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

Abstract

Understanding of human visual perception has historically inspired the design of computer vision architectures. As an example, perception occurs at different scales both spatially and temporally, suggesting that the extraction of salient visual information may be made more effective by attending to specific features at varying scales. Visual changes in the body, due to physiological processes, also occur at varying scales and with modality-specific characteristic properties. Inspired by this, we present BigSmall, an efficient architecture for physiological and behavioral measurement. We present the first joint camera-based facial action, cardiac, and pulmonary measurement model. We propose a multi-branch network with wrapping temporal shift modules that yields efficiency gains and accuracy on par with task-optimized methods. We observe that fusing low-level features leads to suboptimal performance, but that fusing high level features enables efficiency gains with negligible losses in accuracy. We experimentally validate that BigSmall significantly reduces computational cost while achieving comparable results on multiple physiological measurement tasks simultaneously with a unified model.
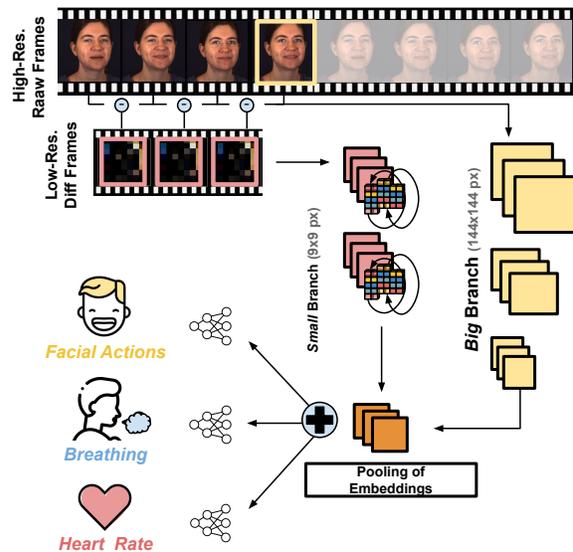
Figure 1. **Overview of the Proposed BigSmall Model.** We present the first joint facial action, cardiac, and pulmonary measurement model from video. By leveraging a dual-branch architecture with wrapped temporal shift modules we achieve strong accuracy with an efficient multi-task implementation.

1. Introduction

Human visual perception occurs at both coarse and fine scales. Attending to coarse *spatial* scales enables a quick estimate of the input to activate scene schemas in memory, while attending to fine scales allows for further refinement [38]. Motion perception is biased towards slower *temporal* motions that are more likely to occur in nature than faster ones [46]. As a result, effective machine learning models for many visual tasks are developed by explicitly constructing networks that harness diverse spatial and temporal scales [6, 15]. Furthermore, global and local features have proven their efficacy in video representation tasks for creating superior models across a wide range of tasks, in-

cluding image object detection [47], sequence classification (e.g., action recognition) [49], and fine-grained temporal understanding (e.g., lip reading) [56].

The measurement of human physiology also requires an understanding of processes with different spatial and temporal features and dynamics. For example, *facial actions* (muscle movements) are idiosyncratic, localized and sporadic, whereas the *human pulse* is almost invariably present in nearly all skin tissue while being highly periodic. *Respiration* or *breathing*, on the other hand, lies somewhere in between, being generally periodic but occasionally irregular, and is only measurable from certain parts of the body (e.g., chest or abdomen). It would seem that the op-

timal spatial and temporal features for measuring these signals would differ. However, facial expressions and cardiopulmonary signals do have shared properties: they are all controlled in part by the autonomic nervous system [13], they are all measured via analysis of the body, and more specifically can be captured through examination of the human face [30, 32]. Thus, even though the low level feature representation of these tasks may seem dissimilar, it may be possible to concurrently learn all these features from a single input modality, and thus suggests that shared information at some scales might benefit performance. Despite these links, there is no empirical evidence to validate or invalidate this hypothesis in video measurement.

Concretely, remote measurement of the human pulse, via photoplethysmography (PPG), leverages aggressive spatial averaging to boost the signal-to-noise ratio of the subtle changes in blood flow present in video pixels [36, 45]. In such situations, leveraging *temporal* information becomes notably advantageous [4, 28], where temporal neural models consistently exhibit superior performance compared to their frame-based counterparts [25, 49, 53]. On the other hand, vision-based facial action recognition requires higher *spatial* resolution features and treats frames as uncorrelated [39, 40]. Limited gains have been observed in facial action recognition through temporal modeling.

While vision-based facial action and physiological measurements have received separate attention, there has been little exploration of multi-task models capable of predicting multiple signals simultaneously. This gap is particularly remarkable, considering the evident similarities between these tasks: both signals originate from the *same* facial regions and exhibit *notable* correlations [3]. Ideally, the goal is to extract all relevant signals from the shared input efficiently, facilitating various downstream tasks [12].

To understand the capacity of a *single* model to generalize across different physiological signals, we train models on a single modality (i.e., PPG, respiration, or AU), and fine-tune the learned embeddings on other modalities. Interestingly, we observe that across different pre-train/fine-tune combinations, **all** models underperform task-optimized models by a large margin (see supplementary materials). This highlights the need for a *unified* and *adaptable* framework that generalizes to diverse signals more efficiently.

To fill this gap, we propose **BigSmall**, the first multi-task neural model for disparate spatial and temporal human physiological measurements. Specifically, BigSmall is comprised of a "Big" branch with high-resolution input for deriving *spatial* texture features, and a "Small" branch, with extremely low-resolution inputs that compress noise from spatial features, which models *temporal* dynamics. We demonstrate empirically that leveraging such properties leads to both state-of-the-art (SOTA) level accuracy and efficiency gains via a unified model. To reduce the compute

overhead, we propose *mixed spatial and temporal scales*, which leverage spatiotemporal properties of branch inputs to improve computational efficiency by more than 60%. Finally, we develop an efficient temporal modeling technique, *Wrapping Temporal Shift Module (WTSM)* to improve temporal feature representation, particularly when only a limited number of frames are available. Extensive evaluations on the tasks of vision-based facial action, respiration, and pulse measurements demonstrate the utility of BigSmall.

To summarize, we make the following **contributions**:

- We present BigSmall, the first *multi-task model for disparate spatial & temporal human physiological measurements*, using a unified two-path spatiotemporal network.
- We propose *mixed spatial and temporal scales* for efficient spatiotemporal modeling while retaining accuracy.
- We develop the *Wrapping Temporal Shift Module* for effective temporal learning, especially when applied to a limited number of input frames.
- We evaluate BigSmall on three physiological vision tasks across multiple real-world video-based human physiology datasets and verify the effectiveness and compute benefits of BigSmall as compared to SOTA methods.

We release our code, trained models, and an interface to simultaneously generate facial action, heart, and breathing measurements from video: github.com/girishvn/BigSmall.

2. Background and Related Work

Multi-Scale Models. Scales in networks can take several forms. Global representations often refer to those for tasks such as classification or a whole video sequence, where as local representations refer to those for detection or localization within video frames. Hjelm and Backman assume that information useful for action classification (i.e., global semantics) should be invariant across space and time within a given video [18, 19]. The concept of leveraging global and local features has drawn attention [47, 50, 56]. Zeng *et al.* argue that feature representations can be learnt that generalize to tasks which require global information and those that require local fine-grained spatio-temporal information (e.g., localization) [56]. SlowFast takes an analogous approach in the temporal domain [15], using two branches to model different frequency scales. Exploiting temporal and spatial scales has been effective in the case of rPPG by implementing a SlowFast transformer network [54] and leveraging global-local spatial features [61]. However, to the best of our knowledge previous research has not employed these principles in the context of diverse multi-task physiological measurements.

Facial Action Recognition. The Facial Action Coding System (FACS) [7, 14] decomposes facial movements into muscle activations called action units (AUs). This coding system has been leveraged to correlate facial ex-
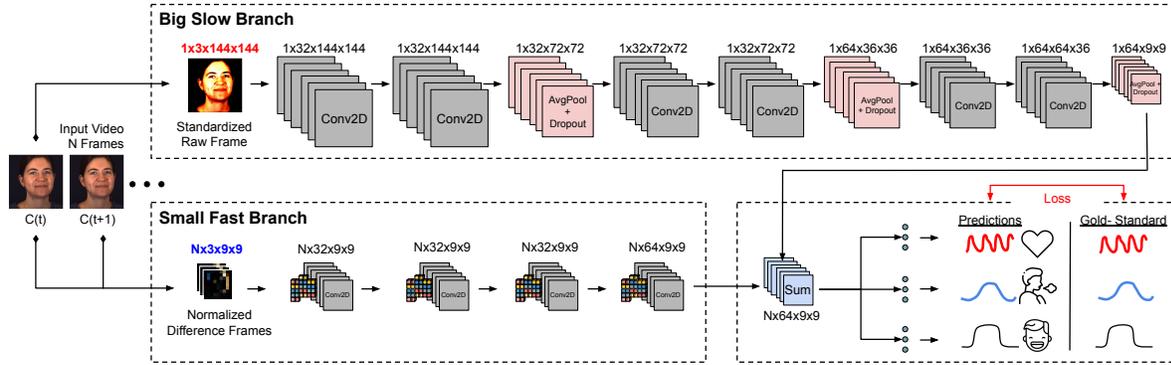
Figure 2. **BigSmall Model Architecture.** By mixing spatial and temporal scales and leveraging Wrapping Temporal Shifts we present an end-to-end efficient multi-task architecture for modeling disparate spatial and temporal physiological signals.

pressions of human emotion (e.g., AU6 - cheek raiser, and AU12 - lip corner puller, together result in a smile - an expression of happiness). Automating FACS using computer vision has a long history due to the laborious nature of manual coding [8, 30]. Recent research has been focused on using deep neural networks for detecting AUs [1, 17, 20, 33]. These task-optimized models make use of high spatial resolution inputs, and often additional pre-processing, or architecture adaptations, utilizing additional features for the representation learning, to achieve SOTA performance [10, 23, 33, 40, 48].

Camera-based Physiological Measurement. Measurement of physiological signals from video is possible as light reflected from the body is modulated by processes such as pulse and breathing [32]. Remote photoplethysmography (rPPG) leverages these subtle modulations to measure the blood volume pulse [36, 43, 45]. Supervised neural networks are the SOTA for rPPG, and similar architectures are often adapted to the breathing task [4, 16, 25, 52–55]. A number of inductive biases inform these model designs. Firstly, as the cardiac pulse is relatively invariant across neighboring skin regions, video frames can be aggressively spatially down-sampled, boosting the pulse signal-to-noise ratio as camera quantization errors average out. Secondly, the pulse signal has characteristic temporal structure and periodicity, therefore implying the benefit of modeling this temporal information [25, 26, 34, 35, 53, 54].

3. Methods

3.1. Modeling Disparate Spatiotemporal Signals

We explore the challenges of learning spatially and temporally disparate tasks, which are perceived at different spatiotemporal scales. For instance, learning periodic physiological signals, such as pulse, requires rich temporal information, relatively high frame rate, and relies on low image resolution to filter irrelevant high-frequency spatial

noise [37, 43]. Conversely, capturing muscle activation features, such as facial actions, demands high spatial resolution to detect subtle texture changes [29]. These activations change more slowly meaning high temporal frequency information is less important. In fact, image-based classification tasks benefit from training with randomized mini-batches to maximize diversity in the data and minimize correlation between individual frames, further underscoring the contrast between spatial and temporal tasks. Finally, breathing, traditionally approached through spatiotemporal methods such as optical flow, can be seen as a time-varying, often periodic task that leverages spatial information (e.g., body motion). Specifically, respiration models leverage higher spatial resolution inputs than rPPG (e.g., Chen *et al.* use 123×123px for respiration and 36×36px for rPPG [5]).

To address the range of temporal and spatial scales needed to simultaneously model these tasks, we propose a multi-task model architecture named "BigSmall" with a high-spatial-resolution branch to capture spatial texture and a low-spatial-resolution temporal branch to capture temporal dynamics. We leverage spatiotemporal scales to reduce the computation of the model, and introduce an efficient technique called Wrapping Temporal Shift Modules (WTSM) to perform temporal modeling within limited temporal context windows. Below we describe the building blocks of this design and our technical contributions, which reduce computation and improve temporal modeling.

3.2. Big: High Resolution Spatial Branch

BigSmall's Big branch is designed to handle tasks that require high spatial fidelity, such as classifying facial actions. To preserve subtle facial edges and textures that comprise many action units, a large input frame resolution is necessary. Unlike the Small branch, the Big branch is not concerned with modeling temporals and generally treats frames as independent. The inputs to the Big pathway are high-resolution standardized raw frames of size $C \times H \times W$
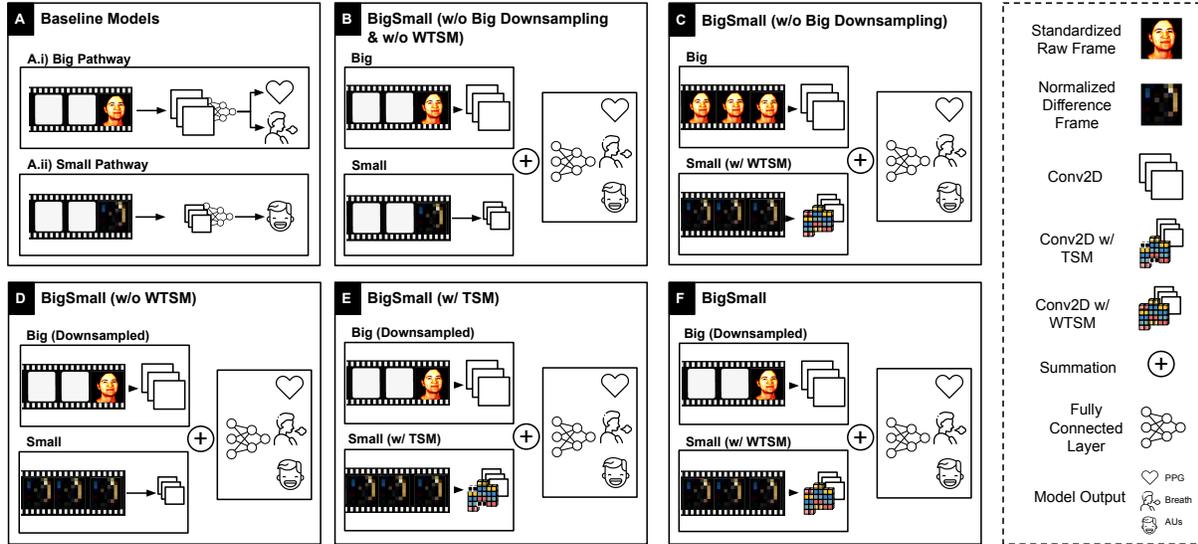
Figure 3. **Model Architecture Iterations of BigSmall.** During our research we designed several candidate networks for multi-task prediction of PPG, breathing, and facial action. These iterations are discussed in Section 3 and in Section 5.3 (ablation studies).

$= 3{\times}H_{\text{big}}{\times}W_{\text{big}}$. In Fig. 3A.i, we summarize the architecture of the Big pathway. The large, raw frames are passed through six convolutional layers with filter depths of [32, 32, 32, 64, 64, 64], and three average pooling layers with dropout, one after every other convolutional layer.

3.3. Small: Low Resolution Temporal Branch

The small branch of the BigSmall model is optimized for tasks that rely on changes between frames rather than specific spatial details. Our proposed Small pathway leverages the fact that temporal tasks require extremely low spatial resolutions, which effectively filters out spatial noise. The Small pathway receives an input image of relatively low resolution defined by $C{\times}H{\times}W = 3{\times}H_{\text{small}}{\times}W_{\text{small}}$. The input is provided in the form of "normalized difference frames", where each input frame represents the difference between a frame K and the subsequent frame $K + 1$. This type of input has been historically used in video-based physiological measurement networks [4, 25] to encode rich temporal information between adjacent time samples. In our evaluation, difference frames encode the pulse signal as changes in color, and motion caused by pulmonary function as edges, as shown in Fig. 3A.ii. These downsampled difference-frames are fed through four convolutional layers with filter depths of [32, 32, 32, 64] in sequence.

3.4. Mixing Temporal and Spatial Scales

Although a model comprised of a fused *Big* pathway and *Small* pathway is able to learn disparate spatio-temporal signals in a multi-task fashion, it provides minimal computational benefit over separate task-optimized networks. For example, such a multi-task network used to predict AU, respiration, and PPG barely improves upon the parameters and floating point operations required to run a big spatial model (for AU) and two small temporal models (for respiration and PPG). Inspired by the SlowFast network [15] which models slow and fast temporal scales in two branches, we propose incorporating different temporal scales on top of our spatial scales to help improve the computational footprint.

Image-based classification models generally assign temporal independence between frames, as slow-changing spatial features result in a lack of interesting temporal dynamics. In the case of facial action, AU activations may stay stagnant for a number of consecutive video frames. Thus, such spatial tasks can be reframed as *slow* temporal tasks, where consecutive frames are highly correlated. Conversely, PPG estimation and similar tasks rely on subtle changes of consecutive low-resolution frames and thus such spatial tasks can be reframed as *fast* temporal tasks.

In BigSmall, as input resolutions of the Big branch are far larger than that of the Small branch, computational load is dominated by convolutions in the Big branch. The ratio of floating point operations (FLOPs) between the Big and Small branches is approximately $(H_{\text{big}}W_{\text{big}})/(H_{\text{small}}W_{\text{small}})$ where H and W denote image height and width. If H_{big} and W_{big} are much larger than H_{small} and W_{small}, the computational burden of the model is driven by the Big branch alone. To address this, we leverage the Big branch to model low-frequency high-spatial resolution signals, while using the small branch to model high-frequency low-resolution signals. By temporally downsampling the inputs to the Big pathway, we reduce the number
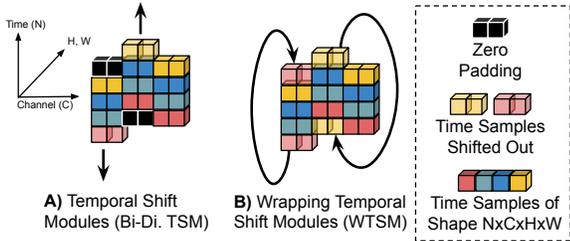
Figure 4. **Wrapping Temporal Shift Modules.** A comparison of temporal shift modules (TSM) and wrapping temporal shift modules (WTSM). For modeling time variant signals with small window sizes, we find that WTSMs provide superior performance.

of frames passed through the Big branch compared to the number of frames passed to the Small branch. By reducing the frames seen by the Big branch to N/M frames passed to Small branch ($M \in \mathbb{Z}^+ > 1$ is the reduction factor and N is the number of Small branch frames) (Figs. 3C and 3F), we reduce the FLOPs executed by the model by approximately M times on average for N frames.

More specifically, the Big pathway receives a single frame ($M = N$) to predict AU activations while the Small branch receives N frames to predict PPG and respiration signals. The Big branch leverages the learned temporal representation and dynamics from the Small branch to infer minor changes in the spatial features of N frames. Heavy reliance on the temporal pathway leads to a small drop in performance, but with the benefit of $1/N$ the original computation cost, where N is the number of frames passed to the Small branch as well as the reduction factor.

3.5. Wrapping Temporal Shift Module

Modeling temporal dynamics and representations beyond consecutive frames are crucial for video-based physiological measurement tasks [25,49,53]. However, the Small branch's difference-frame inputs only allow information sharing between adjacent frames. To address this, temporal shift modules (TSM) [24] are used to learn efficient spatial-temporal representations beyond adjacent frames [25].

However, traditional TSMs fail to build robust temporal representations when the number of input frames (N) is low, as the proportion of zeroed-features increases. As illustrated in Fig. 4, when traditional bi-directional TSMs operate on N frames, $2/(3N)$ of the features are zeroed due to a lack of past and future samples to shift into the first and last time steps. The issue of zero padding becomes particularly problematic when N is **unavoidably small**. Specifically, for AU classification, using a small number of consecutive frames is crucial for achieving optimal performance [40, 57]. However, such small input sequences inevitably introduce challenges in robustly modeling temporal patterns for physiological measurements such as PPG (see Fig. 5).
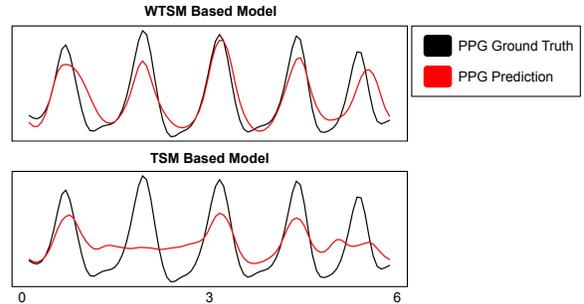

Figure 5. **WTSM vs. TSM for rPPG Prediction.** Models are trained with 9x9px resolution and $N = 3$ consecutive input frames using the TS-CAN [25] backbone. WTSM leads to more robust temporal representations, especially when the N is small.

To address this challenge, we propose the *Wrapping Temporal Shift Module* (WTSM). While traditional TSMs (Fig. 4A) shift out and zero-pad channel-folds from the first and last time-step, WTSMs (Fig. 4B) resolve the problem of zeroed features by wrapping the shifted-out folds to fill the previously zero-padded folds. As confirmed in Fig. 5, WTSM can leverage the inter-frame temporal benefits of TSM without increasing the proportion of zeroed features even with a small N, thereby achieving more robust temporal representations. Note that like TSM, WTSM does not increase parameter or FLOP count incurred by the model.

Also note that unlike long short-term memory networks (LSTMs), the temporal information added by WTSM is not dependent on time-series order. The WTSM helps convolutions learn a time-invariant mapping that maps an input to an output relative to other input-output pairs. As a result, shared features between non-adjacent frames (such as the 1st and Nth frames) do not disturb the temporal representation. Furthermore, wrapping features, as opposed to filling from intermediate frames, best balances the information represented for all N frames.

3.6. The BigSmall Model

By combining the techniques proposed in Section 3.1 to Section 3.5, we present an end-to-end efficient multi-task architecture, called BigSmall, for disparate spatial and temporal signals (see Fig. 2). The proposed architecture leverages a dual pathway system consisting of 1) a Big branch to model fine-grain spatial features from raw high-resolution inputs, and 2) a Small branch optimized for modeling temporal dynamics from low-resolution difference-frames. To achieve computational efficiency, N frames are passed into the Small branch while only 1 frame is passed through the Big branch. This reduction in convolutions in the Big branch leads to a significant reduction in computation by almost a factor of N, as computation of the Small branch is negligible compared to that of the Big branch.

BigSmall benefits from the use of WTSM, which en-

Table 1. **Ablation Studies of BigSmall.** The default BigSmall model is highlighted in gray. **bold** denotes best and **[bold]** denotes second best column results.

Model	Big Branch	Small Branch	Temporal Shift Mechanism	Big Branch Temporal Down Sampling	Heart Rate				Breathing Rate				AU Avg.		Computation	
					MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑	F1↑	Acc↑	FLOPS (M)↓	# Params (M)↓
Small Branch	–	✓	–	–	2.69	6.87	[3.09]	0.86	3.62	[5.18]	17.64	0.15	–	–	3.73	0.70
Big Branch	✓	–	–	–	–	–	–	–	–	–	–	–	45.3	73.8	451.63	0.78
BigSmall	✓	✓	WTSM	–	[2.46]	[6.09]	2.81	[0.88]	3.71	5.28	18.00	0.15	42.5	60.6	456.03	2.14
BigSmall	✓	✓	–	✓	2.47	6.16	2.81	[0.88]	3.65	5.21	17.80	0.16	40.3	62.3	154.01	2.14
BigSmall	✓	✓	TSM	✓	3.03	7.27	3.50	0.85	[3.59]	5.20	[17.63]	[0.17]	43.0	67.3	154.01	2.14
BigSmall	✓	✓	WTSM	✓	**2.38**	**6.00**	**2.71**	**0.89**	**3.39**	**5.00**	**16.65**	**0.21**	[43.3]	[67.4]	154.01	2.14

ables robust derivation of temporal information by passing features between frames. WTSMs are placed before convolutional layers in the Small branch and, when combined with difference-frame inputs, facilitate strong inter-frame feature mapping. This feature is particularly advantageous when training alongside spatial tasks that require high batch-diversity or situations that demand low latency. Additionally, WTSM helps to alleviate the strain put on the temporal branch to infer missing spatial features resulting from temporal down-sampling in the Big branch, by augmenting the temporal representation.

4. Experiments

We evaluate our methods on the tasks of facial action, rPPG, and respiration. We run a series of ablation experiments on the BigSmall model to highlight individual contributions, and compare our model against previously published task-optimized architectures. We train and validate presented models using the BP4D+ dataset [58–60].

Dataset. BP4D+ consists of 10 video tasks from 140 participants (1400 total). Each video is labeled with 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 (200k frames). This AU subset is the only portion of the dataset with concurrent AU, respiration, and PPG labels. We additionally evaluate BigSmall on public rPPG [2, 41], and AU [31] datasets. Additional dataset details are included in the supplementary materials.

Experimental Details. Similar to [21, 39, 40], we use 3-fold cross validation, training on 2 folds and testing on the third, and report performance on the holdout-sets. Due to the sparsity of AU labels and the conflicting nature of the task gradients (explained in Section 5.1), networks are trained on folds from the AU Subset, and validated on data from the entirety of BP4D+ for PPG and breathing tasks. Folds are constructed as to not include subject overlap between train and test sets. Models are trained for 5 epochs, using video chunks of N=3 consecutive frames, an Adam optimizer, and a learning rate of 0.001. The AU multi-label classification task is trained using weighted Binary Cross Entropy Loss. Respiration and PPG are trained with Mean Squared Error Loss. The losses of all 3 tasks are equally weighted and summed to promote equal importance during training. We evaluate binary action unit performance on 12 commonly cited AUs [21, 39, 40] using average F1 and accuracy. PPG and breathing metrics are based on the signal rate (beats/breaths per minute), and for each task we report Mean Average Error (MAE), Root Mean Square Error (RMSE), Mean Average Percent Error (MAPE), and Pearson Correlation (ρ). Additional information regarding SOTA methods, training, metrics, and their derivation can be found in the supplementary material. We adapt our training pipeline from rPPG-Toolbox [27], a toolkit to standardize rPPG deep learning research.

BigSmall Instantiation. The BigSmall model input dimension are chosen to emphasize the different spatial scales of the two branches, and to further highlight the benefits of reducing the computation of the Big branch. The Big branch raw standardized input frames are of shape $C \times H \times W = 3 \times 144 \times 144$. Small branch normalized difference frame inputs are of shape $C \times H \times W = 3 \times 9 \times 9$. The pooling layers of the Big branch are of pool size [2×2, 2×2, 4×4], in order. These pooling sizes are chosen such that the final convolutional output of the Big pathway matches that of the Small pathway, in an effort to balance the feature importance of the branches.

The Big and Small feature maps are combined through upsampling of the Big output and summation in order to prevent extremely large fully connected layers (an artifact of concatenating the Big and Small feature maps before the dense layers) and thus avoid additional model complexity. We explore the use of lateral connections and alternative fusion techniques (discussed in the supplementary material), but find for our tasks, of AU, respiration, and PPG, that mixing high level features, or forcing the combination of low-level features results in performance degradation. As discussed in Section 5.1, this is due to the conflicting gradients of the spatial task (AU) with the temporal task (PPG and breathing). The combined Big and Small feature map is passed to fully connected layers for each learned task.

To match the inputs of the BigSmall model, the PPG and respiration baselines are trained with 9×9 difference-frame inputs, while AU baselines are trained with 144×144 standardized raw inputs. Additionally, following AU SOTA

Table 2. **Comparisons of BigSmall *vs.* SOTA Methods.** BigSmall enables both spatial and temporal human physiological learning simultaneously via a unified model. Supervised models are 3-fold cross validated on BP4D+. **bold** denotes best column results, **[bold]** denotes second best results, and † denotes methods where the "Big" branch input is landmark face aligned and cropped [40].

| Method	Modeling Capability	Heart Rate				Breathing Rate				AU Avg.	
		MAE ↓	RMSE ↓	MAPE ↓	ρ ↑	MAE ↓	RMSE ↓	MAPE ↓	ρ ↑	F1 ↑	Acc ↑
POS [44]	Temporal	10.40	19.53	9.73	0.41	–	–	–	–	–	–
CHROM [11]		5.27	13.28	5.12	0.69	–	–	–	–	–	–
EfficientPhy [26]		8.86	15.91	9.85	0.41	–	–	–	–	–	–
MTTS-CAN [25]		2.86	7.19	3.27	0.85	3.88	5.54	18.88	0.11	–	–
DeepPhys [4]		2.68	6.67	3.07	0.86	**[3.51]**	**[5.16]**	**[17.06]**	**[0.20]**	–	–
Small Branch		2.69	6.87	3.09	0.86	3.62	5.18	17.64	0.15	–	–
DRML [21]	Spatial	–	–	–	–	–	–	–	–	44.0	74.9
AlexNet [22]		–	–	–	–	–	–	–	–	44.2	63.1
Big Branch		–	–	–	–	–	–	–	–	45.3	73.8
DRML † [21]		–	–	–	–	–	–	–	–	51.3	78.6
AlexNet † [22]		–	–	–	–	–	–	–	–	52.5	76.5
JAA-Net † [39]		–	–	–	–	–	–	–	–	**[55.8]**	**[85.9]**
JÂA-Net † [40]		–	–	–	–	–	–	–	–	**57.9**	85.6
Big Branch †		–	–	–	–	–	–	–	–	53.4	79.5
BigSmall	Spatial + Temporal	**2.38**	**6.00**	**2.71**	**0.89**	3.39	**5.00**	**16.65**	**0.21**	43.3	67.4
BigSmall †		**[2.51]**	**[6.10]**	2.88	**[0.88]**	3.93	5.45	18.94	0.12	53.8	80.0
BigSmall++ †		**[2.51]**	6.15	**[2.86]**	**[0.88]**	4.12	5.64	19.62	0.10	54.9	**86.4**
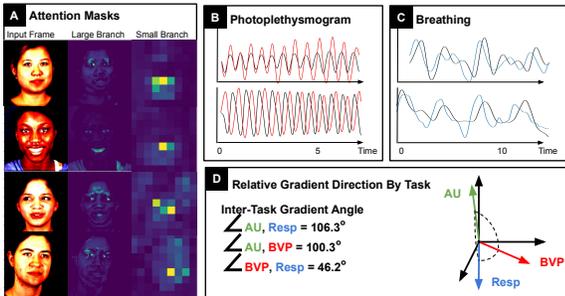
Figure 6. **Examples Outputs.** Attention masks, predicted signals, and relative training gradients. PPG and respiration share some gradient direction. The AU gradient conflicts with these tasks.

models that utilize facial landmarks for face-alignment and cropping [39, 40], we train adaptations of BigSmall that incorporate these added preprocessing steps for the Big branch. We further instantiate a variant of BigSmall, **BigSmall++**, which leverages the JÂA [40] Big backbone. This highlights the efficiency benefits of using task-optimized architectures with the BigSmall framework.

5. Results and Discussion

5.1. Multi-Task AU and Physiological Measurement

The BigSmall model is able to concurrently learn disparate spatiotemporal tasks. We show that the network enables multi-task measurement of facial action units, breathing, and PPG. Table 1 illustrates that BigSmall performs comparatively to the baseline Big and Small single-task-

optimized models, while reducing the computational load needed to run 3 task-specific models by ~66%. Fig. 6A, 6B and 6C show sample attention maps from the Big and Small branch, and PPG and respiration predicted waveforms plotted against the sensor ground truth.

Regression in AU results, as compared to the Big baseline, is explained by an analysis of the multi-tasked signals. Pulse and respiration signals are known to have shared information, in that respiration frequencies can be derived via respiratory sinus arrhythmia (RSA) [36]. Conversely, though AU may leverage some temporal dynamics modeled by the Small branch, the spatial feature representation is expected to share much less information with the PPG and respiration tasks. Indeed, we verify these hypotheses by observing the task-gradients during training [42, 51]. The BigSmall task-gradient-vectors, calculated after the first training epoch, are shown in Fig. 6D. While PPG and respiration signal gradients project onto each other ($\angle$ PPG, Resp = 46.2°), the gradients for the AU signal are much more different ($\angle$ AU, Resp = 106.3°, $\angle$ AU, PPG = 100.3°). This gradient conflict results in a degradation of AU results.

5.2. Computational Efficiency

BigSmall benefits from improved computational efficiency by temporally downsampling the Big slow-spatial signal inputs. Since convolutions of the Big inputs dominate computation, temporal downsampling by a factor of N reduces FLOPs by a factor of N. When paired with the augmenting temporal capacity of WTSMs, which do not increase compute cost, BigSmall functions using only

Table 3. **Evaluation on Public Temporal Signal Datasets: UBFC** [2] and **PURE** [41]. Models are trained on BP4D+. **bold** and **[bold]** denote best and second best column results. Note how TSM-based models struggle when the frame count is low ($N = 3$).

Method	PURE (rPPG)				UBFC (rPPG)			
	MAE↓	RMSE↓	MAPE↓	ρ↑	MAE↓	RMSE↓	MAPE↓	ρ↑
POS [44]	7.89	11.08	10.65	0.89	2.79	4.69	3.25	[0.97]
CHROM [11]	7.29	[10.33]	10.06	[0.90]	3.13	5.11	3.68	[0.97]
EfficientPhy [26]	8.07	24.92	11.81	0.66	9.21	17.11	10.32	0.63
MTTS-CAN [25]	5.99	13.01	7.08	0.74	12.78	22.43	13.90	0.47
DeepPhys [4]	4.73	11.83	5.81	0.78	3.36	12.86	3.37	0.69
BigSmall	**1.97**	**6.48**	**2.56**	**0.93**	**1.03**	**2.55**	**1.14**	**0.99**
BigSmall++	[2.87]	10.44	[3.10]	[0.90]	[1.43]	[3.10]	[1.62]	**0.99**

a fraction of the compute required to run three standalone task-optimized networks (BigSmall: 154M FLOPs vs Big + 2xSmall: 549M FLOPs), while producing comparable predictions as shown in Table 1. Similarly BigSmall++ demonstrates comparable performance to JÂA-Net [40] in Table 2, while utilizing significantly reduced computational costs (e.g., only one-third the computation costs when $N = 3$).

5.3. Ablation Studies

Leveraging Scales to Improve Performance. Table 1 illustrates a 66% reduction in FLOPs in BigSmall as compared to a similar model without Big input downsampling. Interestingly, such design also enjoys notable performance improvements over all metrics, indicating that BigSmall is both computationally efficient and achieves better multi-task performance compared to other design choices.

Improving Temporal Dynamics. The Wrapping Temporal Shift Module assists convolutional layers to better model temporal dynamics even when consecutive input frames are forcibly limited by latency or training conditions. Table 1 illustrates that for an input chunk of $N = 3$ consecutive frames, BigSmall outperforms both a model without temporal shift and a model using traditional bi-direction TSM. Additionally, this demonstrates that the use of traditional TSM, with small N, results in a high proportion of zeroed features and thus a drop in performance. We note that the performance of the spatial AU task improves with the use of WTSM, suggesting that the Small branch temporal dynamics are leveraged to infer missing spatial features caused by temporal downsampling of Big input frames. The ablations of BigSmall are visualized in Fig. 3.

5.4. Comparisons To SOTA Models

We compare our BigSmall models against task-optimized models from the literature. Table 2 demonstrates that our instantiations of BigSmall are comparable to SOTA convolutional AU baselines [21, 22, 39, 40], and exhibits consistent and significant gains when computationally-intensive face-alignment and cropping are utilized. These results further illustrate the performance of BigSmall as

Table 4. **Evaluation on Public Spatial Dataset: DISFA** [31]. Following [39, 40], 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 [39, 40].

Model	DRML [21]	AlexNet [22]	JAA-Net [39]	JÂA-Net [40]	BigSmall	BigSmall++
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

compared to state-of-the-art rPPG and breathing models [4, 25, 26], and unsupervised methods [11, 44]. As shown, existing methods are only capable of performing either the spatial task (i.e., AU detection) or the temporal tasks (i.e., heart and breathing rate) at one time. In contrast, BigSmall enables simultaneous spatiotemporal human physiological measurements with comparable or better performance.

5.5. Cross-Dataset Generalization

We further evaluate the generalization ability of BigSmall to unseen data. We compare BigSmall against other baseline models trained on BP4D+ and tested on two public rPPG datasets: **UBFC** [2] and **PURE** [41]. Table 3 confirms that BigSmall outperforms the other competitors across all evaluated metrics with substantial performance gains. Moreover, these improvements are consistent on both datasets, indicating that BigSmall learns meaningful spatiotemporal information that can generalize to unseen datasets. Following [39, 40] we test AU generalizability by fine-tuning BP4D+ trained embeddings on **DISFA** [31]. Table 4 validates that instantiations of BigSmall perform comparatively to state-of-the-art AU methods.

6. Conclusion

We present BigSmall, the first example of a multi-task architecture for facial action unit, pulse, and respiration measurement from video. BigSmall demonstrates the ability of a unified model to efficiently learn spatially and temporally disparate signals with no significant drop in performance when compared to SOTA task-optimized methods. Additional experiments, AU results, and discussion regarding BigSmall can be found in the supplementary materials. We acknowledge several limitations and potential societal risks of our work. BP4D+ consists of blank-background videos, uncommon in the real word. We do not evaluate our model's performance on compute limited platforms (e.g., embedded devices). Finally, there is the potential for "bad actors" to use these technologies in negligent ways. It is crucial to consider the implications of improving accuracy, availability, and scalability of these methods. We have taken steps to license our methods using responsible behavioral use licenses [9], and look forward to exploring diverse data sources, restricted compute platforms, and additional applications of BigSmall in future work.

References

[1] Carlos Fabian Benitez-Quiroz, Yan Wang, and Aleix M Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, pages 3990–3999, 2017. 3

[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. 6, 8

[3] Mihai Burzo, Daniel McDuff, Rada Mihalcea, Louis-Philippe Morency, Alexis Narvaez, and Verónica Pérez-Rosas. Towards sensing the influence of visual narratives on human affect. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 153–160, 2012. 2

[4] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018. 2, 3, 4, 7, 8

[5] Weixuan Chen and Daniel McDuff. DeepMag: Source-Specific Change Magnification Using Gradient Ascent. *ACM Transactions on Graphics*, 40(1):2:1–2:14, Sept. 2020. 3

[6] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. In *International Conference on Learning Representations*, 2016. 1

[7] Jeffrey F. Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the Facial Action Coding System. In J. A. Coan & J. J. B. Allen, editor, *Series in affective science. Handbook of emotion elicitation and assessment*, chapter Observer-b, pages 203–221. Oxford University Press, 2007. 2

[8] Jeffrey F Cohn and Fernando De la Torre. Automated face analysis for affective computing. In *The Oxford handbook of affective computing*, page 131. Oxford University Press, 2014. 3

[9] 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. 8

[10] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 298–313, 2018. 3

[11] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 7, 8

[12] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015. 2

[13] Paul Ekman, Robert W Levenson, and Wallace V Friesen. Autonomic nervous system activity distinguishes among emotions. *science*, 221(4616):1208–1210, 1983. 2

[14] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 2

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 2, 4

[16] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3995–4004, 2021. 3

[17] Amogh Gudi, H Emrah Tasli, Tim M Den Uyl, and Andreas Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–5. IEEE, 2015. 3

[18] R Devon Hjelm and Philip Bachman. Representation learning with video deep infomax. *arXiv preprint arXiv:2007.13278*, 2020. 2

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[20] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016. 3

[21] 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. 6, 7, 8

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 7, 8

[23] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019. 3

[24] 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. 5

[25] 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. 2, 3, 4, 5, 7, 8

[26] 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. 3, 7, 8

[27] 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

[28] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12404–12413, 2021. 2

[29] Simon Lucey, Ahmed Bilal Ashraf, and Jeffrey F Cohn. *Investigating spontaneous facial action recognition through aam representations of the face*, volume 2. Citeseer, 2007. 3

[30] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017. 2, 3

[31] 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. 6, 8

[32] Daniel McDuff. Camera measurement of physiological vital signs. *arXiv preprint arXiv:2111.11547*, 2021. 2, 3

[33] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11917–11926, 2019. 3

[34] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention. *arXiv:2010.07770 [cs, eess]*, Oct. 2020. arXiv: 2010.07770. 3

[35] Akshay Paruchuri, Xin Liu, Yulu Pan, Shwetak Patel, Daniel McDuff, and Soumyadip Sengupta. Motion matters: Neural motion transfer for better camera physiological sensing. *arXiv preprint arXiv:2303.12059*, 2023. 3

[36] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010. 2, 3, 7

[37] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 3

[38] Philippe G Schyns and Aude Oliva. From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological science*, 5(4):195–200, 1994. 1

[39] 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, 6, 7, 8

[40] 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, 5, 6, 7, 8

[41] 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. 6, 8

[42] Wei-Cheng Tseng. Weichengtseng/pytorch-pcgrad, 2020. 7

[43] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 3

[44] 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. 7, 8

[45] 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, July 2017. Conference Name: IEEE Transactions on Biomedical Engineering. 2, 3

[46] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604, 2002. 1

[47] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 1, 2

[48] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2021. 3

[49] 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. 1, 2, 5

[50] Yuzhe Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[51] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020. 7

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

[53] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 2, 3, 5

[54] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, pages 1–24, 2023. 2, 3

[55] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4186–4196, 2022. 3

[56] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021. 1, 2

[57] Cheng Zhang, Cengiz Öztireli, Stephan Mandt, and Giampiero Salvi. Active mini-batch sampling using repulsive point processes. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, pages 5741–5748, 2019. 5

[58] 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. 6

[59] 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. 6

[60] 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. 6

[61] Changchen Zhao, Menghao Zhou, Zheng Zhao, Bin Huang, and Bing Rao. Learning spatio-temporal pulse representation with global-local interaction and supervision for remote prediction of heart rate. *IEEE Journal of Biomedical and Health Informatics*, 2023. 2