# 1 Expanded Results

We present additional benchmark evaluation results using the rPPG-Toolbox [Liu et al. 2023], utilizing unsupervised methods POS [Wang et al. 2016] and pre-trained supervised neural networks - TS-CAN [Liu et al. 2020], motion-augmented (MA) TS-CAN [Paruchuri et al. 2024], PhysFormer [Yu et al. 2022], and FactorizePhys [Joshi et al. 2024] - shown in Table 1 and representative scatter plots in Figure 1.

**Table 1: Additional Benchmark Evaluation Results. Performance of benchmark methods on our rendered 💓 Heartian avatar videos, evaluated across the UBFC-rPPG [Bobbia et al. 2019], PURE [Stricker et al. 2014], and MMPD [Tang et al. 2023] datasets. Green highlights indicate metrics where signals remain recoverable on par with the benchmark results on the same subsets.**

| | | Test Set | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | UBFC-rPPG | | | | | PURE | | | | | MMPD | | | | |
| **Method** | **Train Set** | MAE | RMSE | MAPE | $\rho$ | SNR | MAE | RMSE | MAPE | $\rho$ | SNR | MAE | RMSE | MAPE | $\rho$ | SNR |
| POS | - | 1.23 ± 0.77 | 2.75 ± 2.43 | 1.01 ± 0.62 | 0.98 ± 0.05 | 0.26 ± 0.92 | 14.23 ± 6.78 | 25.75 ± 18.04 | 29.34 ± 13.98 | 0.48 ± 0.30 | 2.93 ± 0.56 | 1.53 ± 0.63 | 7.43 ± 5.53 | 2.66 ± 1.20 | 0.82 ± 0.04 | 4.20 ± 0.25 |
| TS-CAN | UBFC-rPPG | - | - | - | - | - | 10.01 ± 4.52 | 17.46 ± 12.54 | 17.89 ± 8.68 | 0.68 ± 0.25 | -3.31 ± 0.66 | 2.59 ± 0.52 | 6.60 ± 4.60 | 3.89 ± 0.94 | 0.85 ± 0.04 | -1.71 ± 0.17 |
| | PURE | 9.22 ± 7.47 | 25.37 ± 24.62 | 7.41 ± 5.56 | 0.23 ± 0.34 | -4.9 ± 1.14 | - | - | - | - | - | 2.51 ± 0.57 | 7.02 ± 4.85 | 3.87 ± 1.03 | 0.83 ± 0.04 | -1.76 ± 0.16 |
| TS-CAN (MA) | UBFC-rPPG | - | - | - | - | - | 4.74 ± 4.22 | 14.18 ± 13.80 | 5.23 ± 4.67 | 0.82 ± 0.19 | -0.28 ± 0.96 | 0.97 ± 0.27 | 3.35 ± 2.63 | 1.21 ± 0.31 | 0.96 ± 0.02 | -0.40 ± 0.15 |
| PhysFormer | UBFC-rPPG | - | - | - | - | - | 17.40 ± 6.55 | 27.06 ± 18.36 | 34.29 ± 13.78 | 0.50 ± 0.30 | -2.45 ± 0.69 | 8.69 ± 1.31 | 17.41 ± 7.87 | 13.96 ± 2.21 | 0.25 ± 0.08 | -2.62 ± 0.19 |
| | PURE | 5.53 ± 4.44 | 15.11 ± 14.60 | 4.37 ± 3.32 | 0.64 ± 0.26 | -2.34 ± 1.00 | - | - | - | - | - | 3.12 ± 1.08 | 12.80 ± 8.73 | 5.27 ± 1.95 | 0.58 ± 0.07 | 0.18 ± 0.19 |
| FactorizePhys | UBFC-rPPG | - | - | - | - | - | 14.58 ± 6.82 | 26.06 ± 18.21 | 29.95 ± 14.13 | 0.48 ± 0.30 | 0.58 ± 0.74 | 2.88 ± 0.98 | 11.69 ± 7.62 | 4.83 ± 1.70 | 0.62 ± 0.06 | 1.70 ± 0.22 |
| | PURE | 6.41 ± 5.09 | 17.32 ± 16.77 | 5.05 ± 3.80 | 0.55 ± 0.29 | -1.26 ± 1.00 | - | - | - | - | - | 1.38 ± 0.50 | 5.91 ± 4.55 | 2.40 ± 0.99 | 0.88 ± 0.04 | 1.66 ± 0.19 |

MAE = Mean Absolute Error (BPM), RMSE = Root Mean Square Error (BPM), MAPE = Mean Percentage Error (%), $\rho$ = Pearson Correlation, SNR = Signal-to-Noise Ratio (dB).

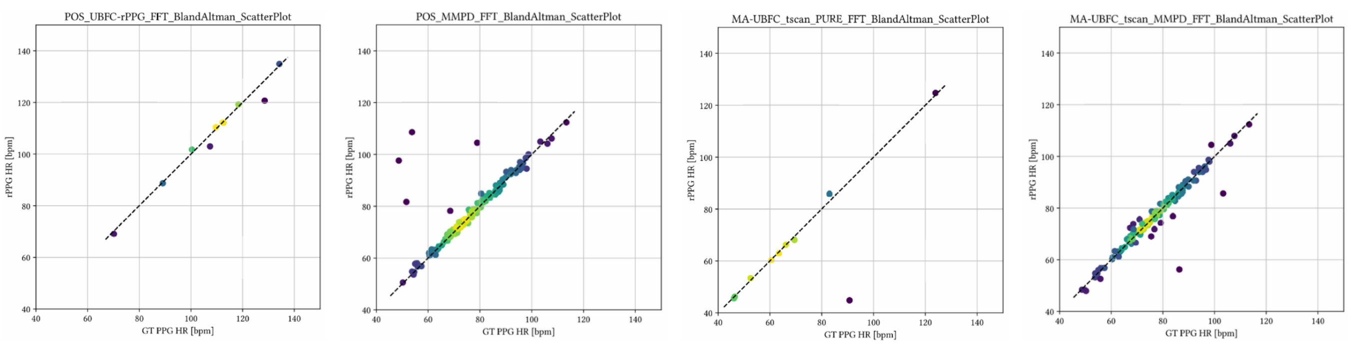


**Figure 1: Visualization of Recovered Heart Rate Accuracy. Scatter plots of rPPG estimates from our rendered 💓 Heartian videos versus ground truth using POS for UBFC-rPPG and MMPD, and motion-augmented TS-CAN for PURE and MMPD.**

# 2 Implementation Details

Facial skin Gaussians are selected based on their absolute distance to the nearest mesh vertex in canonical space, with a per-dataset threshold {0.006, 0.01, 0.016} for UBFC-rPPG [Bobbia et al. 2019], PURE [Stricker et al. 2014], and MMPD [Tang et al. 2023], respectively, to account for scale and Gaussian density differences across optimized avatars by differing camera setups and capture conditions. For POS-based [Wang et al. 2016] weak supervision, we adopt two configurations depending on dataset characteristics. For MMPD [Tang et al. 2023], which involves compressed mobile phone footage, we apply a 1.6s sliding window to obtain locally stable phase estimates. For the other two, which provide uncompressed or high-fidelity recordings, we adopt a simplified global POS formulation, as sliding windows tend to introduce spurious phase shifts. Regarding optimizer settings, all scheduled parameters follow an exponential decay: $\delta_k$ and $\phi_0$ decay over 60,000 frame-level iterations from $9 \times 10^{-3}$ to $3 \times 10^{-4}$ and from $2 \times 10^{-3}$ to $5 \times 10^{-4}$, respectively, while $f_{\text{MLP}}$ decays over 80,000 iterations from $3.8 \times 10^{-3}$ to $5 \times 10^{-4}$. The fundamental waveform parameters use fixed learning rates: $3 \times 10^{-4}$ for $A_1, A_2$; $8 \times 10^{-4}$ and $1 \times 10^{-3}$ for $\mu_1, \mu_2$; $6 \times 10^{-4}$ and $2 \times 10^{-3}$ for $\sigma_1, \sigma_2$; $1 \times 10^{-4}$ for $B$; and $9 \times 10^{-5}$ for $A$. We apply softplus activation to $A_1, A_2, \sigma_1$, and $\sigma_2$ to enforce physical positivity and resist rapid shrinkage caused by photometric loss, which initially tends to suppress albedo modulation by driving them towards zero.

# 3 Ablation Study and Future Work

*Background effects.* The segmented foregrounds are composited onto a black background during avatar rendering by the default setting. To verify that our evaluation results are not confounded by the foreground matting, we add the temporal real backgrounds back onto the rendered avatars via alpha compositing as an ablation study. Under the same evaluation setting, metrics slightly change across methods and selected subsets, without a consistent degradation, as shown in Table 2. In general, pretrained models were trained on data containing real backgrounds. Background composition naturally restores the global image statistics that pre-trained neural networks rely on after normalization. Specifically, PURE's [Stricker et al. 2014] background is more complex and realistic, unlike the single-color backdrops of others. Consequently, models pretrained on PURE tend to benefit more from background composition. This effect also relates to the sensitivity to global spatial context across methods. Meanwhile, we noticed that motion-augmented TS-CAN [Paruchuri et al. 2024] consistently maintains strong and stable performance, highlighting its robustness to the temporal mesh jittering inherently introduced by avatar rendering. This points to a potential direction of combining motion augmentation with physiology-aware avatars to generate robust synthetic rPPG data, which could be incorporated into future systems to improve the real-world and camera-specific rendering for cross-dataset generalization.

**Table 2: Background Compositing Ablation. Metric changes from background compositing over the default black background.**

| | | Test Set | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | UBFC-rPPG | | | | | PURE | | | | | MMPD | | | | |
| **Method** | **Train Set** | MAE | RMSE | MAPE | $\rho$ | SNR | MAE | RMSE | MAPE | $\rho$ | SNR | MAE | RMSE | MAPE | $\rho$ | SNR |
| POS | - | + 0.87 | + 1.15 | + 0.87 | - 0.01 | - 0.94 | + 0.52 | + 0.05 | + 0.62 | - 0.01 | - 0.22 | + 3.04 | + 5.45 | + 5.62 | - 0.24 | - 2.09 |
| TS-CAN (MA) | UBFC-rPPG | - | - | - | - | - | 0.00 | 0.00 | 0.00 | 0.00 | + 0.06 | - 0.30 | - 1.90 | + 0.30 | + 0.01 | + 0.04 |
| FactorizePhys | UBFC-rPPG | - | - | - | - | - | - 0.17 | - 0.30 | - 0.38 | + 0.01 | - 0.11 | - 0.05 | - 0.01 | - 0.06 | 0.00 | - 0.10 |
| | PURE | 0.00 | 0.00 | 0.00 | 0.00 | - 0.05 | - | - | - | - | - | - 1.37 | - 4.47 | - 2.90 | + 0.13 | - 0.07 |

**Figure 2: Relighting Ablation Visualization. From left to right: white, cool, warm, and real-world hospital lighting (Source Link); two subjects from PURE (left) and UBFC-rPPG (right). Solid color environment maps are synthesized with matched luminance.**

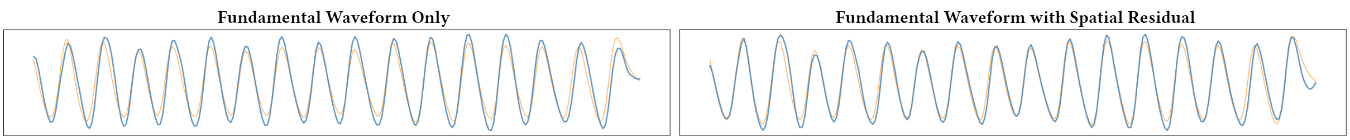


**Figure 3: Ablation on Spatial Residual of Our Modulation Components. Both ours (blue) and GT (orange) are filtered and normalized. Full modulation shows a better alignment of the prescribed signals, with 0.05 MACC improved in this case.**

*Phase Parameterization.* We further experiment with regressing the direct phase coordinate $(x(t), y(t))$ on the unit circle per frame $t$ as an alternative. However, this approach empirically underperformed, as independent per-frame prediction lacks the temporal continuity that the cumulative formulation naturally enforces, making it harder to learn a coherent phase trajectory. While cumulative phase representations are theoretically susceptible to error accumulation across frames, our POS-based [Wang et al. 2016] weak supervision acts as a periodicity anchor that continually corrects drift, mitigating the accumulation effect. Nevertheless, a drift-free phase representation across test frames remains a direction for future work.

*Spatial Variation.* Our framework supports spatially varied rPPG signal intensity, as shown in Figure 4. However, it is not explicitly supervised - the pipeline supervises the mean green-channel albedo as the embedded signal. While the MLP implicitly learns non-uniform modulation patterns, the observed spatial structure remains strongly coupled to the underlying green channel albedo. Explicit spatial supervision could enable more principled control over per-region intensity, which we identify as a future direction.

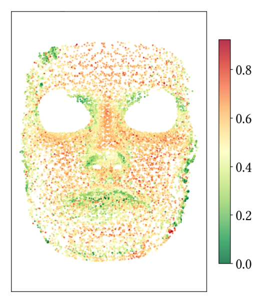


**Figure 4: Mean spatial rPPG intensity across the facial skin region.**

## References


Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. 2019. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters* (2019). doi:10.1016/j.patrec.2017.10.017

Jitesh Joshi, Sos S. Agaian, and Youngjun Cho. 2024. FactorizePhys: Matrix Factorization for Multidimensional Attention in Remote Physiological Sensing. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 96607–96639. doi:10.52202/079017-3063

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. 2020. Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 19400–19411.

Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Soumyadip Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. 2023. rPPG-Toolbox: Deep Remote PPG Toolbox. arXiv:2210.00716

Akshay Paruchuri, Xin Liu, Yulu Pan, Shwetak Patel, Daniel McDuff, and Soumyadip Sengupta. 2024. Motion Matters: Neural Motion Transfer for Better Camera Physiological Measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5933–5942. doi:10.1109/WACV57701.2024.00583

Ronny Stricker, Steffen Müller, and Horst-Michael Gross. 2014. 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 (RO-MAN 2014)*. 1056–1062. doi:10.1109/ROMAN.2014.6926392

Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. 2023. MMPD: Multi-Domain Mobile Video Physiology Dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 1–5. doi:10.1109/EMBC40787.2023.10340857

Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. 2016. Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering* 64, 7 (2016), 1479–1491. doi:10.1109/TBME.2016.2609282

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. 2022. PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. doi:10.1109/CVPR52688.2022.00415