MIT Libraries | DSpace@MIT

MIT Open Access Articles

DeepMag: Source Specific Change Magnification Using Gradient Ascent

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Chen, Weixuan and McDuff, Daniel. 2020. "DeepMag: Source Specific Change Magnification Using Gradient Ascent." ACM Transactions on Graphics.

As Published: http://dx.doi.org/10.1145/3408865

Publisher: ACM

Persistent URL: https://hdl.handle.net/1721.1/146175

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Massachusetts Institute of Technology

DeepMag: Source-Specific Change Magnification Using Gradient Ascent

WEIXUAN CHEN, Massachusetts Institute of Technology
DANIEL MCDUFF, Microsoft Research

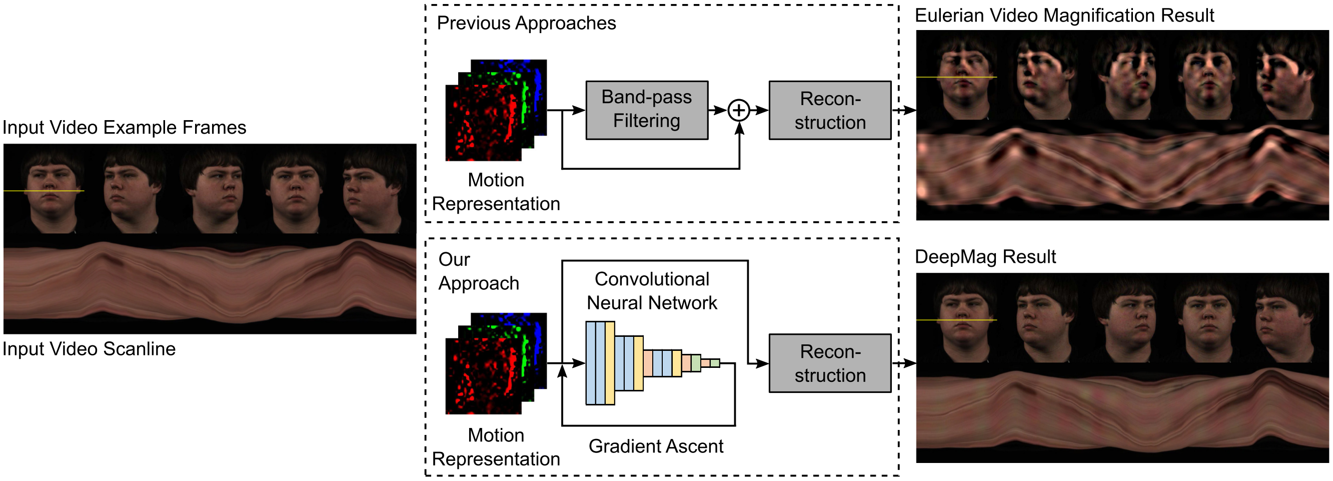


Fig. 1. We present a novel end-to-end deep neural framework for video magnification (DeepMag). Our method allows measurement, magnification and synthesis of subtle color and motion changes from a specific source even in the presence of large motions. We demonstrate this via pulse and respiration manipulation in 2D videos. Our approach produces magnified videos with substantially fewer artifacts when compared to previous methods, such as Eulerian Video Magnification [Wu et al. 2012] shown here. Our method magnifies the red color changes more clearly; otherwise, the video frames show few artifacts.

Many important physical phenomena involve subtle signals that are difficult to observe with the unaided eye, yet visualizing them can be very informative. Current motion magnification techniques can reveal these small temporal variations in video, but require precise prior knowledge about the target signal, and cannot deal with interference motions at a similar frequency. We present DeepMag, an end-to-end deep neural video-processing framework based on gradient ascent that enables automated magnification of subtle color and motion signals from a specific source, even in the presence of large motions of various velocities. The advantages of DeepMag are highlighted via the task of video-based physiological visualization. Through systematic quantitative and qualitative evaluation of the approach on videos with different levels of head motion, we compare the magnification of pulse and respiration to existing state-of-the-art methods. Our method produces magnified videos with substantially fewer artifacts and blurring whilst magnifying the physiological changes by a similar degree.

CCS Concepts: • **Image Processing and Computer Vision** → **Scene Analysis**; **Time-varying Imagery**;

Additional Key Words and Phrases: Video magnification, deep learning

ACM Reference format:
Weixuan Chen and Daniel McDuff. 2020. DeepMag: Source-Specific Change Magnification Using Gradient Ascent. *ACM Trans. Graph.* 40, 1, Article 2 (September 2020), 13 pages.
https://doi.org/10.1145/3408865

1 INTRODUCTION

Revealing subtle signals in our everyday world is important for helping us understand the processes that cause them. Magnifying small temporal variations in video has applications in both basic science (e.g., visualizing physical processes in the world), engineering (e.g., identifying the motion of large structures), and education (e.g., teaching scientific principals). To provide an illustration, physiological phenomena are often invisible to the unaided eye, yet understanding these processes can help us detect and treat negative health conditions. Pulse and respiration magnification, specifically, are good exemplar tasks for video magnification as physiological phenomena cause both subtle color and motion variations. Furthermore, larger rigid and non-rigid motions of the body often mask the subtle variations, which makes the magnification of physiological signals non-trivial.

Authors' addresses: W. Chen, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, Massachusetts, 02139; email: cvx@media.mit.edu; D. McDuff, Microsoft Research, 14820 NE 36th Street, Redmond, Washington, 98052; email: damcduff@microsoft.com.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0730-0301/2020/09-ART2 $15.00
https://doi.org/10.1145/3408865

Several methods have been proposed to reveal subtle temporal variations in video. *Lagrangian* methods for video magnification [Liu et al. 2005] rely on accurate tracking of the motion of particles (e.g., via optical flow) over time. These approaches are computationally expensive and will not work effectively for color changes. *Eulerian* video magnification methods do not rely on motion estimation, but rather magnify the variation of pixel values over time [Wu et al. 2012]. This simple and clever approach allows for subtle signals to be magnified that might otherwise be missed by optical flow. Subsequent iterations of such approaches have improved the method with phase-based representations [Pintea and van Gemert 2016; Wadhwa et al. 2013], matting [Elgharib et al. 2015], second-order manipulation [Zhang et al. 2017], and learning-based representations [Oh et al. 2018]. However, all these approaches use frequency properties to separate the target signal from noise, so they require precise prior knowledge about the signal frequency. Furthermore, if the signal of interest is at a similar frequency to another signal (for example, if head motions are at a similar frequency as the pulse signal), an Eulerian approach will magnify both and cause numerous artifacts (see Figure 1).

To address these problems, we present an approach for magnifying pulse and motion variations in videos that feature other periodic or random motions. Our method leverages a convolutional neural network (CNN) as a video motion discriminator to separate a specific source signal even if it overlaps with other motion sources in the frequency domain. Then, the separated signal can be magnified in video by performing gradient ascent [Erhan et al. 2009] in the input space of the CNN, with the other motion sources untouched. To adapt the gradient ascent method to the video magnification task, several methodological innovations are introduced including adding L1 normalization and sign correction. The whole algorithm proves to work effectively, even in the presence of interference motions with large magnitudes and velocities. Figure 1 shows a comparison between the proposed method and previous approaches.

Magnifying physiological changes on the human body without impacting other aspects of the visual appearance is an especially interesting use case with numerous applications in and of itself. In medicine and affective computing the photoplethysmogram (PPG) and respiration signals are used as unobtrusive measures of cardiopulmonary performance. Visualizing these signals could help in understanding vascular disease, heart conditions (e.g., arterial fibrillation) [Chan et al. 2016], and stress responses. For example, jugular venous pressure (JVP) is analyzed by studying subtle motions of the neck. This is challenging for clinicians, and video-magnification could offer a practical aid. Another application is in the design of avatars [Suwajanakorn et al. 2017]. Synthetic embodied agents may fall into the "uncanny valley" [Mori 1970] or be easily detected as "spoofs" if they do not exhibit accurate physiological responses, including respiration, pulse rates, and blood flow that can be recovered using video analysis [Poh et al. 2010]. Our method presents the opportunity to not only magnify signals but also synthesize them at different frequencies within a video.

The main contributions of this article are to: (1) present our novel end-to-end framework for video magnification based on a deep convolutional neural network and gradient ascent; (2) demonstrate recovery of the pulse and respiration waves and magnification of these signals in the presence of large rigid head motions; (3) systematically quantitatively and qualitatively compare our approach with state-of-the-art motion magnification approaches under different rigid motion conditions.

2 RELATED WORK

2.1 Video Motion Magnification

Lagrangian video magnification approaches involve estimation of motion trajectories that are then amplified [Liu et al. 2005; Wang et al. 2006]. However, these approaches require a number of complex steps including performing a robust registration, frame intensity normalization, tracking and clustering of feature point trajectories, segmentation, and magnification. Another approach, using temporal sampling kernels can aid visualization of time-varying effects within videos [Fuchs et al. 2010]. However, this method involves video downsampling and relies on high framerate input videos.

The neat Eulerian video magnification (EVM) approach proposed by Wu et al. [2012] combines spatial decomposition with temporal filtering to reveal time varying signals without estimating motion trajectories. However, it uses linear magnification that only allows for relatively small magnifications at high spatial frequencies and cannot handle spatially variant magnification. To counter the limitation, Wadhwa et al. [2013] proposed a non-linear phase-based approach, magnifying phase variations of a complex steerable pyramid over time. Replacing the complex steerable pyramid [Wadhwa et al. 2013] with a Riesz pyramid [Wadhwa et al. 2014] produces faster results. In general, the linear EVM technique is better at magnifying small color changes, while the phase-based pipeline is better at magnifying subtle motions [Wu et al. 2012]. Both the EVM and the phase-EVM techniques rely on hand-crafted motion representations. To optimize the representation construction process, a learning-based method [Oh et al. 2018] was proposed, which uses convolutional neural networks as both frame encoders and decoders. With the learned motion representation, fewer ringing artifacts and better noise characteristics have been achieved. In preliminary work, Pintea and van Gemert propose the use of phase-based motion representations in a learning framework that can be applied to the transference (or magnification) of motion Pintea and van Gemert [2016].

One common problem with all the methods above is that they are limited to stationary objects or situations in which the motion of interest is significantly faster than other motions, whereas many realistic applications would involve small motions of interest in the presence of large ones that might be at similar frequencies. After motion magnification, these large motions would result in large artifacts such as halos or ripples, and overwhelm any small temporal variation. A couple of improvements have been proposed including a clever layer-based approach called Dynamic Video Magnification (DVMAG) [Elgharib et al. 2015]. By using matting, it can amplify only a specific region of interest (ROI) while maintaining the quality of nearby regions of the image. However, the approach relies on 2D warping (either affine or translation-only) to discount large motions, so it is only good at diminishing the impact of motions parallel to the camera plane and cannot deal with more complex 3D motions such as the human head rotation.

The other method addressing large motion interferences is video acceleration magnification (VAM) [Zhang et al. 2017]. It assumes large motions to be linear on the temporal scale so that magnifying the motion acceleration via a second-order derivative filter will only affect small non-linear motions. However, the method will fail if the large motions have any non-linear components, and ideal linear motions are rare in real life, especially on living organisms.

Another problem with all the previous motion magnification methods is that they use frequency properties to separate target signals from noise, so they typically require the frequency of interest to be known *a priori* for the best results and, as such, have at least three parameters (the frequency bounds and a magnification factor) that need to be tuned. If there are motion signals from different sources that are at similar frequencies (e.g., someone is breathing and turning their head), it is previously not possible to isolate the different signals.

While in this work, we focus on the use of change magnification in RGB videos, the use of depth information can be used to improve motion magnification results. This is a clear application of an additional modality that naturally helps with segmentation of motions [Kooij and van Gemert 2016]. It is reasonable to think that depth information would similarly help improve our results.

2.2 Gradient Ascent for Feature Visualization

Opposite to gradient descent, gradient ascent is a first-order iterative optimization algorithm that takes steps proportional to the positive of the gradient (or approximate gradient) of a function. Since neural networks are generally differentiable with respect to their inputs, it is possible to perform gradient ascent in the input space by freezing the network weights and iteratively tweaking the inputs toward the maximization of an internal neuron firing or the final output behavior. Early works found that this technique can be used to visualize network features (showing what a network is looking for by generating examples) [Erhan et al. 2009; Simonyan et al. 2013] and to produce saliency maps (showing what part of an example is responsible for the network activating a particular way) [Simonyan et al. 2013].

A recent famous application of gradient ascent in feature visualization is Google DeepDream [Mordvintsev et al. 2015]. It maximizes the L2 norm of activations of a particular layer in a CNN to enhance patterns in images and create a dream-like hallucinogenic appearance. It should be noted that applying gradient ascent independently to each pixel of the inputs commonly produces images with nonsensical high-frequency noise, which can be improved by including a regularizer that prefers inputs that have natural image statistics. Also, following the same idea of DeepDream, not only a network layer but also a single neuron, a channel, or an output class can be set as the objective of gradient ascent. For a comprehensive discussion of various regularizers and different optimization objectives used in feature visualization tasks, see Olah et al. [2017].

None of the previous works have applied gradient ascent to motion magnification or any task related to motions in video. In contrast to DeepDream and similar visualization tools, our method maximizes the output activation of a CNN in motion representations computed from frames instead of in raw images.

2.3 Video-Based Physiological Measurement

Over the past decade video-based physiological measurement using Red-Green-Blue (RGB) cameras has developed significantly [McDuff et al. 2015]. For instance, physiological parameters such as heart rate (HR) and breathing rate (BR) have been accurately extracted from videos of the human body in which subtle changes in light reflected from the skin caused by peripheral blood flow are measured [Chen and McDuff 2018; de Haan and Jeanne 2013; Poh et al. 2010, 2011; Tarassenko et al. 2014; Verkruysse et al. 2008; Wang et al. 2016]. Heart rate has also been measured via subtle body motions associated with blood ejection into the vessels [Balakrishnan et al. 2013], and breathing rate has been measured via more prominent chest volume changes [Janssen et al. 2016; Tan et al. 2010].

Early work on imaging plethysmography identified that spatial averaging of skin pixel values from an imager could be used to recover the blood volume pulse [Takano and Ohta 2007]. The strongest pulse signal was observed in the green channel [Verkruysse et al. 2008], but a combination of color channels provides improved results [McDuff et al. 2014; Poh et al. 2010]. Combining these insights with face tracking and signal decomposition enables a fully automated recovery of the pulse wave and heart rate [Poh et al. 2010].

In the presence of dynamic lighting and motion, advancements were needed to successfully recover the pulse signal. Leveraging models grounded in the optical properties of the skin has improved performance. Using a linear weighting of the chrominance signals (CHROM) [de Haan and Jeanne 2013] makes assumptions about the skin color profile to white-balance the video frames [McDuff 2018]. Another method, named the Pulse Blood Vector (PBV) approach, [de Haan and van Leest 2014] makes use of variations in light absorption by blood across the frequency spectrum to weight the color channels. As peripheral blood flow is not uniform across the body, adapting the ROI can also improve the performance of iPPG measurements [Tulyakov et al. 2016]. Both these methods assume the weighting of color channel information is uniform across the skin region.

Most of the methods described above use unsupervised learning and assumptions based on the physical properties of skin and blood. Formulating the problem in the form of a supervised learning task is non-trivial and performance in early explorations was modest [Monkaresi et al. 2014; Osman et al. 2015]. Recent advances in deep neural video analysis offer opportunities for recovering accurate physiological measurements. Recently, Chen and McDuff [2018] presented a supervised method using a convolutional attention network that provided state-of-the-art measurement performance and generalized across people. Our video magnification algorithm is based on a novel framework that allows recovery of pulse and respiratory waves using such a convolutional architecture.

3 METHODS

3.1 Video Magnification Using Gradient Ascent

Figure 2 shows the workflow of the proposed video magnification algorithm using gradient ascent. Similar to previous video magnification algorithms, it reads a series of video frames

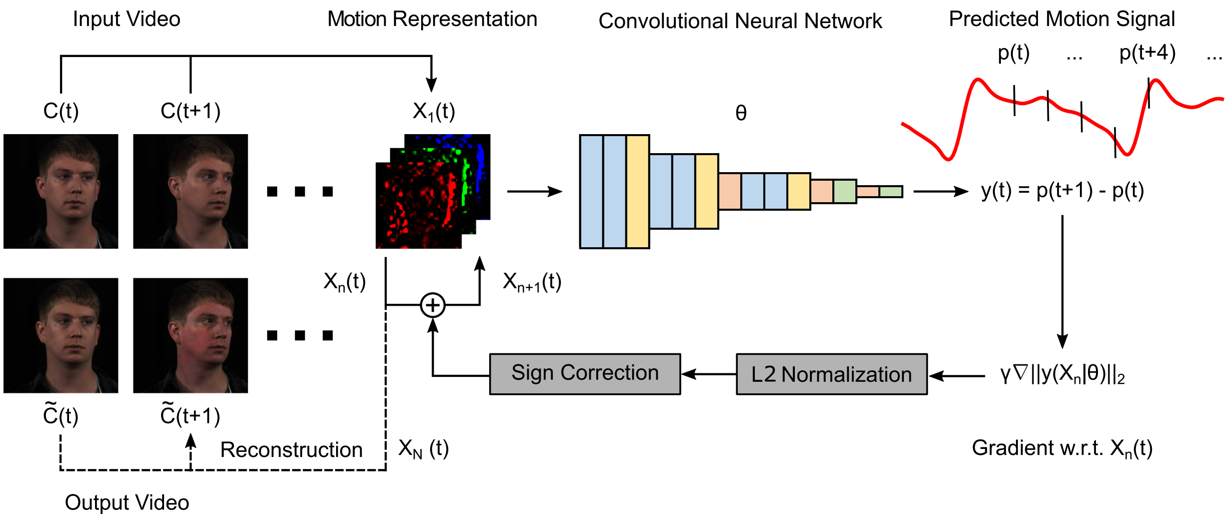


Fig. 2. The architecture of DeepMag. The CNN model predicts the motion signal of interest based on a motion representation computed from consecutive video frames. Magnification of the motion signal in video can be achieved by amplifying the L2 norm of its first-order derivative and then propagating the changes back to the motion representation using gradient ascent.

$C(t), t = 1, 2, \ldots, T$, magnifies a specific subtle motion in them, and outputs frames of the same dimension $\widetilde{C}(t), t = 1, 2, \ldots, T$.

The first step of our algorithm is computing the input motion representation $X_1(t)$ from the original video frames $C(t), t = 1, 2, \ldots, T$. $X_1(t)$ represents any change happening between two consecutive frames $C(t)$ and $C(t+1)$. Common motion representations include frame difference and optical flow. Different motion representations can emphasize different aspects of motions. For example, the physio-logy-based motion representation called normalized frame difference [Chen and McDuff 2018] was proposed to capture skin absorption changes robustly under varying rigid motions. On the other hand, optical flow based on the brightness constancy constraint is good at representing object displacements, but largely ignores the light absorption changes of objects.

In realistic videos the motion representations are comprised of multiple motions from different sources. For example, unconstrained facial video recordings commonly contain not only respiration movements and pulse-induced skin color changes but also head rotations and facial expressions. As we are only interested in magnifying one of these motions at a time, a video magnification algorithm should have the ability to separate the target motion from the others in the motion representation. Previous methods have typically used frequency-domain characteristics of the target motion in separation, so they rely on precise prior knowledge about the motion frequency (e.g., the exact heart rate). Furthermore, if any other motion overlaps with the target motion in frequency, it will still be magnified and cause artifacts. To improve the specificity of magnification and reduce the dependence on prior knowledge, we propose to use a deep convolutional neural network (CNN) to model the relationship between the motion representation and the motion of interest. As shown in Figure 2, the CNN has the input motion representation $X_1(t)$ as its input, and the first-order derivative $y(t)$ of the target motion signal $p(t)$ as its output. For many motion types, there are available datasets with paired videos and ground truth motion signals (e.g., facial videos with pulse and respiration signals measured from medical devices). Therefore, the weights θ of the CNN can be determined by training it on one of these datasets. It has been shown in Chen and McDuff [2018] that CNNs trained in this way have good generalization ability over different human subjects, different backgrounds, and different lighting conditions. We use the motion representation and CNN architecture presented in Chen and McDuff [2018] as our starting point. However, it is non-trivial to extend this for the purposes of magnification—as described below.

As the CNN has established the relationship between the input motion representation $X_1(t)$ and the target motion signal $p(t)$, magnification of $p(t)$ in $X_1(t)$ can be achieved by amplifying the L2 norm of its first-order derivative $y(t)$ and then propagating the changes back to $X_1(t)$ using gradient ascent. A hyperparameter, γ, reflects the step size applied to the gradient ascent. The process is performed iteratively and N reflects the total number iterations; this can then be expressed as

$$X_{n+1} = X_n + \gamma \nabla \|y(X_n|\theta)\|_2, \quad n = 1, 2, \ldots, N-1 \tag{1}$$

in which N is the total number of iterations and γ is the step size. θ is the weights of the CNN, which are frozen during gradient ascent. $\nabla\|y(X_n|\theta)\|_2$ is the gradient of $\|y(t)\|_2$ with respect to $X_n(t)$, which is the direction to which $X_n(t)$ can be modified to specifically magnify the target motion rather than the other motions. Note that both X_n and y correspond to time point t in Equation (1), but t is omitted for conciseness.

The vanilla gradient ascent in Equation (1) is appropriate for magnifying a single motion representation $X_1(t)$ at time t. However, for video magnification, a series of motion representations $X_1(t), t = 1, 2, \ldots, T$ need to be processed and magnified to the same level. Since the magnitude of the gradient is sensitive to the surface shape of the objective function (i.e., a point on a steep surface will have high magnitude whereas a point on the fairly flat surface will have low magnitude), it is not guaranteed that the accumulated gradient will be proportional to the original motion amplitude. Therefore, we apply L1 normalization to the gradient

$$X_{n+1} = X_n + \gamma \frac{\nabla \|y(X_n|\theta)\|_2}{\|\nabla \|y(X_n|\theta)\|_2\|_1} \tag{2}$$

so that only the gradient direction is kept and the gradient magnitude is controlled by the step size γ.

Another problem with Equation (1) is that motions in opposite directions contribute equivalently to the L2 norm of $y(t)$. As a result, the target motion might be amplified in terms of the absolute amplitude but 180-degrees out of phase. To address the problem, we correct the signs of the gradient to always match the signs of the input motion representation

$$X_{n+1} = X_n + \gamma \frac{\nabla \|y(X_n|\theta)\|_2 \odot sgn(X_n \odot \nabla \|y(X_n|\theta)\|_2)}{\|\nabla \|y(X_n|\theta)\|_2\|_1}, \tag{3}$$

in which $sgn(\cdot)$ is the sign function and $\odot$ is element-wise multiplication.

Summing up the changes of $X_n(t)$ in all the iterations, we get the final expression of the magnified motion representation:

$$X_N = X_1 + \sum_{n=1}^{N-1} \gamma \frac{\nabla \|y(X_n|\theta)\|_2 \odot sgn(X_n \odot \nabla \|y(X_n|\theta)\|_2)}{\|\nabla \|y(X_n|\theta)\|_2\|_1} \tag{4}$$

There are only two hyper-parameters γ and N, which can be tuned to change the magnification factor. Finally, the magnified motion representation can be combined with previous frames to iteratively generate the output video. The complete algorithm is summarized in Algorithm 1.

ALGORITHM 1: DeepMag video magnification

```
Require: $C(t), t = 1, 2, \ldots, T$ is a series of video frames, $\mathcal{M}$ is a motion representation estimator, $\theta$ is the pre-trained CNN weights for predicting a target motion signal $y$, $\gamma$ is the step size, and $N$ is the number of iterations
for $t = 1$ to $T - 1$ do
    Compute motion representation: $X_1(t) \leftarrow \mathcal{M}(C(t), C(t+1))$
    for $n = 1$ to $N - 1$ do
        Compute gradient: $G_n(t) \leftarrow \nabla \|y(X_n(t)|\theta, t)\|_2$
        L1 normalization: $G_n(t) \leftarrow G_n(t) / \|G_n(t)\|_1$
        Sign correction: $G_n(t) \leftarrow G_n(t) \odot sgn(G_n(t) \odot X_n(t))$
        Gradient ascent: $X_{n+1}(t) \leftarrow X_n(t) + \gamma G_n(t)$
    end for
end for
$\widetilde{C}(1) = C(1)$
for $t = 1$ to $T - 1$ do
    Reconstruct magnified frame $\widetilde{C}(t+1) \leftarrow \mathcal{M}^{-1}(\widetilde{C}(t), X_N(t))$
end for
return $\widetilde{C}(t), t = 1, 2, \ldots, T$
```

3.2 Example I: Pulse Magnification

One example of applying our proposed algorithm is in the magnification of subtle skin color changes associated with the cardiac cycle. As blood flows through the skin, it changes the light reflected from it. A good motion representation for these color changes is normalized frame difference [Chen and McDuff 2018], which is summarized below.

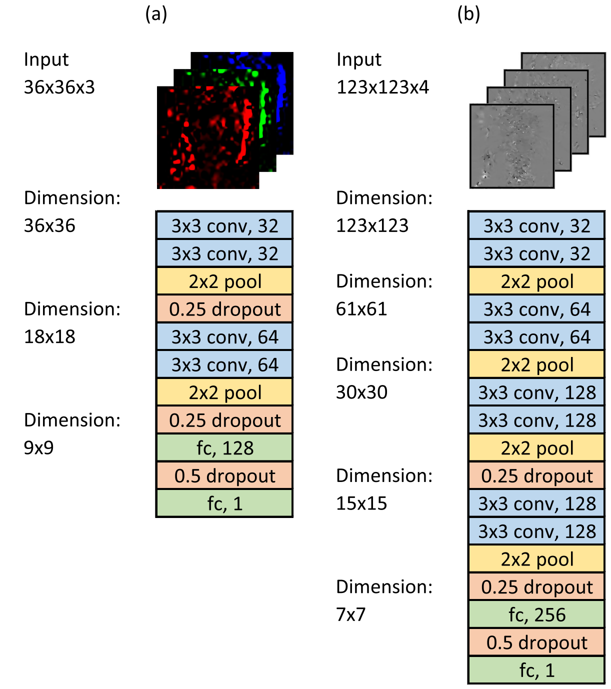


Fig. 3. We used two exemplar tasks to illustrate the benefits of DeepMag. (a) Color change (Blood flow) magnification. (b) Motion (respiration) magnification. These two tasks require different input motion representations (frame differences) and CNN architectures due to the nature of the motion signals. The color change magnification input is a normalized frame difference computed from the downsampled RGB frames. The motion magnification input is the phase variations in a complex steerable pyramid; we computed a pyramid with octave bandwidth and four orientations ($\theta = 0°, 45°, 90°, 135°$).

For modeling lighting, imagers, and physiology, previous works used the Lambert-Beer law (LBL) [Lam and Kuno 2015; Xu et al. 2014] or Shafer's dichromatic reflection model (DRM) [Wang et al. 2016]. We build our motion representation on top of the DRM as it provides a better framework for separating specular reflection and diffuse reflection. Based on the approach introduced by Chen and McDuff [2018], if we assume that the illumination has an invariant spectral composition but varying intensity, the RGB values of the k-th skin pixel in an image sequence can then be represented by a time-varying function:

$$C_k(t) = I(t) \cdot (\boldsymbol{v}_s(t) + \boldsymbol{v}_d(t)) + \boldsymbol{v}_n(t), \tag{5}$$

where $C_k(t)$ is a matrix of the color values; $I(t)$ is the luminance intensity level; $I(t)$ is modulated by two components: specular reflection $\boldsymbol{v}_s(t)$, mirror-like light reflection from the skin surface, and diffuse reflection $\boldsymbol{v}_d(t)$, the absorption and scattering of light in skin-tissue; $\boldsymbol{v}_n(t)$ denotes the noise in the camera sensor (i.e., quantization noise). We assume $I(t)$, $\boldsymbol{v}_s(t)$, and $\boldsymbol{v}_d(t)$ can all be decomposed into a time-invariant and a time-dependent part through a linear transformation [Wang et al. 2016]. The unit color vector of the skin-tissue $\boldsymbol{u}_d$ multiplied by the stationary reflection strength d_0 and the relative pulsatile strengths caused by hemoglobin and melanin absorption $\boldsymbol{u}_p$ multiplied by the blood volume pulse (BVP) signal $p(t)$.

$$\boldsymbol{v}_d(t) = u_d \cdot d_0 + \boldsymbol{u}_p \cdot p(t) \tag{6}$$

$\boldsymbol{v}_s(t)$ is a product of the unit color vector of the light source spectrum $\boldsymbol{u}_s$ multiplied by the sum of the stationary and varying parts of specular reflections, s_0 and $s(t)$, respectively.

$$\boldsymbol{v}_s(t) = \boldsymbol{u}_s \cdot (s_0 + s(t)) \tag{7}$$

$I(t)$ is a function of the stationary part of the luminance intensity I_0 and the intensity variation observed by the camera $I_0 \cdot i(t)$.

$$I(t) = I_0 \cdot (1 + i(t)) \tag{8}$$

The stationary components from the specular and diffuse reflections can be combined into a representation of the stationary skin reflection:

$$C_k(t) = I_0 \cdot (1 + i(t)) \cdot (\boldsymbol{u}_c \cdot c_0 + \boldsymbol{u}_s \cdot s(t) + \boldsymbol{u}_p \cdot p(t)) + \boldsymbol{v}_n(t), \tag{9}$$

where $\boldsymbol{u}_c$ denotes the unit color vector of the skin reflection and c_0 denotes the reflection strength.

The time-varying components are much smaller than the stationary components in Equation (9); therefore, we can neglect any product between the time varying terms and approximate $\boldsymbol{c}_k(t)$ as:

$$\begin{aligned} C_k(t) \approx \boldsymbol{u}_c \cdot I_0 \cdot c_0 \cdot (1 + i(t)) + \\ \boldsymbol{u}_s \cdot I_0 \cdot s(t) + \boldsymbol{u}_p \cdot I_0 \cdot p(t) + \boldsymbol{v}_n(t) \end{aligned} \tag{10}$$

We downsample every frame to L pixels by L pixels using bicubic interpolation. Based on the finds of Wang et al. [2015], we select $L = 36$, which was emperically found to work well for face videos. The resulting values still obey the DRM model only without the camera quantization error:

$$\begin{aligned} C_l(t) \approx \boldsymbol{u}_c \cdot I_0 \cdot c_0 + \boldsymbol{u}_c \cdot I_0 \cdot c_0 \cdot i(t) + \\ \boldsymbol{u}_s \cdot I_0 \cdot s(t) + \boldsymbol{u}_p \cdot I_0 \cdot p(t), \end{aligned} \tag{11}$$

where $l = 1, \ldots, L^2$ is the new pixel index in every frame.

Then, we need to reduce the dependency of $C_l(t)$ on the stationary skin reflection color $\boldsymbol{u}_c \cdot I_0 \cdot c_0$, resulting from the light source and subject's skin tone. In Equation (11), $\boldsymbol{u}_c \cdot I_0 \cdot c_0$ appears twice. It is difficult to eliminate the second term as it interacts with the unknown $i(t)$. However, the first time-invariant term, which is usually dominant, can be removed by taking the first-order derivative of both sides of Equation (11) with respect to time:

$$C_l'(t) \approx \boldsymbol{u}_c \cdot I_0 \cdot c_0 \cdot i'(t) + \boldsymbol{u}_s \cdot I_0 \cdot s'(t) + \boldsymbol{u}_p \cdot I_0 \cdot p'(t) \tag{12}$$

As identified by Chen and McDuff [2018], a problem with the frame difference representation is the spatially heterogeneous nature of the intensity level. We follow the same normalization strategy by dividing $\boldsymbol{C}_l'(t)$ by the temporal mean of $\boldsymbol{C}_l(t)$ to remove I_0:

$$\begin{aligned} \frac{\boldsymbol{C}_l'(t)}{\overline{\boldsymbol{C}_l(t)}} \approx [1\ 1\ 1]^T \cdot i'(t) + diag^{-1}(\boldsymbol{u}_c)\boldsymbol{u}_s \cdot \frac{s'(t)}{c_0} + \\ diag^{-1}(\boldsymbol{u}_c)\boldsymbol{u}_p \cdot \frac{p'(t)}{c_0} \end{aligned} \tag{13}$$

We compute $\overline{\boldsymbol{C}_l(t)}$ over two consecutive frames to minimize occlusion problems and prevent the propagation of errors. The normalized frame difference we used as the motion representation is expressed as:

$$\boldsymbol{X}_1(l,t) = \frac{\boldsymbol{C}_l'(t)}{\overline{\boldsymbol{C}_l(t)}} \sim \frac{\boldsymbol{C}_l(t+1) - \boldsymbol{C}_l(t)}{\boldsymbol{C}_l(t+1) + \boldsymbol{C}_l(t)} \tag{14}$$

The CNN we used for extracting pulse signals from the motion representation is shown in Figure 3(a). The pooling layers are 2x2 average pooling, and the convolution layers have a stride of one. All the layers use ReLU as the activation function. Note that bounded activation function such as tanh and sigmoid are not suitable for this task, as they will limit the extent to which the motion representation can be magnified in the gradient ascent.

After the gradient ascent, the input motion representation $\boldsymbol{X}_1(l,t)$ was magnified as $\boldsymbol{X}_N(l,t)$, from which we could reconstruct the magnified video. The first step of reconstruction is to denoise the output motion representation by filtering the accumulated gradient:

$$\widetilde{\boldsymbol{X}_N}(l,t) = \boldsymbol{X}_1(l,t) + \mathcal{F}(\boldsymbol{X}_N(l,t) - \boldsymbol{X}_1(l,t)), \tag{15}$$

in which $\mathcal{F}$ is a zero-phase band-pass filter. Note that unlike previous motion magnification methods the function of the filter here is not to select the target motion but to remove low- and high-frequency noise so the filter bands do not need to precisely match the motion frequency in the video and can be chosen conservatively. Specifically, a sixth-order Butterworth filter with cut-off frequencies of 0.7 and 2.5 Hz was used to generally cover the normal heart rate range (42 to 150 beats per minute). Then, we applied the inverse operation of Equation (14) to reconstruct the downsampled version of the frames $\widetilde{\boldsymbol{C}_l(t)}$:

$$\widetilde{\boldsymbol{C}_l}(t+1) = \frac{1 + \widetilde{\boldsymbol{X}_N}(l,t)}{1 - \widetilde{\boldsymbol{X}_N}(l,t)} \cdot \widetilde{\boldsymbol{C}_l}(t), \widetilde{\boldsymbol{C}_l}(1) = \boldsymbol{C}_l(1) \tag{16}$$

Finally, $\boldsymbol{C}_l(t)$ was upsampled back to the original video resolution:

$$\widetilde{\boldsymbol{C}_k}(t) = \boldsymbol{C}_k(t) - \mathcal{U}(\boldsymbol{C}_l(t)) + \mathcal{U}(\widetilde{\boldsymbol{C}_l}(t)), \tag{17}$$

in which $\mathcal{U}$ is an image upsampling operator.

3.3 Example II: Motion Magnification

Our second example is amplifying subtle motions on the human body induced by respiration. We used phase variations in a complex steerable pyramid [Portilla and Simoncelli 2000; Simoncelli et al. 1992] to represent the local motions in a video. The basis functions of the pyramid are scaled and oriented Gabor-like wavelets with both cosine- and sine-phase components. Each pair of filters can be used to separate the amplitude and phase of local wavelets. Specifically, each scale r and orientation θ is a complex image that can be expressed as:

$$A(r,\theta,t)e^{i\phi(r,\theta,t)}, \tag{18}$$

where A and ϕ are amplitude and phase, respectively. We use the first-order derivative of the local phases ϕ as our input motion representation:

$$X_1(r,\theta,t) = \phi(r,\theta,t+1) - \phi(r,\theta,t) \tag{19}$$

Based on prior work [Gautama and Van Hulle 2002], we find that these phase variations are approximately proportional to displacements of image structures along the corresponding orientation and scale. To lower computational cost, we computed a pyramid with octave bandwidth and four orientations ($\theta = 0°, 45°, 90°, 135°$). Using half-octave or quarter-octave bandwidth and more orientations would enable our algorithm to amplify more motion details, but would require significantly greater computational resources. In theory, $X_1(r, \theta, t)$ contains $r = 1, 2, \ldots, R$ scales of representations in different spatial resolutions, and extracting the target respiration motion from them would need R different CNNs to fit different input dimensions. However, we found that $X_1(r, \theta, t)$ and the amplified $X_N(r, \theta, t)$ on different scales were approximately proportional to 0.5^r, so it is possible to only process one scale $r = r_0$ and interpolate the other scales with it.

The CNN we used for extracting respiration signals from the motion representation is shown in Figure 3(b). The neural network is deeper than the one used for pulse magnification because the input motion representation for respiration has a higher dimension. The pooling layers and convolution layers are of the same type as in Figure 3(a). As we met the dying ReLU problem (ReLU neurons were stuck in the negative side and always output 0) in our experiments, the activation functions of all the layers were replaced with scaled exponential linear units (SELU) [Klambauer et al. 2017].

After gradient ascent, the input motion representation $X_1(r_0, \theta, t)$ was magnified as $X_N(r_0, \theta, t)$, from which we could reconstruct the magnified video. Unlike in Equation (1), the phase variations were reconstructed by reversing Equation (19) before denoising:

$$\widetilde{\phi}(r_0, \theta, t+1) = X_N(r_0, \theta, t) + \widetilde{\phi}(r_0, \theta, t),$$
$$\widetilde{\phi}(r_0, \theta, 1) = \phi(r_0, \theta, 1) \qquad (20)$$

Then, the reconstructed phase was denoised by band-pass filtering and 2π phase clipping:

$$\widetilde{\phi}(r_0, \theta, t) = \phi(r_0, \theta, t) + \mathcal{F}(\widetilde{\phi}(r_0, \theta, t)) \cdot \frac{sgn(2\pi - |\phi(r_0, \theta, t)|) + 1}{2} \qquad (21)$$

The filter $\mathcal{F}$ is a sixth-order zero-phase Butterworth filter with cut-off frequencies of 0.16 and 0.5 Hz for generally covering the normal breathing rate range (10 to 30 beats per minute). The magnified phase of the other scales can be interpolated by exponentially scaling the filtered term:

$$\widetilde{\phi}(r_0, \theta, t) = \phi(r_0, \theta, t) + \mathcal{F}(\widetilde{\phi}(r_0, \theta, t)) \cdot \frac{sgn(2\pi - |\phi(r_0, \theta, t)|) + 1}{2} \cdot \left(\frac{1}{2}\right)^{r - r_0} \qquad (22)$$

Finally, the magnified video frame $\widetilde{C}(t)$ can be reconstructed from all the scales of the complex steerable pyramid with their phase updated as Equation (22).

4 DATA

We used the dataset collected by Estepp et al. [2014] for testing our approach. Videos were recorded with a Basler Scout scA640-120gc GigE-standard, color camera, capturing 8-bit, 658x492 pixel images, 120 fps. The camera was equipped with 16 mm fixed focal length lens. Twenty-five participants (17 males) were recruited to participate for the study. Nine individuals were wearing glasses, eight had facial hair, and four were wearing makeup on their face and/or neck. The participants exhibited the following estimated Fitzpatrick Sun-Reactivity Skin Types [Fitzpatrick 1988]: I-1, II-13, III-10, IV-2, V-0. Gold-standard physiological signals were measured using a BioSemi ActiveTwo research-grade biopotential acquisition unit.

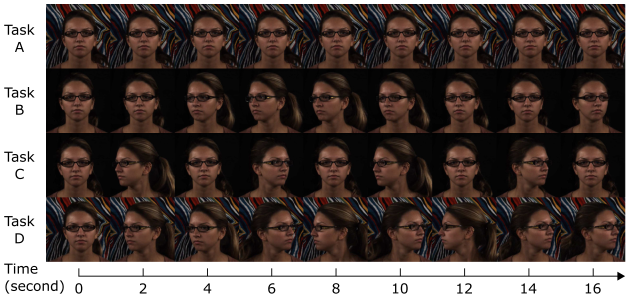


Fig. 4. Exemplary frames from the four tasks of our video dataset. Note the different backgrounds and head rotation speeds.

We used videos of participants during a set of four, five-minute tasks for our analysis. Two of the tasks (A and D) were performed in front of a patterned background and two (B and C) were performed in front of a black background. The four tasks were designed to capture different levels of head rotation about the vertical axis (yaw). Examples of frames from the tasks can be seen in Figure 4.

Task A: Participants stayed still, allowing for small natural motions.

Task B: Participants performed a 120-degree sweep centered about the camera at a speed of 10 degrees/sec.

Task C: Similar to Task B but with a speed of 30 degrees/sec.

Task D: Participants were asked to reorient their head position once per second to a randomly chosen target positioned in 20-degree increments over a 120-degree arc, thus, simulating random head motion.

5 EVALUATION

We compare the color change magnification results to Eulerian video magnification [Wu et al. 2012] and video acceleration magnification [Zhang et al. 2017], and compare the motion magnification results to phase-based Eulerian video magnification [Wadhwa et al. 2013], video acceleration magnification, and learning-based video motion magnification [Oh et al. 2018] (EVM and phase-based EVM perform poorly for motion magnification and color change magnification, respectively). In each case, we perform qualitative evaluations similar to that presented in prior work. In addition, we perform a quantitative evaluation by assessing the image quality of the resulting videos. Prior work has generally not considered quantitative evaluations.

For obtaining our own results, the CNN model was either trained and tested on different time periods of the same videos (participant-dependent) or trained and tested on videos of different human participants (participant-independent), both using a 20% holdout rate for testing. Note, the auxiliary pulse and respiration signals are used at training time for learning but are not used at test time. The qualitative and quantitative results we show in the following sections are always from video excerpts in the test set. To achieve a fair comparison, all the compared methods used the same filter bands: [0.7 Hz, 2.5 Hz] for pulse color change

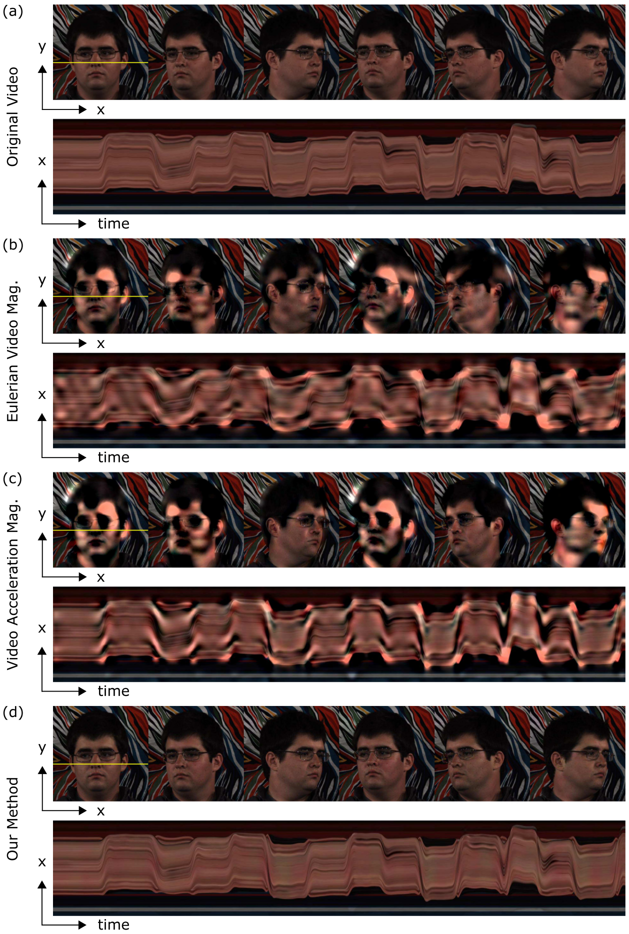


Fig. 5. Scan line comparisons of color change magnification methods for a Task D video: (a) original video, (b) Eulerian video magnification [Wu et al. 2012], (c) video acceleration magnification [Zhang et al. 2017], and (d) our method. The yellow line shows the source of the scan line in the frames. The section of video shown was 15 seconds in duration. Our method produces clearer magnification of the color change due to blood flow and significantly fewer artifacts.

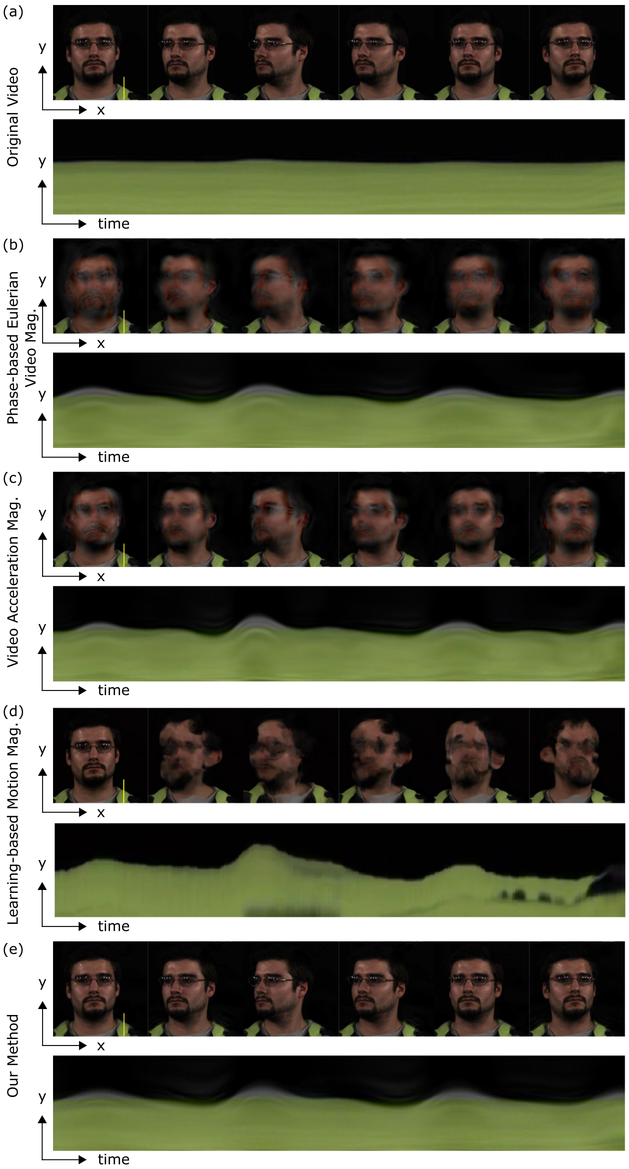


Fig. 6. Scan line comparisons of motion magnification methods for a Task B video: (a) original video, (b) phase-based Eulerian video magnification [Wadhwa et al. 2013], (c) video acceleration magnification [Zhang et al. 2017], (d) learning-based motion magnification [Oh et al. 2018], and (e) our method. The yellow line shows the source of the scan line in the frames. The section of video shown was 15 seconds in duration. Our method produces comparable magnification of the respiration motion and significantly fewer artifacts and blurring.

magnification, and [0.16 Hz, 0.5 Hz] for respiration motion magnification. Since VAM uses difference of Gaussian (DoG) filters defined by a single pass-band frequency, we adopted the center frequencies of the physiology frequency bands ($\sqrt{0.7 \times 2.5} = 1.3\ Hz$ for pulse, and $\sqrt{0.16 \times 0.5} = 0.28\ Hz$ for respiration) as its filtering parameters. In the color change magnification baselines, video frames were decomposed into multiple scales using a Gaussian pyramid with the intensity changes in the fourth level amplified (following the source code released by Wu et al. [2012]). All the motion magnification baselines used complex steerable pyramids with octave bandwidth and four orientations. The magnification factors of all the methods were tuned to be visually the same on Task A without head motion interferences.

5.1 Color Change Magnification

We apply our method to the task of magnifying the photoplethysmogram. In this task, the target variable for training the CNN was the gold standard contact PPG signal. The input motion representation was 36 pixels × 36 pixels × 3 color channels. In terms of the hyper-parameters of gradient ascent, the number of iterations N was chosen to be 20, and the step size γ was chosen to be 6×10^{-5}. We found these choices provided a moderate magnification level, equivalent to the magnification using EVM. Different choices of these hyper-parameters will be discussed in the following sections.

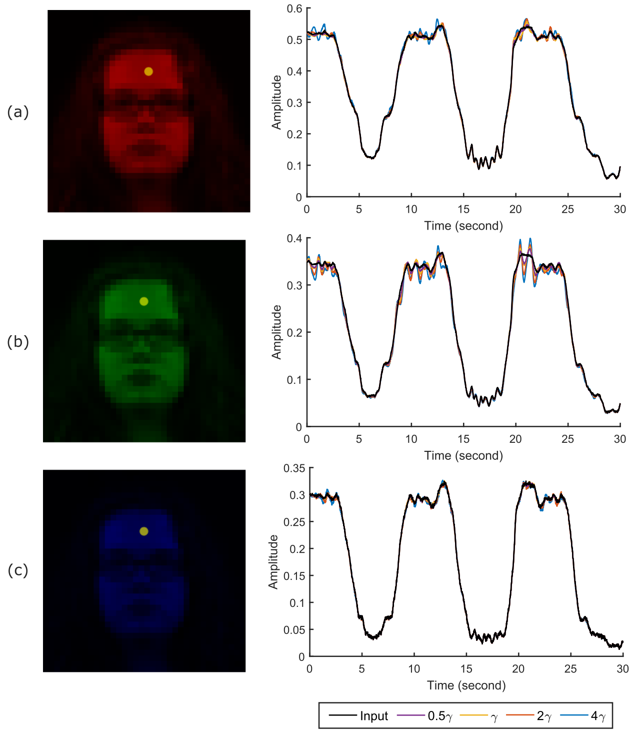


Fig. 7. Original and magnified traces of a pixel (the yellow dot) in three color channels of a Task B video (a) red channel, (b) green channel, and (c) blue channel. Magnified traces using different step sizes γ are shown in different colors. The notches (large variations in image intensity) in the traces correspond to when the participant rotated her head to the far left/right, and the pixel was no longer on the skin. Our method amplified the subtle color changes of the pixel only when it was on the skin, and kept the relative magnitudes of the pulse in three color channels with the green channel one being the strongest.

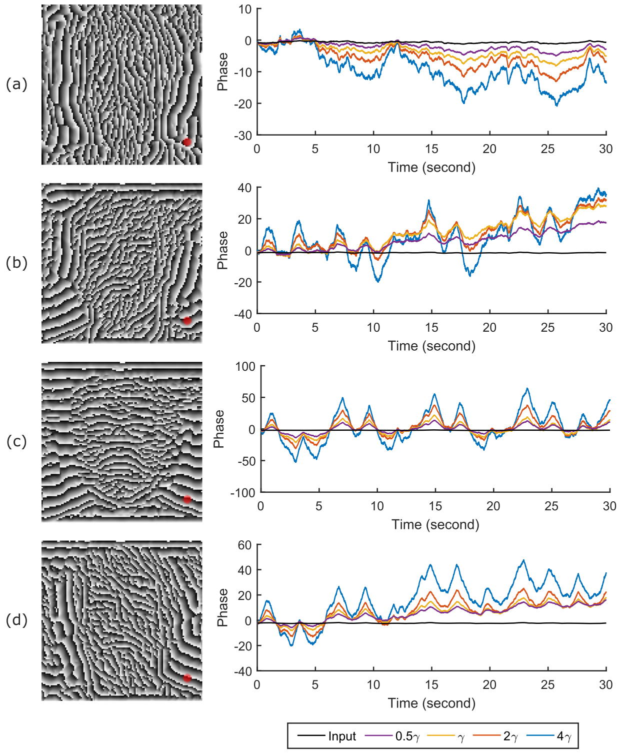


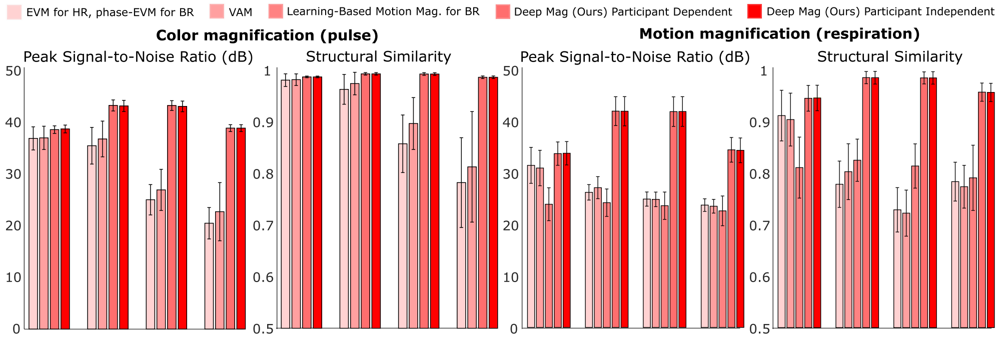
Fig. 8. Original and magnified traces of a pixel (the red dot) in the phase representation $\phi(r_0, \theta, t)$ of a Task C video along four orientations (a) $\theta = 0°$, (b) $\theta = 45°$, (c) $\theta = 90°$, and (d) $\theta = 135°$. Magnified traces using different step sizes γ are shown in different colors. The pixel exhibits a respiration movement mainly in the vertical direction, so its magnified phase traces have the highest amplitude along the $\theta = 90°$ orientation.

Figure 5 shows a qualitative comparison between our method and the baseline methods. The human participant in the video reoriented his head once per second to a random direction. In the horizontal scan line of the input video, only the head rotation is visible and the subtle color changes of the skin corresponding to pulse cannot be seen with the unaided eye. In the results of the baseline methods, strong motion artifacts are introduced. This is because the complex head motion is not distinguishable from the pulse signal in the frequency domain, so it is amplified along with the pulse. Since the pulse-induced color changes are several orders of magnitude weaker than the head motion, they are completely buried by the motion artifacts in the amplified video. The VAM scan line (Figure 5(c)) shows slightly fewer artifacts than the EVM scan line (Figure 5(b)) as the head rotation was occasionally semi-linear. On the other hand, our algorithm uses a deep neural network to separate the pulse signal from the head motion, and uses gradient ascent to specifically amplify it. Consequently, its scan line (Figure 5(d)) preserves the morphology of the head rotation while revealing the periodic color changes clearly on the skin.

To show the magnification effects on different colors and different object surfaces, we drew the original and magnified traces of a pixel in three color channels of a video in Figure 7. The human participant in the video rotated her head left and right, so the selected pixel was on her forehead half of the time and was on the black background in the other half (corresponding to the notches in the traces). First, the pulse-induced color changes were only magnified when the pixel was on the skin surface, which proved the good spatial specificity of our algorithm. Second, the magnified pulse signal has much higher amplitude in the green channel than in the other channels. This is consistent with previous findings that the amplitude of the human pulse is approximately 0.33:0.77:0.53 in RGB channels under a halogen lamp [de Haan and Jeanne 2013], and verifies that our algorithm faithfully kept the original physiological property in magnification. Third, we changed the chosen step size γ to its multiples (0.5γ, 2γ, and 4γ) with the number of iterations N unaltered, and visualized the resulting pixel traces also in Figure 7. There is a clear trend that longer step sizes lead to higher amplitudes of the magnified pulse.

To perform a quantitative evaluation of video quality, we used two metrics: peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). In both cases, we calculated the metrics on every frame of the tested videos, and took their averages across all participants within each task. The reference frame in each case was the corresponding frame from the original, unmagnified video. Table 1 shows a comparison of the video quality metrics for the baselines and our method. Although the magnified blood flow or respiration

Table 1. Video Quality Measured via Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) for the Magnified Videos



Task	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
EVM [Wu et al. 2012b]	36.5	35.1	24.8	20.3	.975	.957	.853	.779	-	-	-	-	-	-	-	-
Phase-EVM [Wadhwa et al. 2013]	-	-	-	-	-	-	-	-	31.1	25.9	24.6	23.5	.907	.775	.726	.780
VAM [Zhang et al. 2017]	36.6	36.4	26.7	22.5	.976	.969	.892	.809	30.6	26.8	24.6	23.2	.900	.800	.720	.770
Learning-based MM [Oh et al. 2018]	-	-	-	-	-	-	-	-	23.6	23.9	23.3	34.1	.807	.821	.810	.787
DeepMag (Ours) - P. Dep.	38.2	**42.8**	**42.8**	**38.5**	**.981**	**.987**	**.987**	**.981**	33.3	**41.5**	41.4	**34.1**	**.940**	**.980**	**.979**	**.952**
DeepMag (Ours) - P. Ind.	**38.3**	42.7	42.6	**38.5**	**.981**	**.987**	**.987**	**.981**	**33.4**	**41.5**	**41.4**	34.0	**.940**	.979	**.979**	.951

The baselines for color change magnification were EVM [Wu et al. 2012] and VAM [Zhang et al. 2017], and for motion magnification were phase-EVM [Wadhwa et al. 2013], VAM, and learning-based motion magnification [Oh et al. 2018]. The table shows the average metrics among all videos within each task, while the bar charts also show the standard deviations as error bars. Our models (both participant-dependent and participant-independent) produce videos with higher PSNR and SSIM compared to the baselines for all tasks. The benefit of our model is particularly strong for videos with greater levels of head rotation. We observed that while the magnification causes changes to the video that artifacts dominated, PSNR and SSIM still provide a reasonable quantitative measure of overall magnified video quality.

will naturally cause the metrics to be lower, and this makes the quantitative metric imperfect, we found that artifacts in the generated videos had a much more significant impact on their values than the magnified physiology. Thus, overall, the PSNR and SSIM scores do capture the performance of the magnification methods from one perspective with lower PSNR, and SSIM values indicate more artifacts and lower quality. According to the table, our methods achieve both higher PSNR and SSIM than the baseline methods, which verifies the ability of our methods to magnify subtle color changes with motion artifact suppressed. On Task A containing limited head motions, the metrics of the baseline methods are very close to those of our method. However, as the head rotation becomes faster and random on more difficult tasks, the video quality of the baseline outputs dramatically decreases. This is because their algorithms amplify any motion lying in the filter band and does so indiscriminately. The magnification thus leads to significant artifacts when large head motions are present. On the other hand, using our method, the video quality is maintained at almost the same level on different tasks. Both PSNR and SSIM are only slightly lower on Task A and Task D because the patterned background is more vulnerable to artifacts than the black one. The difference between the participant-dependent results and the participant-independent results is also very small, suggesting that our algorithm has good generalization ability and can be successfully applied to new videos containing different human participants without additional tuning.

5.2 Respiration Magnification

We apply our method to the task of magnifying respiration motions. In this task, the target variable for training the CNN was the gold standard respiration signal measured via the chest strap. Given the subtle nature of the motions, we found that a higher dimension input motion representation was needed than for the PPG magnification. As shown in Figure 3, the motion representation was in 123 pixels × 123 pixels × 4 orientations. The gradient ascent hyper-parameters N and γ were chosen to be 20 and 3.6×10^{-3} to produce moderate magnification effects.

Figure 6 shows a qualitative comparison between our method and the baseline methods. The human participant in the video rotated his head at a speed of 10 degrees/sec. A vertical scanline on his shoulder was drawn along with time to show the respiration movement. In the input video, the respiration movement is very subtle. Both our method and the baseline methods greatly increased its magnitude (Figure 6(b)–(d)). However, the baseline methods cannot clearly distinguish the phase variations caused by respiration and by head rotation, so it also amplified the head rotation and blurred the participant's face. Our method is based on a better motion discriminator learned via the CNN so that the head motions are not amplified.

To show the intermediate phase variations and different magnification effects along different orientations, we drew the original and magnified traces of a pixel in the phase representation $\phi(r_0, \theta, t)$ (Figure 8). Since the selected pixel is on the shoulder of the human participant, the respiration movement is mainly in the vertical direction. As a result, the amplified phase variations corresponding to breathing have the highest amplitude along $\theta = 90^\circ$ (Figure 8(c)) and the lowest amplitude along $\theta = 0^\circ$ (Figure 8(a)). We also changed the chosen step size γ to its multiples (0.5γ, 2γ, and 4γ) with the number of iterations N unaltered, and visualized the resulting phase traces in Figure 8. The figure suggests that the magnification level always increases along with the step size.

The same quantitative metrics as those for color change magnification were computed and shown in Table 1. They also generally follow the same pattern as in the color change magnification

Table 2. Video Quality Measured via Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) for Task C Videos Magnified to Different Levels

	PSNR (dB)				SSIM			
Step size	0.5γ	γ	2γ	4γ	0.5γ	γ	2γ	4γ
Pulse	43.2	42.6	41.6	39.9	0.987	0.987	0.986	0.986
Respiration	42.0	41.4	40.6	39.6	0.982	0.979	0.974	0.965

analysis: The video quality of the baseline methods is impacted by the level of head motions, while our method is considerably more robust. There is no significant difference between our participant-dependent results and participant-independent results.

5.3 Magnification Factors

The magnification factor of our algorithm is controlled by two hyper-parameters, the number of iterations N and the step size γ. In Figures 7 and 8, we chose the same N and tuned γ to be different multiples. The resulting magnification levels were always higher when γ was longer. However, there is a tradeoff in the selection of γ, as a higher magnification factor also introduces more artifacts. Table 2 shows the average video quality metrics PSNR and SSIM for our output videos on an exemplary task (Task C) with different choices of γ. For both the pulse and respiration magnification tasks, the video quality decreases to different extents with the increase of γ. Given that artifacts considerably reduce the PSNR and SSIM metrics (as shown in Table 1), the fact that the values do not change dramatically with γ shows that few artifacts are introduced with increasing magnification.

To quantitatively analyze the effects of N and γ on the magnification factor, we drew exemplary learning curves for one of our videos in Figure 9(a) with different choices of parameters. The curves show the changes of our CNN loss, the L2 norm of the differential motion signal, which is a good estimate of the target motion magnitude. According to the learning curves, both N and γ positively correlate with the motion magnitude, and the relationship between N and the motion magnitude is semi-linear. However, a longer step size with fewer iterations is not equivalent to a shorter step with more iterations. In Figure 9(b), we show how the loss changes along with the product of N and γ, which suggests that relatively small step sizes and more iterations can increase the magnification factor more efficiently.

5.4 Gradient Ascent Mechanisms

Compared with traditional gradient ascent, we added two new mechanisms to adapt the approach to the task of video magnification: L1 normalization and sign correction. Here, we show experimental results to support the necessity of these mechanisms.

The goal of applying L1 normalization is to make sure every frame in a video is magnified to the same level. To achieve the best results, we found the gradient $\nabla\|y(X_n|\theta)\|_2$ in Equation (1) should be approximately proportional to the motion representation X_n. However, it was not the case without L1 normalization. Figure 10 shows the time series and histograms of the L1 norms of X_1 and $\nabla\|y(X_1|\theta)\|_2$ for a 30-second video. It is obvious that the distribution of the motion representation is Gaussian, while the distribution of the gradient is highly skewed. To correct the distribution of the gradient to match the motion representation, it needs to be L1 normalized.

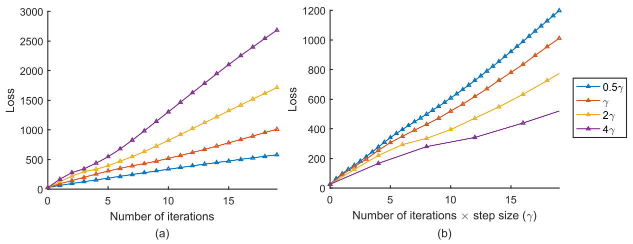


Fig. 9. Learning curves: (a) The change of the CNN loss with different numbers of iterations N and different step sizes γ. (b) The change of the CNN loss with different products of N and γ.

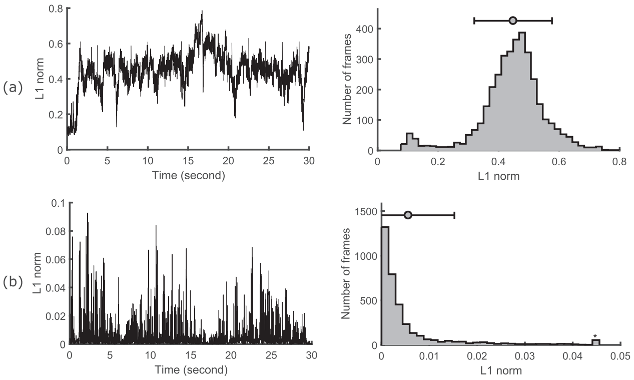


Fig. 10. (a) Time series and histograms of the L1 norms of the input motion representation X_1 for a 30-second video. (b) Time series and histograms of the L1 norms of the motion gradient $\nabla\|y(X_1|\theta)\|_2$ for the same video.

In Figure 11, we show the pixel-wise correlation coefficients between the input and the magnified motion representations (after bandpass filtering) in the respiration magnification task, with and without the sign correction mechanism. When there is no sign correction, the correlation coefficients have both positive and negative values (Figure 11(b)). As introduced in Section 3.1, the negative values appear because the target motion could be amplified with its direction reversed. In the example in Figure 11(b), most of the negative values happen on the background, which are negligible as the background has nearly no motion to amplify, but some of them are on the human body, which will cause the output video to be blurry on magnification. After sign correction is applied, all the correlation coefficients become positive (Figure 11(c)). Many of the pixels in the background have a correlation close to one as there was little texture or motion in the background. We have masked these pixels in Figure 11(d).

5.5 Generalizability to Other Videos and Datasets

We applied our method to videos used in Wu et al. [2012]. For these examples, we used the network trained on the previously described dataset and did not retrain it or use an auxiliary signal from the test videos. The impact of motion magnification on the "face" video is shown in Figure 12(a) and (b). The impact of color

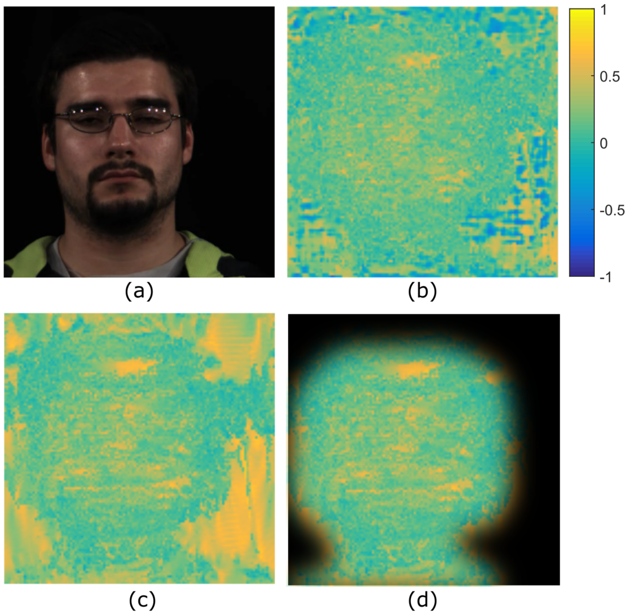


Fig. 11. Pixel-wise correlation coefficients between the input and magnified motion representations in the respiration magnification task, without the sign correction mechanism (b) and with the sign correction mechanism (c). A masked version of (c) is shown with black pixels in the background of the original input video colored black. Note: Due to the rotation of the head in the video, the region is larger than the face in the example frame, some of the pixels were black from only some of the frames.

change magnification on the "baby2" video is shown in Figure 12(c) and (d). Our method generalized well to these new videos, despite the "baby2" video being quite different in composition (although the task is the same). We did have to change the step-size γ in order to increase the amplification factor to get the results on these videos. In these cases, the step size γ was set to 1×10^{-2} for motion and 1×10^{-3} for color change magnification. See the supplementary video for examples of the motion and color change magnified videos.

5.6 Limitations

Our method does not result in entirely artifact-free magnified videos. For example, in the respiration magnification case (see Figure 12 and supplementary video), the edges of the shoulders become blurred and some definition is lost. In pulse magnification, the artifacts are a little more subtle and typically manifest as blotchy changes in color. DeepMag is a supervised method and, therefore, for optimal results, representative training data is required. We have demonstrated a level of generalizability; taking the example of the baby in Figure 12, it is possible to magnify the pulse color changes effectively using a model trained only on adults. This is presumably because the skin of the baby resembles that of adults and the effect of the pulse on the color changes are similar (albeit at a different frequency). In our participant independent training, we used a corpus of videos of 20 subjects and found that this generalized fairly well for many videos similar to those in Figure 12. However, it would not be effective to train a model on pulse data and then expect it to be able to accurately magnify respiration motions.

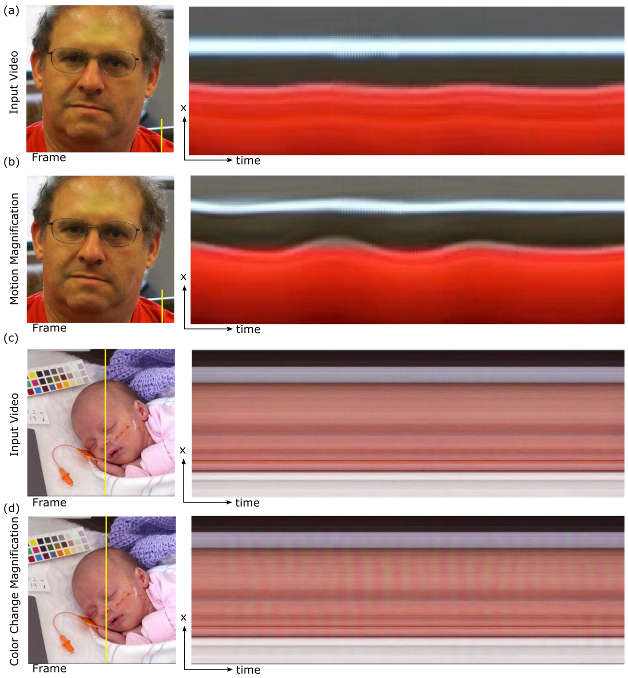


Fig. 12. Scan lines for motion (respiration) magnification method applied to the "head" video and color change (pulse) magnification applied to the "baby2" video from Wu et al. [2012]. In these cases, the step size γ was set to 1×10^{-2} for motion and 1×10^{-3} for color change magnification. The yellow line shows the source of the scan line in the frames. (a) Original input video of head. (b) Motion magnified video. (c) Input video video of baby. (d) Color change video magnified video. See the supplementary video for examples. Note: Our method does not eliminate all artifacts in the magnified videos. For example, in (b), the edges of the shoulders become blurred and some definition is lost.

6 CONCLUSIONS

Revealing subtle signals in our everyday world is important for helping us understand the processes that cause them. We present a novel single deep neural framework for video magnification that is robust to large rigid motions. Our method leverages a CNN architecture that enables magnification of a specific source signal even if it overlaps with other motion sources in the frequency domain. We present several methodological innovations in order to achieve our results, including adding L1 normalization and sign correction to the gradient ascent method.

Pulse and respiration magnification are good exemplar tasks for video magnification as these physiological phenomena cause both subtle color and motion variations that are invisible to the unaided eye. Our qualitative evaluation illustrates how the PPG color changes and respiration motions can be clearly magnified. Comparisons with baseline methods show that our proposed architecture dramatically reduces artifacts when there are other rotational head motions present in the videos.

In a systematic quantitative evaluation our method improves the PSNR and SSIM metrics across tasks with different levels of rigid motion. By magnifying a specific source signal, we are able to maintain the quality of the magnified videos to a greater extent.

We focused our attention on pulse and respiration magnification in this article. We have tested our approach qualitatively on video examples from other datasets. We cannot guarantee that the performance improvements will be universal across other domains. We feel that physiological signal magnification is a particularly useful application of video magnification and, hence, we chose pulse and respiration as exemplar tasks that have different properties (color changes and motions); however, we see no reason to believe that our approach could not be applied successfully in other domains.

REFERENCES

Guha Balakrishnan, Fredo Durand, and John Guttag. 2013. Detecting pulse from head motions in video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2013), 3430–3437.

Pak-Hei Chan, Chun-Ka Wong, Yukkee C. Poh, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Ming-Zher Poh, Daniel Wai-Sing Chu, and Chung-Wah Siu. 2016. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *Journal of the American Heart Association* 5, 7 (2016), e003428.

Weixuan Chen and Daniel McDuff. 2018. DeepPhys: Video-based physiological measurement using convolutional attention networks. *arXiv preprint arXiv:1805.07888* (2018).

Gerard de Haan and Vincent Jeanne. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2878–2886.

Gerard de Haan and Arno van Leest. 2014. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological Measurement* 35, 9 (2014), 1913.

Mohamed Elgharib, Mohamed Hefeeda, Fredo Durand, and William T. Freeman. 2015. Video magnification in presence of large motions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4119–4127.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. *University of Montreal* 1341, 3 (2009), 1.

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

Thomas B. Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology* 124, 6 (1988), 869–871.

Martin Fuchs, Tongbo Chen, Oliver Wang, Ramesh Raskar, Hans-Peter Seidel, and Hendrik P. A. Lensch. 2010. Real-time temporal shaping of high-speed video streams. *Computers & Graphics* 34, 5 (2010), 575–584.

Temujin Gautama and M. A. Van Hulle. 2002. A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Transactions on Neural Networks* 13, 5 (2002), 1127–1136.

Rik Janssen, Wenjin Wang, Andreia Moço, and Gerard de Haan. 2016. Video-based respiration monitoring with automatic region of interest detection. *Physiological Measurement* 37, 1 (2016), 100–114. http://stacks.iop.org/0967-3334/37/i=1/a=100?key=crossref.be9e80b618c48e376025e318d84dff96.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*. 972–981.

Julian F. P. Kooij and Jan C. van Gemert. 2016. Depth-aware motion magnification. In *Proceedings of the European Conference on Computer Vision*. Springer, 467–482.

Antony Lam and Yoshinori Kuno. 2015. Robust heart rate measurement from video using select random patches. In *Proceedings of the IEEE International Conference on Computer Vision*. 3640–3648.

Ce Liu, Antonio Torralba, William T. Freeman, Frédo Durand, and Edward H. Adelson. 2005. Motion magnification. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 519–526.

Daniel McDuff. 2018. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Daniel McDuff, Justin R. Estepp, Alyssa M. Piasecki, and Ethan B. Blackford. 2015. A survey of remote optical photoplethysmographic imaging methods. In *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6398–6404.

Daniel McDuff, Sarah Gontarek, and Rosalind Picard. 2014. Improvements in remote cardio-pulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering* 61, 10 (2014), 2593–2601.

Hamed Monkaresi, Rafael A. Calvo, and Hong Yan. 2014. A machine learning approach to improve contactless heart rate monitoring using a webcam. *IEEE Journal of Biomedical and Health Informatics* 18, 4 (2014), 1153–1160.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. 2015. Deepdream-a code example for visualizing neural networks. *Google Res* 2 (2015).

Masahiro Mori. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.

Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Frédo Durand, William T. Freeman, and Wojciech Matusik. 2018. Learning-based video motion magnification. *arXiv preprint arXiv:1804.02684* (2018).

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill* 2, 11 (2017).

Ahmed Osman, Jay Turcot, and Rana El Kaliouby. 2015. Supervised learning approach to remote heart rate estimation from facial videos. In *Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, 1–6.

Silvia L. Pintea and Jan C. van Gemert. 2016. Making a case for learning motion representations with phase. In *Proceedings of the European Conference on Computer Vision*. Springer, 55–64.

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

Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. 2011. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering* 58, 1 (2011), 7–11.

Javier Portilla and Eero P. Simoncelli. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision* 40, 1 (2000), 49–71.

E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. 1992. Shiftable multiscale transforms. *IEEE Transactions on Information Theory* 38, 2 (March 1992), 587–607. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=119725.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 95.

Chihiro Takano and Yuji Ohta. 2007. Heart rate measurement based on a time-lapse image. *Medical Engineering & Physics* 29, 8 (2007), 853–857.

K. S. Tan, R. Saatchi, H. Elphick, and D. Burke. 2010. Real-time vision based respiration monitoring system. In *Proceedings of the 7th International Symposium on Communication Systems Networks and Digital Signal Processing (CSNDSP)* (2010), 770–774.

L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh. 2014. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement* 35, 5 (2014), 807–831.

Sergey Tulyakov, Xavier Alameda-Pineda, and Elisa Ricci. 2016. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the Computer Vision Pattern Recognition* (2016), 2396–2404.

Wim Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. 2008. Remote plethysmographic imaging using ambient light. *Optics Express* 16, 26 (2008), 21434–21445.

Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. 2013. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 80.

Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. 2014. Riesz pyramids for fast phase-based video magnification. In *Proceedings of the 2014 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–10.

Jue Wang, Steven M. Drucker, Maneesh Agrawala, and Michael F. Cohen. 2006. The cartoon animation filter. In *ACM Transactions on Graphics (TOG)*, Vol. 25. ACM, 1169–1173.

Wenjin Wang, Albertus Den Brinker, Sander Stuijk, and Gerard De Haan. 2016. Algorithmic principles of remote-PPG. *IEEE Transactions on Biomedical Engineering* PP, 99 (2016), 1–12.

Wenjin Wang, Sander Stuijk, and Gerard de Haan. 2015. Exploiting spatial redundancy of image sensor for motion robust rPPG. *IEEE Transactions on Biomedical Engineering* 62, 2 (2015), 415–425.

Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John V Guttag, Frédo Durand, and William T Freeman. 2012. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics* 31, 4 (2012), 65.

Shuchang Xu, Lingyun Sun, and Gustavo Kunde Rohde. 2014. Robust efficient estimation of heart rate pulse from video. *Biomedical Optics Express* 5, 4 (2014), 1124.

Yichao Zhang, Silvia L. Pintea, and Jan C. Van Gemert. 2017. Video acceleration magnification. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition* (2017), 502–510.

Received July 2018; revised March 2020; accepted May 2020