

國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Institute of Computer Science and Information Engineering

College of Electrical and Computer Science

National Taiwan University

Master Thesis

利用注意力機制的輕量深度學習模型進行非接觸式生理徵象偵測

Non-contact vital sign monitoring  
with attention-based lightweight deep learning model

蘇柏燁

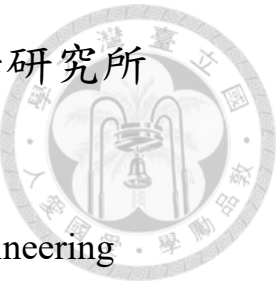
Po-Yeh Shu

指導教授: 李明穗 博士

Advisor: Ming-Sui Lee Ph.D.

中華民國 111 年 9 月

September, 2022





# Acknowledgements

終於要從台大資工所畢業了，不知道是不是疫情的關係，總覺得時間過得特快，感覺準備考研究所彷彿還是昨天的事。這段時間首先要感謝我的父母，提供我各種生活上的支援，讓我從準備考試到念研究所的期間都不需要為了生計而煩惱。以及李明穗老師這兩年來的指導，從碩一時閱讀論文的指導到碩二接了台大醫院的計畫後給我的研究方向跟內容上的建議，讓我可以順利完成研究和口試。

關於研究的部分真的特別感謝台大醫院蘇東弘醫師的團隊還有嘉瑋學長拍攝了大量的資料讓我進行研究，多虧了這些資料才能有這篇論文。每次開會時蘇醫師和周承復老師的建議也給了我很大的啟發。

再來要感謝實驗室的夥伴，謝謝世耘、偉綸、昱霖，這兩年一起做 DIP 作業，一起當助教，一起在實驗室射飛鏢跟打遊戲聊油宅的話題還有討論學術方面的東西真的是非常美好的一個回憶。還有謝謝釋翔、立淞、柏維、皓廷，你們的加入也讓實驗室變得更加熱鬧，雖然我碩二後因為疫情跟通勤的關係變得比較少直接來學校了哈哈。

另外也要謝謝一起修課的其他實驗室的夥伴們，謝謝婕吟跟如芬，修 ADA 的時候真的有妳們一起討論才能順利面對各種作業；謝謝瑋安開的 ML 群讓我們在修機器學習的時候更加安心了；謝謝恬儀在修 NLP 和高等計網時都在同一組 Carry 我，真的感謝這兩年來一起修課的夥伴們。還有要特別感謝的就是日文歌曲

社的朋友們，放學後可以去社團和大家一起唱日 K 還有聊各種二次元的話題，不管是一起聊 Vtuber 或是 Lovelive 或是動漫新番的東西，還有晚上掛在 DC 上玩垃圾遊戲，真的是在研究所生涯中最放鬆的時刻。



要感謝的太多了，那就謝天吧，希望接下來當兵跟求職也能順利。



## 摘要

生理徵象 (vital signs) 是一組表示生命現象的重要訊號，包括呼吸、心跳、血壓等內容。在新冠肺炎 (Covid-19) 的疫情下，遠距醫療的需求正在增加。基於影片對患者的呼吸心跳進行生理徵象分析，將可為遠距醫療帶來幫助。過去的研究指出從人類影片分析其呼吸及心跳是可行的。最近幾年，深度學習技術更為這項任務帶來更好的表現，但目前的深度學習方法仍有成本上的限制。訓練深度學習模型需要錄製呼吸或心跳的波型，而支援此功能的設備通常較難以取得，大部分生理徵象偵測的設備僅支援紀錄每分鐘呼吸或心跳次數。另一方面，使用深度學習進行影片處理往往需要使用較大的神經網路模型以及龐大的運算資源進行訓練，這也使的這個任務所需要的運算成本較高。本研究提出了弱監督式學習方法，讓訓練過程不再需要錄製生理訊號的波型，只要有頻率即可進行訓練。另外，結合了傳統電腦視覺演算法與深度學習，大幅降低了模型的大小及訓練過程需要的運算資源，本研究中提出的基於注意力機制的方法，可自動選擇適合偵測的區域，提升了呼吸心跳偵測的準確度。本研究使用台大醫院錄製的資料集來訓練提出的模型，並在公開的資料集上進行測試，在該公開資料集上的表現超越了目前表現最好的研究。

**關鍵字：**生理徵象、遠程光體積描計圖法、呼吸偵測、電腦視覺、深度學習、弱監督式學習、注意力機制



# Abstract

Vital signs are a group of the most important medical signs, including respiration, heart rate, and blood pressure. During the Covid-19 pandemic, the demand for remote telemedicine has increased. It would be helpful if vital signs such as heart rate or respiration rate can be detected from the patients' videos. Previous researches show that detecting heart rate and respiration rate is feasible. In recent years, deep learning approaches have improved the performance of this task. However, there are still some limitations to this task. To train such a deep learning model, recording vital signs signals is needed, while most devices do not support this function. Most devices such as oximeters can only record the average heart rate or respiration rate. Besides, training such a video processing model costs a lot of computational resources and makes the cost of computational resources high. This research tends to solve these limitations and improve performance. The proposed weakly-supervised training methods make training a vital signs detection network more easily. Only an average heart or respiration rate label for a video is sufficient. Also, the

proposed work combines traditional computer vision algorithms and deep learning. It makes the deep learning model extremely lightweight compared with current end-to-end models. The proposed channel attention architecture can wisely select a proper signal for detecting vital signs. The models are trained on a private dataset recorded by National Taiwan University Hospital and tested on a public dataset. This research has reached a better performance than current state-of-the-art works on the public dataset.

**Keywords:** Vital signs, Remote photoplethysmography, Respiration rate detection, Computer vision, Deep learning, Weakly-supervised learning, Attention



# Contents

	<b>Page</b>
<b>Acknowledgements</b>	<b>i</b>
<b>摘要</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
<b>Chapter 2 Related Work</b>	<b>4</b>
2.1 Video based heart rate detection . . . . .	4
2.2 Video based respiration rate detection . . . . .	5
2.3 Deep learning approaches . . . . .	5
<b>Chapter 3 Method</b>	<b>7</b>
3.1 Vital Sign Detection . . . . .	7
3.1.1 Heart rate . . . . .	7
3.1.2 Respiration rate . . . . .	10
3.2 Vital Sign Detection via Deep Learning . . . . .	12

<b>Chapter 4</b>	<b>Experiment and result</b>	<b>18</b>
4.1	Experiment Environment . . . . .	18
4.2	Dataset . . . . .	18
4.3	Heart Rate . . . . .	20
4.4	Respiration Rate . . . . .	22
4.5	Comparison . . . . .	24
4.6	Discussion . . . . .	27
<b>Chapter 5</b>	<b>Conclusion</b>	<b>28</b>
<b>References</b>		<b>29</b>







# List of Figures

1.1	Overview of the proposed work . . . . .	3
3.1	Proposed framework for heart rate detection . . . . .	8
3.2	The ROIs for heart rate detection . . . . .	9
3.3	Proposed framework for respiration rate detection . . . . .	10
3.4	ROIs for respiration rate detection . . . . .	11
3.5	An overview of the deep learning model architecture . . . . .	12
3.6	The proposed fusion model . . . . .	13
3.7	The proposed channel attention module . . . . .	15
3.8	Relevant and irrelevant frequency range . . . . .	16
4.1	Cohface Dataset . . . . .	19
4.2	NTUH-21 Dataset . . . . .	20
4.3	The respiration wave from different subjects . . . . .	22
4.4	The improvement of channel attention . . . . .	23



# List of Tables

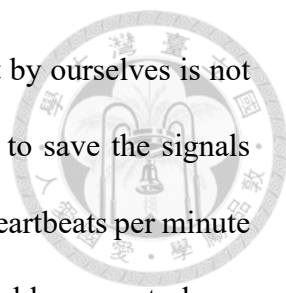
4.1	Heart rate result from each ROI . . . . .	20
4.2	Heart rate result with models . . . . .	21
4.3	Respiration result from each ROI . . . . .	25
4.4	Respiration rate result with models . . . . .	25
4.5	Result compare with SOTA works . . . . .	26
4.6	Model parameter numbers and GPU usage . . . . .	26
4.7	Performance of different models . . . . .	27



# Chapter 1 Introduction

Contact devices are widely used to measure vital signs like pulse or respiration rate. Electrocardiogram machines record heart electrical activity using electrodes placed on the skin. Oximeters record changes in light absorption from the skin which is caused by blood volume changes. The contact devices can provide precise and reliable vital sign signals. However, wearing and taking off those devices are inconvenient. The Sars-CoV-2(Covid-19) pandemic has changed the lifestyle of humans recently. People are asked to keep social distance to avoid infection. Thus, the demand for remote telemedicine increased. Video-based non-contact physiological signals monitoring methods became more significant to remote telemedicine. It would be helpful if the vital signs such as heart rate or respiration rate can be detected from the patients' videos. Previous research has shown that it is possible to detect heart rate from human face video with small intensity variation caused by the blood volume changes. Detecting respiration rate is also possible because of the motion caused by breathing.

In recent years, deep learning-based approaches have improved the performance and robustness of video-based vital signs detection. However, there are some constraints in training such a deep learning model. First, to train a vital sign extractor network, we need the videos as the input and vital signs recorded by contact sensors as the ground truth, but there are only a few datasets that provide videos and corresponding signals. Moreover,



most of those datasets are not freely available. Collecting the dataset by ourselves is not easy work, either. Most vital signs measuring devices are not able to save the signals recorded from the sensors. They provide only calculated values like heartbeats per minute or breaths per minute. The devices whose sensor signals are available are not cheap. Designing a training method that can train a model only with average frequency may be helpful for data collection. Another problem is that the current deep learning approaches are based on an end-to-end model architecture. Those models take a sequence of frames as their input and need huge GPU resources for training. This work is trying to find a method that is able to avoid the huge cost of GPU resources and remain the robustness of deep learning. Our work provides a video-based vital signs detecting framework, which is able to detect heart rate and respiration rate from video. The combination of non-learning-based feature extraction and deep-learning-based algorithms lets our neural network model be very lightweight and trainable with normal personal computers. The weakly-supervised learning method makes the vital signs model trainable with only heart rate or respiration rate labels, without a vital sign signal ground truth. Also, our channel attention method can choose a proper region from different ROIs, and improve the robustness of vital sign detection.

The experiments and model training is run on a dataset recorded by National Taiwan University Hospital. A public dataset called Cohface with pulse and respiration wave ground truth is used to evaluate the performance and compare our result with current state-of-the-art works and compare our predicted vital signs with ground truth.

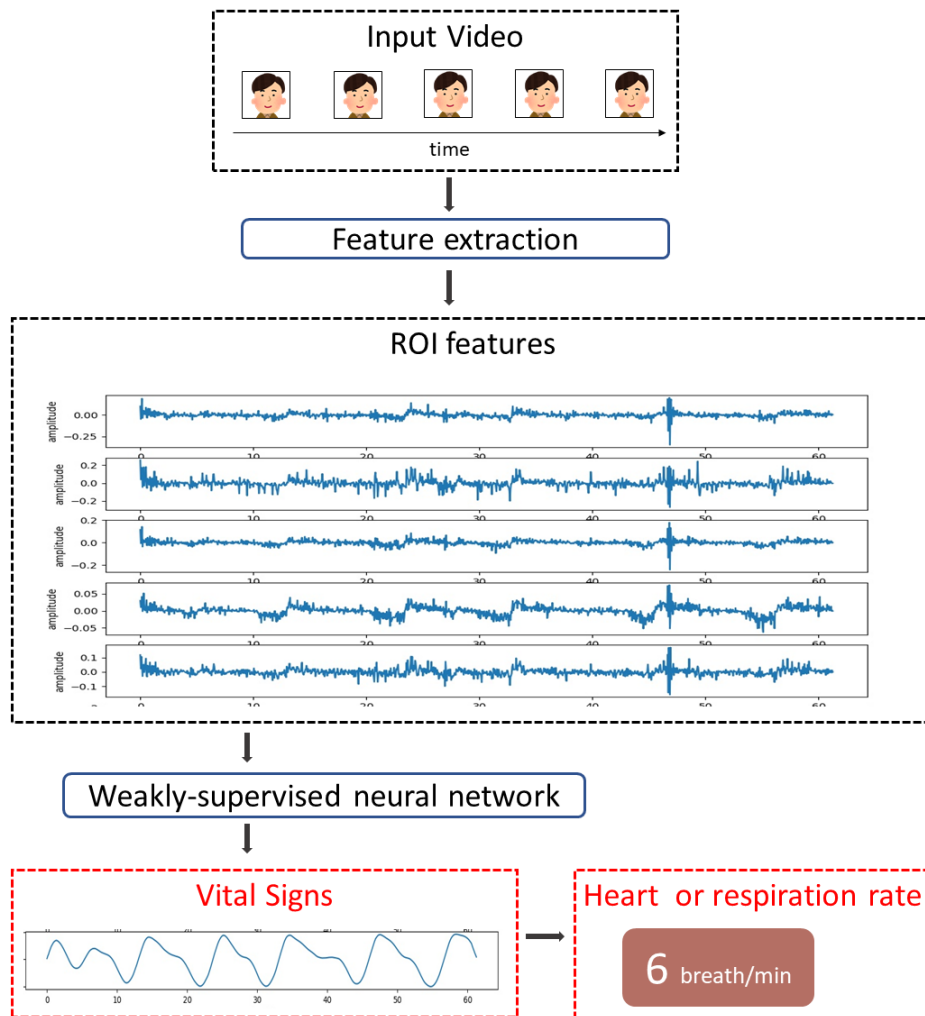


Figure 1.1: Overview of the proposed work



## Chapter 2 Related Work

The previous works about video-based heart rate and respiration rate detection are reviewed in this chapter. Both algorithm-based methods and deep-learning-based methods are mentioned along with some constraints about the current state-of-the-art works.

### 2.1 Video based heart rate detection

Remote photoplethysmography (rPPG) is a non-contact method for video-based heart rate detection. The reflected light from skin changes due to the variation of blood volume, and photoplethysmography (PPG) is a method that detects such variation[3][1]. Verkrusse et al.'s works showed that it is possible to detect pulse signals from an RGB facial video and the signal is called remote PPG (rPPG). The green channel intensity can be used to extract rPPG[18]. Some early works for video-based heart rate detection are based on recording the changes of color on facial skin[20][6][14][12].

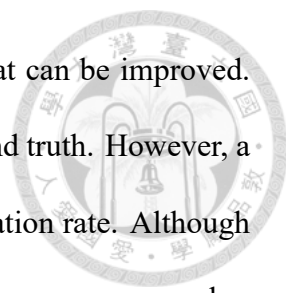
## 2.2 Video based respiration rate detection



The motion caused by breathing is recognizable by human vision. Thus, motion-based signals can be used to detect respiration rates. Tveit et al. [4] record the changes in each frame's local phase as their motion signal and computes the respiration rate with Fourier transform. There is another way to detect respiration rate. Chen et al.[17] mentioned that the human heart rate becomes faster when inhaling, and becomes slower when exhaling. Their work detects the rPPG signal first and computes respiration rate from heart rate variation.

## 2.3 Deep learning approaches

In recent years, some deep learning approaches for vital sign detection were proposed. Early works are simple but the signals can be corrupted by noises easily. Deep learning methods improved the performance by learning the relation between video and label[8]. Deepphys[5] predicts each frame's signal value with 2D-CNN and the spatial attention method helps the model find important regions. Nowara et al.'s work[13] is based on the model proposed by Deepphys. The signal extracted from the background region is useful for denoising because it may contain some information about noises. PhysNet[21] detects rPPG with an end-to-end model. A 3D-CNN and a model which consists of 2D-CNN and LSTM are proposed in PhysNet. Although the deep learning methods have reached the state-of-the-art and shown the robustness of noises, they need the ground truth signal that is hard to collect. Gideon et.al proposed a self-supervised training method[8]. With their training tips, The vital sign detecting model can be trained with unlabeled videos.



In video-based vital sign detection, there are still some problems that can be improved. First, the supervised training methods need expansive vital sign ground truth. However, a lot of devices like oximeters can record average heart rate and respiration rate. Although labeling a video with a vital sign wave is difficult, labeling a video with an average number like “The heart rate of the person in this video is 70 per minute” is much easier. Recording such kinds of ground truth does not need an expansive device. It would be helpful if there exists a weakly-supervised training method that only needs the average numbers as ground truth. This weakly-supervised method may reach a balance point between supervised and self-supervised methods.

Another problem is that training video-based deep learning models cost a lot of usage of GPU resources. The length of the videos for training can not be too long due to the constraints of GPU memory. In PhysNet-3DCNN, the length of training clips is about 32 to 256 frames. It is shorter than 10 seconds. The question is: Do we really need to use a lot of GPU resources to train an end-to-end model? There are 2 main tasks for the deep learning models, One is compressing a frame to a feature vector or a single value, and the other is denoising the features with temporal information. If an algorithm-based method or a pre-trained model is sufficient for compressing the frames, we can compress the frames with the method first and train a model just for denoising. The denoising model can be trained with low GPU memory usage, and the model can also see a longer time.





## Chapter 3 Method

The proposed works are described in this chapter. The method of heart rate and respiration rate detection will be described separately. To further improve the performance of the above methods, a lightweight neural network model is proposed. The architecture and training strategy of the neural network model will be described later.

### 3.1 Vital Sign Detection

#### 3.1.1 Heart rate

A framework shown in Figure 3.1 for video heart rate detection is proposed. First, we detect and select certain regions on face as our ROIs. Then we calculate the average pixel value within the ROIs in each frame as our rPPG signal.

- **ROI selection**

In previous works, the cheek region is popular for ROI selection. The average color of pixels in the cheek region can be used to extract rPPG signals and predict heart rate[12][18][17]. In Convolution Neural Network (CNN) based approaches, ROI selection is not needed. The models detect important parts of the face and extract rPPG signals automatically. Recent research shows that spatial attention is useful

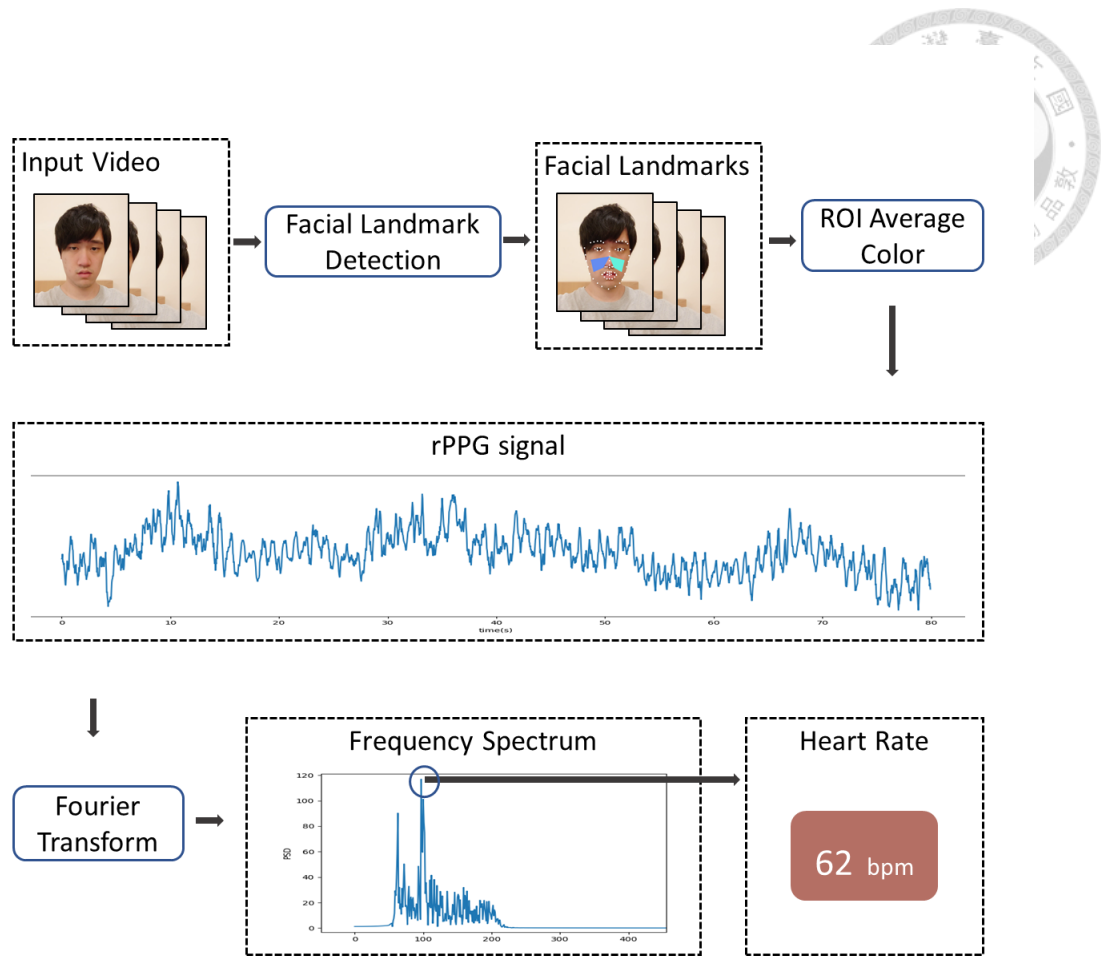
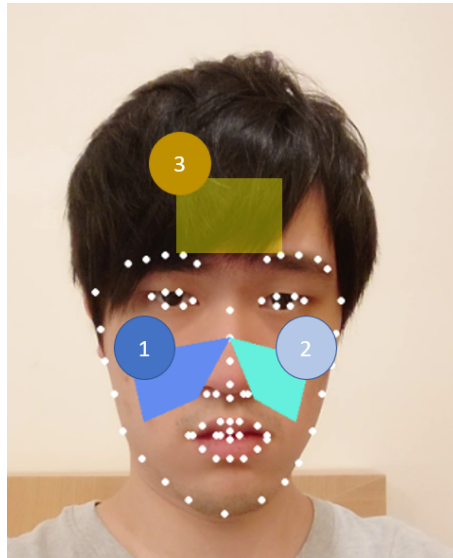


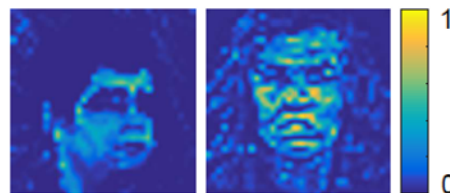
Figure 3.1: Proposed framework for heart rate detection

for CNN based rPPG extraction. Observing the spatial attention masks, we can find that the models tend to assign the cheek and forehead region higher attention score as the images shown in figure 3.2.(b)[5] [13]. We can speculate that after training with face videos and heartbeat signals, the models “think” that cheek and forehead region is more important for rPPG extraction.

According to previous works’ ROI selection and attention map for CNN based method, we take 3 regions as our ROIs: left cheek, right cheek, and forehead. We calculate the average color of each ROI as rPPG signals. To obtain the ROI positions, we use a pretrained facial landmark predicting model[2]. Each frame will be input to the model and the output of it is the position of 68 facial landmarks. Then



(a) The facial landmarks and the ROIs for heart rate detection



(b) An visualization of DeepPhys's attention mask[5]

Figure 3.2: The ROIs for heart rate detection

the boundary of ROIs is calculated based on those facial landmarks as the image shown in figure 3.2.(a).

#### • rPPG signal

For each frame in the input video sequence, the average RGB color of each pixel in ROI will be calculated. The green channel intensity of the average color is what we want for rPPG signals. Since we can obtain a single value from a frame, an intensity-time signal can be generated from a video sequence. We call this intensity-time signal rPPG and it will be used to estimate heart rate.

#### • Heart Rate Calculation

Given a rPPG signal, we perform Fast Fourier Transform (FFT) on the signal to obtain the frequency domain of the PPG signal. Then we find the frequency with

maximum magnitude within the frequency range of human heart rate.



### 3.1.2 Respiration rate

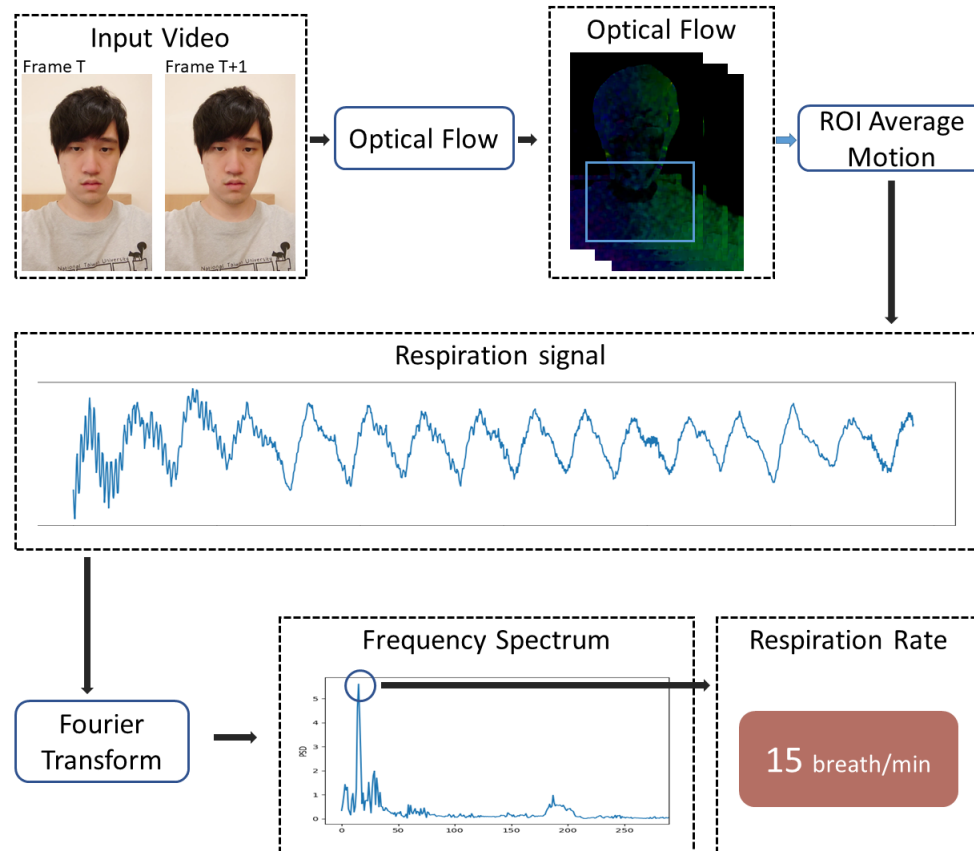


Figure 3.3: Proposed framework for respiration rate detection

The proposed framework for respiration rate detection is shown in Figure 3.3. The method is similar to heart rate detection since the respiration wave is also a periodic signal.

#### ROI selection

Watching the videos with breathing humans, we can easily observe that there exists some motion near the chest region and those motions are caused by breathing.

We speculate that motion or interframe information can provide useful features for

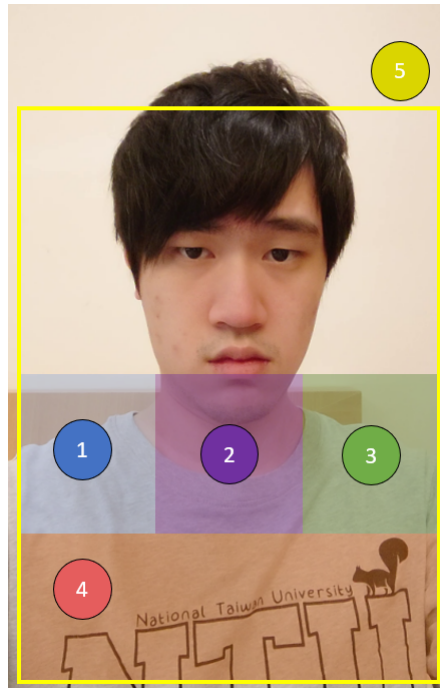


Figure 3.4: ROIs for respiration rate detection

video-based respiration detection.

Previous works show that motion near the chest can be used to predict respiration rate[17][16]. To obtain the region of the chest, we use the Haar feature-based cascade classifier. Haar cascade is a popular algorithm for object detection[19]. It is widely used for face or body detection because of its precision and low complexity. The Haar feature-based cascade classifier we use is provided by OpenCV. We detect the chest bounding box first, then the bounding box is splitted into 5 ROIs as shown in Figure 3.4. The 5 ROIs are left shoulder, neck, right shoulder, chest and the bounding box from head to upper body.

- **Motion signal**

To obtain the motion of each frame for respiration rate analysis, an optical flow algorithm is applied. Optical flow is a computer vision that estimates the motion between the pixels of 2 images. For each frame and its next frame in a video, the optical flow is predicted with an algorithm provided by OpenCV[7]. The summation

within the ROI on the vertical flow map is the motion signal for respiration rate prediction.



- **Respiration rate calculation**

The Respiration rate is also calculated from the frequency domain of the motion signal, while it is slightly different from heart rate. Because the respiration rate changes more frequently than the heart rate, a 12-second-long sliding window is applied to calculate local respiration rate. Then the average of all local respiration rates is computed as the average respiration rate of the input video.

### 3.2 Vital Sign Detection via Deep Learning

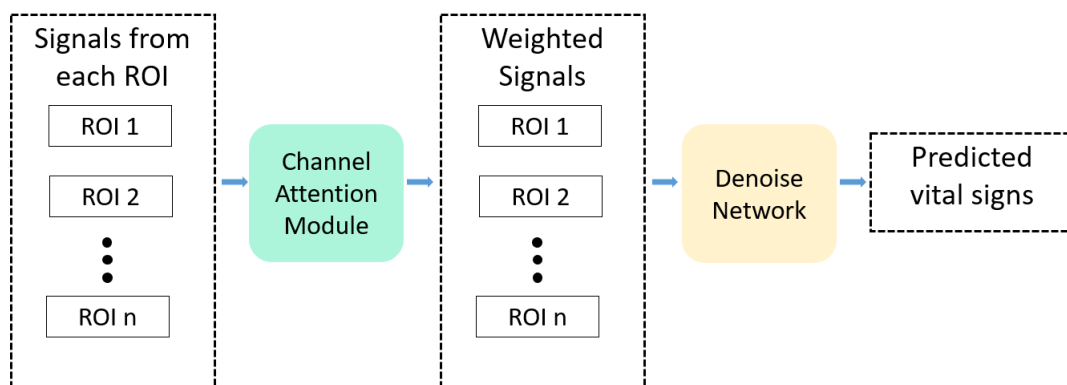


Figure 3.5: An overview of the deep learning model architecture

Although the algorithm-based methods that we proposed above can detect vital signs easily under well controlled recording environments, the vital sign signal can be corrupted easily by noises such as illumination variations, extreme light and shadow, or subject’s moving. The corruption on signals may make us choose the wrong frequency in the frequency domain and predict the wrong heart rate or respiration rate. To improve the robustness of video vital sign detection, we proposed some 1D-CNN based models. The aim of the CNN models is to extract the “real” vital sign signals from noisy signals. The model’

s input is the 1D signals obtained from each ROIs. Then the model will combine those signals and output a 1D signal as the denoised result. The architecture of the proposed deep learning model is shown in Figure 3.5. The model is consist of two modules. The channel attention module attempts to assign each signal an attention weight. The signal with higher attention weight is assumed to have higher discriminating power. The denoise network attempts to extract the "real" vital signs from the noisy input signals.

• **Denoise model**

The model architecture is shown in Figure 3.6. The input of the model is the signals

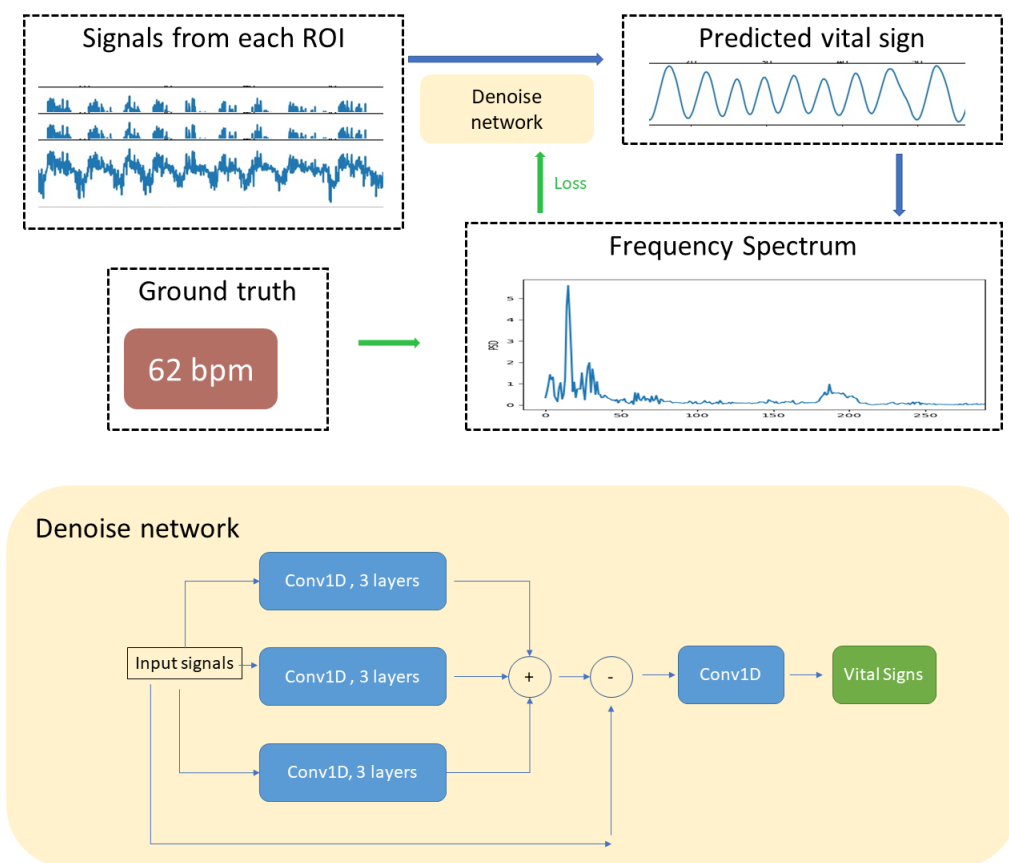
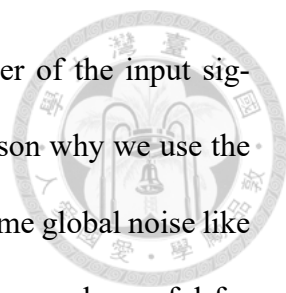


Figure 3.6: The proposed fusion model

output from each ROIs and the signal from the background region. Those signals



are concatenated as a  $C \times L$  array. The  $C$  denotes the number of the input signals and the  $L$  denotes the length of the input signals. The reason why we use the background signal is that the background signal may contain some global noise like camera moving or illumination variations. Such global noises may be useful for denoising[13]. The input is fed into three convolution blocks with different kernel sizes and added. Those three convolution blocks have different receptive fields, and we hope that they can process the signals on different scales[11]. The output of the model is a 1D vital sign signal that is denoised. The heart rate or respiration rate will be calculated by the method mentioned in 3.1 or 3.2.

- **Channel Attention Module**

Although the denoise model generated a cleaner signal and reached better performance than the original ROI signal, we noticed some problems with this model. For example, in respiration rate detection, there are a lot of videos in our training data that have cleaner signals in certain regions. In other regions, the respiration waves are not obvious. It would let our model tend to choose the signals from the former regions, and ignore the signals from the latter. To avoid such an overfitting problem, a method based on channel attention is proposed to solve it. Channel attention is a technique that multiplies each channel by attention weights, and improves the performance of a deep learning model[10]. In our channel attention method, each channel is fed into a score predictor independently. The score predictor consists of convolution layers, global max pooling layer, and fully connected layers. The output of the score predictor is a single value called “attention weight”. The values are transformed between 0 and 1 with the Softmax function. Each channel is multiplied by its “attention weight” first and processed by the denoise model. In



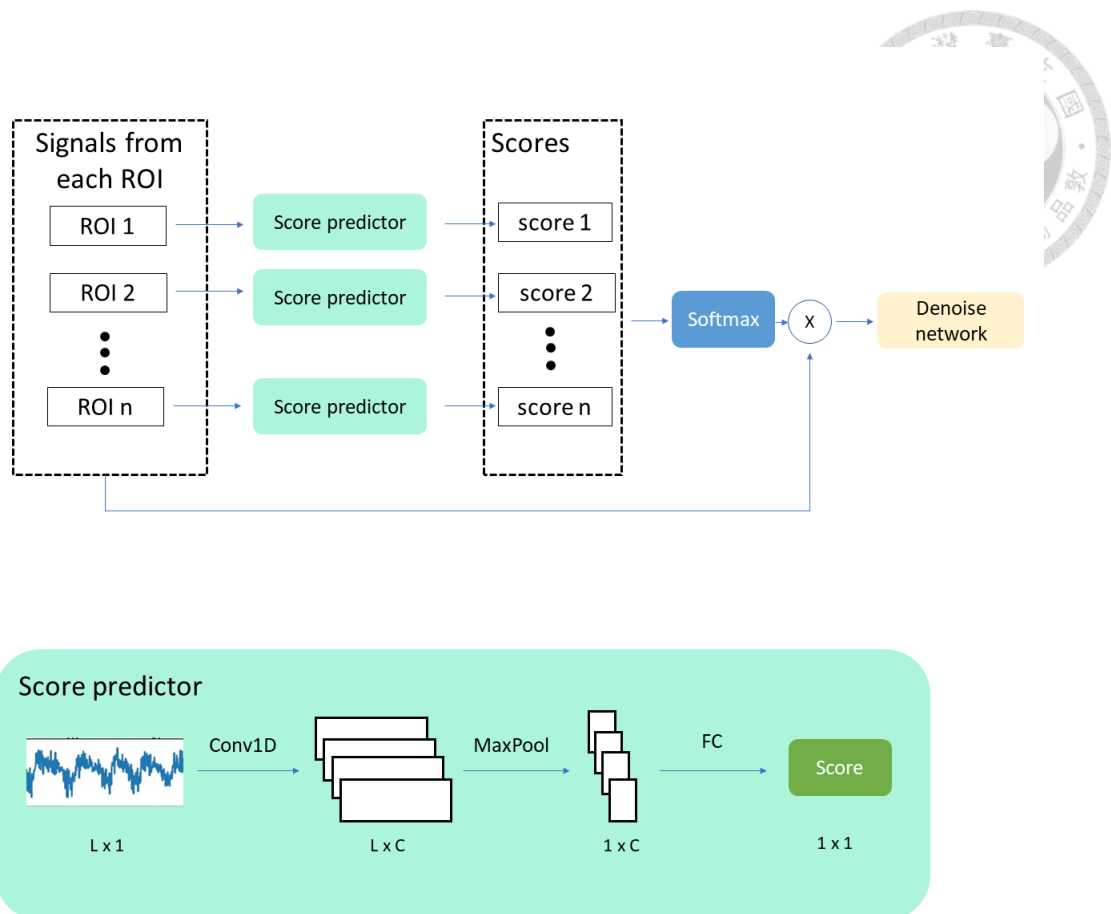


Figure 3.7: The proposed channel attention module

another word, the proposed attention method will choose an ROI automatically.

- **Loss functions**

We use several loss functions to train the deep learning models. All of them are weekly or self-supervised. That is, we don't need a pulse or respiration wave signal to train our models. Only a single number (BPM) is needed for the input video sequence.

- **Irrelevant Power Ratio**

This self-supervised loss function was proposed by John et.al[8], we assume that a “clean” signal should have more energy distributed between a certain

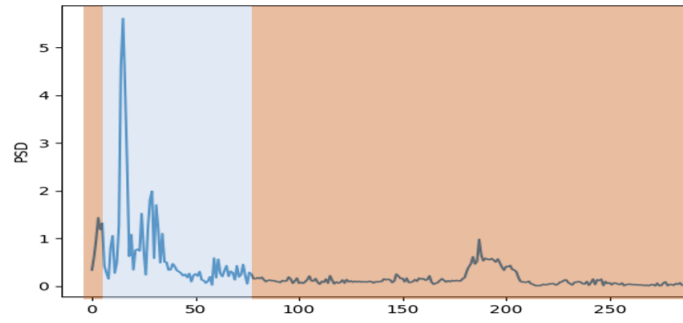


Figure 3.8: Relevant and irrelevant frequency range

frequency range. For example, a heartbeat signal should not contain too much energy that does not belong to the human heart rate. We divide the frequency into relevant (40-180 BPM for heart rate, 4-30 for respiration rate) and irrelevant (other frequency) ranges. The ratio between the total energy in irrelevant frequency range and the whole frequency spectrum is the loss.

Given a vital sign signal  $x$ , the power spectral density (PSD)  $\mathbf{X}$  can be obtained by taking fast Fourier transform (FFT) of  $x$ . This loss can be computed by the equation :

$$L_{IPR} = \frac{\sum_{f \in F^+ \setminus F} \mathbf{X}(f)}{\sum_{f \in F} \mathbf{X}(f)} \quad (3.1)$$

Where  $F^+$  means the relevant heart rate or respiration rate frequency range and  $F$  means all frequency of  $\mathbf{X}$ .

#### • Frequency Power Ratio Loss

To train a model with a frequency label value, this weakly supervised loss function is proposed. It is similar to the Irrelevant Power Ratio. The relevant frequency range is a small window centered at the ground truth frequency. We set the width of the window 10Hz for heart rate, 6Hz for respiration rate detection model. This loss can be computed by the equation :

$$L_{FPR} = \frac{\sum_{f \in F^+ \setminus F} \mathbf{X}(f)}{\sum_{f \in F} \mathbf{X}(f)} \quad (3.2)$$

The  $F^+$  in this loss means the frequencies within the window.

$$\forall f \in F^+, \|f - \Delta\| < f^{GT} \quad (3.3)$$

Where  $\Delta$  means half width of the window and  $f^{GT}$  means the ground truth frequency.

The weighted sum of those two losses is the final loss function for training the neural network.

$$L_{ALL} = \alpha L_{IPR} + \beta L_{FPR} \quad (3.4)$$

The weights  $\alpha$  and  $\beta$  are set at 0.1 and 1 during the experiment.





## Chapter 4 Experiment and result

The dataset we used will be introduced and the result will be shown in this chapter. For each video, a heart rate and a respiration rate value are predicted. To compare the performance of different experiments, the mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and Pearson correlation (PC) of the predicted rate and the ground truth rate are calculated.

### 4.1 Experiment Environment

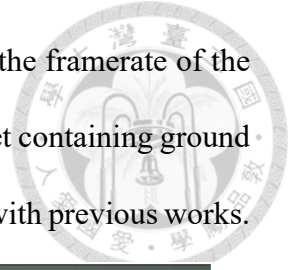
The experiments are run on a personal computer with normal equipment. It's a PC with Ryzen 5800x CPU and Nvidia Geforce RTX3080 GPU. During training, a batch size of 8 examples is used. The model is trained for 30 epoches with Adam optimizer.

### 4.2 Dataset

#### Cohface [9]

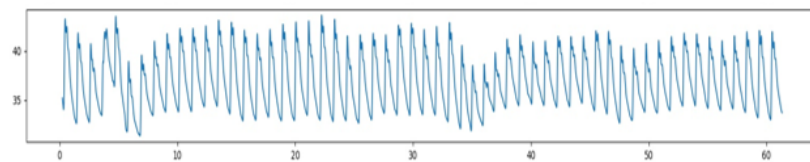
Cohface is a public dataset provided by Idiap Research Institute. It consists of 160 videos from 60 subjects. Each data consists of a 1 minute video, a blood volume pulse (BVP) sensor signal and a respiration belt signal. We think that it is a chal-

lenging dataset because the videos are highly compressed and the framerate of the videos are low (20FPS). This dataset is a freely-available dataset containing ground truth signals. We will compare the performance on this dataset with previous works.

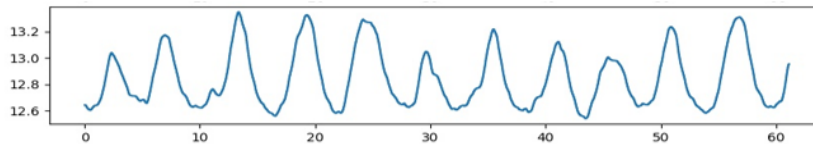


(a) Well controlled conditions

(b) Natural conditions



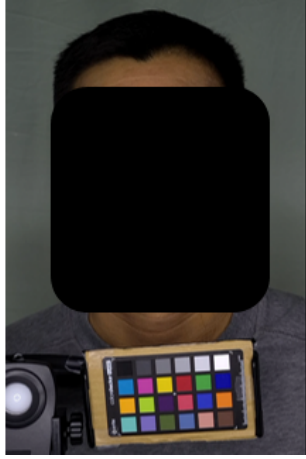
(c) Pulse



(d) Respiration

Figure 4.1: Cohface Dataset

**NTUH-21** This dataset is our private dataset recorded by National Taiwan University Hospital. 111 videos with vital signs labels are recorded. During recording the videos, the vital signs are recorded by the Masimo oximeter. The average heart rate and respiration is recorded every 3 seconds. The oximeter does not provide a waveform signal of pulse or breathing, but the average number is enough for our weekly-supervised training method.



(a) Videos



Date	SpO2 %	PR bpm	PI	PVI	RRp rpm
2021/3/2	98	71	3.5	27	14
2021/3/2	98	71	3.7	27	14
2021/3/2	98	71	4.5	27	15
2021/3/2	98	71	5.3	27	15

(b) Ground truth recorded by Masimo oximeter

Figure 4.2: NTUH-21 Dataset

### 4.3 Heart Rate

Heart Rate								
Region	NTUH-21				Cohface			
	MAE	RMSE	MAPE	PC	MAE	RMSE	MAPE	PC
Left Cheek	<b>2.21</b>	<b>6.06</b>	<b>0.026</b>	<b>0.89</b>	<b>15.91</b>	24.15	<b>0.2</b>	-0.2
Right Cheek	3.41	10.66	0.036	0.64	16.94	<b>23.45</b>	0.22	<b>-0.07</b>
Forehead	7.27	15.9	0.08	0.34	19.58	25.41	0.26	-0.08
Not face	15.4	22.1	0.18	0.17	28.7	31.14	0.39	-0.02

Table 4.1: Heart rate result from each ROI

**Different ROI performance** The result of heart rate prediction on each ROI is shown in Table 4.1. In NTUH-21, the predicted heart rates from the cheek are already close to the ground truth, and the heart beat is easily seen from the rPPG signal. In Cohface, the performance is not as good as our dataset. The predictions of heart rate are wrong in a lot of cases, and the heart beats are not obvious in the rPPG signal. The reason the performance of those two datasets are different is the resolution of the videos. In NTUH-21 dataset, the videos have over 2k resolution and 30 or 60 frames per minute. However, in Cohface dataset, the videos are highly compressed

<b>Heart Rate</b>									
Region	<b>NTUH-21 (Test set)</b>				<b>Cohface</b>				
	MAE	RMSE	MAPE	PC	MAE	RMSE	MAPE	PC	
Left Cheek	3.64	8.76	0.04	0.81	15.91	24.15	0.2	-0.2	
Right Cheek	6.24	15.79	0.07	0.39	16.94	23.45	0.22	-0.07	
Forehead	7.06	16.68	0.08	0.37	19.58	25.41	0.26	-0.08	
Denoise model	<b>2.26</b>	6.44	0.03	<b>0.89</b>	4.89	10.19	0.07	0.61	
channel attention	5.27	14.23	0.05	0.51	15.59	22.95	0.2	-0.02	
Denoise model with attention	2.27	<b>6.43</b>	<b>0.03</b>	0.64	<b>1.69</b>	<b>3.9</b>	<b>0.02</b>	<b>0.94</b>	

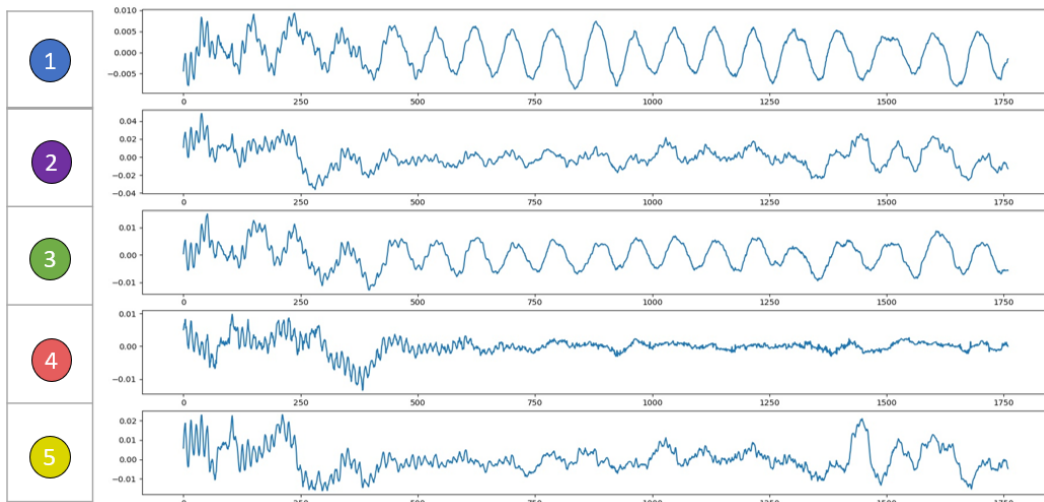
Table 4.2: Heart rate result with models

and have only 20 frames per minute. Those reasons may increase the difficulty of heart rate detection.

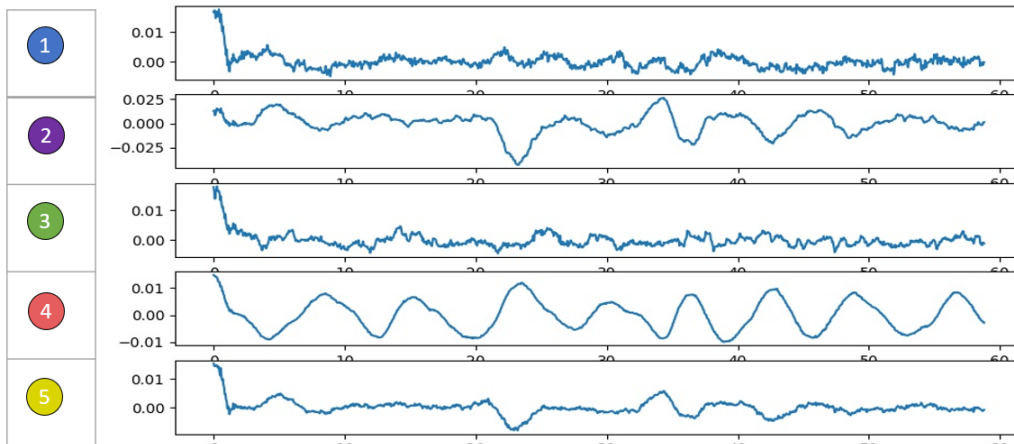
**Denoise network** 68 cases in the NTUH-21 dataset are split as our training set for the neural network approach. The results of different deep learning models are shown in Table 4.2. The performance on the Cohface test set improved a lot. The model seems to be able to deal with the noises caused by the compression algorithms.

**Channel attention** In heart rate detection, the performance does not improve after only applying the channel attention method. Denoising is still important after choosing a proper ROI.

**Denoise network and channel attention** After applying the channel attention method and the denoise network, the performance on Cohface improved from only using the denoise model. Although the channel attention did not improve the performance without applying the denoise model, choosing a proper ROI helps the denoise model work better.



(a) Subject 1



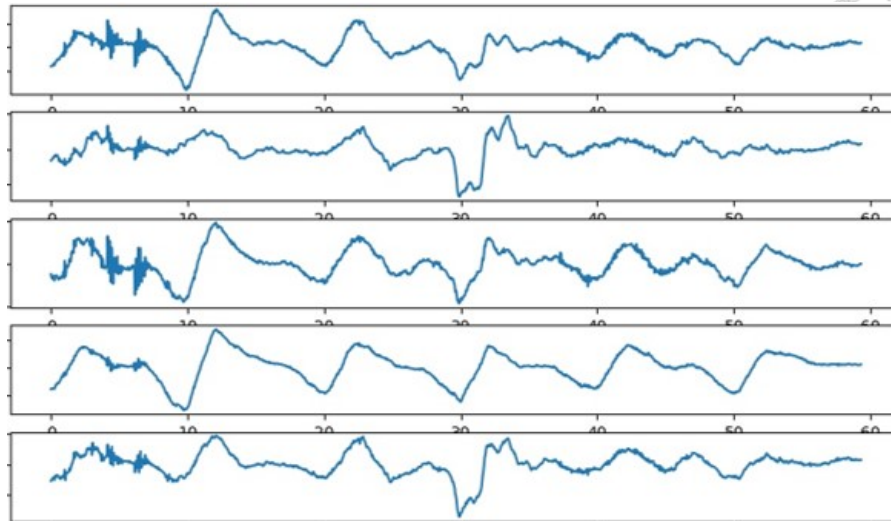
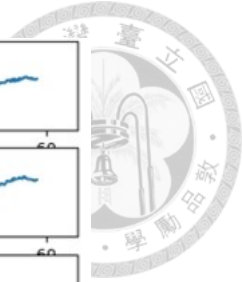
(b) Subject 2

Figure 4.3: The respiration wave from different subjects

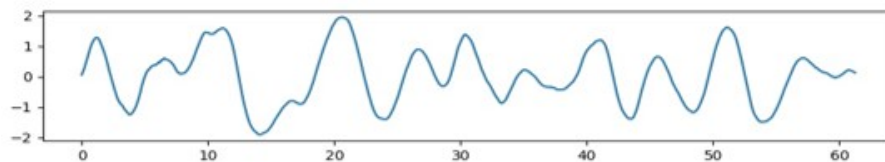
## 4.4 Respiration Rate

**Different ROI performance** The result of respiration rate prediction on each ROI is shown in Table 4.3. Observing the signal from each ROI, we can notice an interesting thing. As the respiration signals shown in Figure 4.3, in some cases, the motion signals from the shoulder region can provide us with the correct respiration rate, while the other regions can't. In some cases, the motion signals from the chest region perform better than those from the shoulder region. If we focus on only a

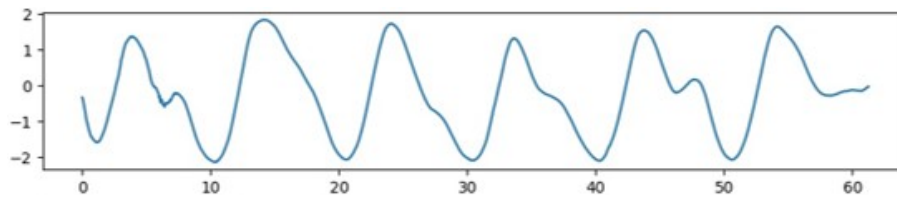




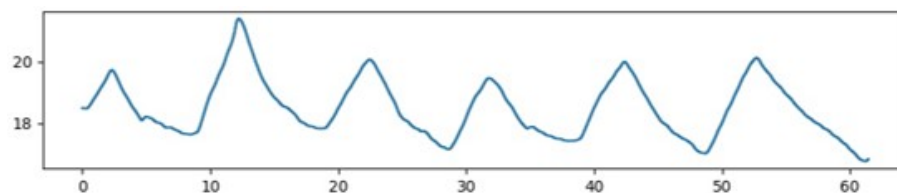
(a) Respiration signals of a subject



(b) Result only using denoise network



(c) Result with channel attention and denoise network

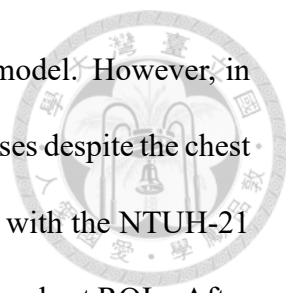


(d) Ground truth signal

Figure 4.4: The improvement of channel attention

certain region, we will get wrong respiration rates in some cases. However, the correct answer may be available from another region. That is the reason we proposed some deep learning models to fusion those signals.

**Denoise network** The result of respiration rate prediction with different models is shown in Table 4.4. The MAE and RMSE decreased after applying the denoise network.



Also, in NTUH-21, the signal became cleaner after using the model. However, in the Cohface dataset, the model outputs wrong signals in some cases despite the chest signals being similar to breathing waves. The model is trained with the NTUH-21 dataset. In this dataset, the shoulder ROIs perform better than the chest ROIs. After training, the model tends to choose the shoulder region and ignore the chest region.

**Channel attention** After applying the channel attention method, the MAE and RMSE decreased on the NTUH-21 dataset, and the MAE decreased on the Cohface dataset. In figure 4.3.(a), the left shoulder region (channel 1) gets the max attention score; in figure 4.3.(b), the chest region (channel 4) gets the max score. The attention scores shows the ability of this architecture for choosing the proper ROI automatically.

**Denoise network and channel attention** After applying the channel attention method and the Denoise network, it reached the best performance on the Cohface dataset. The performance is better than all ROIs and applying the denoise network only. Figure 4.4 shows the improvement of channel attention. The respiration signal from the fourth ROI (chest) is similar to the ground truth signal, while the signal from the first and third ROIs (shoulder) are not. With the channel attention module, the model can judge each channel independently, and choose the best channel automatically. As long as there exists a channel that is similar to a breathing wave, the model will be able to choose it.

## 4.5 Comparison

The performance of our channel attention model on the Cohface dataset is compared with other works that also used this dataset. In heart rate detection, our heart rate detection



Respiration Rate								
Region	NTUH-21				Cohface			
	MAE	RMSE	MAPE	PC	MAE	RMSE	MAPE	PC
Left Shoulder	1.68	2.52	0.11	0.71	1.97	3.97	0.22	0.47
Neck	2.3	2.92	0.14	0.54	2.06	2.98	0.19	0.75
Right Shoulder	<b>1.66</b>	<b>2.42</b>	<b>0.11</b>	<b>0.74</b>	1.9	3.74	0.21	0.53
Chest	2.26	3.52	0.14	0.53	<b>0.71</b>	<b>1.18</b>	<b>0.07</b>	<b>0.95</b>
Upper body	2	2.63	0.13	0.64	1.2	2.08	0.13	0.84

Table 4.3: Respiration result from each ROI

Respiration Rate									
Region	NTUH-21 (Test set)				Cohface				
	MAE	RMSE	MAPE	PC	MAE	RMSE	MAPE	PC	
Left Shoulder	1.97	2.94	0.14	0.72	1.97	3.97	0.22	0.47	
Neck	2.98	3.86	0.21	0.53	2.06	2.98	0.19	0.75	
Right Shoulder	2.08	3.24	0.16	0.65	1.9	3.74	0.21	0.53	
Chest	2.54	3.47	0.18	0.63	0.71	1.18	0.07	0.95	
Denoise model	<b>1.41</b>	<b>1.88</b>	0.1	<b>0.87</b>	0.8	1.35	0.07	0.93	
channel attention	1.71	2.69	0.11	0.72	0.69	1.2	0.07	0.95	
Denoise model with attention	1.44	2.02	<b>0.09</b>	0.85	<b>0.53</b>	<b>0.9</b>	<b>0.05</b>	<b>0.97</b>	

Table 4.4: Respiration rate result with models

performance has less MAE, RMSE, and higher Pearson correlation than Gideon et al.’s supervised and unsupervised results[13]. In respiration rate detection, we compare our result with MT-TS-DAN[15]. The model they proposed is modified from Deepphys. Our MAE is less, too.

We also compare the parameter numbers and the GPU memory usage with previous works while training with batch size 2, 128 frames of 128x128 video sequence. The memory usage is shown in Table 4.6. Our model has the least parameter numbers and costs the least GPU resources. Note that CAN does not need too much GPU resource for training because its input is a single frame. Table 4.7 shows the performance of heart rate detection on the Cohface dataset with different models. The result of the 3D-CNN is provided by Gideon et al.’s work and the result of the CAN is reproduced by ourselves. The proposed work has a light-weight deep learning model without sacrificing accuracy.

Cohface								
Region	HR				RR			
	MAE	RMSE	MAPE	PC	MAE	RMSE	MAPE	PC
Gideon[13] (supervised)	2.5	7.8		0.84				
Gideon (unsupervised)	1.8	5.5		0.37				
MT-TS-DAN[15]						5.72		
Ours	<b>1.69</b>	<b>3.9</b>	<b>0.02</b>	<b>0.94</b>	<b>0.53</b>	<b>0.9</b>	<b>0.05</b>	<b>0.97</b>

Table 4.5: Result compare with SOTA works

Model weight		
Model	Parameters	GPU memory
CAN[5]	795811	15306KB
2DCNN+LSTM[21]	264149	4220MB
3D-CNN[21][13]	858497	7119 MB
Ours	<b>2614</b>	<b>979 KB</b>

Table 4.6: Model parameter numbers and GPU usage



<b>Cohface-HR</b>				
Model	MAE	RMSE	MAPE	PC
CAN[5]	9.46	10.33	0.15	0.68
3D-CNN[13]	1.8	5.5		0.37
Ours	<b>1.69</b>	<b>3.9</b>	<b>0.02</b>	<b>0.94</b>

Table 4.7: Performance of different models

## 4.6 Discussion

In most videos in the NTUH-21 dataset, the quality of the signals extracted from the ROIs is good enough to predict a correct heart rate or respiration rate. However, in the Cohface dataset, the noises caused by video compression and the recording environment make the heart rate prediction fail. The proposed denoise network successfully solved this problem by learning to eliminate the noise with weakly-supervised training methods. The performance of heart rate detection in the Cohface dataset increased a lot after adopting the denoise network. The performance of respiration rate detection did not increase obviously since the compression did not corrupt the motion signals too much like the rPPG signals. The proposed channel attention method further improved the performance of both heart and respiration rate detection. Like the signals shown in figure 4.3 and figure 4.4, choosing a proper ROI first can make the denoise network work better. The denoise network and the channel attention network make the vital sign detection system more robust in different recording situations. The performance of the proposed work is better than the current state-of-the-art works in the Cohface dataset. Furthermore, compared to the end-to-end deep learning models, the parameter numbers and the memory usage for training are extremely low without sacrificing accuracy. Such a lightweight model makes training a video-based vital sign detection model on a personal device possible.



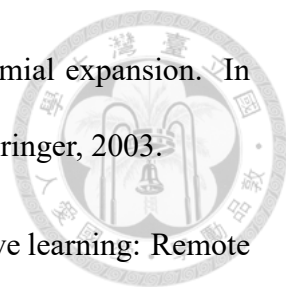
## Chapter 5 Conclusion

Although Deep learning models have improved the performance of video-based vital sign detection, collecting the ground truth signals is not easy. Most devices can only record the average heart rate or respiration rate. The proposed weakly supervised method makes training a model only with an average rate label possible. The proposed method extracts signals from video with handcraft algorithms and denoises the signals with deep learning models. The combination of traditional computer vision methods and deep learning makes the model need fewer parameters and cost less computational resources than an end-to-end model. Despite the model is lightweight, the performance of the proposed method is still better than the current end-to-end models. The proposed channel attention module can choose a proper region in video where a usable vital sign signal can be extracted and improve the performance of the denoise model.

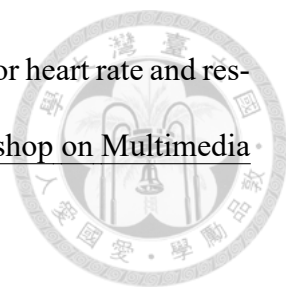


## References

- [1] J. Allen. Photoplethysmography and its application in clinical physiological measurement. Physiological measurement, 28(3):R1, 2007.
- [2] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In 2017 IEEE International Conference on Computer Vision (ICCV), pages 1021–1030, 2017.
- [3] A. V. Challoner and C. A. Ramsay. A photoelectric plethysmograph for the measurement of cutaneous blood flow. Phys Med Biol, 19(3):317–328, May 1974.
- [4] M. Chen, Q. Zhu, H. Zhang, M. Wu, and Q. Wang. Respiratory rate estimation from face videos. In 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), pages 1–4, 2019.
- [5] W. Chen and D. McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the european conference on computer vision (ECCV), pages 349–365, 2018.
- [6] G. De Haan and V. Jeanne. Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering, 60(10):2878–2886, 2013.

- 
- [7] G. Farneback. Two-frame motion estimation based on polynomial expansion. In Scandinavian conference on Image analysis, pages 363–370. Springer, 2003.
- [8] J. Gideon and S. Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 3975–3984, 2021.
- [9] G. Heusch, A. Anjos, and S. Marcel. A reproducible study on remote heart rate measurement. arXiv preprint arXiv:1709.00962, 2017.
- [10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.
- [11] X. Li, W. Wang, X. Hu, and J. Yang. Selective kernel networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 510–519, 2019.
- [12] J. Moreno, J. Ramos-Castro, J. Movellan, E. Parrado, G. Rodas, and L. Capdevila. Facial video-based photoplethysmography to detect hrv at rest. International journal of sports medicine, 36(06):474–480, 2015.
- [13] E. M. Nowara, D. McDuff, and A. Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4935–4944, 2021.
- [14] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE transactions on biomedical engineering, 58(1):7–11, 2010.



- 
- [15] Y. Ren, B. Syrnyk, and N. Avadhanam. Dual attention network for heart rate and respiratory rate estimation. In 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), pages 1–6, 2021.
- [16] Q.-V. Tran, S.-F. Su, C.-C. Chuang, V.-T. Nguyen, and N.-Q. Nguyen. Real-time non-contact breath detection from video using adaboost and lucas-kanade algorithm. In 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS), pages 1–4, 2017.
- [17] D. M. Tveit, K. Engan, I. Austvoll, and □. Meinich-Bache. Motion based detection of respiration rate in infants using video. In 2016 IEEE International Conference on Image Processing (ICIP), pages 1225–1229, 2016.
- [18] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. Opt. Express, 16(26):21434–21445, Dec 2008.
- [19] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, volume 1, pages I–I. Ieee, 2001.
- [20] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic principles of remote ppg. IEEE Transactions on Biomedical Engineering, 64(7):1479–1491, 2017.
- [21] Z. Yu, X. Li, and G. Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In K. Sidorov and Y. Hicks, editors, Proceedings of the British Machine Vision Conference (BMVC), pages 29.1–29.12. BMVA Press, September 2019.