SUPPLEMENTARY MATERIALS OF

BigSmall: Efficient Multi-Task Learning for Disparate Spatial and Temporal Physiological Measurements

Girish Narayanswamy[1], Yujia Liu[1], Yuzhe Yang[2], Chengqian Ma[1],
Xin Liu[1], Daniel McDuff[1], Shwetak Patel[1]
[1]University of Washington, [2]Massachusetts Institute of Technology
{girishvn, nyjliu, cm74}@uw.edu, yuzhe@mit.edu
{xliu0, dmcduff, shwetak}@cs.washington.edu

A. Overview of Appendices

Here we present additional experimental results and discussion that solidify the findings we discuss in the main publication. We include additional experiments and analysis regarding the facial action unit task and model architecture and ablation studies in Section B. Example waveforms are included in Section C. Details of pre and post processing are included in Section D and Section E. Details of SOTA methods and datasets are included in Section F. Additional discussions regarding broader impacts and future work can be found in Section G. Other discussions may be found in Section H. Code, pre-trained models, and a video figure can be found at our github repository: github.com/girishvn/BigSmall. Additional information can be found at our website: girishvn.github.io/BigSmall.

B. Additional Experiments, Discussions

Here we cover experiments that motivate the need for a unified multi-task physiological model, full and additional AU results for cross-dataset generalization, full AU results for SOTA model comparisons, ablation results regarding BigSmall branch information sharing and fusion, AU results using gray scale inputs, experiments regarding optimal chunk length, and additional experiment detail not outlined in the main publication.

B.1. Cross Modality Pre-training and Fine-tuning

As discussed in the main paper, we run an experiment where we pre-train adaptations of BigSmall on a single physiological signal (either PPG, respiration, or AU), and fine-tune the resulting embedded (resetting the final dense layers, and freezing the remainder of the embedding), on a different signal. These results, relative to single-physiological-task trained and validated results, can be seen in Fig.2. Interestingly, we see a degradation of results in **all** pre-train fine-tune pairs by no less that 23%, and as much as 459%. Note that in the figure lower metrics for PPG and respiration are better, and the higher measures are better for AU. This suggests that the embeddings from one modality do not adapt well to another, and thus that SOTA performance requires individual task-optimized networks. This illustrates the utility of a unified general framework, like BigSmall, that is able to simultaneously learn these disparate signals while making efficiency gains over task optimized models.

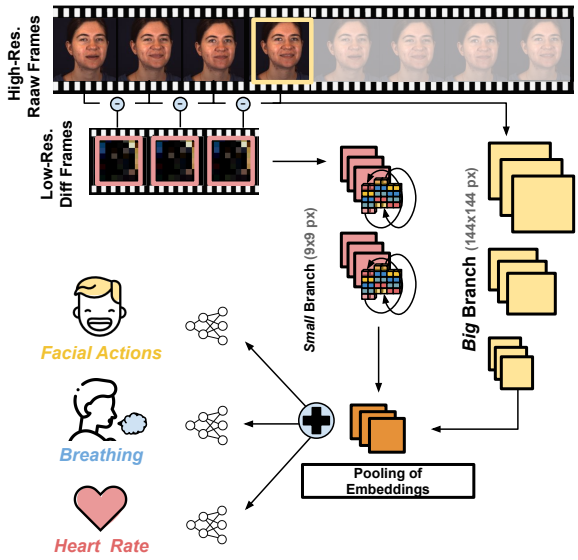

Figure 1. **Overview of the Proposed BigSmall Model.** We present the first joint facial action, cardiac, and pulmonary measurement model from video. By leveraging a dual-branch architecture with wrapped temporal shift modules we achieve strong accuracy with an efficient multi-task implementation.

Table 1. **AU comparisons of the BigSmall model *vs.* literature baselines.** Models trained/tested on BP4D+. † denotes the use of landmark face alignment for the Big input.

Metrics		DRML [5]	AlexNet [6]	Big Pathway	DRML† [5]	AlexNet† [6]	JAA-Net† [12]	JÂA-Net† [13]	BigSmall (Ours)	BigSmall† (Ours)	BigSmall++† (Ours)
AU (F1)	AU01	16.3	24.3	20.7	24.8	30.3	43.2	43.5	22.1	30.0	42.4
	AU02	12.0	19.5	16.5	20.2	26.4	34.7	37.9	18.6	25.7	35.3
	AU04	8.0	12.3	11.4	18.7	17.7	22.9	28.9	12.6	22.0	24.2
	AU06	73.9	72.4	75.6	80.7	81.6	81.7	83.1	70.2	82.7	82.5
	AU07	78.4	79.8	76.4	82.3	84.2	83.6	84.6	73.3	83.1	85.8
	AU10	80.9	82.0	81.6	88.6	88.6	88.0	89.7	74.7	88.7	89.2
	AU12	80.1	78.9	81.6	87.2	87.1	86.6	88.0	73.6	86.4	87.6
	AU14	70.9	72.8	68.5	77.6	79.0	74.2	80.5	67.7	75.6	79.6
	AU15	21.3	13.8	24.0	34.3	30.1	35.5	35.7	26.2	34.0	33.1
	AU17	32.6	24.3	34.4	36.7	37.8	42.9	45.8	29.6	40.7	36.5
	AU23	35.4	36.0	37.1	43.9	42.8	49.0	51.8	38.3	50.2	43.6
	AU24	18.4	14.3	15.6	20.5	23.8	27.0	25.4	12.1	26.0	18.6
AU (Avg)	F1	44.0	44.2	45.3	51.3	52.5	55.8	57.9	43.3	53.8	54.9
	Acc. (%)	74.9	63.1	73.8	78.6	76.5	85.9	85.6	67.4	80.0	86.4

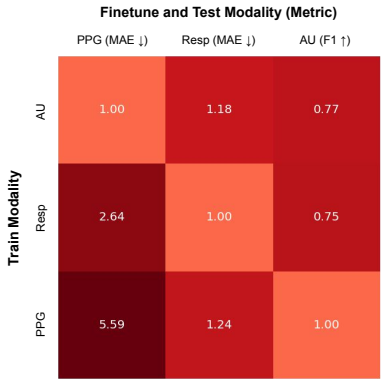

Figure 2. **Cross-Physiological Signal Pre-Training.** Pre-training a BigSmall esc. model on one modality and fine-tuning on another leads to a drop in relative performance for all modalities.

B.2. BP4D+ SOTA Comparison: Full AU Results

We report individual AU results for BigSmall iterations and AU SOTA models presented in the main publication. These models are trained and validated (via 3-fold cross validation) on BP4D+. These results are shown in Table 1.

B.3. Cross-AU-Dataset Generalization

Like prior work [12, 13] we evaluate AU model generalizability by fine-tuning BP4D+ pre-trained models the on DISFA [11] dataset (for 8 common) action units. These action units include the following AUs: 1, 2, 4, 6, 9, 12, 25, 26. In Table 2 we provide AU-level results for the DISFA generalization results presented in the main publication. For these experiments models are trained on BP4D+ for 5 epochs and then refined on DISFA for 2 epochs. We see that iterations of BigSmall out perform common AU baselines and perform similarly to AU SOTA models.

B.4. Fusion / Data Sharing Ablation Experiments

We explore the type of connections used to fuse the Big and Small branches of BigSmall. These results are shown in Table 3.

We first test the use of concatenation of the Big and Small feature maps (as opposed to summing). Concatenation of the features maps results in a negligible difference in performance while significantly increasing the number of parameters in the model due to large output dense layers.

We further explore the use of lateral information sharing of high-level features between the Big and Small branches. These lateral connection occur after the first pooling layer of the Big branch and after the second convolutional layer of the Small branch. We test Big-to-Small, Small-to-Big, and bi-directional lateral connections. Big-to-Small lateral connections temporally upsample and spatial downsample the Big feature map to match the dimensions of the Small branch, and then concatenate these features with the Small branch feature map (along the channel dimension). Small-to-Big lateral connections temporally downsample and spatially upsample the Small branch feature map to match the dimensions of the Big branch, and then concatenate these features with the Big feature map (along the channel dimension). Bi-direction lateral connections utilize both the aforementioned Big-to-Small and Small-to-Big lateral connections.

We find that all methods of high-level information sharing benefit the PPG task. AU performance benefits from Big-to-Small fusion, but regresses considerably with Small-to-Big fusion. Respiration benefits from Small-to-Big fusion, but regresses considerably with Big-to-Small fusion. This suggests that though high level information sharing

Table 2. **Evaluation on Public Spatial Dataset: DISFA** [11]. Following [12, 13], we fine-tune models trained on BP4D+ on DISFA, and evaluate using a 3-fold cross-validation across 8 AUs. All inputs are face-aligned following [12, 13].

Model	DRML [5]	AlexNet [6]	JAA-Net [12]	JÂA-Net [13]	BigSmall	BigSmall++
AU01	18.	11.1	19.9	28.7	18.4	27.6
AU02	15.6	9.7	4.2	35.2	18.4	27.1
AU04	31.7	26.2	36.8	49.3	35.4	45.8
AU06	35.4	39.2	28.5	42.3	45.0	33.7
AU09	24.6	21.8	24.7	23.7	23.6	22.8
AU12	60.8	53.1	60.5	65.8	64.7	63.9
AU25	72.9	69.3	75.8	86.1	84.2	82.6
AU26	46.2	34.3	42.7	43.9	49.5	38.5
Avg. F1 ↑	38.2	33.1	36.6	46.9	42.4	42.7
Avg. Acc. ↑	81.4	74.1	80.9	86.0	80.5	86.7

may benefit all tasks independently, the high level features of interest differ between respiration and AU, preventing a unified lateral connection system that benefits all tasks simultaneously.

B.5. Gray Scale Big Input

Some previous works [5] train AU models using gray scale images which preserve texture information and reduce the number parameters which may cause overfitting. We find that using gray scale Big inputs results in reduced performance for BigSmall. This is likely as the Big branch of BigSmall is able to leverage color-channel-dependent variations embedded in the 3-color-channel Small input difference frames. Results in Table 4.

B.6. Optimal Input Frame Number for Spatial Task

As detailed in the main paper, spatial task performance degrades when trained with a high of number consecutive frames which reduces variance in the training mini batches. We train the AU task-optimized Big branch model using a number of chunked data lengths to empirically illustrate how performance degrades as the number of consecutive frames increases. We observe that there is significant degradation in AU task performance after N exceeds 9. For our experiments we use N = 3 to highlight the abilities of BigSmall and the Wrapping Temporal Shift Modules in situations that necessitate small N due to training or latency considerations. This is highlighted in Fig. 3.

B.7. Additional Experiment Details

Face Aligned AU Inputs. AU SOTA [5, 12, 13], use face-aligned images, which drastically improve AU multilabel classification results. This face alignment, as implemented by [12, 13] involves a "similarity transformation including in-plane rotation, uniform scaling, and translation... This transformation is shape-preserving and brings no change to the expression" [13]. It should be noted that this transformation requires pre-annotated facial landmarks and additional preprocessing, reducing the efficiency of AU networks.

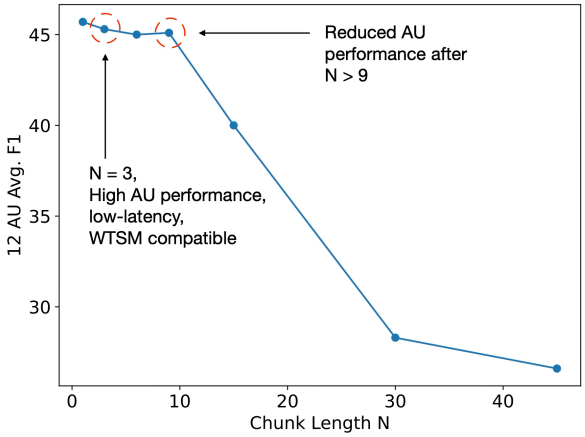

Figure 3. **Consecutive Frames N vs Avg. 12 AU F1.** These 12 AU average F1 scores, from the Big branch model trained with a number of different consecutive frames N, shows that AU performance degrades as the N increases.

BigSmall Adaptations For Face Aligned Inputs. We find that BigSmall performs significantly better with face aligned inputs when batch normalization layers are added to the Big Branch. We add these batch norm layers after the first, third, and fifth convolutional layers in the Big branch.

JAA-Net Loss Functions. Though BigSmall and other AU networks use weighted binary cross entropy as the loss function, we use the custom loss functions described in [12, 13] for JAA-Net and JÂA-Net.

C. Example Waveforms

Fig. 4 illustrates additional PPG and Respiration predictions from BigSmall plotted against the sensor ground truth. NOTE, PPG predictions are plotted against the Blood Pressure waveform (BP4D+ pulse ground truth). This accounts for the similar waveform frequency content but phase-misalignment. Similar animated waveform plots may be found in our video figure.

D. Preprocessing

D.1. Video Frame Inputs

Raw and normalized difference inputs are processed to match the preprocessing of [8]. The described transforms are performed per-video before the videos are chunked. Before each frame is transformed, the frames are center cropped, along the vertical axis, in order to produce square frames.

Table 3. **BigSmall Branch Data Sharing Ablation:** We compare different designs of data sharing between the two branches of BigSmall.

Model	Fusion Method	Lateral Connection	Heart Rate				Breathing Rate				AU Avg.		Computation	
			MAE	RMSE	MAPE	ρ	MAE	RMSE	MAPE	ρ	F1	Acc	FLOPS (M)	# Params (M)
BigSmall	Sum	Bi-Directional	**2.21**	**5.46**	**2.55**	**0.91**	3.93	5.54	18.98	0.10	**46.9**	**72.3**	172.35	2.16
BigSmall	Sum	Big-To-Small	2.32	5.84	2.62	0.89	3.80	5.39	18.42	0.12	46.0	69.5	154.76	2.15
BigSmall	Sum	Small-To-Big	2.37	5.96	2.70	0.89	**3.37**	**4.99**	**16.48**	0.19	40.6	61.4	171.60	2.15
BigSmall	Concat	–	2.28	5.68	2.58	0.90	3.72	5.28	17.94	0.15	43.5	67.3	156.00	4.13
BigSmall	Sum	–	2.38	6.00	2.71	0.89	3.39	5.00	16.65	**0.21**	43.3	67.4	**154.01**	**2.14**

Table 4. **Comparison of BigSmall With Gray Scale Big Input.** Best results of each row are in **bold**.

Metrics		BigSmall w/ Gray Scale Big Pathway Input	BigSmall (Ours)
Heart Rate	MAE	**2.29**	2.38
	RMSE	**5.75**	6.00
	MAPE	**2.59**	2.71
	ρ	**0.89**	**0.89**
Resp. Rate	MAE	3.62	**3.39**
	RMSE	5.26	**5.00**
	MAPE	17.63	**16.65**
	ρ	0.18	**0.21**
AU (F1)	AU01	19.6	**22.1**
	AU02	18.1	**18.6**
	AU04	11.5	**12.6**
	AU06	65.0	**70.2**
	AU07	71.3	**73.3**
	AU10	71.2	**74.7**
	AU12	68.9	**73.6**
	AU14	**68.0**	67.7
	AU15	25.2	**26.2**
	AU17	24.8	**29.6**
	AU23	35.1	**38.3**
	AU24	8.7	**12.1**
AU (Avg)	F1	40.6	**43.3**
	Acc. (%)	61.3	**67.4**

Small Inputs (Normalized Difference Frames). Normalized difference frames are derived by taking the difference of a frame $k[n]$ and a frame $k[n+1]$ such that $k_{diffnorm}[n] = (k[n+1] - k[n])/(k[n+1] + k[n])$. This denominator normalization factor helps to reduce dependence on per-frame-skin brightness and appearance [8]. The resulting frames are mean and standard deviation standardized. These frames are then downsampled to 9x9px.

Big Input (Raw Frames). The raw frames are mean and standard deviation standardized. The resulting frames are then downsampled to 144x144px. As described above as well, we also generate a version of the BigSmall dataset where the Big inputs are land-mark face aligned.

D.2. Data Labels

Label Preparation. Following previous work [8, 19], the respiration and PPG labels are difference noramlized, to match the format of the Small branch difference frame inputs. This is done such that for a sample $k[n]$, $k_{diffnorm}[n] = (k[n+1] - k[n])/(k[n+1] + k[n])$. The resulting samples are mean and standard deviation standardized. AU labels are not difference normalized as the spatial branch (Big branch) inputs are not difference normalized.

PPG Pseudo Labels. Early explorations indicated an ineptitude of BigSmall to effectively learn the PPG signal when trained on blood pressure waveform labels (the BP4D+ ground truth heart signal). Thus, we train the PPG task using "pseudo" PPG labels derived using the Plane Orthogonal-to-Skin (POS) [18] method. These POS-derived signals are then aggressively filtered using a 2nd Order Butterworth filter centered on the mean heart rate (accounting for 20 BPM variation), derived by using a FFT-based calculation on the sensor-ground truth blood pressure waveform. The minimum and maximum filter cutoff frequencies were set to normal heart-rate frequencies of [0.70, 3] Hz. The amplitude of the resulting signals are then normalized using the Hilbert envelope. Although these "pseudo" labels are used to train, all models are still evaluated against BP4D+'s ground truth blood pressure waveform which shares the PPG signal's heart rate frequency.

AU Labels. BP4D+ has labels for 34 AU activations. We choose to use 12 of these AUs for training and evaluation based off previously published literature [5, 12, 13] and as these 12 AUs (1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 24) have sufficient positive occurrences in the dataset. Some AU activations in both BP4D+ and DISFA are labeled as intensities [0-5], where 0 is no activation and 5 is maximum activation. For DISFA, following previous work [5, 12, 13] we use 8 AUs (1, 2, 4, 6, 9, 12, 25, 26) for fine-tuning and evaluation. Following previously published work we train and test using binarized AU activation (0 for inactive, 1 for activate regardless of intensity) for both AU datasets.

BP4D+ Data Splits. We split the BP4D+ dataset into the following 3 subject-independent splits, used for 3-fold cross-validation. Note that all splits have approximately equal participants, and approximately equal subjects of each

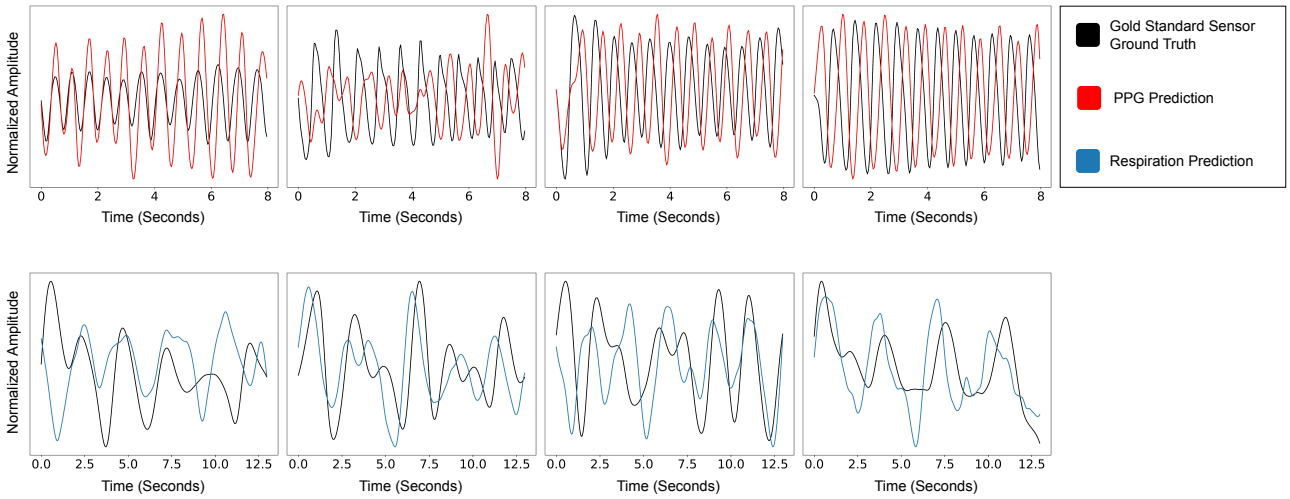

Figure 4. **Sample PPG and Respiration Waveforms.** BigSmall PPG ad respiration waveforms plotted against the sensor ground truth. Note that PPG predictions are plotted again the blood pressure waveform, the BP4D+ heart-signal ground truth.

biological sex. "F" denotes female subjects, while "M" denotes male subjects.

Split 1: F003, F004, F005, F009, F017, F022, F028, F029, F031, F032, F033, F038, F044, F047, F048, F052, F053, F055, F061, F063, F067, F068, F074, F075, F076, F081, M003, M005, M006, M009, M012, M019, M025, M026, M031, M036, M037, M040, M046, M047, M049, M051, M054, M056

Split 2: F001, F002, F008, F018, F021, F025, F026, F035, F036, F037, F039, F040, F041, F042, F046, F049, F057, F058, F060, F062, F064, F066, F070, F071, F072, F073, F077, M001, M002, M007, M013, M014, M022, M023, M024, M027, M029, M030, M034, M035, M041, M042, M043, M048, M055

Split 3: F078, M008, F080, M011, F014, M033, F020, M010, M052, M057, M017, M038, F030, F051, M032, F013, F011, F015, F016, F065, M015, M020, F007, F050, F010, M021, F012, F045, F059, M045, F023, M004, F069, M044, M053, M018, M058, M050, F019, F024, F034, F079, M039, F056, F054, F027, F043

Excluded Samples BP4D+. We exclude the following samples from the BP4D+ dataset due labeling issues (mismatch length, missing data, etc.):

F001T8, F010T10, F013T6, F014T8, F015T6, F016T6, F019T4, F022T7, F024T4, F024T9, F027T4, F028T8, F029T9, F030T7, F030T9, F033T6, F033T7, F033T8, F036T6, F038T1, F041T7, F043T1, F043T10, F043T7, F047T7, F048T7, F051T4, F054T7, F059T4, F061T4, F061T7, F062T4, F062T8, F067T4, F068T7, F072T4, F073T4, F077T4, F078T9, F081T4, M005T5, M005T7, M009T10, M009T4, M009T7, M011T8, M014T4, M014T7, M017T10, M017T7, M019T3, M023T10, M024T1, M024T2, M030T4, M033T1, M033T9, M035T6, M041T4, M041T7, M042T7, M046T1, M047T10, M047T7, M049T6, M049T7, M051T4, M055T8.

DISFA Fine-Tuning Splits. We split the DISFA dataset into the following 3 subject-independent splits, used for 3-fold cross-validation model fine-tuning. Note that all splits have approximately equal participants.

Split 1: SN001, SN002, SN003, SN004, SN005, SN006, SN007, SN008, SN009

Split 2: SN010, SN011, SN012, SN013, SN016, SN017, SN018, SN021, SN023

Split 3: SN024, SN025, SN026, SN027, SN028, SN029, SN030, SN031, SN032

E. Postprocessing

E.1. Heart and Respiration Rate From Waveform

PPG and respiration waveform labels are difference normalized to match the temporal branch inputs. Thus predictions are also in a difference normalized form. PPG and respiration waveforms are derived from the difference normalized waveforms by taking the cumulative sum of the waveform at every sample and then detrending the resulting vector.

Signal rates are then derived by applying a 2nd Order Butterworth filter with cut-off frequencies of [0.75, 2.5] Hz for heart rate and [0.08, 0.5] Hz for respiration rate to the signal waveforms and using a peak detection algorithm on the Fourier spectrum of the filtered signals.

E.2. AU Model Prediction Thresholding

AU outputs from the final model layer are passed through a sigmoid function to bound the output (0,1). We use a threshold of 0.5 to binarize the output of the sigmoid such that AU sigmoid output < 0.5 = 0 (inactive) and AU sigmoid output ≥ 0.5 = 1 (active).

E.3. Heart and Respiration Rate Evaluation Metrics

Mean Average Error (MAE). The MAE as defined between the predicted signal rate R_{pred} and the ground truth signal rate R_{gt} for a total of T instances:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|R_{gt} - R_{pred}|$$

Root Mean Square Error (RMSE). The RMSE as defined between the predicted signal rate R_{pred} and the ground truth signal rate R_{gt} for a total of T instances:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(R_{gt} - R_{pred})^2}$$

Mean Average Percent Error (MAPE). The MAPE as defined between the predicted signal rate R_{pred} and the ground truth signal rate R_{gt} for a total of T instances:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{R_{gt} - R_{pred}}{R_{gt}}\right|$$

Pearson Correlation (ρ). The Pearson correlation as defined between the predicted signal rate R_{pred} and the ground truth signal rate R_{gt} for a total of T instances, and $\overline{R}$ the mean of R over T instances:

$$\rho = \frac{\sum_{t=1}^{T}\left(R_{gt.t} - \overline{R_{gt}}\right)\left(R_{pred.t} - \overline{R_{pred}}\right)}{\sqrt{\left(\sum_{t=1}^{T}R_{gt.t} - \overline{R_{gt}}\right)^2\left(\sum_{t=1}^{T}R_{pred.t} - \overline{R_{pred}}\right)^2}}$$

E.4. AU Evaluation Metrics

F1. The F1 as defined between a list of predictions and ground truth labels, where TP is the true positive count, FP is the false positive count, and FN is the false negative count:

$$100 * \frac{2TP}{2TP + FP + FN}$$

Accuracy (%). The accuracy as defined between a list of predictions and ground truth labels, where TP is the true positive count, TN is the true negative count FP is the false positive count, and FN is the false negative count:

$$100 * \frac{TP + TN}{TP + TN + FP + FN}$$

F. SOTA Methods and Dataset Descriptions

F.1. Temporal Task Baselines

An implementation of these rPPG baseline methods may be found in [10].

DeepPhys [2]. A dual pathway convolutional neural network for PPG estimation. The network utilizes attention from the "Appearance Branch" which models the location of skin pixels, to assist the "Motion Branch" which models changes in skin color correlated to the pulse signal.

MTTS-CAN [8]. An efficient dual pathway convolutional neural network for PPG and respiration multi-tasking. The network utilizes attention from the "Appearance Branch" which models the location of skin pixels, to assist the "Motion Branch" which models changes in skin color correlated to the pulse signal. The "Motion Branch" makes use of Temporal Shift Modules [7] to share information between time samples.

EfficientPhys [9]. An efficient implementation of a convolutional rPPG network that utilizes a single-branch architecture. The network makes use of normalization and learnable normalization modules as well as self attention.

POS [18]. A signal processing method that utilizes the individual color channel (R, G, B) signals. These signals are split into overlapping window segments. For each window segment each color channel signal is normalized by its mean. The PPG signal for that window is then calculated through a relationship between the original color channel signals and mean signals. The final PPG signal is reconstructed by piecing together the overlapping window segments.

CHROM [4]. A signal processing method that utilizes chrominance signals to derive the PPG signal. The method filters the individual color channel (R, G, B) signals around the normal heart rate frequency, and then windows the signals into overlapping segments. A relationship between the color-channel-based signals is then used to derive the PPG signal windows. The resulting segments are further Hanning-windowed and pieced together using an overlapping add technique to obtain the final PPG signal.

F.2. Spatial Task Baselines

JAÂ-Net [13]. A network that achieves SOTA performance. This convolutional network simultaneously learns action unit activations and facial landmarks. This multi-tasking results in stabilized learning of action units and facial features.

JAA-Net [12]. An earlier iteration of [13]. This network has a similar architecture to its successor with a slightly difference in layers used for local AU extraction and the definition of the loss function.

DRML [5]. Deep Region and Multi-Label Learning is a convolutional network that utilizes region learning to better

isolate regions of the face in which different AUs activate. The use of a "region layer" helps the model learn spatial information regarding individual AU's without incurring the computational cost of needing to isolate individual pixels as is done by [17].

AlexNet [6]. A convolutional network used to baseline image classification tasks. It consists of a number of convolutional and pooling layers before a number of fully connected layers.

F.3. Multi-Task (PPG + Resp + AU) Datasets

BP4D+ [20–22] The BP4D+, a large multimodal emotion dataset, consists of face video (25fps) from 140 participants (82 female, 58 male). Each participant records 10 trials, each of which is meant to elicit a specific emotional response: *happiness*, *surprise*, *sadness*, *startle*, *skepticism*, *embarrassment*, *fear*, *pain*, *anger*, *disgust*. These trials are labeled with the following signals: blood pressure (systolic/diastolic/mean/bp wave), heart rate, respiration (rate/wave), electrodermal activity. Trials 1/6/7/8 are FACs encoded for the most "facially expressive" portion. We refer to the portion of the dataset with AU labels as the AU subset (consisting of 200k frames). This AU subset is the only portion of the dataset with concurrent AU, respiration, and PPG labels.

F.4. PPG Datasets

PURE [16]. A dataset comprised of RGB video recordings (30fps) from 10 participants (2 female, 8 male). Participants are seated and front lit with ambient light from a window. Each subject participates in 6 recordings, each with the individual performing different motion tasks. The dataset contains ground truth, contact-sensor-based, PPG and SpO2 measurements.

UBFC [1]. A dataset comprised of RGB video recordings (30fps). Participants are seated and lit with ambient light. The dataset contains ground truth, contact-sensor-based, PPG measurements.

F.5. AU Datasets

DISFA [11]. A dataset comprised of 4 minutes of RGB video recordings (20fps) per 27 subjects. Each frame of the dataset is manually FACS coded for 12 AUs (AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU17, AU20, AU25, AU26) with an intensity measure [0-5].

G. Broader Impacts and Future Work

Potential Risks and Mitigation Strategy Physiological sensing has a wide range of potentially positive applications in health sensing. However, there is also the potential for "bad actors" to use these technologies in negative or negligent ways. Therefore, it is crucial to consider the implications of improving the accuracy, availability, and scalability of sensing methods of this kind. To mitigate negative outcomes, we have taken steps to license our models and code using responsible behavioral use licenses [3].

Application To Other Domains. Though BigSmall is evaluated on physiological sensing tasks, we believe that such a model may allow multi-tasking in other domains in which modeling disparate spatiotemporal signals may be of interest. We hypothesize that a BigSmall-esc. model may show significant benefit in situations where the modeled signals are more related (shared task-gradient direction) than those presented in this work.

COVID-19. The COVID-19 pandemic has catalyzed interest in remote medicine and health sensing via ubiquitous technologies (e.g., a mobile phone) [14, 15]. However, the sensitive nature of biometrics often dictates that these models run on-device. Mobile sensing requires the use of efficient networks that can be run in near-real-time without significant computational limitations.

Future Work. Future work entails the evaluation of BigSmall on resource constrained platforms such as mobile devices and embedded processors. We also plan to train BigSmall on videos with dynamic backgrounds (as BP4D+ has blank background), and utilize additional data augmentation techniques to help build a more robust embedding. Finally, we intend to explore the use of different model backbones for both the Big and Small branches.

H. Other Discussions

H.1. Task Specifications of WTSM

We find that an early, preprint version of [7] (arxiv.org/abs/1811.08383v1) also explored the use of a temporal shift module that wraps features to fill zeroed fields. This "circulant shift" TSM was found to underperform the zero-padded TSM for full video understanding tasks (e.g., activity recognition). In contrast, our Wrapping Temporal Shift Module (WTSM) is designed to build a time-invariant mapping of input-output pairs (e.g., the regression mapping from a frame to the PPG value at that frame). Furthermore, the "circulant shift" was validated on video-level understanding, where the number of consecutive frames, N, is high. In contrast, our WTSM is designed to build robust embeddings when N is extremely low - case in which zero-padded would result in a detrimental proportion of zeroed features.

References

[1] 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. 7

[2] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional atten-

tion networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 6

[3] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 778–788, 2022. 7

[4] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 6

[5] Zhao Kaili, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016. 2, 3, 4, 6

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2, 3, 7

[7] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 6, 7

[8] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020. 3, 4, 6

[9] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5008–5017, 2023. 6

[10] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Yuntao Wang, Soumyadip Sengupta, Shwetak Patel, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. *arXiv preprint arXiv:2210.00716*, 2022. 6

[11] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2, 3, 7

[12] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 705–720, 2018. 2, 3, 4, 6

[13] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021. 2, 3, 4, 6

[14] Anthony C Smith, Emma Thomas, Centaine L Snoswell, Helen Haydon, Ateev Mehrotra, Jane Clemensen, and Liam J Caffery. Telehealth for global emergencies: Implications for coronavirus disease 2019 (covid-19). *Journal of telemedicine and telecare*, 26(5):309–313, 2020. 7

[15] Xuan Song, Xinyan Liu, and Chunting Wang. The role of telemedicine during the covid-19 epidemic in china—experience from shandong province, 2020. 7

[16] 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. 7

[17] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 7

[18] 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. 4, 6

[19] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. Simper: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations*, 2023. 4

[20] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013. 7

[21] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 7

[22] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 7