Combining Magnification and Measurement for Non-Contact Cardiac Monitoring

Ewa M. Nowara*, Daniel McDuff†, Ashok Veeraraghavan*
*Rice University, Houston, TX
†Microsoft Research, Redmond, WA
{emn3,vashok}@rice.edu, damcduff@microsoft.com

Abstract

Deep learning approaches currently achieve the state-of-the-art results on camera-based vital signs measurement. One of the main challenges with using neural models for these applications is the lack of sufficiently large and diverse datasets. Limited data increases the chances of overfitting models to the available data which in turn can harm generalization. In this paper, we show that the generalizability of imaging photoplethysmography models can be improved by augmenting the training set with "magnified" videos. These augmentations are specifically designed to reveal useful features for recovering the photoplethysmogram. We show that using augmentations of this form is more effective at improving model robustness than other commonly used data augmentation approaches. We show better within-dataset and especially cross-dataset performance with our proposed data augmentation approach on three publicly available datasets.

1. Introduction

Imaging photoplethysmography (iPPG) [1] is a set of approaches to measure vital signs from videos without directly touching the skin. Contactless measurements of vital signs are advantageous in several scenarios, including patients with injured or sensitive skin (e.g., premature babies or burn victims), long-term measurements where wearing a contact sensor may hinder the participants, or sleep monitoring where wearing a contact sensor might make it difficult to fall asleep naturally.

Deep learning methods achieve state-of-the-art results on many computer vision tasks, including iPPG [2, 3, 4, 5, 6]. However, many of the best performing deep learning architectures require large training datasets to achieve good performance. Unfortunately, many computer vision applications have limited availability of such large datasets required to train these large models. This lack of appropriate training datasets limits the performance of deep learning models and often leads to overfitting to the small training set.

Video datasets used for traditional computer vision tasks, such as action recognition have hundreds of thousands of videos. For example, the 20BN-something-something Dataset V2 [7] contains 220,847 videos and the Kinetics-700 dataset has 650,000 video clips [8]. This is two orders of magnitude more than the number of videos in the largest available physiology datasets. For example, the VIPL-HR dataset has 3,130 videos [5] and the AFRL dataset only has 300 [9]. Therefore, it is very hard to train machine learning models, and especially complex deep learning models, on these physiology datasets. Consequently, most existing iPPG work has used heuristics-based non-machine learning methods [10]. Public video datasets for physiological measurements are usually very small because of several challenges associated with dataset collection. They require large storage because the images usually have to be uncompressed. They require complicated synchronization of the ground truth contact sensor with the video capture and access to such a medical-grade ground truth sensor. Moreover, there are often privacy issues with recording and publicly releasing face videos and physiological information. But perhaps the biggest challenge is that each time we want to explore a new application area in iPPG, we have to collect a new large dataset to train a model to work in that setting. For example, if we are interested in sleep monitoring or driver monitoring, we have to collect a dedicated dataset for this task that would be large enough to train a deep learning model. The data collection process is very slow and expensive, making it hard to make progress on problems in new applications.

To address the challenge of limited data, several data augmentation approaches have been proposed in the computer vision and deep learning communities. These augmentation methods can involve simple image manipulations such as rotating the image by varying degrees, translating it by a different number of pixels horizontally and verti-

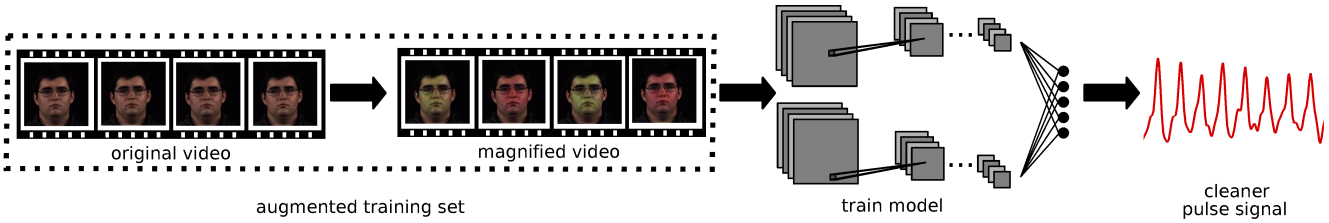

Figure 1. By magnifying the pulse variations in the video, we can create a larger training set. These physiologically relevant augmentations improve the quality of the estimated iPPG signal and increase the generalizability of deep learning models to different datasets.

cally, flipping the image horizontally or vertically, cropping, zooming in, or changing the color of the image. However, the existing data augmentation methods were intended for improving performance on tasks where the signal of interest is the dominant information in the video. For example, in action recognition, often the dominant motion and intensity variations in the video are directly related to the action of interest. These augmentation methods do not generalize well to the iPPG task where the signal of interest is not the dominant signal in the video and is buried in much larger irrelevant variations, such as large motion or ambient light variation. When existing data augmentation approaches are applied to iPPG videos, the model tends to focus on the large, obvious variations highlighted by the augmentation, which are not related to the physiological signal of interest. This leads to the same or even worse performance of the model than without this data augmentation.

In this work, we propose a data augmentation approach appropriate for iPPG videos by using video magnification methods, as illustrated in Fig. 1. We demonstrate that video magnification methods can be successfully used as data augmentation by selectively magnifying the physiological signal without visibly affecting the other motions or variations in the video. Video magnification has been used in the past to improve the performance of iPPG methods by magnifying the physiological variations in the video [11, 12, 12, 13, 14]. However, different from these approaches, we only magnify the videos of the training set and leave the test set videos intact. This ensures that the model is able to learn useful features for iPPG measurement and implicitly learn that videos may have different amplitudes of the iPPG signal. This approach leads to especially large improvements on videos with a low signal-to-noise ratio (SNR), e.g., videos of participants with darker skin types or videos with larger motion.

We use an existing end-to-end convolutional attention neural network architecture which takes a video as input and outputs a predicted iPPG signal [2]. We present results on three publicly available datasets using the most recent state-of-the-art video magnification method called DeepMag [14] which was specifically developed for magnifying physiological signals. We achieve large improvements in heart rate (HR) estimation with our proposed data augmentation, especially when training and testing on very different datasets.

Our results demonstrate that this approach not only improves the overall performance of the deep learning model but it also improves the model's ability to generalize to new and more challenging data. We achieve better performance on videos of participants with dark skin types whose videos have lower iPPG SNR due to higher melanin concentration in the skin which absorbs more light inside the tissue [15]. We also observe improvements when training and testing on videos with different motion and different compression levels, showing that the proposed data augmentation approach can help the model generalize to new and diverse sources of variations in the test set.

2. Related Work

2.1. Imaging Photoplethysmography

IPPG approaches have achieved high accuracy in measuring heart rate (HR) [16, 2], breathing rate (BR) [17, 2], and heart rate variability (HRV) [17, 18] from video recordings. They also show promising results in measuring blood oxygenation (SpO2) [19] and blood pressure (BP) [20]. As the iPPG technology has matured, many approaches have achieved robustness to challenging motion [21, 22, 23, 24, 25, 26, 27], low light settings and varying illumination [28, 29, 30], and video compression [31, 32, 4, 3, 6]. Currently, end-to-end deep learning approaches outperform existing unsupervised methods and achieve state-of-the-art performance in vital sign estimation from video [2, 33, 34, 32, 35, 5, 3, 36, 4, 6]. However, deep learning methods work best when they are trained on data that is similar to the test set data. Hence, these methods often struggle with generalizing to new data that may have different motion [2, 37], different video compression [6], or even participants with different skin types or genders [15]. In this work, we show that the difficulty of cross-dataset generalizability can be overcome by using data augmentation that is appropriate for the physiological measurement.

2.2. Data Augmentation

Data augmentation is commonly used in computer vision tasks [38, 39], such as object classification or object detection [40]. Data augmentation is a solution to avoid over-

fitting when training a model on a limited training dataset. Overfitting occurs when the model perfectly fits the training data but is unable to generalize well to the unseen test samples. Making the training set larger and more diverse with data augmentation can alleviate the overfitting issue without having to alter the network architecture. Data augmentation improves the performance when it can create additional training instances that better resemble the test set. For example, translating and cropping a face in an image may help a face detection or recognition network if the face in a test image is not centered [38]. Commonly used data augmentation involves simple geometric transformations of the image, alterations of the color space, kernel filters, and mixing multiple images. Geometric transformations may involve rotating an image clockwise or counterclockwise within a specified angle range or flipping an image horizontally or vertically [39]. Color space augmentations may involve isolating a single R, G, or B color channel, or manipulating the RGB values to increase or decrease the brightness of an image [39]. Images may also be converted from RGB to a different color space, such as YUV, CMY, HSV, or grayscale [41]. Images can be cropped within a patch of interest or translated up, down, left, or right to create instances with different positions of the object in the frame. Noise can also be injected to the image (e.g., Gaussian noise) to help the network learn more robust features [42]. Similarly, images can be blurred or sharpened by convolving them with an appropriate kernel [43]. Finally, multiple images can be combined together by cropping and rearranging patches together [44] or by adding and averaging pixel intensities from several images [45]. Image classification or object detection models must be robust to different viewpoints, lighting, occlusions, background, or scale. Therefore, these kinds of data augmentations make sense for tasks, such as image classification. However, such augmentations do not necessarily help iPPG algorithms where the model should be robust to motion and illumination variations which affect the amplitude of the signal itself, not only the appearance of the videos.

2.3. Video Magnification

Video magnification methods have been used to amplify and reveal subtle color variations or motions in the video. Early works relied on Lagrangian methods which required accurate tracking of the motion over time using optical flow [46]. But, they were computationally expensive and often worked poorly on objects with varying intensity. Later, Eulerian methods were developed which linearly magnified the pixel intensity variations over time in a fixed video location [11]. These methods were more efficient than Lagrangian methods and were able to magnify very subtle color changes which would not have been possible by using optical flow. The Eulerian approaches first decompose the video spatially using filtering and steerable pyramids [47]. Then, they temporally filter the signal in the video to only magnify the selected frequency band. These methods were later improved by magnifying the phase information obtained with complex steerable filters, instead of the amplitude of the intensity, which worked better for magnifying subtle motions [48]. A more recent approach improved the phase-based Eulerian method by magnifying accelerations, that is deviations of intensity change, instead of the intensity change itself [12]. A learning-based method, similar to the Eulerian approach, offered even more improved magnification [13].

However, all of these approaches require precisely knowing the narrow frequency range of the signal to be magnified, which is not always possible. For example, the human heart rate can vary between 30 and 300 beats per minute (BPM). Providing this frequency range is too broad for these color or motion magnification approaches to work well. Moreover, if the signal of interest and other variations, such as motion, are present in the video within a similar frequency range, these magnification methods are not able to separate the two signals and will result in visible artifacts. To address these challenges, Chen et al. used a deep learning method, called DeepMag, trained to specifically amplify only the pulse signal [14]. It does not require knowing the pulse frequency in advance and it is able to separate the pulse-induced intensities from motion variations if the model was trained on videos with similar motion.

3. Proposed Approach

In this section, we present our proposed approach of data augmentation for physiological signals. We describe the different kinds of data augmentation we have evaluated, the details about training the networks, and the datasets we used for training and evaluation.

3.1. Data Augmentation

We compared the results of the model trained on the original dataset without any augmentations ("No Augmentation" in Table 1) to several data augmentation approaches to increase the size of the training set. Examples of augmentations we used are shown in Fig. 2.

Standard Augmentation. First, we used standard data augmentation approaches commonly used in other areas of computer vision ("Standard Augmentation" in Table 1) to create 10 times more training data. These augmentations included random rotations clockwise and counterclockwise by up to 20 degrees, flipping the video frames horizontally and vertically, and translating the video frames horizontally and vertically by 10% of the frame's width and height. When the images were rotated or translated, we repeated the pixel values at the boundary to preserve the dimensions of the images.

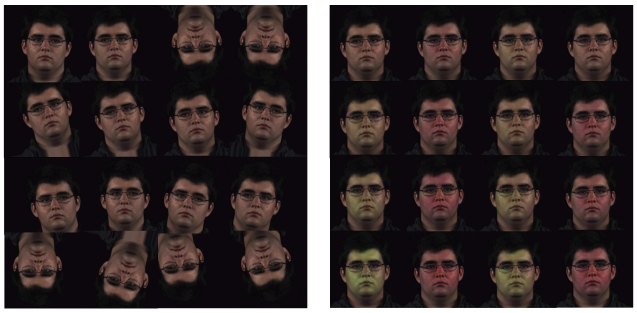

Figure 2. Examples of data augmentation with standard computer vision augmentations (on the left) and the proposed augmentations using iPPG magnification (on the right). The magnification levels illustrated in each row on the right are 4 X, 6 X, 8 X, and 14 X magnification, from top to bottom.

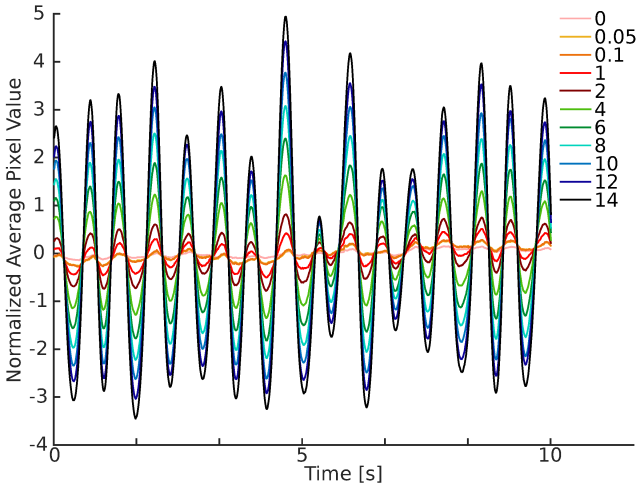

Figure 3. Examples of the green channel waveforms obtained from videos at 10 different magnification levels. Magnification of 0 corresponds to the original video without any magnifications. Using video magnification as data augmentation directly changes the amplitude of the iPPG signals. This allows the network to explicitly learn useful features for iPPG signals. On the other hand, standard computer vision augmentations do not have any impact on the amplitude of the iPPG signal as they do not change the temporal pulsatile intensities in the video.

Pulse Magnification Augmentation. Next, we compared an approach of data augmentation using a magnification of the physiological signals to create 10 times more training data as well. The advantage of these augmentations over standard augmentation approaches is that they selectively alter the iPPG signal and allow the network to pay special attention to the physiologically relevant features. Fig. 3 shows examples of the green channel waveforms obtained from videos at 10 different magnification levels compared to the original video (magnification = 0).

Most existing video magnification approaches magnify motion instead of color variations because they were intended for demonstrating mechanical phenomena, such as the motion of the camera shutter, eye saccades, the motion of a crane, etc. [12, 13]. Few methods were used to magnify the color variations that could be used to reveal the blood flow in the skin [11, 14]. In this work, we use a DeepMag approach [14] for magnifying the iPPG signals.

DeepMag is an end-to-end network which takes an original video as input and outputs a magnified video. It does not need the user to provide the frequency of the pulse signal to be magnified because the CAN model is able to automatically find the correct frequency range corresponding to the pulse. DeepMag is based on the same convolutional attention neural network (CAN) architecture that we use to extract the iPPG signals [2]. Therefore, it is able to correctly find facial regions with strong iPPG signals and the correct iPPG frequency to make the magnified video not only appear realistic but also to be useful for later training a model with the augmented dataset with magnified videos. The magnification is achieved via gradient ascent to visualize the pulse signals on the face. The weights of the CAN model pre-trained on the iPPG extraction task are frozen during the gradient ascent to perform video magnification. We selected a range of 10 magnification levels of 0.05 X, 0.1 X, 1 X, 2 X, 4 X, 6 X, 8 X, 10 X, 12 X, and 14 X. We found that these magnification levels were large enough to visibly magnify the iPPG signal in the video but not too large to avoid artifacts at higher magnifications. We scaled the magnification levels by multiplying them by the temporal standard deviation of the pixel intensities of the input video to be magnified. The results with these magnification experiments are referred to as "Heuristic iPPG Augmentation" in Table 1.

We also compared the video magnification with DeepMag to the non-deep-learning Eulerian Video Magnification (EVM) method [11]. However, we found that the performance with EVM was not as good as with DeepMag. EVM magnifies the entire image frame within a provided frequency range of interest resulting in false pulsatile variations in the background that likely confuse the network during training and the network is unable to learn which regions to focus on in the video.

Joint Training of Pulse Extraction and Magnification. We also tested whether we can jointly train the network to learn how to magnify the videos in the training set while learning to extract the iPPG signal from the videos ("Interleaved iPPG Augmentation" in Table 1). In this approach, we first train the network on videos without any magnifications for 25 epochs to make sure the model has converged sufficiently to extract a good iPPG signal so that it can magnify the videos well. We found that when we began magnifying the videos at earlier epochs, the magnification was erroneous since the model used to magnify the signals has not converged yet and it lead to poor performance. After 25 epochs, we alternated between using the magnified and the

original videos every other epoch. For each epoch where we magnified the videos, we randomly sampled from the same 10 magnification levels used in the "Heuristic iPPG Augmentation" to train the network. We did not change the amplitude of the ground truth pulse signal for any of the data augmentation experiments.

Compared Non-Deep Learning Methods. We compared the performance of the deep learning models trained with and without different data augmentation strategies to several non-deep learning methods, including POS [27], CHROM [26], and ICA [16]. Methods which do not use deep learning do not suffer from overfitting. Therefore, even though these methods are older than the compared deep learning approaches, they often perform better on the cross-dataset experiments, where the deep learning model was trained on a very different dataset from the test set. In order to extract the iPPG signal using POS, CHROM, and ICA, we detected the face in the first video frame using MATLAB's face detection (`vision.CascadeObjectDetector()`). Face detection wasn't necessary for the deep learning method because it operates end-to-end and is able to implicitly learn which regions in the video are likely to contain the iPPG signal. We spatially averaged all facial pixels in the red, green, and blue channels. We then used the three channel traces to apply POS, CHROM, and ICA methods to extract the iPPG signal using the iPhys toolbox [49].

Computing HR: We computed the HR by taking the Fourier transform of the output iPPG signal from each method, finding the frequency with the maximum energy in the power spectrum and multiplying the frequency by 60 to convert it from Hertz (Hz) to beats per minute (BPM). We estimated HR for each non-overlapping 30 second time window and averaged the errors over all time windows and all videos in each dataset. For each compared method, we normalized the extracted iPPG signals by subtracting the temporal mean, dividing by the standard deviation, and we bandpass filtered the signals with pass-band frequencies of [0.7 2.5] Hz.

3.2. Training Details

We used an existing convolutional attention neural network (CAN) to extract the iPPG signal from a video in an end-to-end fashion [2]. The CAN architecture contains the appearance and motion branches joined through an attention mechanism. The appearance branch takes as input a single RGB image and uses it to learn which regions are likely to contain strong iPPG signals, so that the network can selectively focus on those regions and ignore the remaining regions. The motion branch takes as input a normalized difference of two frames and its role is to learn to separate the intensity variations induced by the physiological signal from other variations, e.g., caused by motion. The attention mechanism allows the network to place higher weights on pixels which contain a strong iPPG signal and lower weights on pixels which do not. We trained all models, with and without data augmentation, for 32 epochs and we used a mean squared error (MSE) loss between the predicted and the ground truth iPPG waveforms. Please see [2] for the architecture details.

We trained the CAN [2] on the stationary videos (Task 2) of the AFRL dataset [9]. We used a subject-independent cross-validation, where we trained the model on 40 videos of 20 subjects and tested it on 10 videos of 5 different subjects. The videos were downsampled to 30 frames per second (fps) from the original 120 fps for the efficiency of the training. We chose to train the network only on the easier stationary videos, free of major corruption sources to create a large domain gap between the very easy training set and the very hard test set. In these stationary experiments, the subjects sat still without the headrests to allow for small natural head motion. We tested the trained model on several very challenging datasets, in order to illustrate the benefits of data augmentation for generalizing to different and more difficult datasets.

First, we tested the model on the 10 left-out videos of the stationary AFRL videos (Task 2) [9] to evaluate the within dataset performance. Then we tested the cross-dataset generalizability of differently trained models to 10 AFRL videos (the same test subjects as in the stationary experiments) with very large head motion where the subjects reoriented their heads randomly once every second (Task 6). We also tested the model on all videos of the MMSE-HR dataset [33] which contained different motion and different subjects from the AFRL dataset, and all NIR and RGB videos of the MR-NIRP dataset [28] which contained both stationary and motion experiments. Both MMSE-HR and MR-NIRP contain several dark skin type subjects which makes these datasets additionally challenging.

We report the results with mean absolute error (MAE) between the ground truth and the estimated heart rate within 30 second time windows without overlap, and SNR for each time window. SNR was computed as the area under the power spectrum curve around the first and second harmonic of the ground truth heart rate frequency divided by the rest of the spectrum within 0.7 to 4 Hz. We converted the magnitude of the SNR values to decibels on the log scale.

3.3. Datasets

AFRL [9] contains 300 videos of 25 participants recorded at 120 fps as 8-bit, 658×492 pixel images with a Scout scA640-120gc GigEstandard color camera. Each subject was recorded during 12 experiments with varying head motion, each lasting five minutes. Each motion experiment was recorded with a solid black background and patterned background. The ground truth signals were

recorded using fingertip reflectance photoplethysmograms and electrocardiograms. We used the photoplethysmograms as ground truth to train the network and the electrocardiograms to compute the HR estimation errors. We center-cropped the ARFL video frames to 492×492 pixels to remove the background areas.

MMSE-HR [50] contains 102 videos of 40 participants recorded at 25 fps as 1040×1392 resolution images. The ground truth physiological signals were recorded as blood pressure (BP) wave at 1000 fps and an average HR which was updated after every heart beat. 19 videos had noisy ground truth average HR. We recomputed the HR for those videos by detecting peaks in the blood pressure waveform and computing the interbeat interval (IBI) between them. We estimated HR as $\frac{1}{\mu(IBI)}$ where $\mu(IBI)$ is the mean IBI. We trained the network using the blood pressure waveforms as ground truth signals and the average HR to compute the HR estimation errors. The MMSE-HR recordings were captured during spontaneous emotion elicitation experiments with sudden and uncontrolled motion and facial expressions. This makes the MMSE-HR dataset more challenging than AFRL because there are large and sudden variations in the motion and in the pulse of the subjects. Moreover, this dataset contains subjects with darker skin types which leads to lower SNR of the iPPG signals and especially affects the performance of deep learning models [15].

MR-NIRP [28] contains 15 videos of eight participants simultaneously recorded in RGB and NIR at 30 fps as 10-bit images with 640×640 resolution. FLIR Grasshopper3 GS3-PGE-23S6C-C camera was used to record the RGB videos and Point Grey Grasshopper GS3-U3-41C6NIR-C camera with a 940 nm bandpass filter with 10 nm passband was used to record the NIR videos. For each recording, the exposure was fixed, gamma correction was turned off, and gain was set to zero. The dataset contains stationary experiments where the subjects were asked to sit still and motion experiments where the participants were asked to talk and move their head. We detected the face in each frame and cropped a region around it of 110% width and height of the detected bounding box because the background was not uniform and could affect the performance of the deep learning model. This dataset is challenging because the iPPG signals have lower SNR in NIR [28, 30] and because this dataset contains several subjects with darker skin types which also leads to lower SNR of the iPPG signals [15].

4. Results

We have tested the generalizability of the model trained on the stationary videos of the AFRL dataset [9] with and without the different kinds of data augmentation to different datasets with more challenging conditions. We tested the model on videos with different and larger motion than the motion present in the training set, videos with subjects with darker skin types, and differently compressed videos. These results are summarized in Table 1.

We found that increasing the training set with data augmentations improved the performance, especially when testing on videos that are significantly different from the training set. Standard data augmentations improved the performance on within-dataset experiments but they did not offer improvements on cross-dataset experiments. This is likely because rotating or flipping the images introduced in the training set, did not make the training videos appear more similar to the test set videos. On the other hand, augmentations using magnifications of the iPPG signals, create additional training data with a varying amplitude of the physiological signal. This better resembles the test set videos which also may have a higher or a lower amplitude of the iPPG signal. By using signal magnification as data augmentation, we can train the network to be robust to videos with different SNR and different amplitudes of the iPPG signal.

Different Motion. When we trained the network on stationary videos only, the network did not generalize well to videos with large head motion (**AFRL Motion**) or different kind of motion (**MR-NIRP (RGB)**). Augmenting the training set with pulse magnifications significantly improved the results (Table 1). Unsupervised methods which do not use deep learning do not suffer from overfitting and they can perform reasonably well on different datasets. Therefore, our augmentation approach did not always outperform the compared non-deep-learning benchmark methods (POS, CHROM, or ICA). However, it did consistently perform better than the compared deep learning method without data augmentations ("No Augmentation") and the deep learning method with standard computer vision augmentations ("Standard Augmentation"). This shows that our approach is able to reduce overfitting and it can help the network generalize to new challenging data. Both augmentation approaches using heuristic and interleaved magnification lead to better performance on videos with different motion. However, we obtained the largest improvements when we trained the network interleaved with magnifying the videos during training of the network ("Interleaved iPPG Augmentation"). Our interleaved magnification approach reduced the MAE by as much as 46 % (6.66 BPM) on AFRL videos with large motion and by as much as 42 % (0.52 BPM) on MR-NIRP (RGB) videos over the compared "No Augmentation" baseline. However, it is possible that augmenting one training dataset can potentially lead to further overfitting to that dataset, resulting in worse performance on different test datasets. This could be the reason why we achieve slightly worse performance on the MMSE-HR dataset [50] which has different facial motion.

Different Imaging Modality — NIR. NIR videos are more challenging than RGB for two reasons. First, the SNR

Table 1. Cross-dataset generalizability

	AFRL Still		AFRL Motion		MMSE		MR-NIRP (NIR)		MR-NIRP (RGB)	
	MAE	SNR	MAE	SNR	MAE	SNR	MAE	SNR	MAE	SNR
No Augmentation	1.49	4.20	14.39	-9.11	**3.08**	1.16	2.89	-2.53	1.23	7.91
Standard Augmentation	1.43	3.75	15.34	-10.93	4.84	-1.36	11.67	-5.30	5.01	5.46
Heuristic iPPG Augmentation	1.42	4.41	10.72	-8.53	3.59	0.93	**2.85**	**-2.35**	0.71	7.56
Interleaved iPPG Augmentation	1.41	2.18	7.73	-8.32	3.59	-0.96	6.52	-4.36	0.79	**9.36**
POS [27]	1.28	5.93	**7.23**	**-3.05**	3.90	2.33	-	-	**0.68**	4.98
CHROM [26]	**1.27**	3.97	10.70	-3.32	3.74	1.90	-	-	1.75	3.59
ICA [16]	**1.27**	**6.27**	12.82	-4.87	5.44	**3.03**	-	-	1.57	5.32

of the iPPG signals is an order of magnitude lower in NIR compared to RGB [28, 30, 51]. Second, NIR videos look very different from RGB videos. Therefore, deep learning models only trained on RGB videos will struggle to generalize to the different looking and monochrome NIR videos. We obtained modest improvements in MAE and SNR with the heuristic iPPG augmentation on NIR videos (**MR-NIRP (NIR)** in Table 1). The results could be likely improved if we could include some NIR videos during training with magnification augmentations. However, there are few NIR video datasets which are sufficiently large for training a deep learning model. We could not evaluate the baseline methods, POS, CHROM, and ICA, on the NIR videos, because these methods require three camera channels.

Darker Skin Types. People with darker skin types have a higher melanin content in the skin. This leads to more light being absorbed inside the skin and less light returning to the camera, causing lower SNR of the iPPG signals and less robustness to motion and other sources of variations [15]. Deep learning methods are especially susceptible to worse performance on videos of subjects with darker skin types if the model was trained on videos of predominantly light skin type subjects. Publicly available iPPG video datasets contain very few subjects with darker skin types. Therefore, we combined the videos of subjects with darker skin types V and VI on the Fitzpatrick scale [52] from the MMSE-HR [50] and MR-NIRP (RGB) [28] datasets to create a larger test set. We compared the results on these dark skin type videos to the remaining ones with lighter skin types I - IV. We observed improvements in performance on videos of the more challenging darker skin types with both heuristic and interleaved iPPG augmentations. The improvements are the largest with the interleaved approach, reducing the MAE by almost 6 % (0.14 BPM) and increasing the SNR by as much as 1.18 dB (Table 2). The results on the light skin type videos are already very good with all methods and the improvements with any augmentations are not as apparent.

Video Compression. Video compression removes subtle information, negatively affecting the iPPG signals [31, 53, 3]. Obtaining iPPG signals from compressed videos is particularly challenging for deep learning models because the networks tend to overfit to the compression of the training set videos [6, 4]. We test the performance of different methods on stationary AFRL videos [9] compressed with constant compression rate factors (CRF) = 18, 24, 30, and 36. The original videos used for training the deep learning model and to evaluate the methods reported in Table 1 were already slightly compressed with CRF = 12. All methods are negatively affected by compression, and the results, especially SNR, become consistently worse with increasing compression (Table 3). We notice improvements in performance at higher compression levels (CRF = 24, 30) with our proposed iPPG augmentation methods. The heuristic iPPG augmentation provides the largest improvements on these experiments. However, the interleaved iPPG augmentation approach does not lead to better performance on the experiments with different compression. At high compression of CRF = 36 all methods already perform poorly because the compression artifacts are very large [6].

Comparison to Baseline Methods. Sometimes the unsupervised baseline methods, including POS, CHROM, and ICA performed better than the deep learning method, especially on the challenging cross-dataset results. The reason could be that these methods do not use machine learning and they are not prone to overfitting to the training set. Moreover, ICA uses detrending which often removes a lot of the noise and leads to higher SNR, despite having a larger MAE than the compared methods.

5. Discussion

We only magnify the training set videos as a part of data augmentation and we do not manipulate the test set videos in any way. This justifies using a deep learning magnification method which was trained on the same training set that will be used to train a model to extract the iPPG signals. However, magnifying the test set videos could potentially further improve the performance on videos with very low SNR. Perhaps the model could be trained to learn how much to magnify each video to obtain a reliable signal and it would magnify lower SNR videos more.

Even well-performing video magnification may accidentally magnify other intensity variations in the video in a similar frequency range as the physiological signal. By only training on clean, stationary videos and testing on very chal-

Table 2. Generalizability to videos with different skin types of MMSE-HR and MR-NIRP (RGB)

	Light skin types (I - IV)		**Dark skin types (V - VI)**	
	MAE	SNR	MAE	SNR
No Augmentation	1.86	**4.67**	2.50	5.26
Standard Augmentation	3.99	2.35	6.95	2.42
Heuristic iPPG Augmentation	**1.85**	4.40	2.38	4.90
Interleaved iPPG Augmentation	1.91	3.87	**2.36**	**6.44**
POS [27]	2.31	3.82	4.88	0.95
CHROM [26]	2.75	3.03	4.38	0.46
ICA [16]	3.25	4.40	7.51	1.57

Table 3. Generalizability to Compressed Stationary AFRL Videos

	CRF 18		**CRF 24**		**CRF 30**		**CRF 36**	
	MAE	SNR	MAE	SNR	MAE	SNR	MAE	SNR
No Augmentation	2.67	2.88	1.87	-0.52	4.98	-7.01	11.57	-10.51
Standard Augmentation	3.11	2.71	2.13	-2.02	6.44	-8.41	12.85	-11.66
Heuristic iPPG Augmentation	3.80	2.63	**1.42**	1.56	**3.90**	-5.46	13.47	-12.07
Interleaved iPPG Augmentation	7.14	-3.72	3.72	-6.15	8.30	-9.71	**10.25**	-11.22
POS [27]	1.47	3.25	2.99	**2.45**	15.49	-5.01	19.51	-7.19
CHROM [26]	**1.40**	1.62	1.94	1.53	9.09	-4.37	16.91	**-6.07**
ICA [16]	1.66	**3.38**	2.25	1.58	8.12	**-4.00**	17.46	-6.40

lenging videos with motion and other variations, we avoid the problem of accidentally magnifying both the signal and the noise during training. This allows the network to focus primarily on the skin pixels and facial regions which contain strong iPPG signals.

While our interleaved iPPG augmentation approach achieves promising results, more work is needed to finalize the best way to jointly train the network to magnify the training set videos and to obtain the iPPG estimates. We have tested several combinations and we found that we obtained the best results when we began the magnifications after training for 25 epochs and by magnifying the signals only in every other training epoch. We have also only sampled the amount of magnification from a fixed range of magnification factors that we found to work well. As part of future work, we would like to train the model end-to-end to learn what the best magnification amount is for a given set of training videos instead of sampling from a fixed range.

6. Conclusions

We have demonstrated that augmenting the training set with magnifications of the iPPG signal improved the performance, especially on very challenging cross-dataset experiments where the test set videos are very different from the training set videos. This augmentation approach helps especially on videos with lower SNR, such as videos with large motion or large video compression.

We presented two approaches of iPPG data augmentation, the heuristic and interleaved approaches. The interleaved approach shows promise to perform better on many challenging tasks. However, more work is needed to understand when this approach is helpful and how we can improve its performance by changing the way we magnify videos during training of the network, the amount of magnification, and the training epoch at which we begin the magnification.

There are few publicly available iPPG datasets with sufficiently diverse participants (e.g., with different skin types and genders). Therefore, it is difficult to close the performance gap on such videos with deep learning models and to avoid overfitting. Our augmentation approach is a promising step in this direction to reduce overfitting and to improve cross-dataset generalizability without the need to collect new data. We hope that these results will inspire new training approaches for iPPG applications to alleviate the challenges with collecting large datasets.

Acknowledgments

Ewa Nowara and Ashok Veeraraghavan were partially supported by the NFS SaTC Award CNS-1801372, NSF Expeditions Award CCF-1730574, and NSF PATHS-UP Award EEC-1648451.

References

[1] Daniel J McDuff, Justin R Estepp, Alyssa M Piasecki, and Ethan B Blackford. A survey of remote optical photoplethysmographic imaging methods. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 6398–6404. IEEE, 2015. 1

[2] 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. 1, 2, 4, 5

[3] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly

compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 151–160, 2019. 1, 2, 7

[4] Ewa Nowara and Daniel McDuff. Combating the impact of video compression on non-contact vital sign measurement using supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2, 7

[5] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018. 1, 2

[6] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. Systematic analysis of video-based pulse measurement from compressed videos. *Biomedical Optics Express*, 12(1):494–508, 2021. 1, 2, 7

[7] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017. 1

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[9] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1462–1469. IEEE, 2014. 1, 5, 6, 7

[10] Christoph Hoog Antink, Simon Lyra, Michael Paul, Xinchi Yu, and Steffen Leonhardt. A broader look: Camera-based vital sign estimation across the spectrum. *Yearbook of medical informatics*, 28(1):102, 2019. 1

[11] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8, 2012. 2, 3, 4

[12] Yichao Zhang, Silvia L Pintea, and Jan C Van Gemert. Video acceleration magnification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–537, 2017. 2, 3, 4

[13] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 633–648, 2018. 2, 3, 4

[14] Weixuan Chen and Daniel McDuff. Deepmag: Source-specific change magnification using gradient ascent. *ACM Transactions on Graphics (TOG)*, 40(1):1–14, 2020. 2, 3, 4

[15] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 284–285, 2020. 2, 6, 7

[16] Ming-Zher Poh, Daniel 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. 2, 5, 7, 8

[17] Ming-Zher Poh, Daniel 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

[18] Amruta Pai, Ashok Veeraraghavan, and Ashutosh Sabharwal. Camerahrv: robust measurement of heart rate variability using a camera. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*, volume 10501, page 105010S. International Society for Optics and Photonics, 2018. 2

[19] L Tarassenko, M Villarroel, A Guazzi, J Jorge, DA Clifton, and C Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014. 2

[20] Mohamed Elgendi, Richard Fletcher, Yongbo Liang, Newton Howard, Nigel H Lovell, Derek Abbott, Kenneth Lim, and Rabab Ward. The use of photoplethysmography for assessing hypertension. *NPJ digital medicine*, 2(1):1–11, 2019. 2

[21] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[22] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015. 2

[23] Richard Macwan, Yannick Benezeth, and Alamin Mansouri. Heart rate estimation using remote photoplethysmography with multi-objective optimization. *Biomedical Signal Processing and Control*, 49:24–33, 2019. 2

[24] Richard Macwan, Serge Bobbia, Yannick Benezeth, Julien Dubois, and Alamin Mansouri. Periodic variance maximization using generalized eigenvalue decomposition applied to remote photoplethysmography estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1332–1340, 2018. 2

[25] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2396–2404, 2016. 2

[26] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 2, 5, 7, 8

[27] 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, 2017. 2, 5, 7, 8

[28] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1353–135309. IEEE, 2018. 2, 5, 6, 7

[29] Weixuan Chen, Javier Hernandez, and Rosalind W Picard. Estimating carotid pulse and breathing rate from near-infrared video of the neck. *Physiological measurement*, 39(10):10NT01, 2018. 2

[30] Ewa M Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 2, 6, 7

[31] M. Rapczynski, P. Werner, and A. Al-Hamadi. Effects of video encoding on camera-based heart rate estimation. *IEEE Transactions on Biomedical Engineering*, 66(12):3360–3370, 2019. 2, 7

[32] Daniel McDuff. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1367–1374, 2018. 2

[33] Qi Zhan, Wenjin Wang, and Gerard de Haan. Analysis of cnn-based remote-ppg to understand limitations and sensitivities. *arXiv preprint arXiv:1911.02736*, 2019. 2, 5

[34] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference, Newcastle, UK*, pages 3–6, 2018. 2

[35] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3580–3585. IEEE, 2018. 2

[36] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. *arXiv preprint arXiv:2007.06786*, 2020. 2

[37] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising remote vitals measurements using inverse attention. *arXiv preprint arXiv:2010.07770*, 2020. 2

[38] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019. 2, 3

[39] Luke Taylor and Geoff Nitschke. Improving deep learning using generic data augmentation. *arXiv preprint arXiv:1708.06020*, 2017. 2, 3

[40] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *European Conference on Computer Vision*, pages 566–583. Springer, 2020. 2

[41] Aranzazu Jurio, Miguel Pagola, Mikel Galar, Carlos Lopez-Molina, and Daniel Paternain. A comparison study of different color spaces in clustering based image segmentation. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 532–541. Springer, 2010. 3

[42] Francisco J Moreno-Barea, Fiammetta Strazzera, José M Jerez, Daniel Urda, and Leonardo Franco. Forward noise adjustment scheme for data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 728–734. IEEE, 2018. 3

[43] Guoliang Kang, Xuanyi Dong, Liang Zheng, and Yi Yang. Patchshuffle regularization. *arXiv preprint arXiv:1707.07103*, 2017. 3

[44] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931, 2019. 3

[45] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018. 3

[46] Ce Liu, Antonio Torralba, William T Freeman, Frédo Durand, and Edward H Adelson. Motion magnification. *ACM transactions on graphics (TOG)*, 24(3):519–526, 2005. 3

[47] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000. 3

[48] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013. 3

[49] Daniel McDuff and Ethan Blackford. iphys: An open non-contact imaging-based physiological measurement toolbox. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6521–6524. IEEE, 2019. 5

[50] 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, 7

[51] Vytautas Vizbara. Comparison of green, blue and infrared light in wrist and forehead photoplethysmography. *BIOMEDICAL ENGINEERING 2016*, 17(1), 2013. 7

[52] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988. 7

[53] Daniel McDuff, Ethan B Blackford, and Justin R Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 63–70. IEEE, 2017. 7