deng. rppg-hiba: Hierarchical balanced framework for remote physiological measurement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2982–2991, 2024. 3, 6

[57] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20587–20596, 2022. 6

[58] C.-Y. Wu, J. Johnson, J. Malik, C. Feichtenhofer, and G. Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9065–9075, 2023. 2

[59] Lin Xi, Weihai Chen, Changchen Zhao, Xingming Wu, and Jianhua Wang. Image enhancement for remote photoplethysmography in a low-light environment. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 1–7. IEEE, 2020. 6

[60] Y. Xu, Z. Shi, W. Yifan, H. Chen, C. Yang, S. Peng, Y. Shen, and G. Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 3

[61] Bryan P Yan, William HS Lai, Christy KY Chan, Alex CK Au, Ben Freedman, Yukkee C Poh, and Ming-Zher Poh. High-throughput, contact-free detection of atrial fibrillation from video with deep learning. *JAMA cardiology*, 5(1):105–107, 2020. 6

[62] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. Pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[63] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement. *international conference on computer vision*, 2019. 1, 2

[64] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Processing Letters*, 28:1290–1294, 2021. 1

[65] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H. S. Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. *arXiv: Computer Vision and Pattern Recognition*, 2021. 1, 2, 7

[66] K. Zhang, S. Bi, H. Tan, Y. Xiangli, N. Zhao, K. Sunkavalli, and Z. Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. 3

[67] Xinyu Zhang, Weiyu Sun, Hao Lu, Ying Chen, Yun Ge, Xiaolin Huang, Jie Yuan, and Yingcong Chen. Self-similarity prior distillation for unsupervised remote physiological measurement. *IEEE Transactions on Multimedia*, 2024. 2

[68] Yuting Zhang, Hao Lu, Xin Liu, Yingcong Chen, and Kaishun Wu. Advancing generalizable remote physiological measurement through the integration of explicit and implicit prior knowledge. *arXiv preprint arXiv:2403.06947*, 2024. 1, 2

[69] Yuting Zhang, Hao Lu, Qingyong Hu, Yin Wang, Kaishen Yuan, Xin Liu, and Kaishun Wu. Period-llm: Extending the periodic capability of multimodal large language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29237–29247, 2025. 2

[70] Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. Rhythmmamba: Fast remote physiological measurement with arbitrary length videos. 2025. 2