Self-supervised ECG Representation Learning for Emotion Recognition

Paper by Pritam Sarkar and Ali Etemad
Presented by Jörg Simon

About me

- PhD on using DeepLearning to detect Human Factors from BioSignals

- Prof. Eduardo Veas and Herbert Danzinger

Self-supervised ECG Representation Learning for Emotion Recognition

Paper by Pritam Sarkar and Ali Etemad

Presented by Jörg Simon

Self-supervised ECG Representation Learning for Emotion Recognition

Paper by Pritam Sarkar and Ali Etemad
Presented by Jörg Simon

Queens University
Ambient Intelligence & Interactive Machines Laboratory (Aiim Lab)

Repositories

- https://code.engineering.queensu.ca/17ps21/SSL-ECG
  - Tensorflow 1.14
  - Problem: They assume a folder with preprocessed ECG, and did not provide that. You would still have to code that by yourself.
- https://github.com/grazai/SSL-ECG-Paper-Reimplementaton
  - I could not parse SWELL, so I replaced that with DEAP

- 2 Versions of the Paper: Conference and longer Journal Version. I **use the Journal Version** (also linked in the GitHub)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3014842, IEEE Transactions on Affective Computing

Self-supervised ECG Representation Learning for Emotion Recognition

Pritam Sarkar, Ali Etemad

Abstract—We exploit a self-supervised deep multi-task learning framework for electrocardiogram (ECG) -based emotion recognition. The proposed solution consists of two stages of learning *a)* learning ECG representations and *b)* learning to classify emotions. ECG representations are learned by a signal transformation recognition network. The network learns high-level abstract representations from unlabeled ECG data. Six different signal transformations are applied to the ECG signals, and transformation recognition is performed as pretext tasks. Training the model on pretext tasks helps the network learn spatiotemporal representations that generalize well across different datasets and different emotion categories. We transfer the weights of the self-supervised network to an emotion recognition network, where the convolutional layers are kept frozen and the dense layers are trained with labelled ECG data. We show that the proposed solution considerably improves the performance compared to a network trained using fully-supervised learning. New state-of-the-art results are set in classification of arousal, valence, affective states, and stress for the four utilized datasets. Extensive experiments are performed, providing interesting insights into the impact of using a multi-task self-supervised structure instead of a single-task model, as well as the optimum level of difficulty required for the pretext self-supervised tasks.

Index Terms—Self-supervised Learning, ECG, Emotion Recognition, Multi-task Learning.

1 Introduction

Affective computing is a field of study that deals with understanding human emotions, intelligent human-machine interaction, and computer-assisted learning among others [1], [2]. The goal of affective computing is to equip machines with the ability to model and interpret the emotional states of humans. Emotion is considered a *physiological* and *psychological* expres-

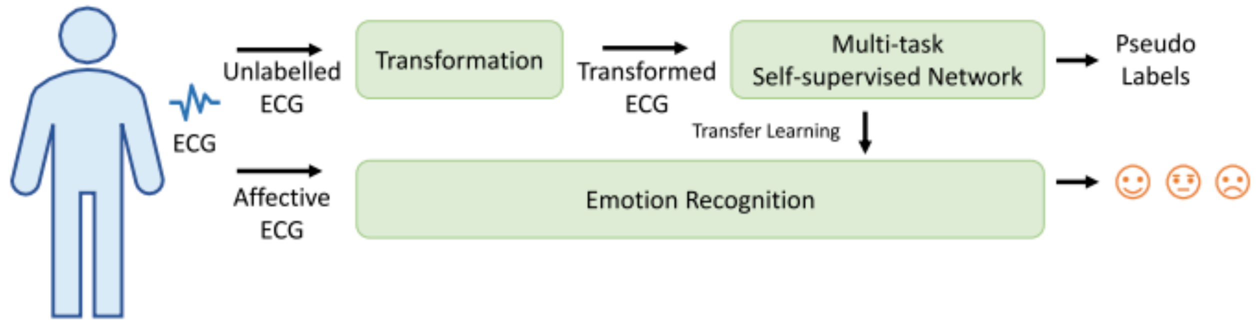


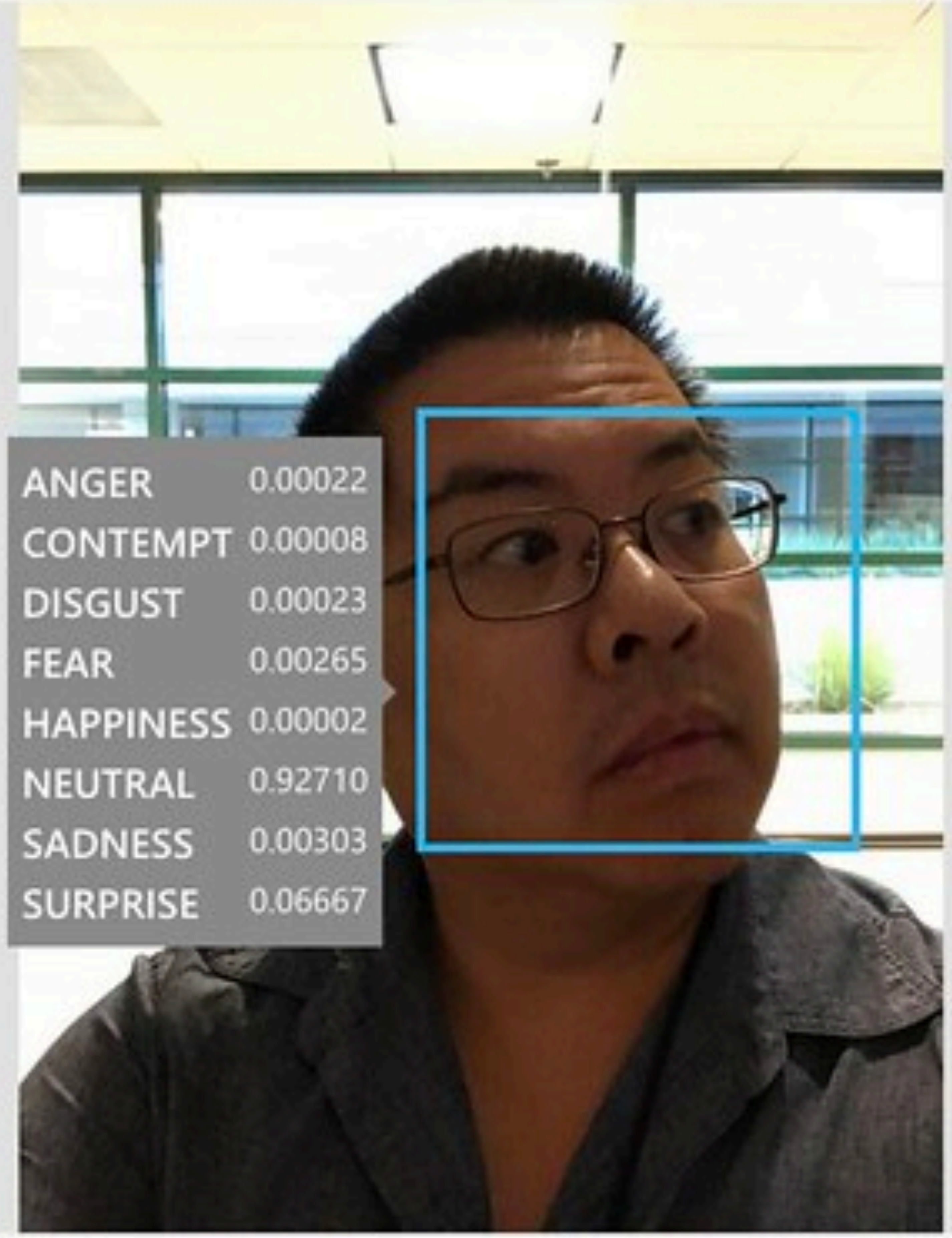
Fig. 1: An overview of the proposed framework for self-supervised emotion recognition is presented.

Self-supervised ECG Representation Learning for Emotion Recognition

Emotion Recognition

- part of Affective Computing (1970s onward)

Emotion Recognition

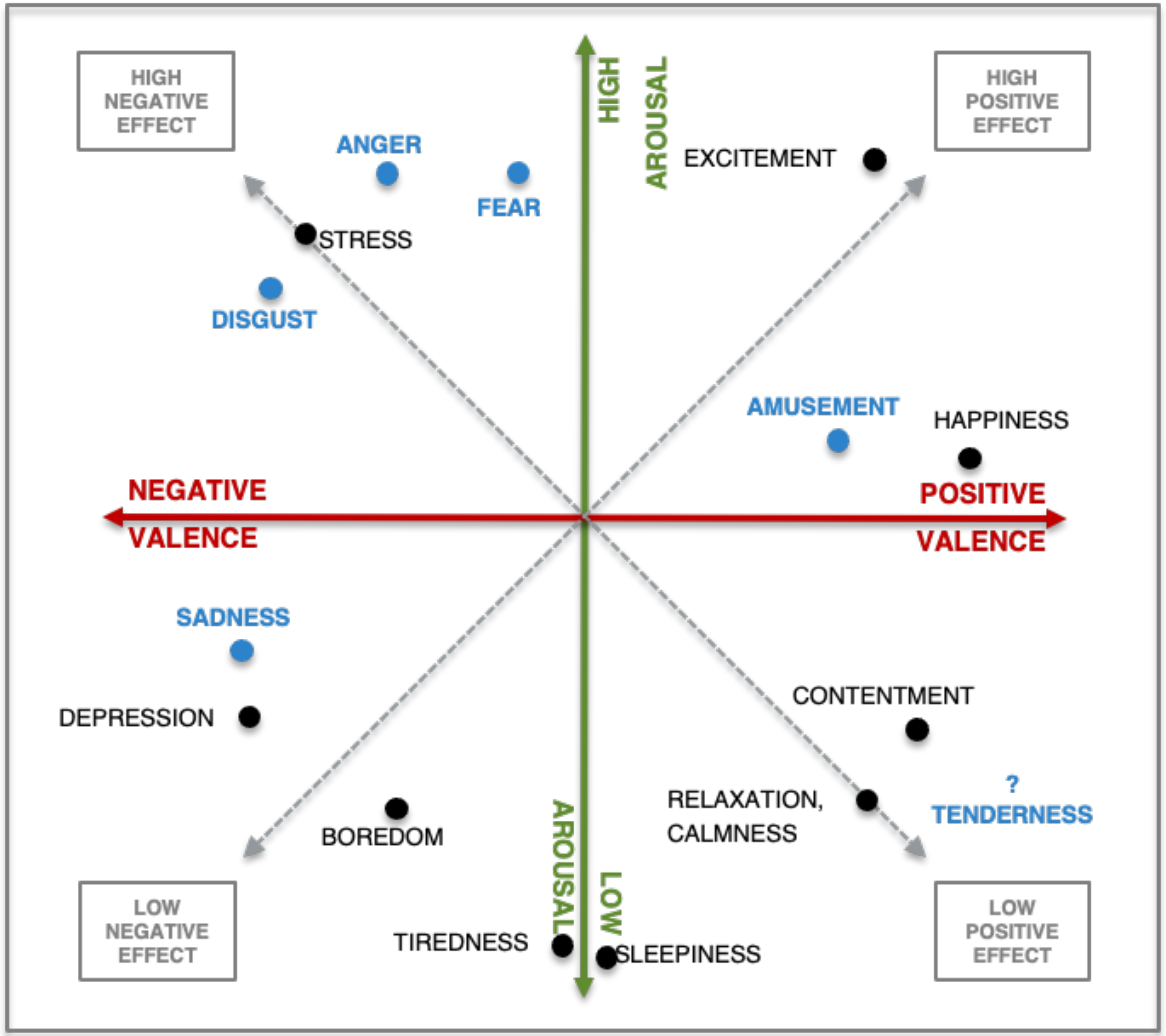


Detection result:
1 faces detected

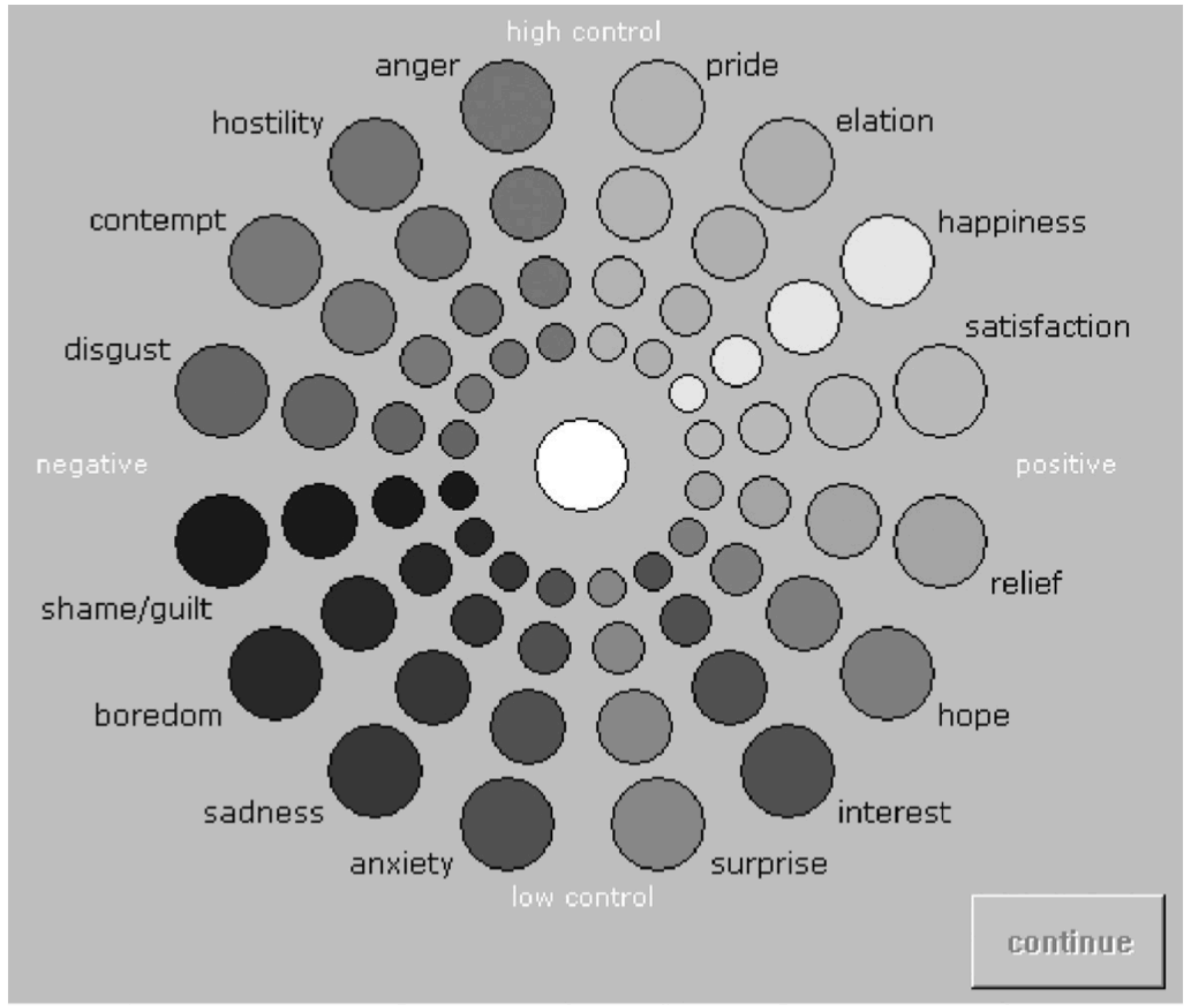
JSON:

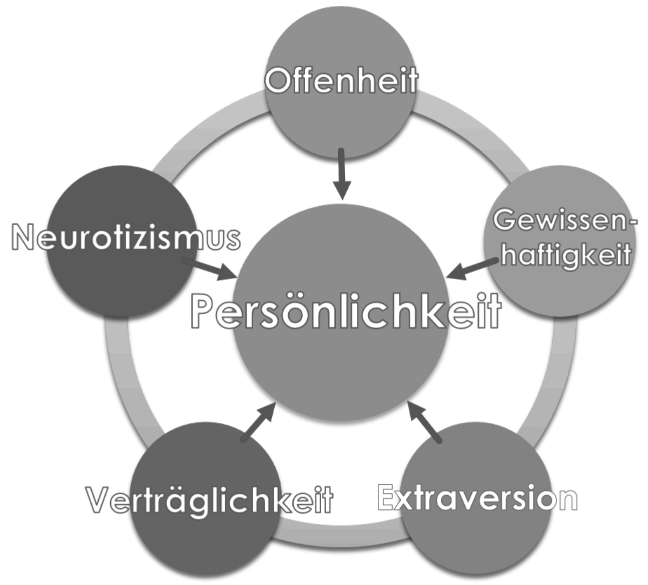
```
[
  {
    "faceRectangle": {
      "top": 211,
      "left": 208,
      "width": 232,
      "height": 232
    },
    "scores": {
      "anger": 0.000222139541,
      "contempt": 8.289404E-05,
      "disgust": 0.000231529353,
      "fear": 0.00265476666,
      "happiness": 1.599525E-05,
      "neutral": 0.9270953,
      "sadness": 0.00302594039,
      "surprise": 0.06667146
    }
  }
```

Representations of Emotions



Affect categories	Pertinent words or word stems
Admiration/Awe	admir*, ador*, awe*, dazed, dazzl*, enrapt*, enthrall*, fascina*, marveli*, rapt*, reveren*, spellbound, wonder*, worship*
Amusement	amus*, fun*, humor*, laugh*, play*, rollick*, smil*
Anger	anger, angr*, cross*, enrag*, furious, fury, incens*, infuriat*, irate, ire*, mad*, rag*, resent*, temper , wrath*, wrought*
Anxiety	anguish*, anxi*, apprehens*, diffiden*, jitter*, nervous*, trepida*, wari*, wary, worried*, worry*
Being touched	affect*, mov*, touch*
Boredom	bor*, ennui, indifferen*, languor*, tedi*, wear*
Compassion	commiser*, compass*, empath*, pit*
Contempt	contempt*, denigr*, deprec*, deris*, despi*, disdain*, scorn*
Contentment	comfortabl*, content*, satisf*
Desperation	deject*, desolat*, despair*, desperat*, despond*, disconsolat*, hopeless*, inconsol*
Disappointment	comedown, disappoint*, discontent*, disenchant*, disgruntl*, disillusion*, frustrat*, jilt*, letdown, resign*, sour*, thwart*





Environmental influence

Demographic

Relating to the characteristics of human populations

e.g. age

Personal

Relating to individuals' behavioral patterns

e.g. Big-five personality traits

Affect

e.g. arousal, valence, negative emotion

Sociocultural

Relating to social interaction and cultural aspects

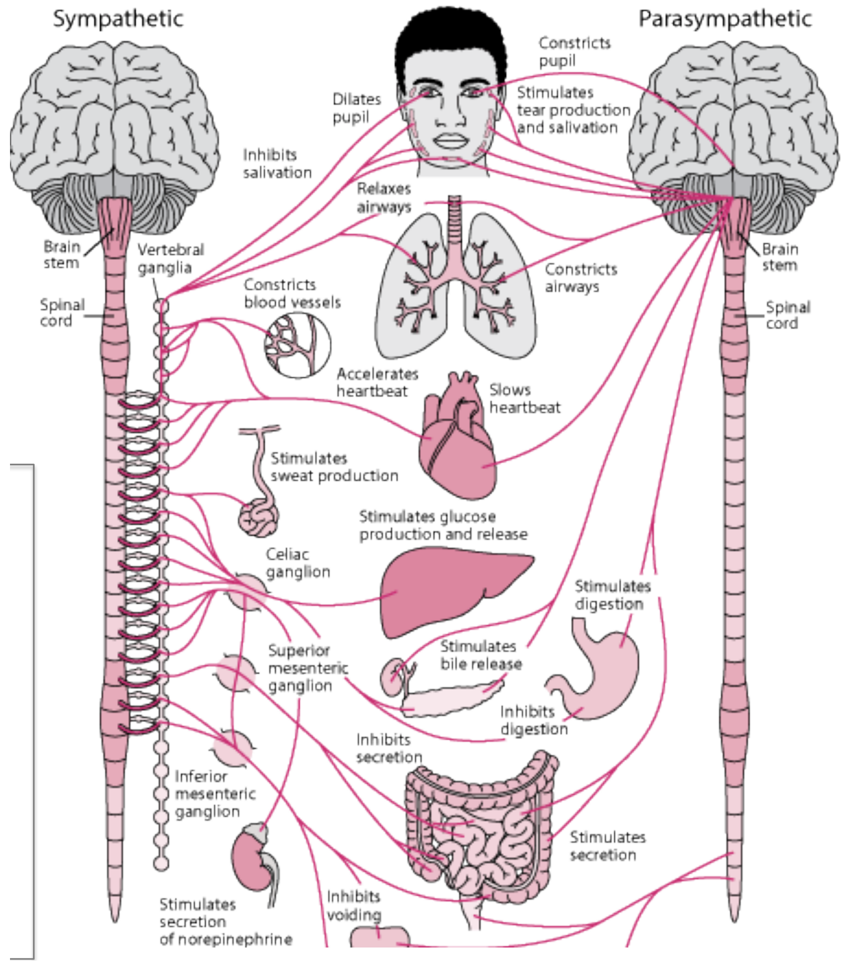
e.g. conflict, interest, sincerity

Psychophysiological

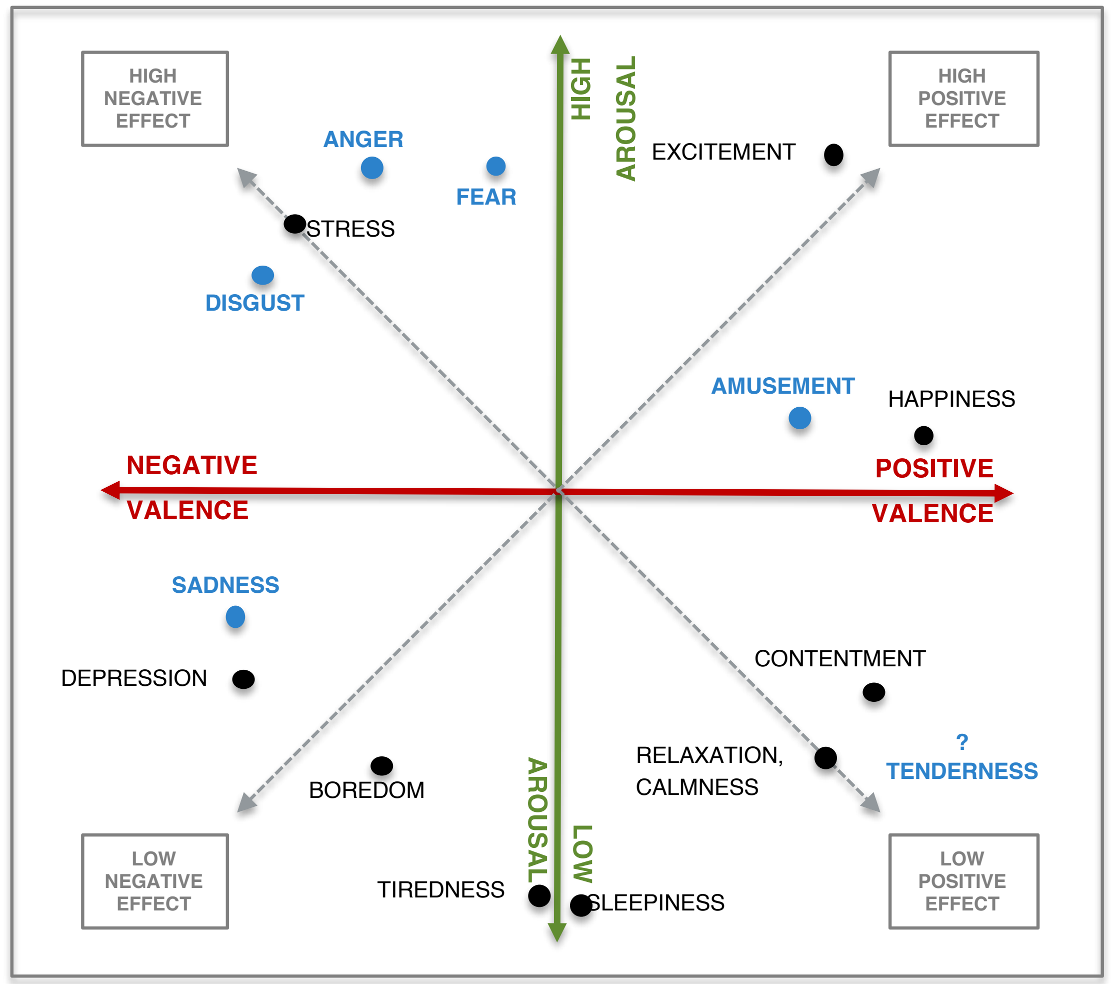
Relating to mental and physical conditions

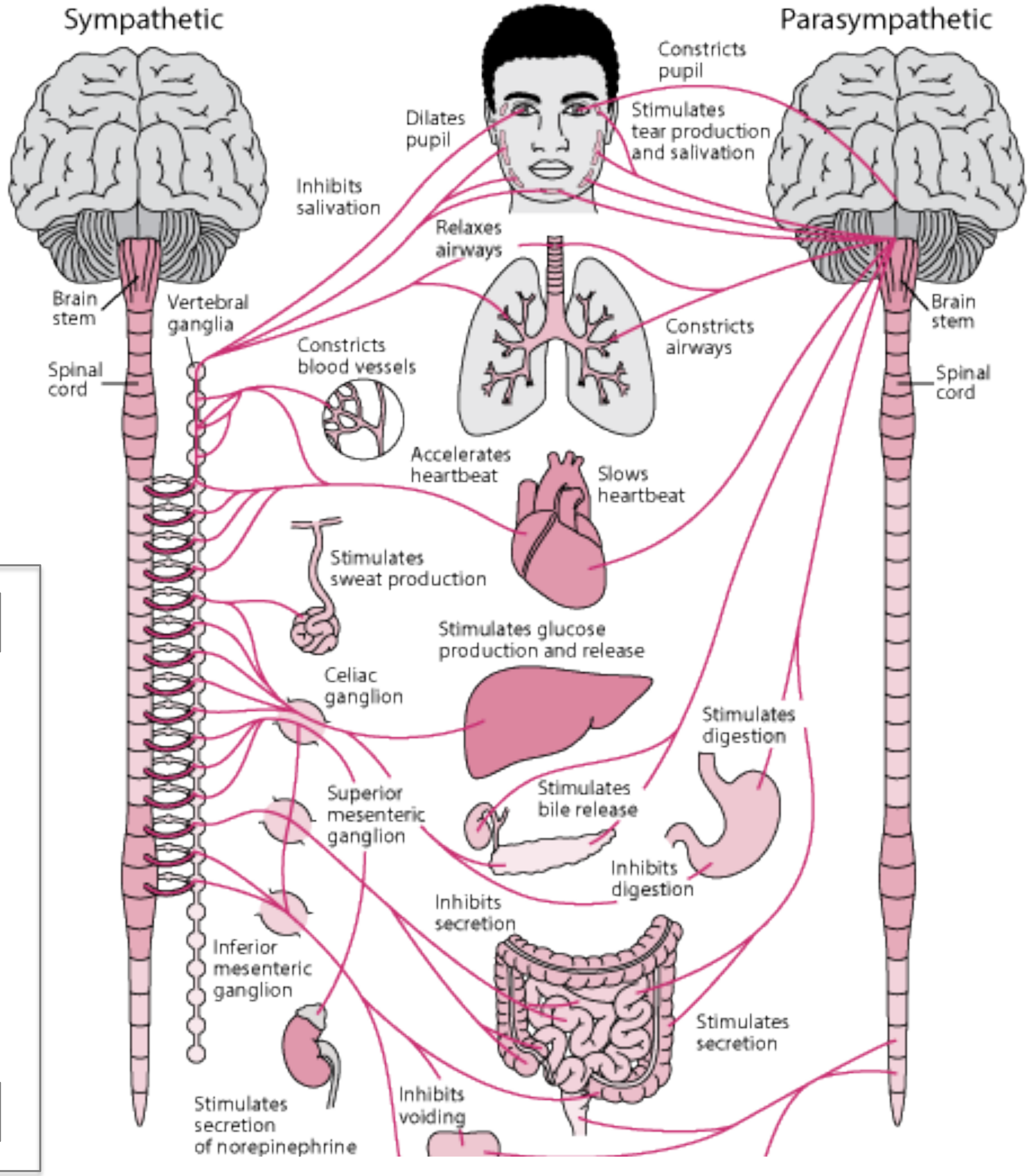
e.g. illness, physical/ cognitive load stress, intoxication, sleepiness

externalize



Physiological Aspect





The Autonomous Nervous System (**ANS**) triggers specific **organic reactions** to Emotions

Physiological Aspect

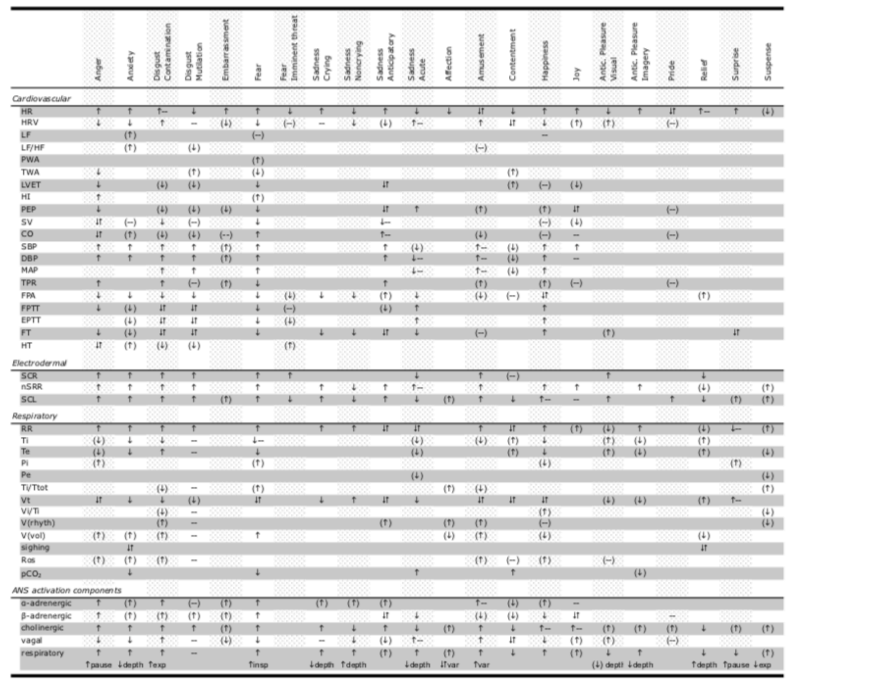
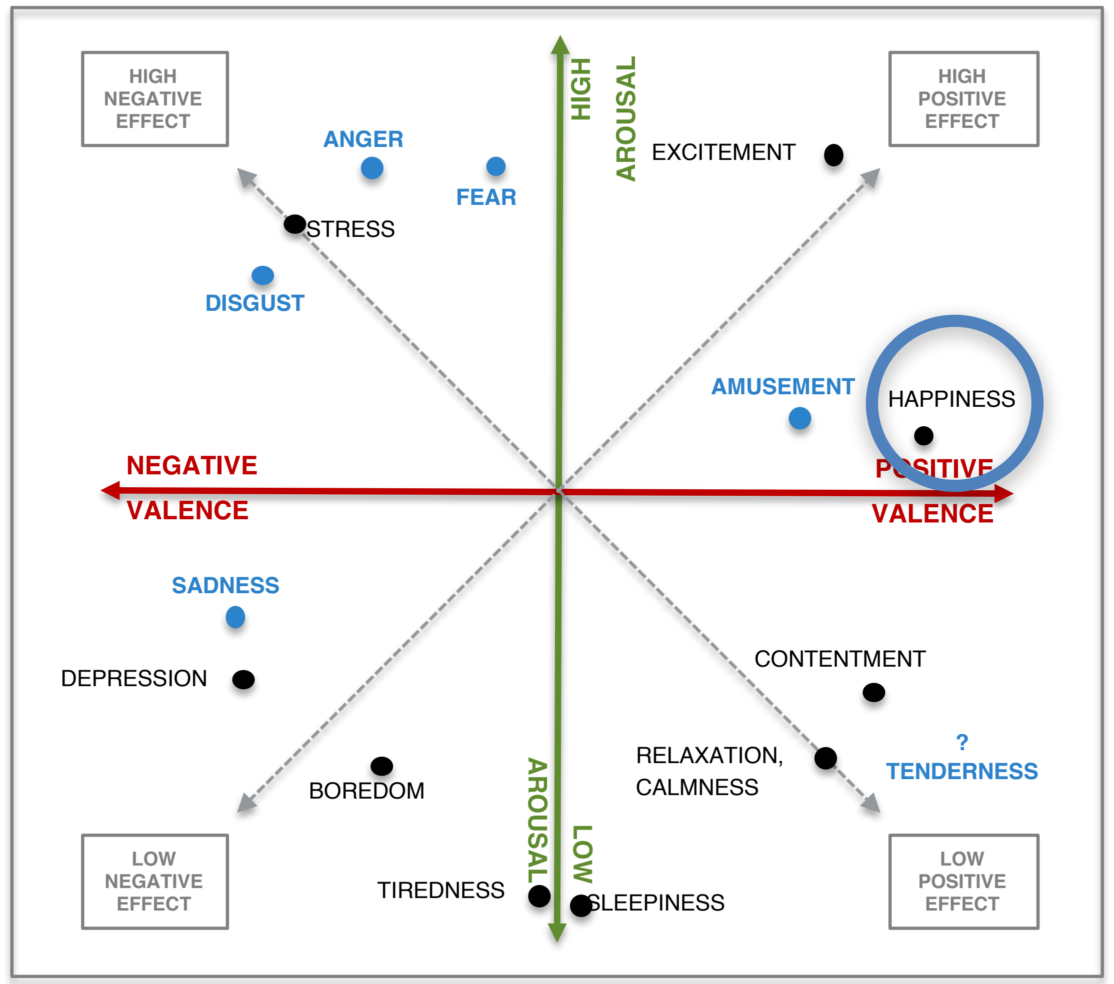
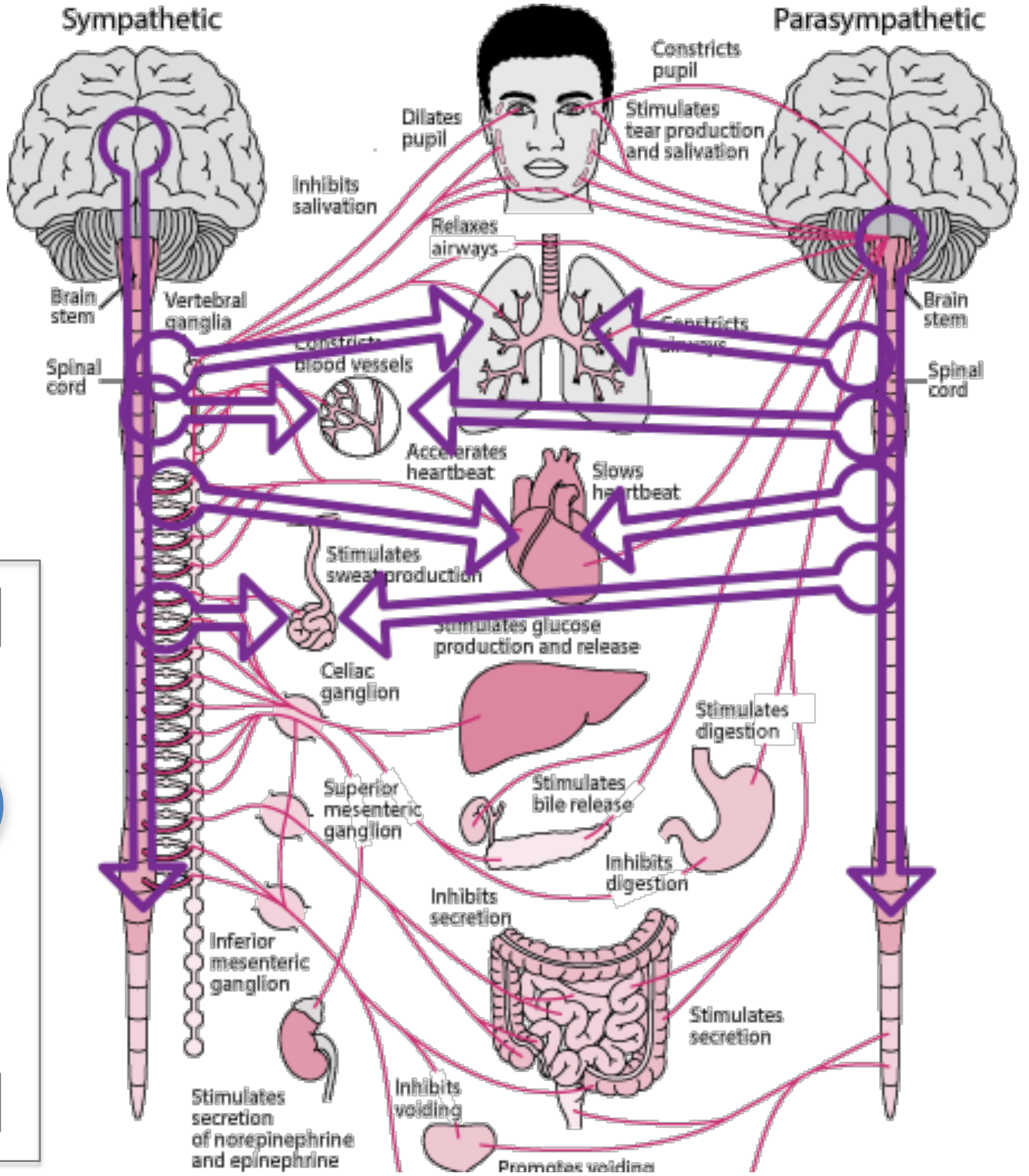


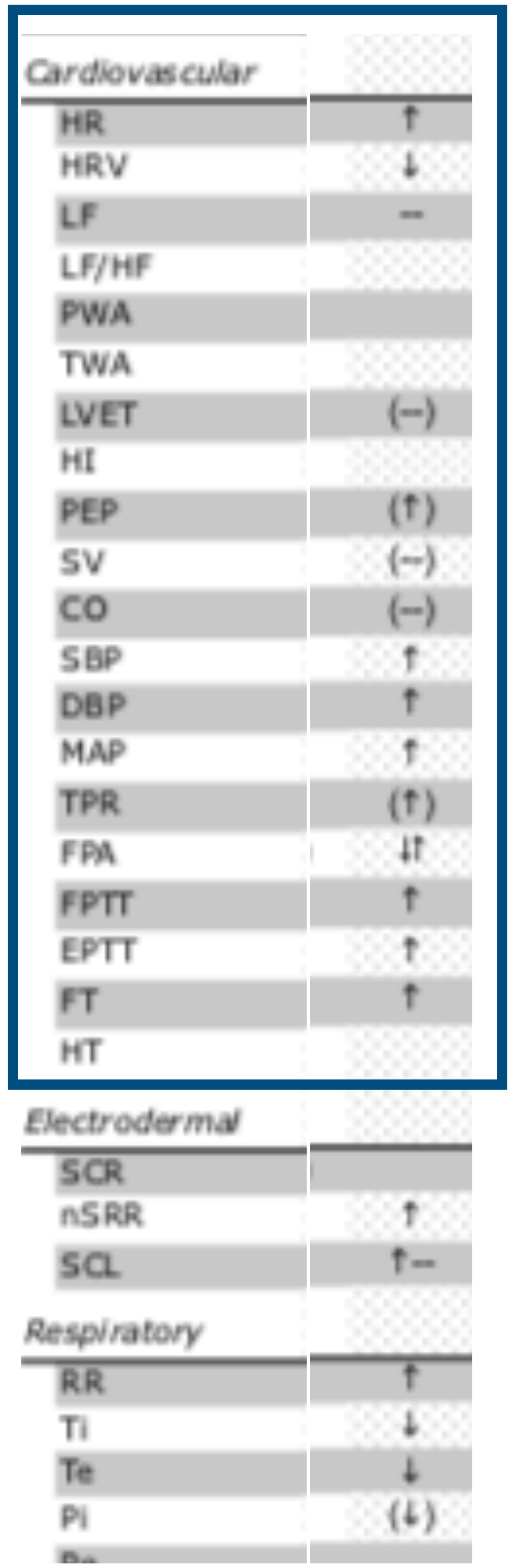
Table 2 Overview of Modal* ANS Responses Found for Reviewed Emotions.

Note. *Modal responses were defined as the response direction reported by the majority of studies (unweighted), with at least three studies indicating the same response direction. Arrows indicate increased (↑), decreased (↓), or no change in activation from baseline (–), or both increases and decreases between studies (↓↑). Arrows in parentheses indicate tentative response direction, based on fewer than three studies. *Abbreviations*: pause – respiratory pause time; depth – respiratory depth; exp – respiratory expiration time; insp – respiratory inspiration time; var – respiratory variability. For abbreviations of other physiological measures, see Table 3.

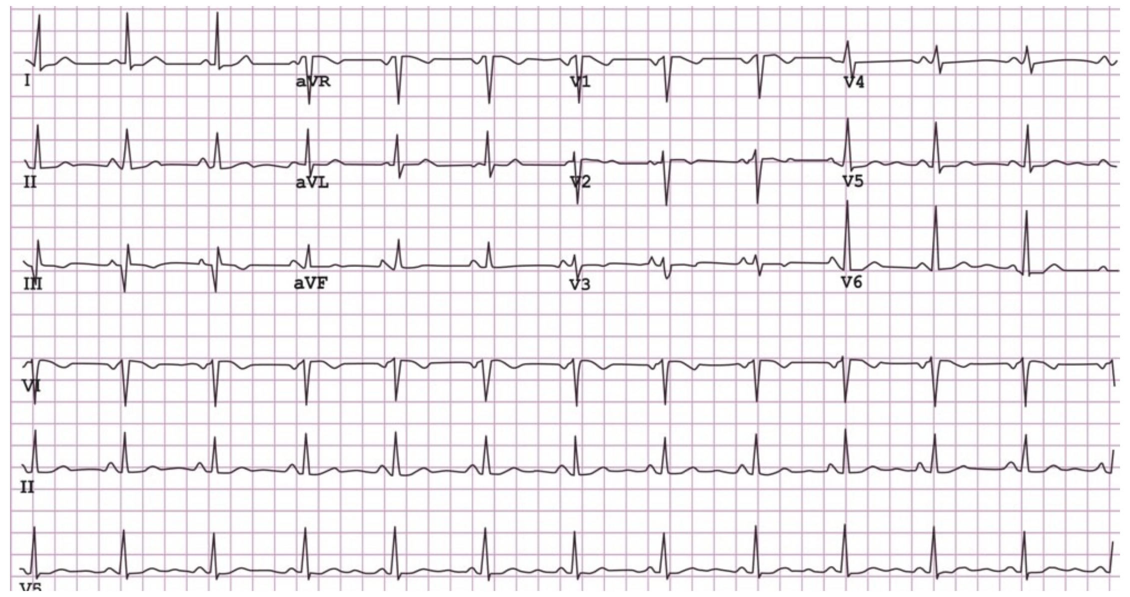




Happy

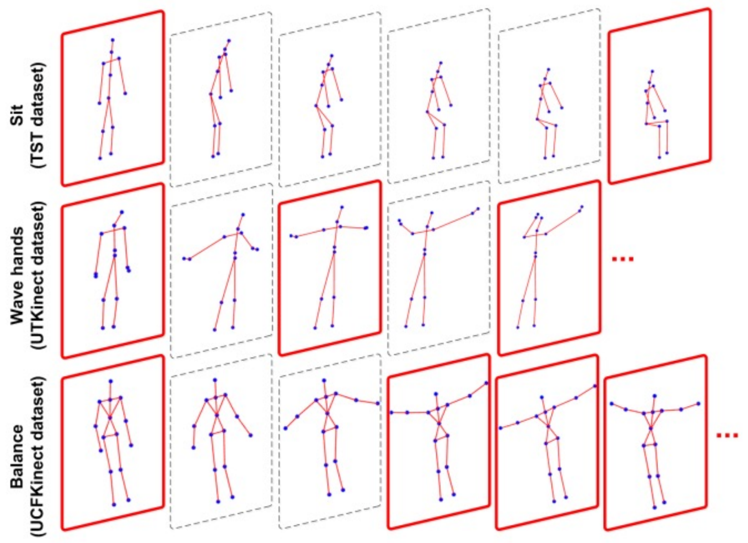


	Happiness
Cardiovascular	
HR	↑
HRV	↓
LF	–
LF/HF	
PWA	
TWA	
LVET	(–)
HI	
PEP	(↑)
SV	(–)
CO	(–)
SBP	↑
DBP	↑
MAP	↑
TPR	(↑)
FPA	↓↑
FPTT	↑
EPTT	↑
FT	↑
HT	
Electrodermal	
SCR	
nSRR	↑
SCL	↑–
Respiratory	
RR	↑
Ti	↓
Te	↓
Pi	(↓)

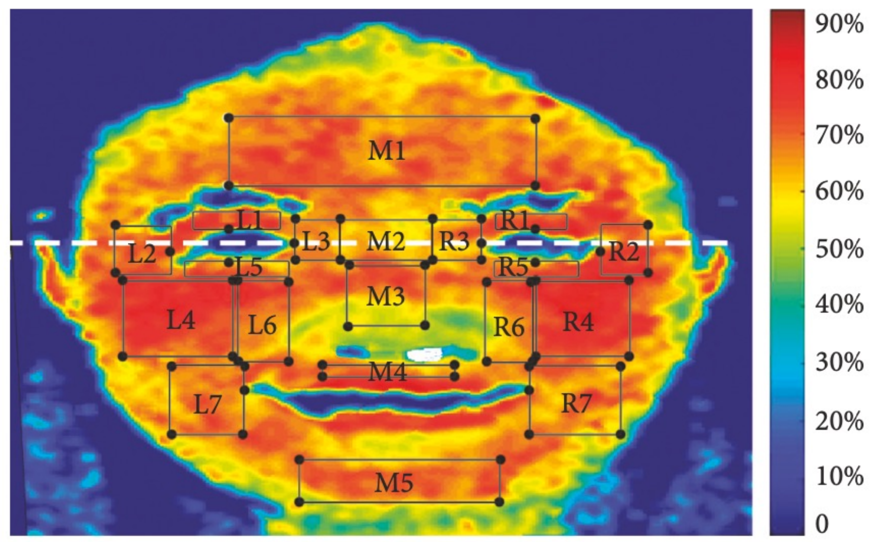


ECG

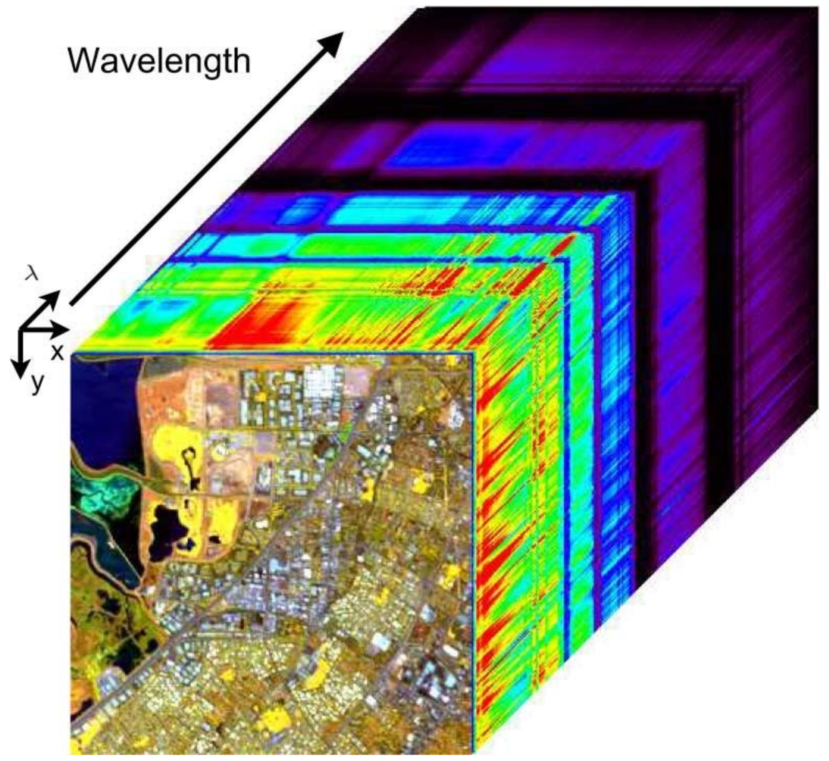
EEG

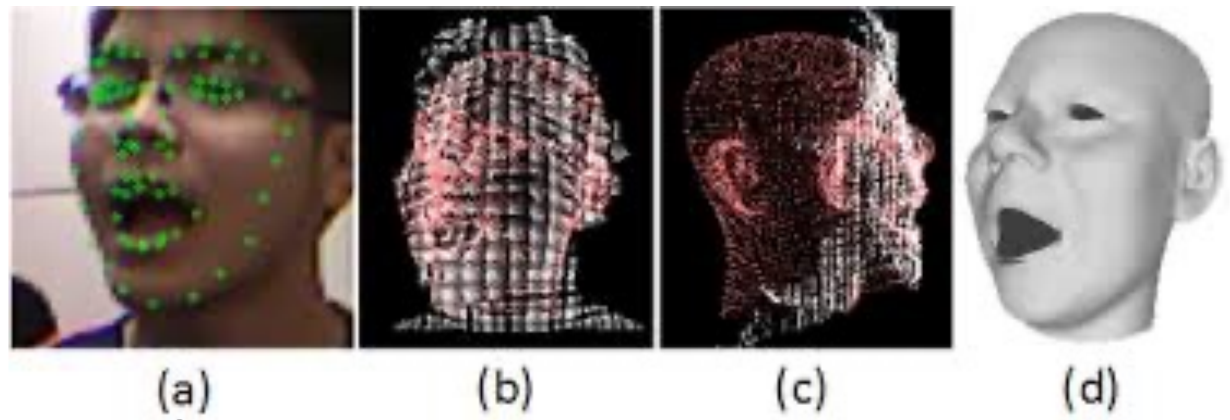


Posture

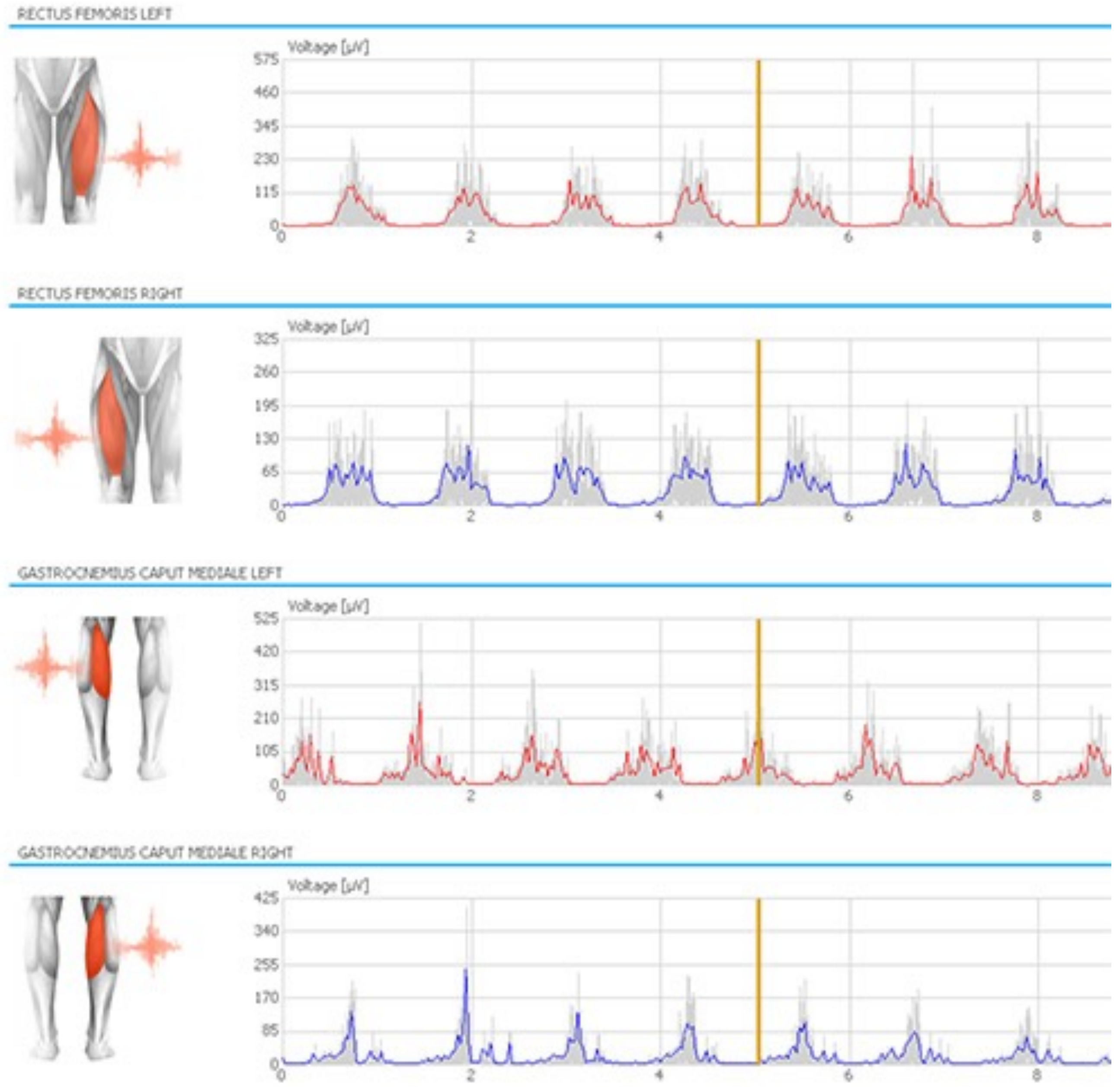


Spectral Imaging

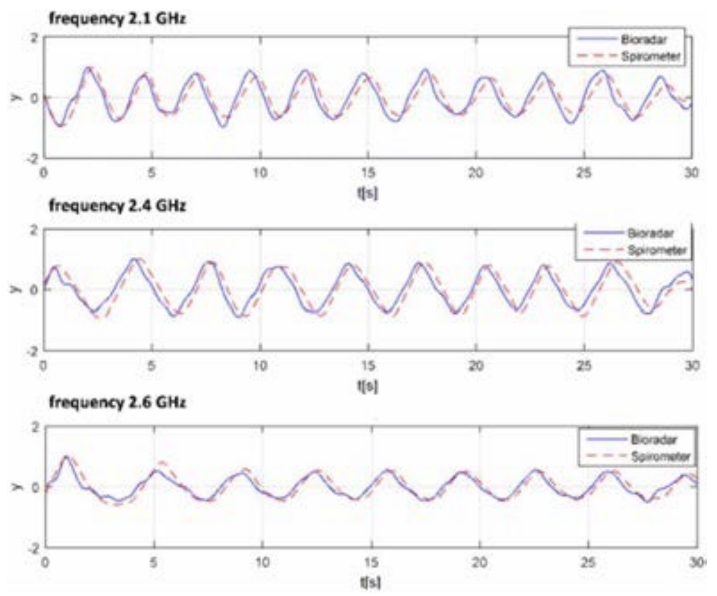




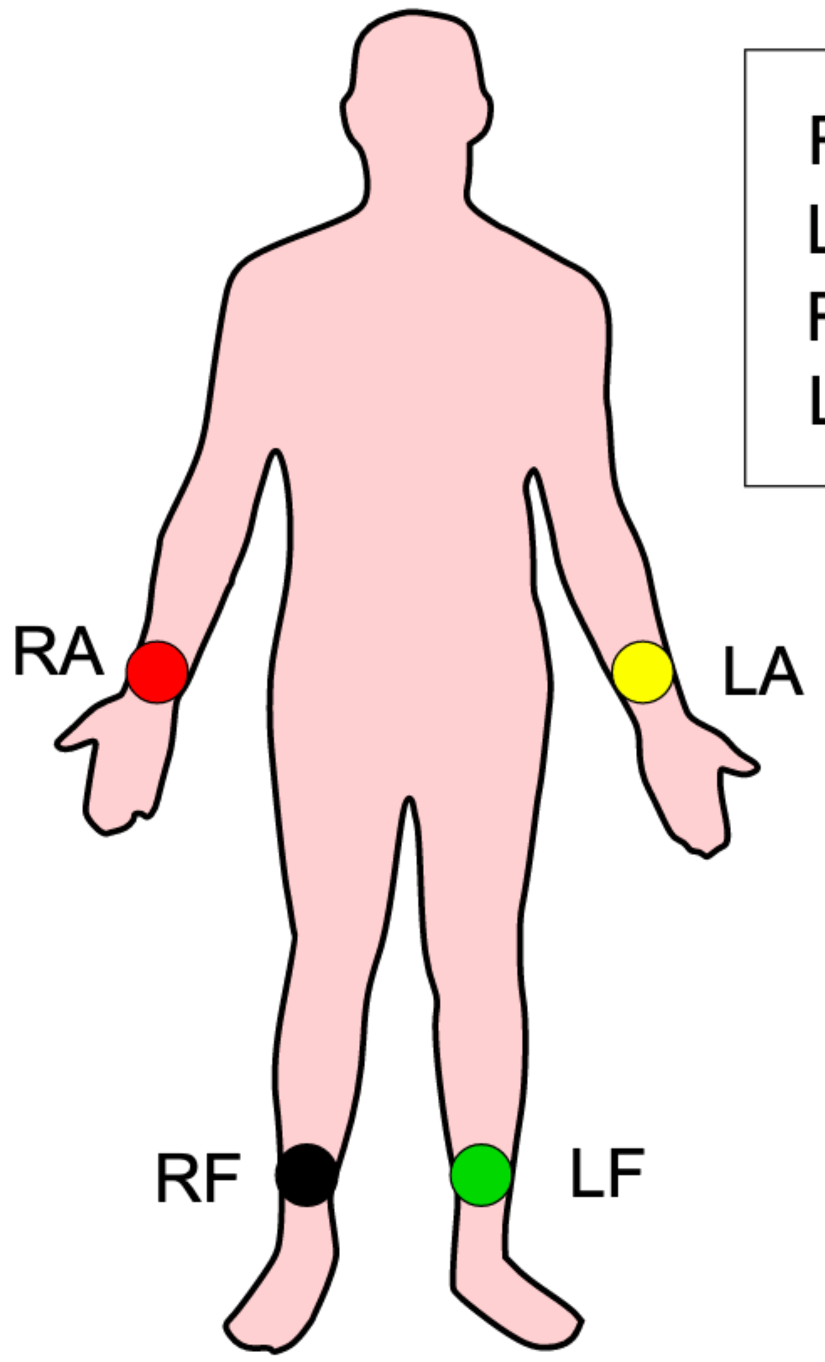
Depth Images



EMG

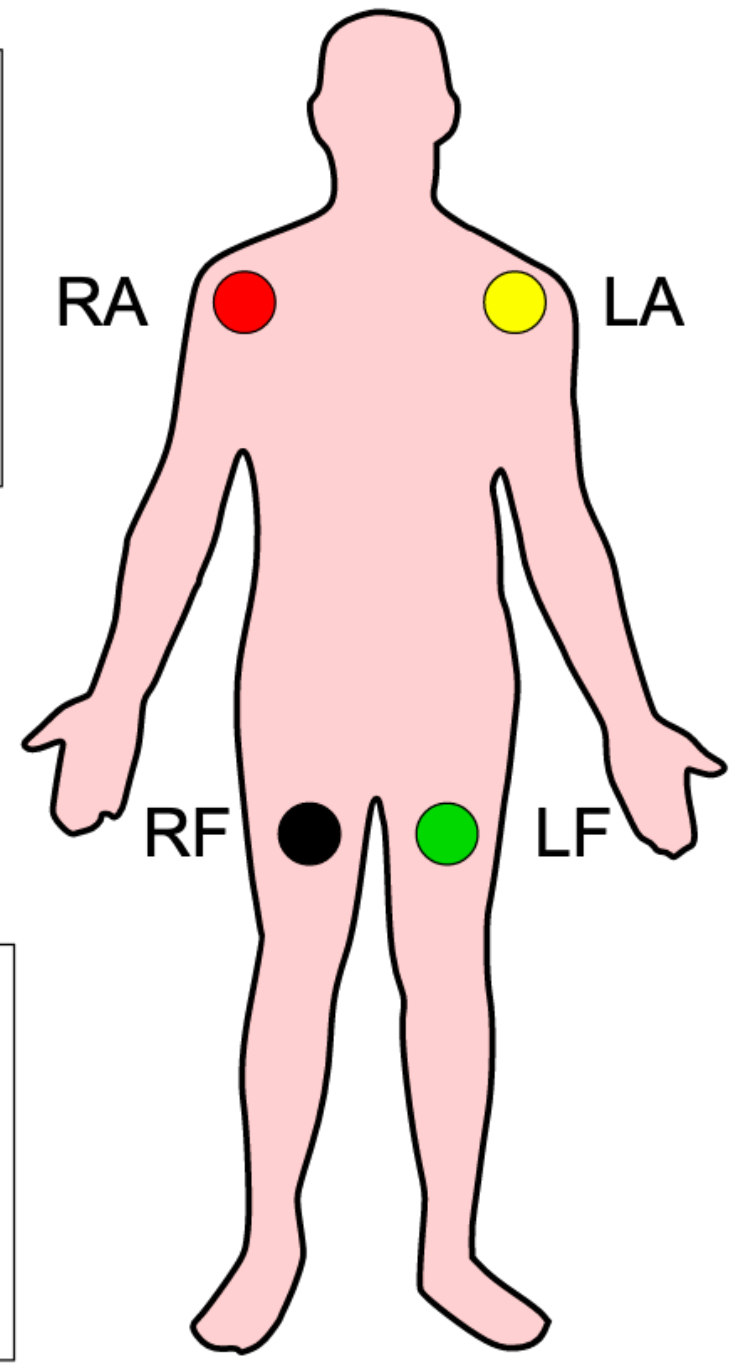


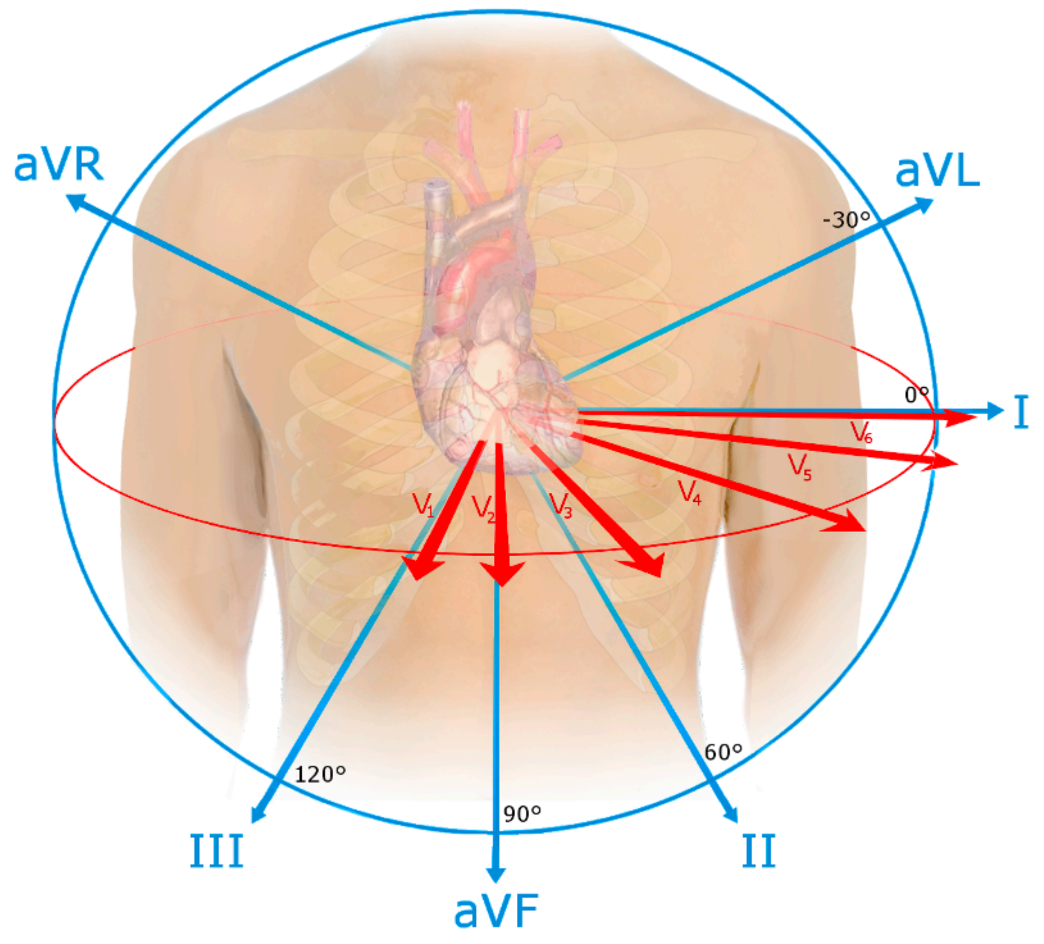
Radar



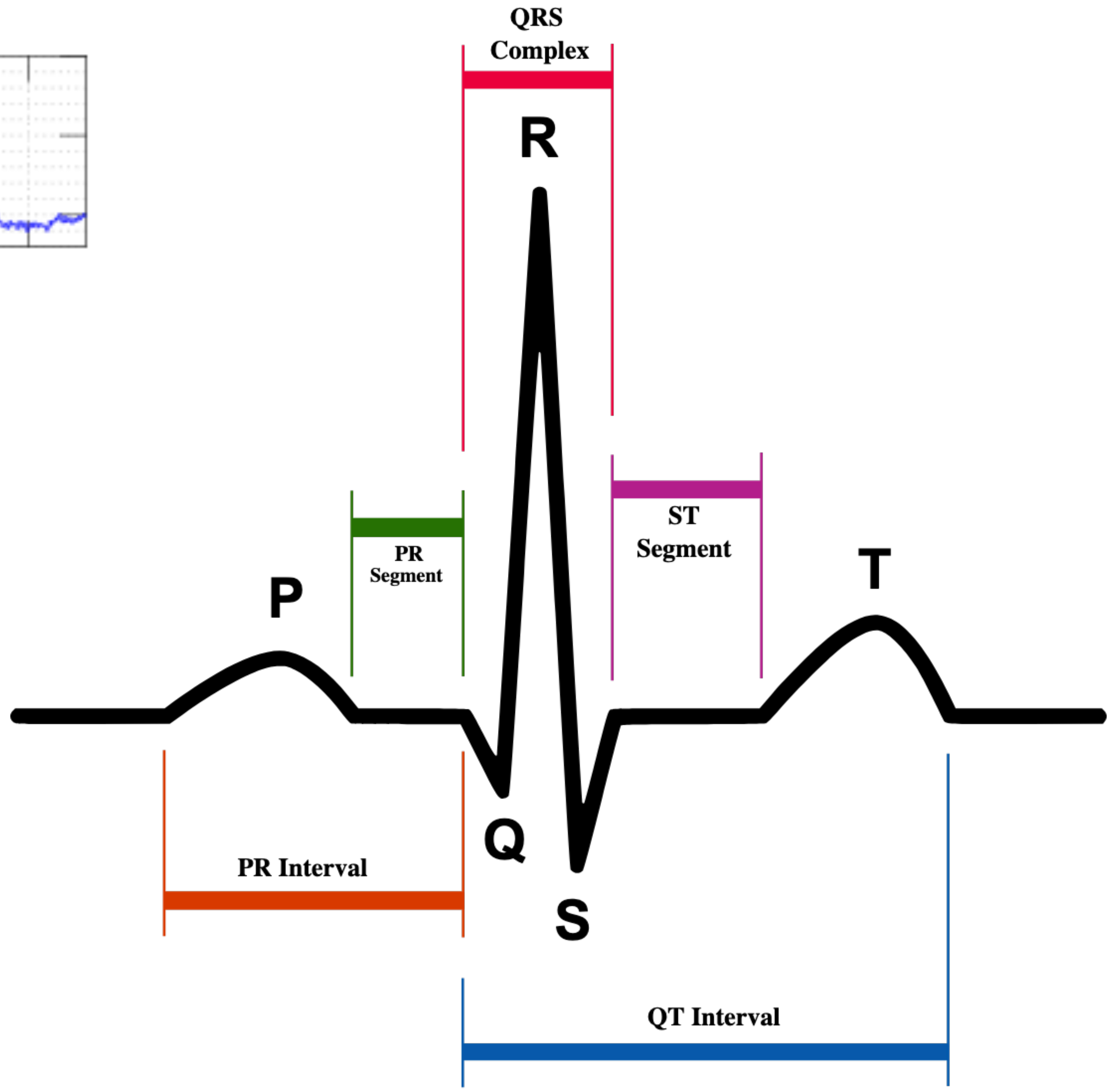
RA = Right Arm
LA = Left Arm
RF = Right Foot
LF = Left Foot

RA - Vermell
LA - Groc
RF - Negre
LF - Verd



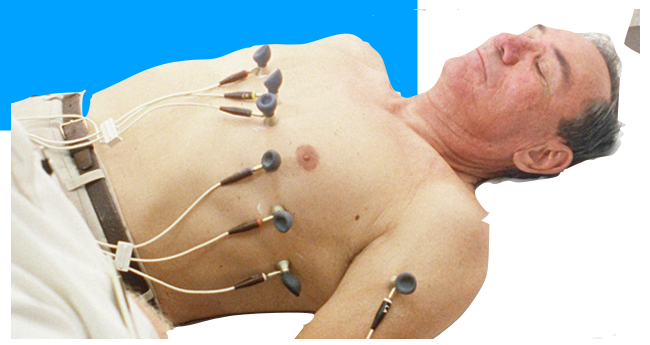


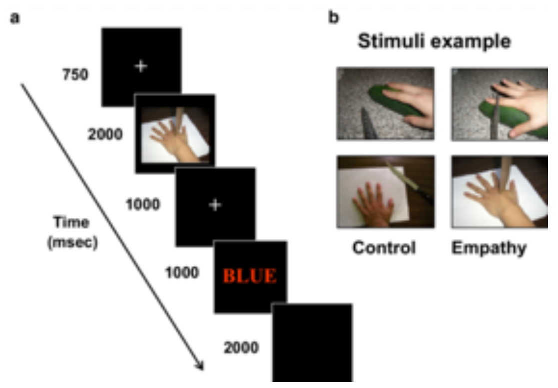

10
5
0

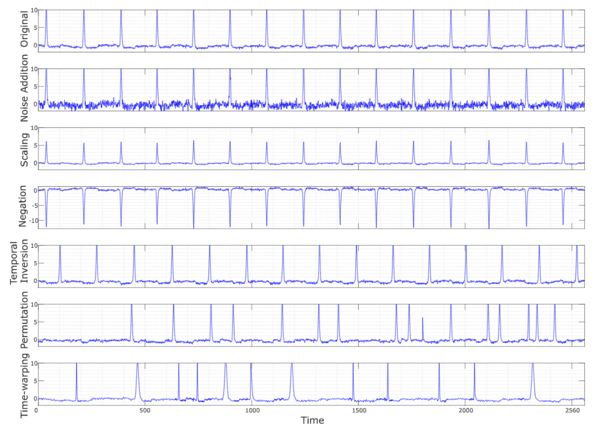

QRS
Complex
R
PR
Segment
ST
Segment
P
T
Q
S
PR Interval
QT Interval

Process

Instrumented Humans → Controlled Experiment → Experiment Data → Model → Verification

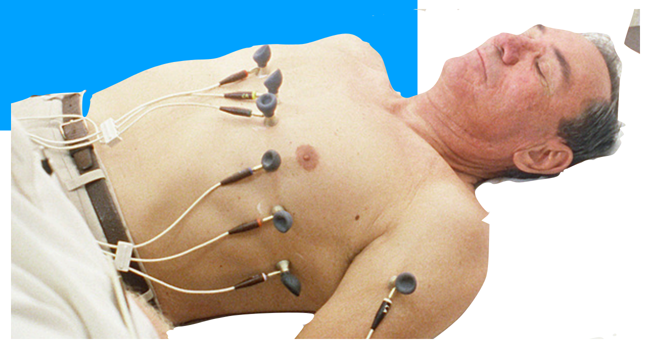


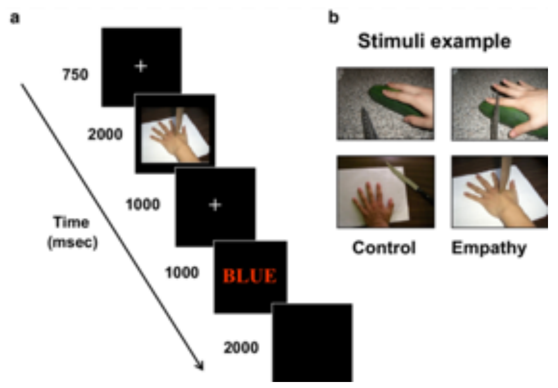


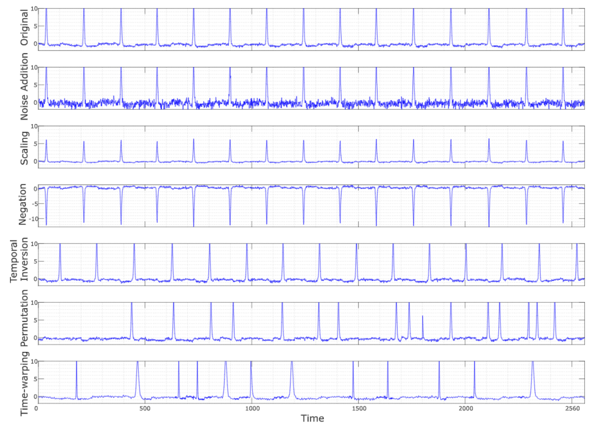


Process

Instrumented Humans → Controlled Experiment → Experiment Data → Model → Verification







Problem: Deep Learning needs a lot of Data

Problem: non Deep Learning Models do not perform good (a bit above chance)

Process

Instrumented Humans → Controlled Experiment → Experiment Data → Model → Verification

Combine Datasets for Deep Learning

Problem: Deep Learning needs a lot of Data

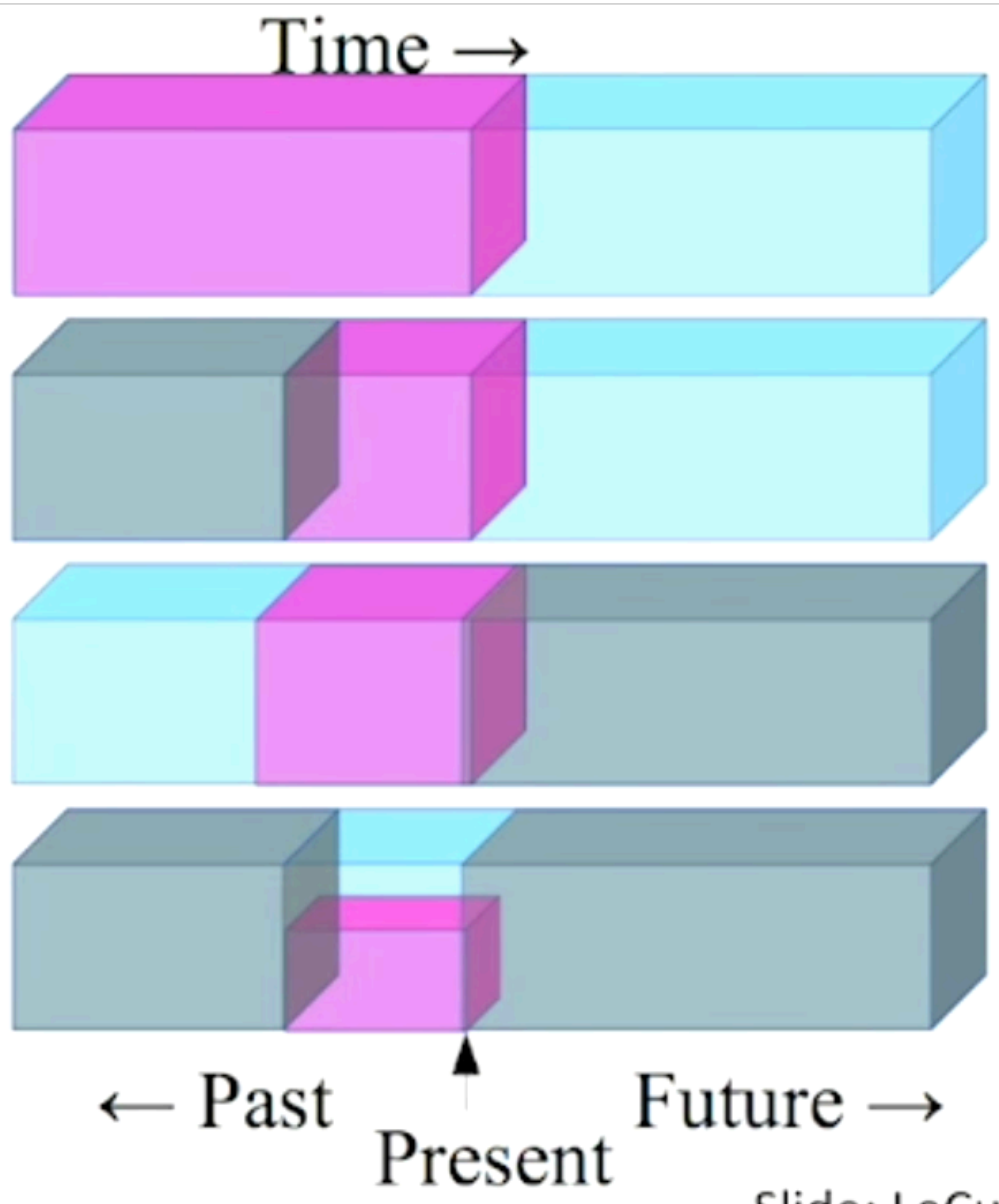
Problem: non Deep Learning Models do not perform good (a bit above chance)

Self-Supervised Deep Learning

- Somehow use **data itself**, without any explicit labels given, to **provide labels** to a supervised learning process. It uses the *supervised learning paradigm in an unsupervised manner*.

Self-Supervised Deep Learning

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



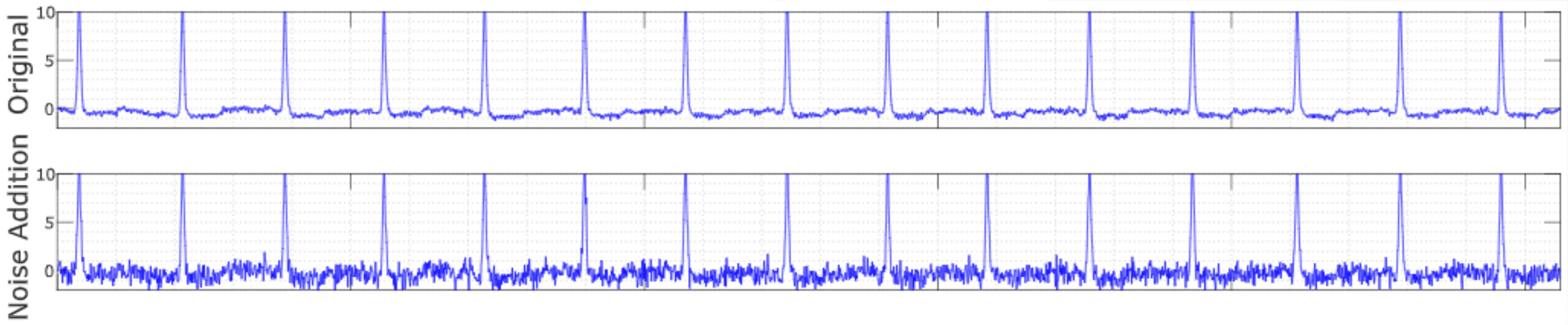
Slide: LeCun

Self-Supervised Deep Learning

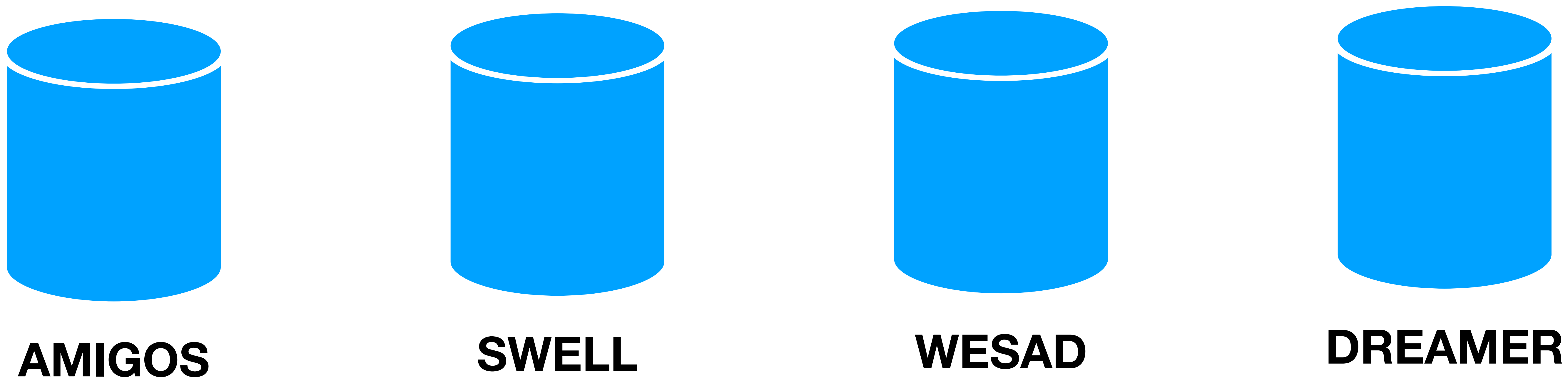
- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the p
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.

Slide: LeCun

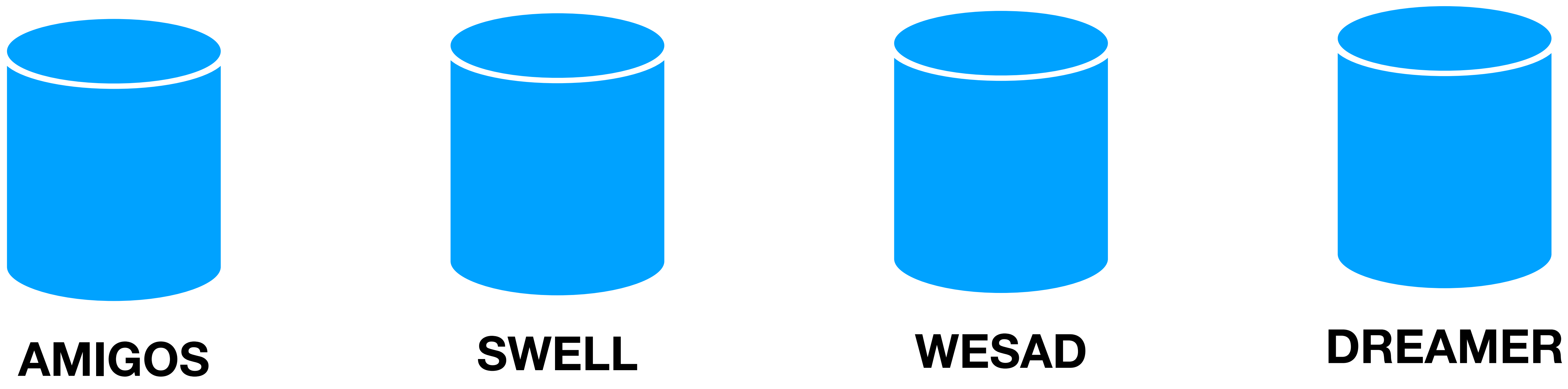
Distinguish real data from distorted data



Self-supervised ECG Representation Learning for Emotion Recognition



Self-supervised ECG Representation Learning for Emotion Recognition



4 Experiments

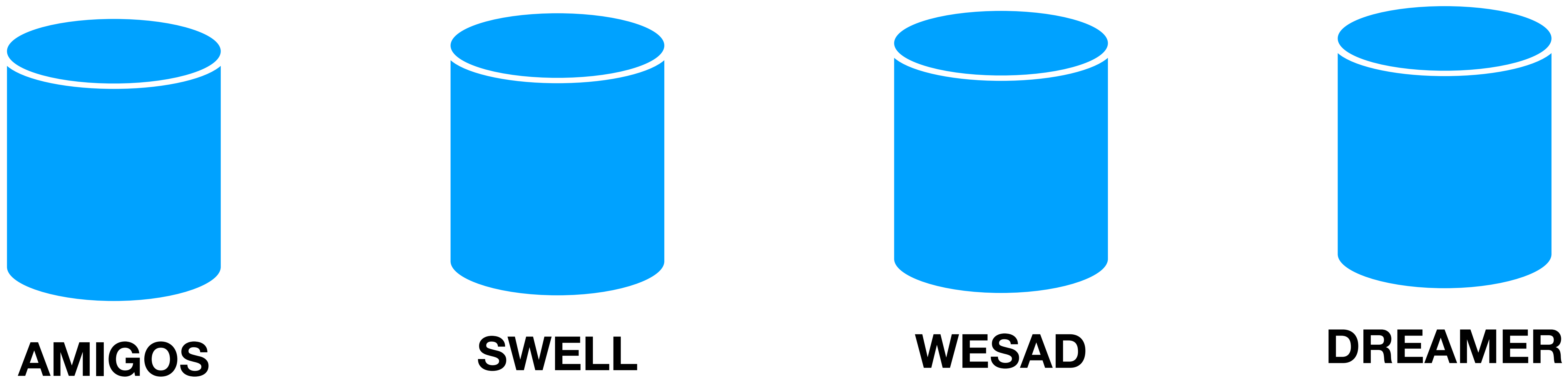
4.1 Datasets

We used four publicly available datasets to evaluate the proposed solution in depth on a large number of different subjects, in different circumstances and under different data collection protocols, and using different hardware. The important characteristics of the datasets are summarized in Table 2 and a brief description of each dataset is provided bellow.

TABLE 2: The summary of the four datasets used are presented.

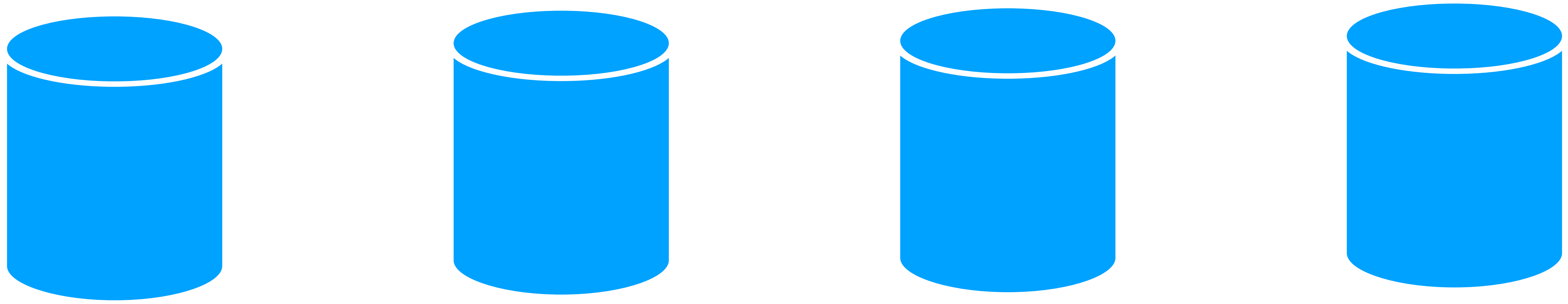
Dataset	Participants	Attributes	Classes
AMIGOS	40	Arousal	9
		Valence	9
DREAMER	23	Arousal	5
		Valence	5
WESAD	17	Affect State	4
SWELL	25	Stress	3
		Arousal	9
		Valence	9

Self-supervised ECG Representation Learning for Emotion Recognition



- 40 Participants
- Watching 16 + 4 Movie Clips
- Many Emotion Representations including Valance & Arousal
- 2x Shimmer ECG on left and right hand + one Base
- 256Hz Sampling frequency

Self-supervised ECG Representation Learning for Emotion Recognition

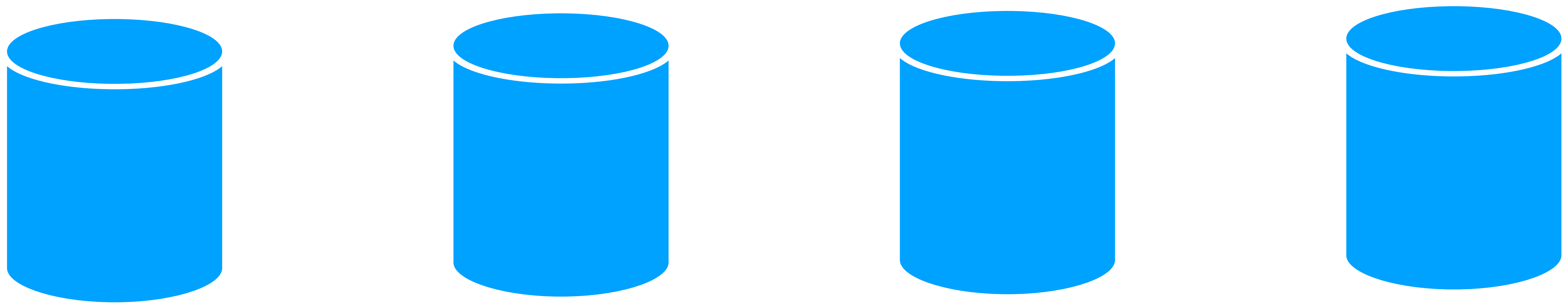


AMIGOS | **SWELL** | **WESAD** | **DREAMER**

SWELL

- 25 Participants
- Office Tasks to induce Stress
- Affect Score (1-9) and Stress
- TMSI MOBI
- 2048Hz Sampling frequency

Self-supervised ECG Representation Learning for Emotion Recognition



AMIGOS

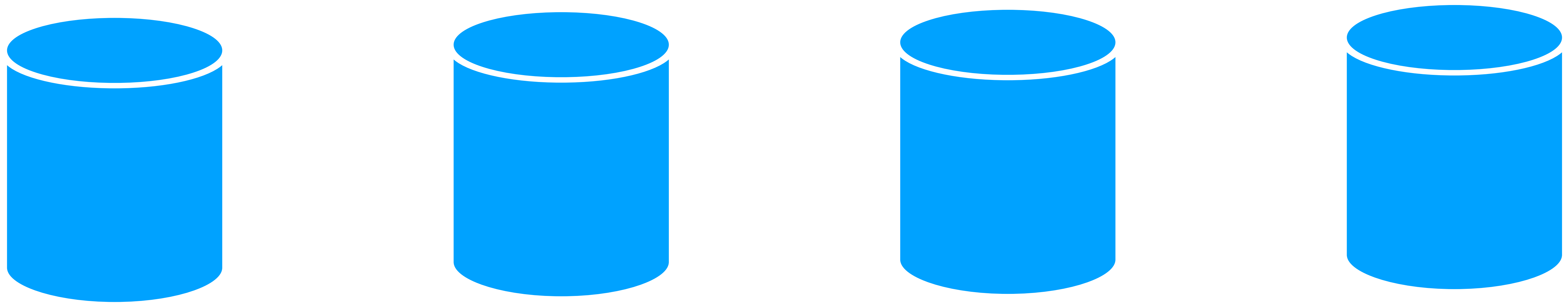
SWELL

WESAD

- 17 Participants
- Different Tasks
- Valance & Arousal
- RespiBAN
- 700Hz Sampling frequency

DREAMER

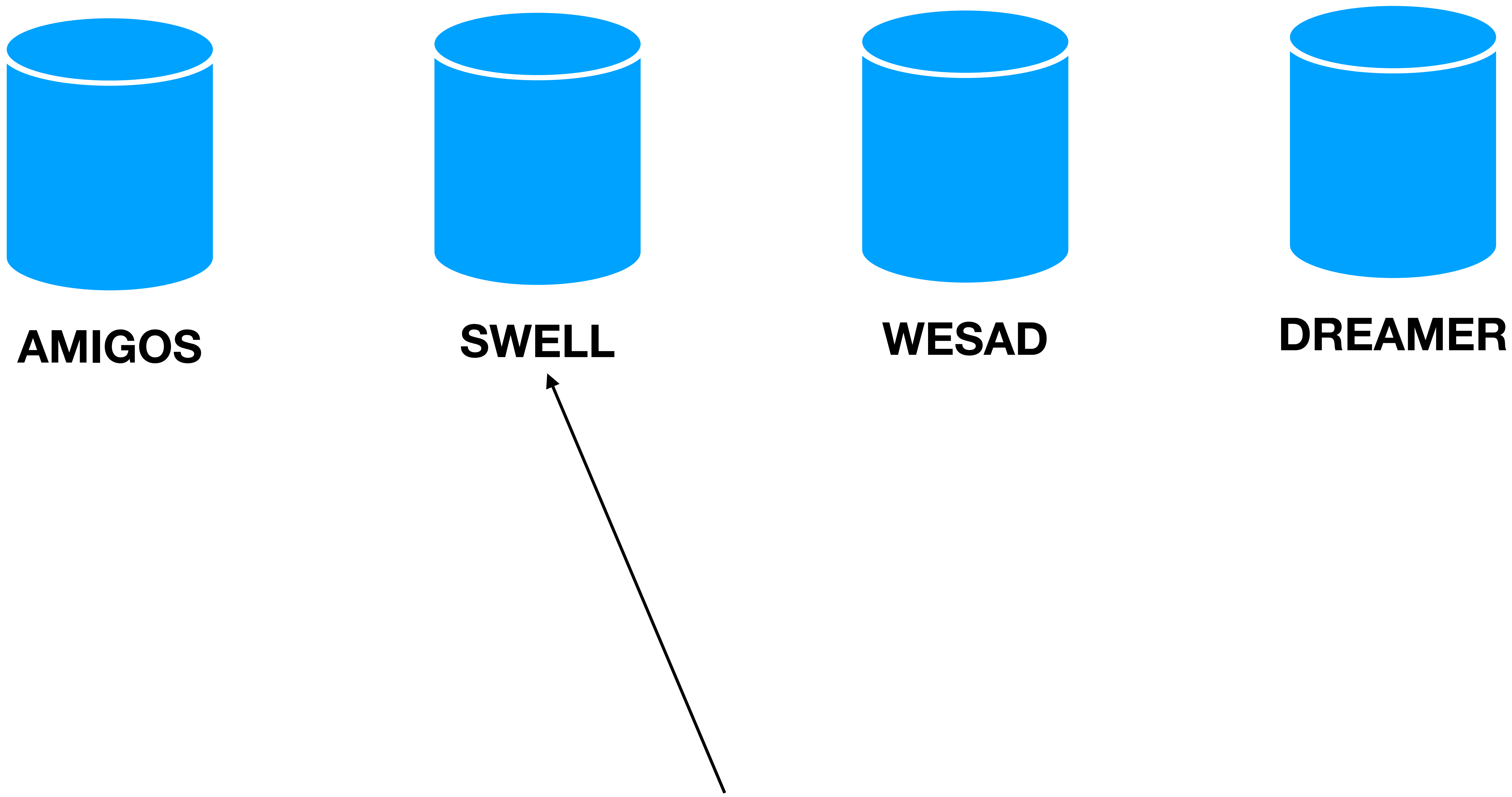
Self-supervised ECG Representation Learning for Emotion Recognition



AMIGOS SWELL WESAD DREAMER

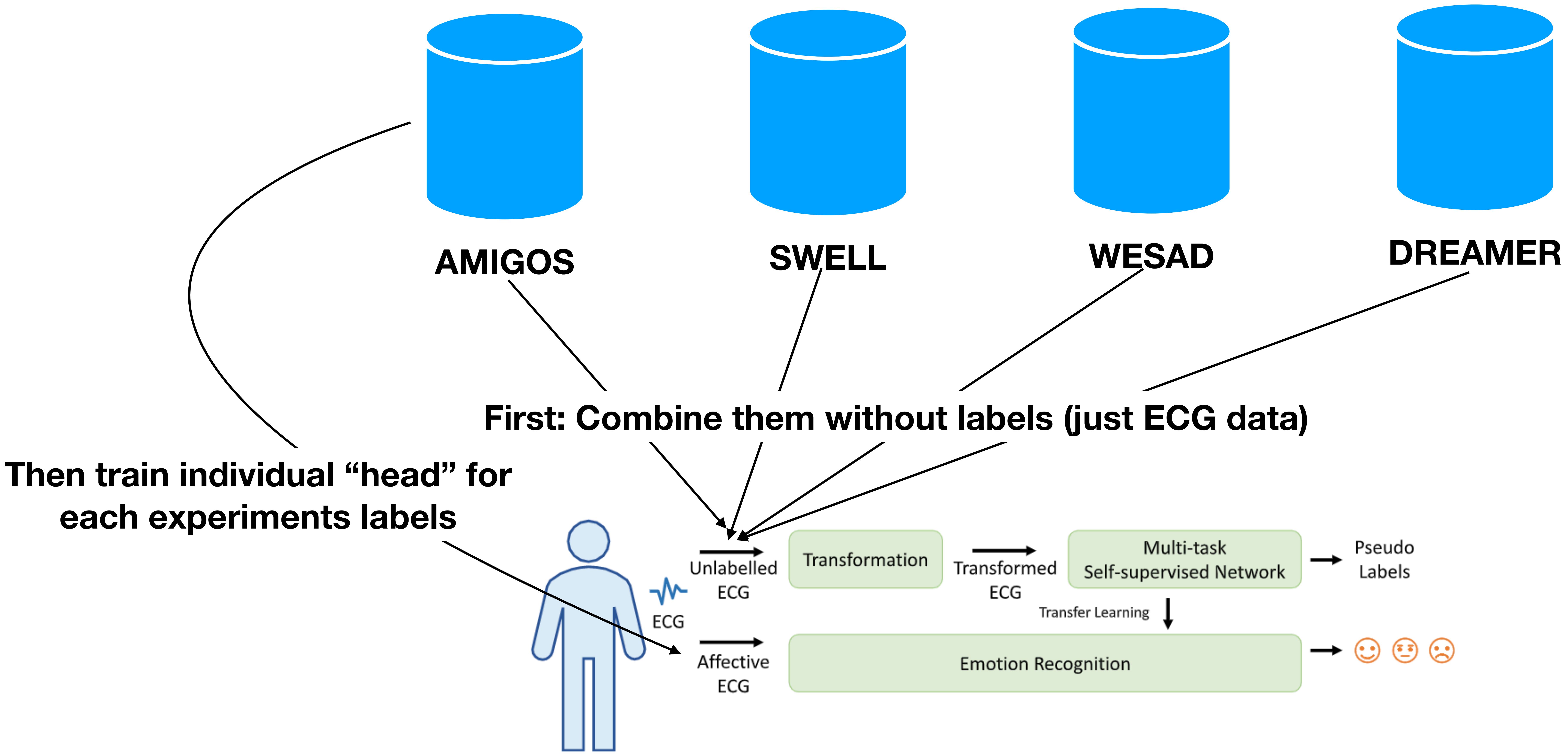
- 23 Participants
- 18 video clips
- Valance & Arousal and Basic Emotions
- 2x Shimmer + base
- 256Hz Sampling frequency

Self-supervised ECG Representation Learning for Emotion Recognition

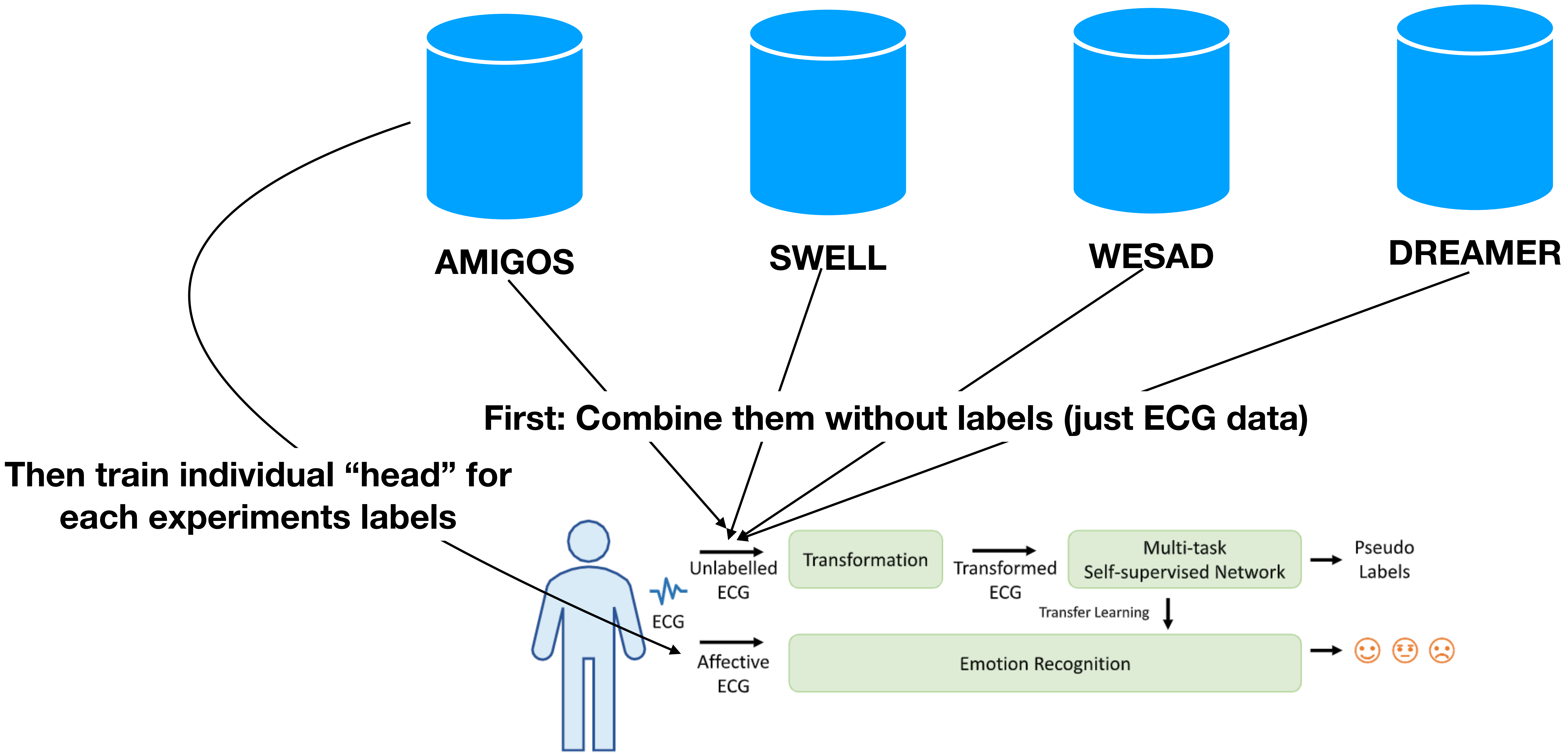


Swell only has some Matlab parser, I don’t have Matlab

Self-supervised ECG Representation Learning for Emotion Recognition



Self-supervised ECG Representation Learning for Emotion Recognition



4.2 Data Pre-processing

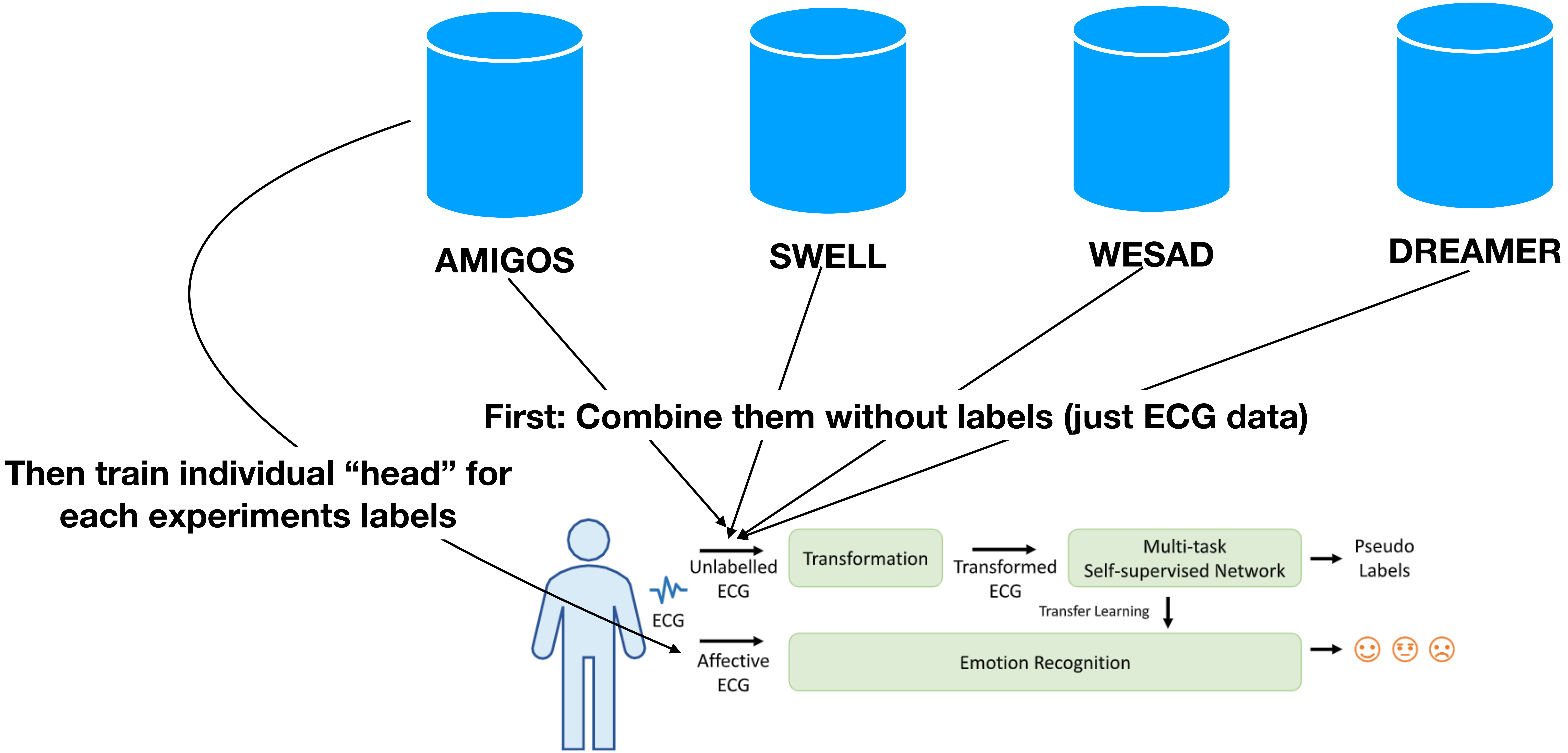
Since the above-mentioned datasets have been collected using different hardware, the signals have different spatiotemporal properties such as spatial range and sampling rate. To minimize the effects of such inter-dataset variations and discrepancies, three simple pre-processing steps were taken. First, SWELL and WESAD ECG signals are downsampled to 256 Hz to be consistent with AMIGOS and DREAMER. Next, we remove baseline wander from all the four datasets by applying a high-pass IIR filter with a pass-band frequency of 0.8 Hz. Lastly, we perform user-specific z-score normalization. While a number of other pre-processing operations such as feature extraction could have been done, we intentionally kept the pre-processing minimal and simple in order to better understand the impact of the proposed model on learning important ECG representations based on almost raw input. Finally, the filtered ECG signals are segmented into a fixed window size of 10 seconds and stacked into an array. No overlap is designated between segments to avoid any potential data leakage between training and testing data. It should be noted that the selection of the window size was empirical. Prior works utilizing these datasets have selected a wide range of different window sizes. For example, [4] and [32] selected a window length of 20 seconds for AMIGOS, whereas [22] and [35] selected a window of 60 seconds for both DREAMER and SWELL. As other examples, [23] has selected 5-second windows for WESAD, while [34] has selected 1-second windows for the same dataset.

4.3 Implementation and Training

In order to train the model, successive to pre-processing, each segment is used to generate the 6 transformations described earlier. Finally, the original ECG signals along with the transformed signals are used to train the signal transformation recognition network. We implement the proposed architecture using TensorFlow. We share the implementation of the self-supervised network[1].

Similar to other works in this area [4], [22], [23], [32], [34], [35], we use 10-fold cross-validation to evaluate the performance of the model successive to shuffling of the pre-processed dataset. We randomly select 90% of the data for training and the remaining 10% is used for testing (this process is repeated 10 times without repeating the shuffling step). To train both the signal transformation recognition network and the emotion recognition network, Adam optimizer [49] is used with a learning rate of 0.001 and batch size of 128. The signal transformation recognition network is trained for 100 epochs, while the emotion recognition network is trained for 250 epochs. The number of training epochs for each network is selected to enable the training reach a steady state. Figure 4(A) and (B) show the loss vs. training epoch for the transformation recognition and emotion recognition networks respectively during training. Figure 4(A) shows that the loss for

Self-supervised ECG Representation Learning for Emotion Recognition



4.2 Data Pre-processing

Since the above-mentioned datasets have been collected using different hardware, the signals have different spatiotemporal properties such as spatial range and sampling rate. To minimize the effects of such inter-dataset variations and discrepancies, three simple pre-processing steps were taken. First, SWELL and WESAD ECG signals are downsampled to 256 Hz to be consistent with AMIGOS and DREAMER. Next, we remove baseline wander from all the four datasets by applying a high-pass IIR filter with a pass-band frequency of 0.8 Hz. Lastly, we perform user-specific z-score normalization. While a number of other pre-processing operations such as feature extraction could have been done, we intentionally kept the pre-processing minimal and simple in order to better understand the impact of the proposed model on learning important ECG representations based on almost raw input. Finally, the filtered ECG signals are segmented into a fixed window size of 10 seconds and stacked into an array. No overlap is designated between segments to avoid any potential data leakage between training and testing data. It should be noted that the selection of the window size was empirical. Prior works utilizing these datasets have selected a wide range of different window sizes. For example, [4] and [32] selected a window length of 20 seconds for AMIGOS, whereas [22] and [35] selected a window of 60 seconds for both DREAMER and SWELL. As other examples, [23] has selected 5-second windows for WESAD, while [34] has selected 1-second windows for the same dataset.

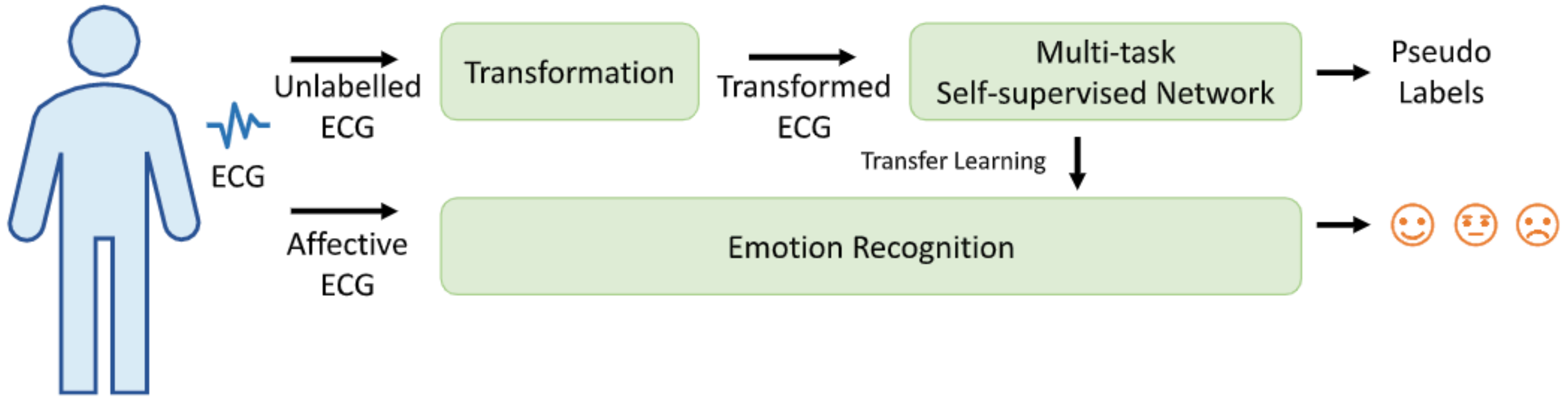
4.3 Implementation and Training

In order to train the model, successive to pre-processing, each segment is used to generate the 6 transformations described earlier. Finally, the original ECG signals along with the transformed signals are used to train the signal transformation recognition network. We implement the proposed architecture using TensorFlow. We share the implementation of the self-supervised network[1].

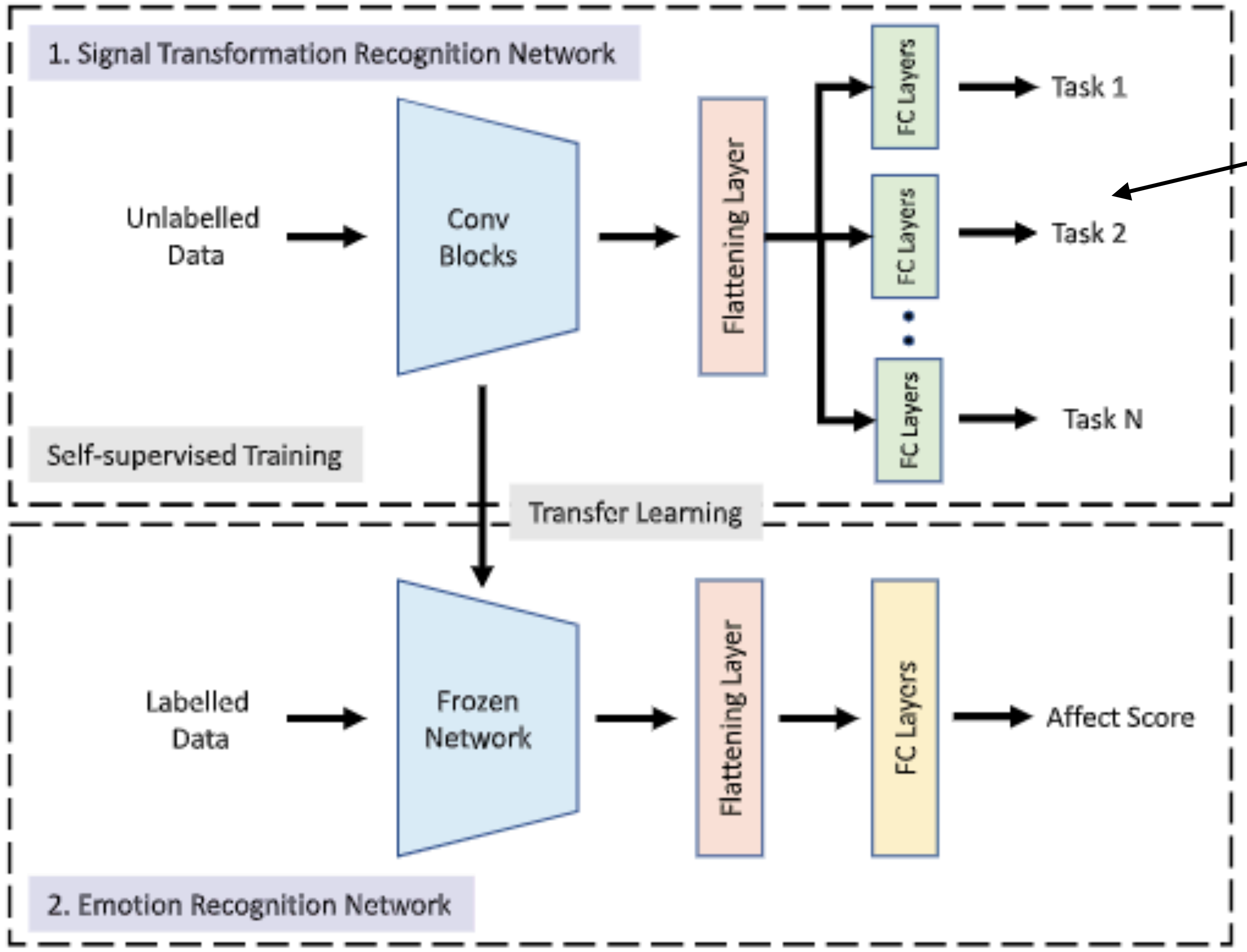
Similar to other works in this area [4], [22], [23], [32], [34], [35], we use 10-fold cross-validation to evaluate the performance of the model successive to shuffling of the pre-processed dataset. We randomly select 90% of the data for training and the remaining 10% is used for testing (this process is repeated 10 times without repeating the shuffling step). To train both the signal transformation recognition network and the emotion recognition network, Adam optimizer [49] is used with a learning rate of 0.001 and batch size of 128. The signal transformation recognition network is trained for 100 epochs, while the emotion recognition network is trained for 250 epochs. The number of training epochs for each network is selected to enable the training reach a steady state. Figure 4(A) and (B) show the loss vs. training epoch for the transformation recognition and emotion recognition networks respectively during training. Figure 4(A) shows that the loss for

https://github.com/grazai/SSL-ECG-Paper-Reimplementaton/tree/main/src/datasets

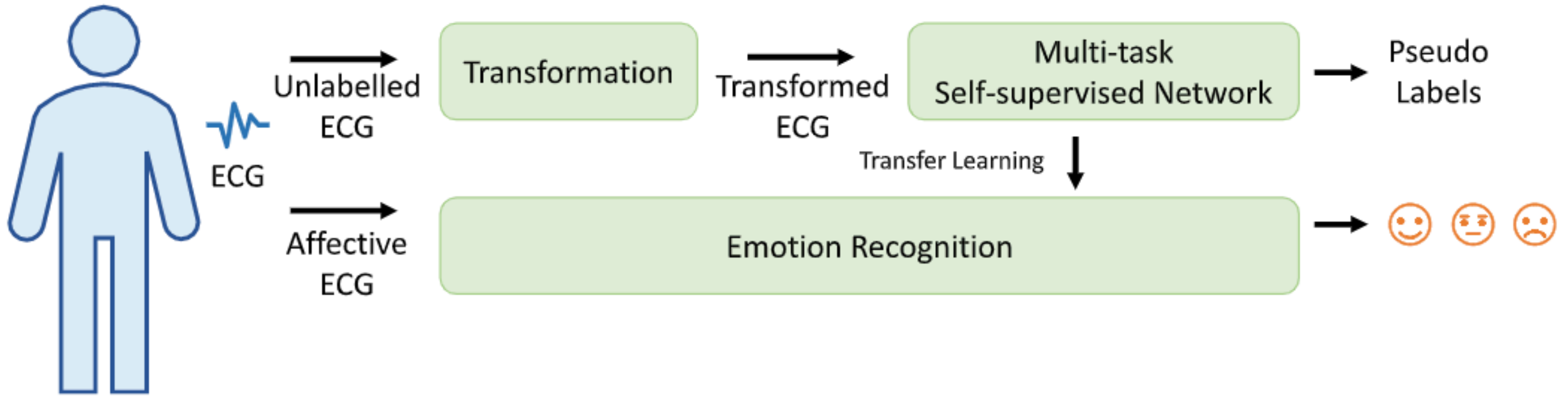
Self-supervised ECG Representation Learning for Emotion Recognition

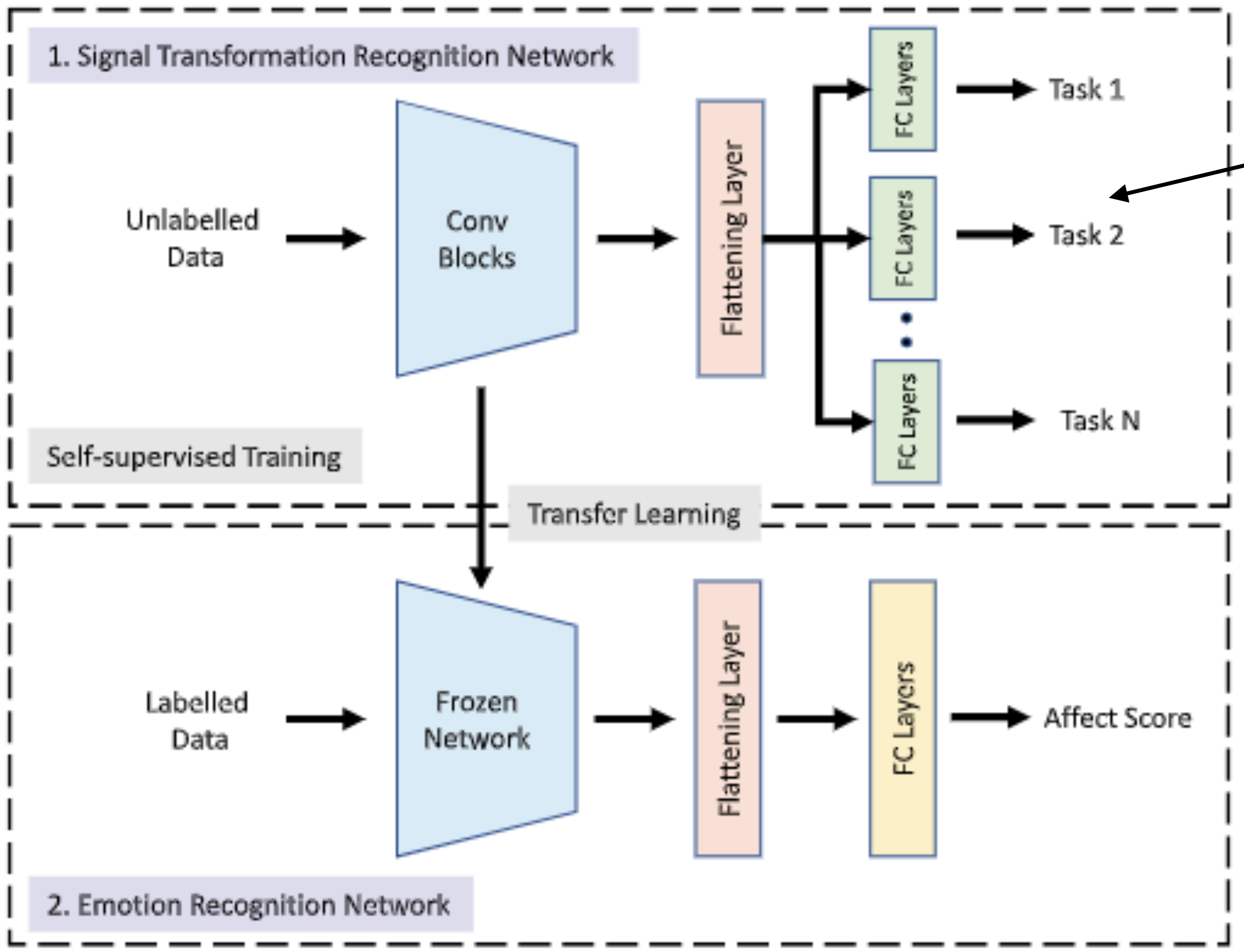


What are these Tasks?



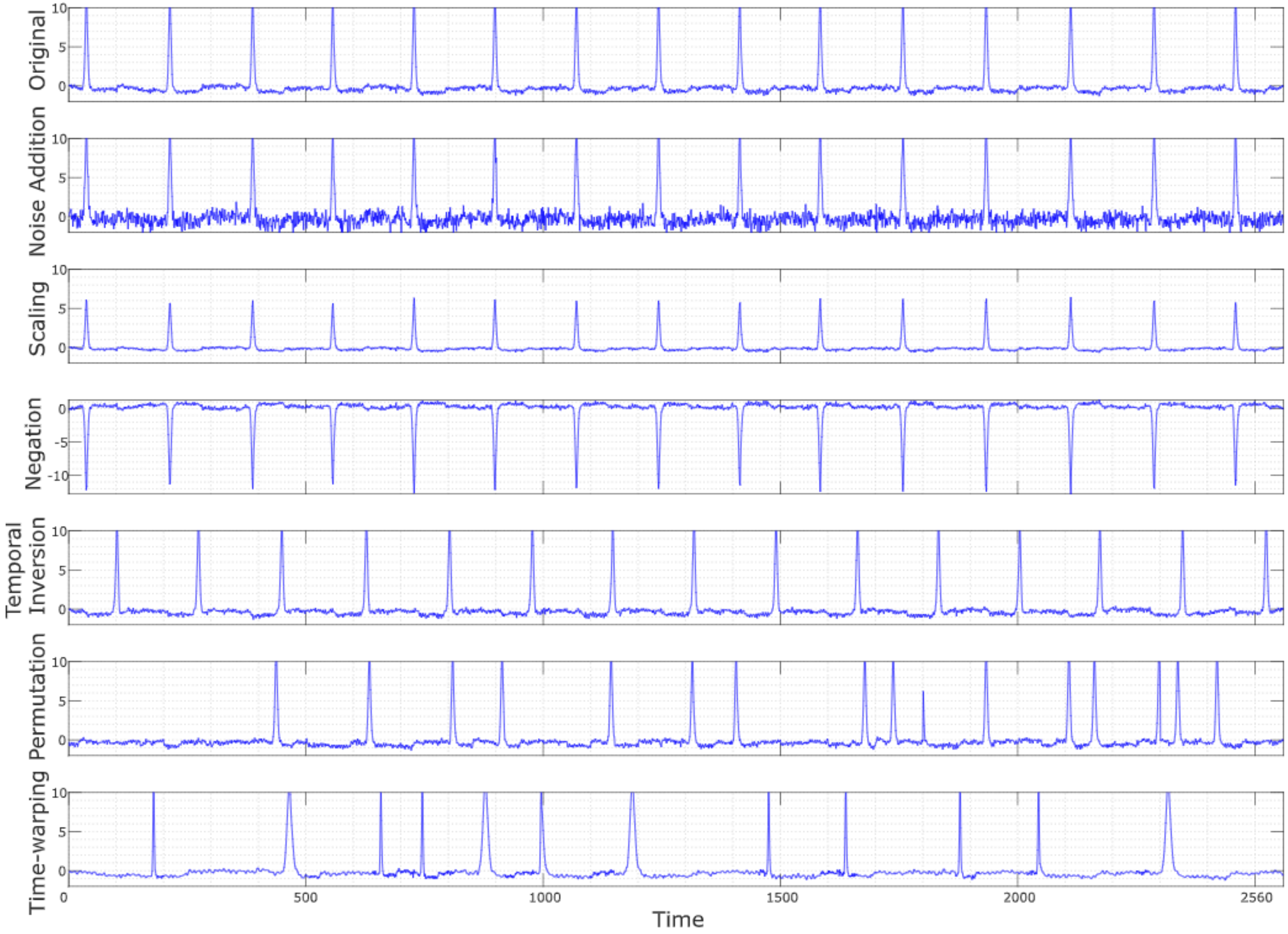
Self-supervised ECG Representation Learning for Emotion Recognition



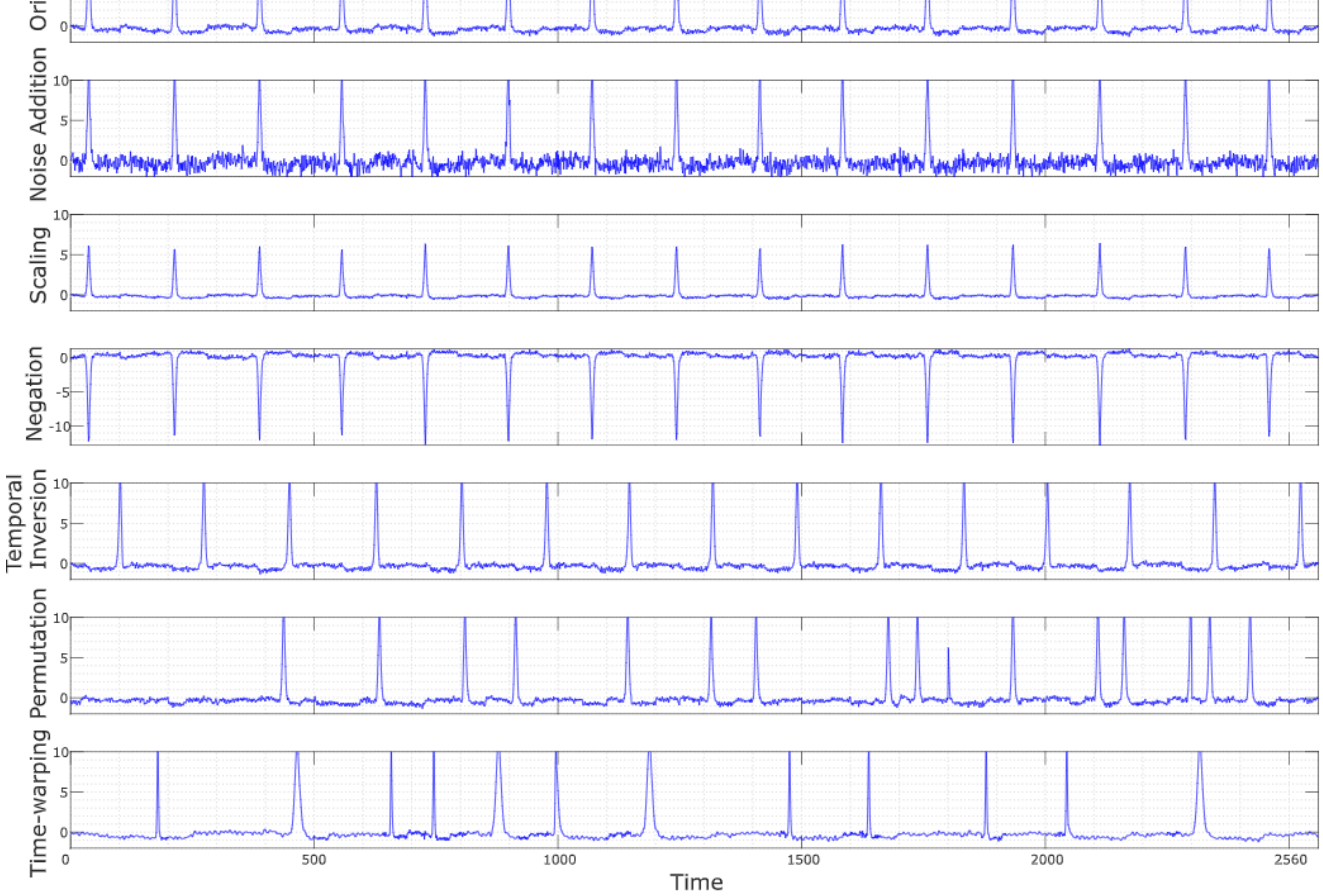


What are these Tasks?

Each Task is a binary classification to detect if a specific signal data augmentation has been applied to the ECG signal



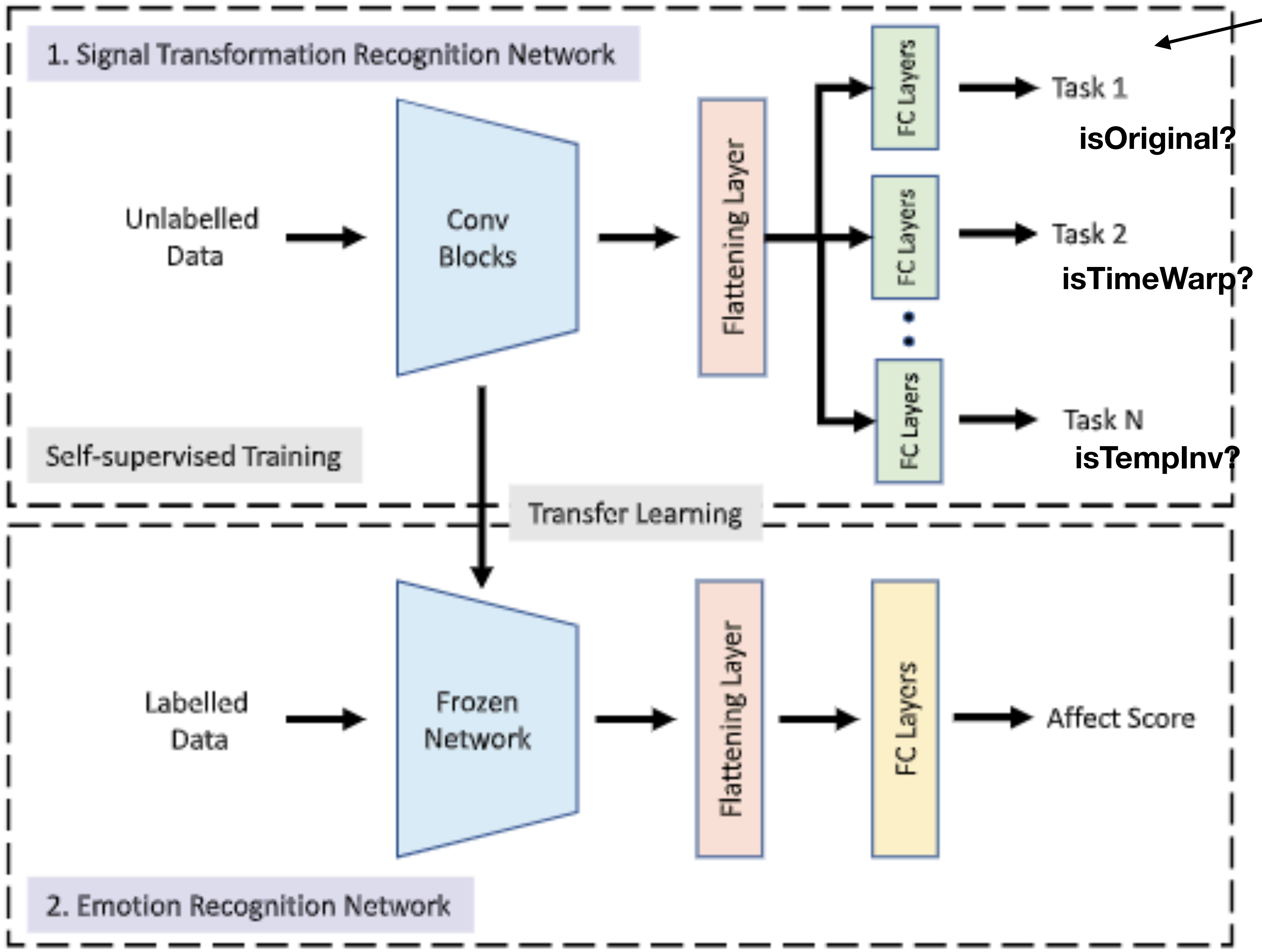
Original
Noise Addition
Scaling
Negation
Temporal Inversion
Permutation
Time-warping
10
5
0
-5
-10
0
500
1000
1500
2000
2560
Time



https://github.com/grazai/SSL-ECG-Paper-Reimplementaton/blob/main/src/augmentations.py
https://github.com/grazai/SSL-ECG-Paper-Reimplementaton/blob/main/src/data.py

Self-supervised ECG Representation Learning for Emotion Recognition

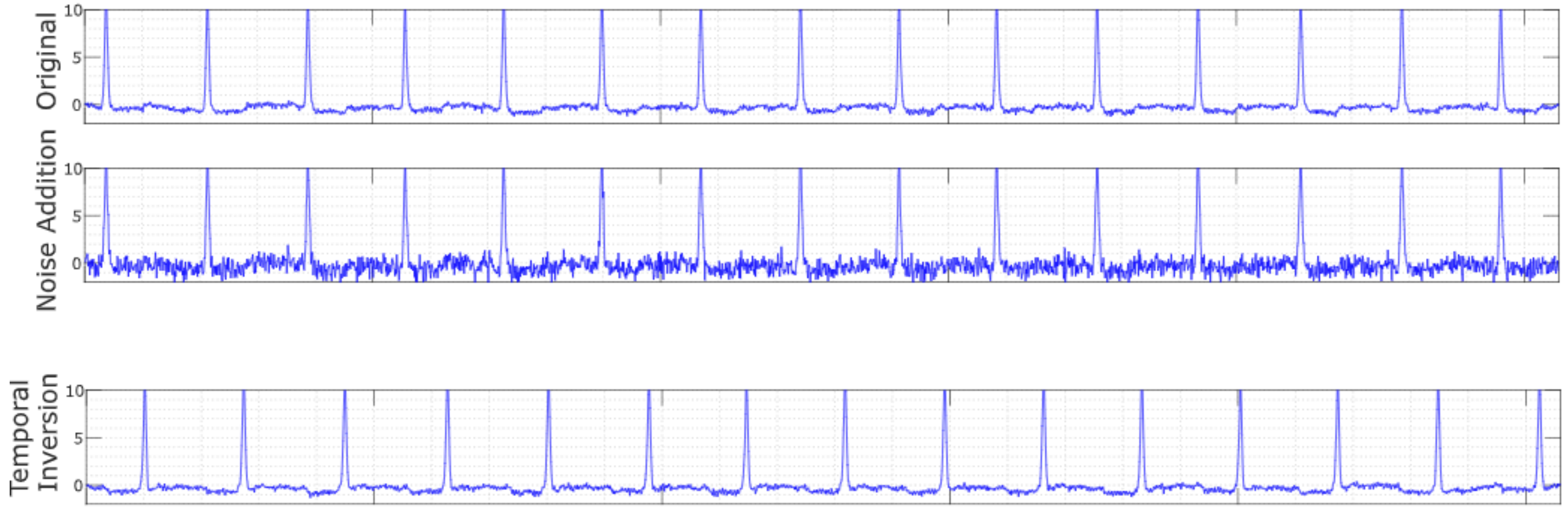
What are these Tasks?



True

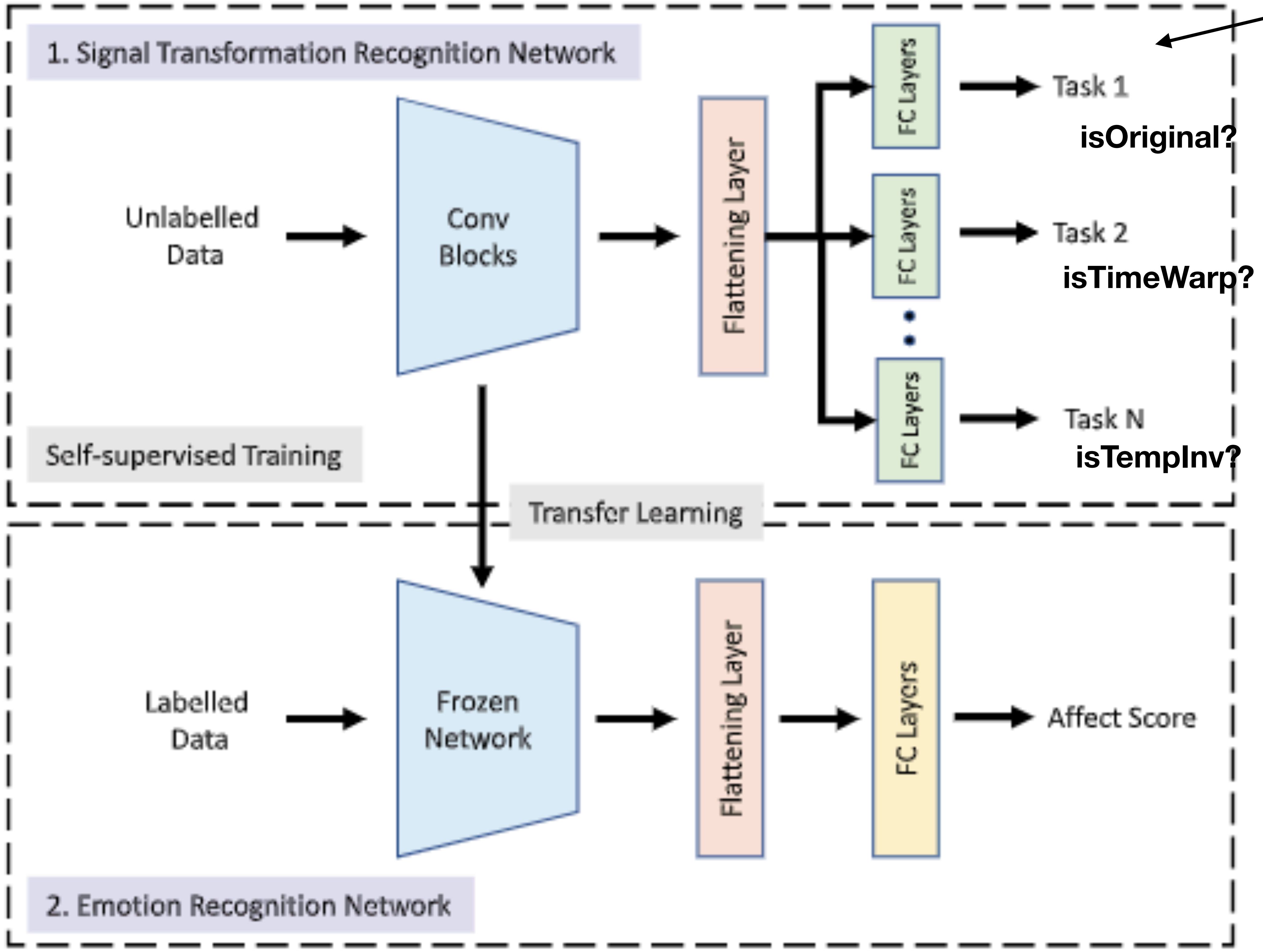
False

True



Self-supervised ECG Representation Learning for Emotion Recognition

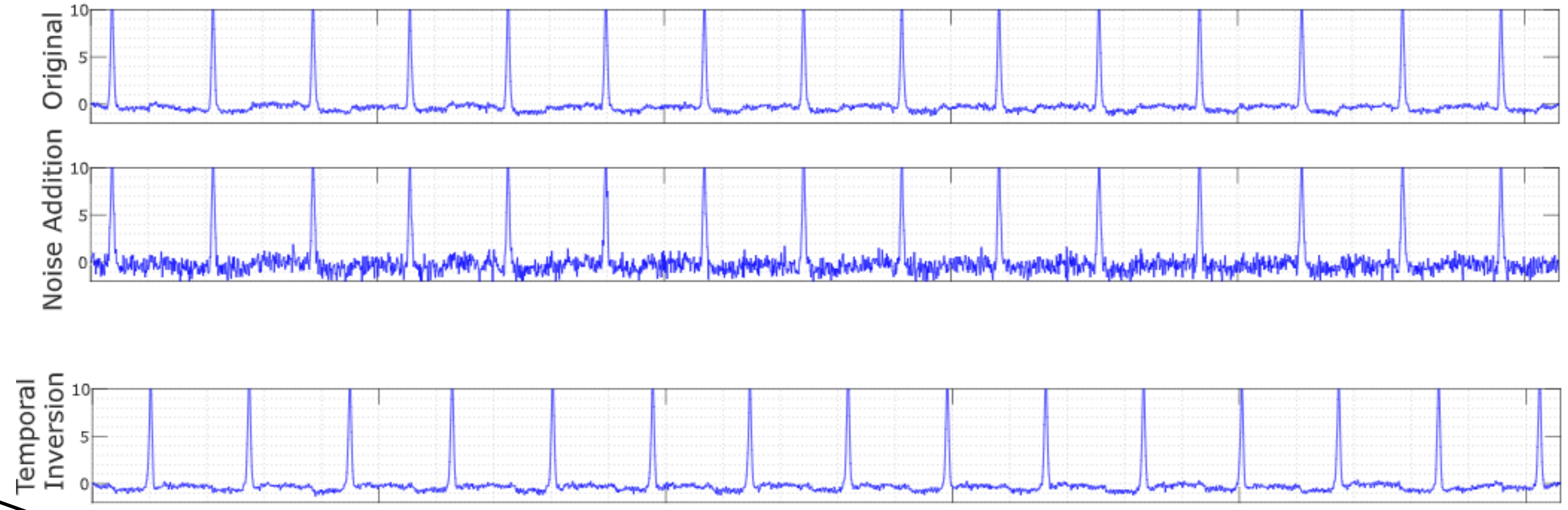
What are these Tasks?



True

False

True

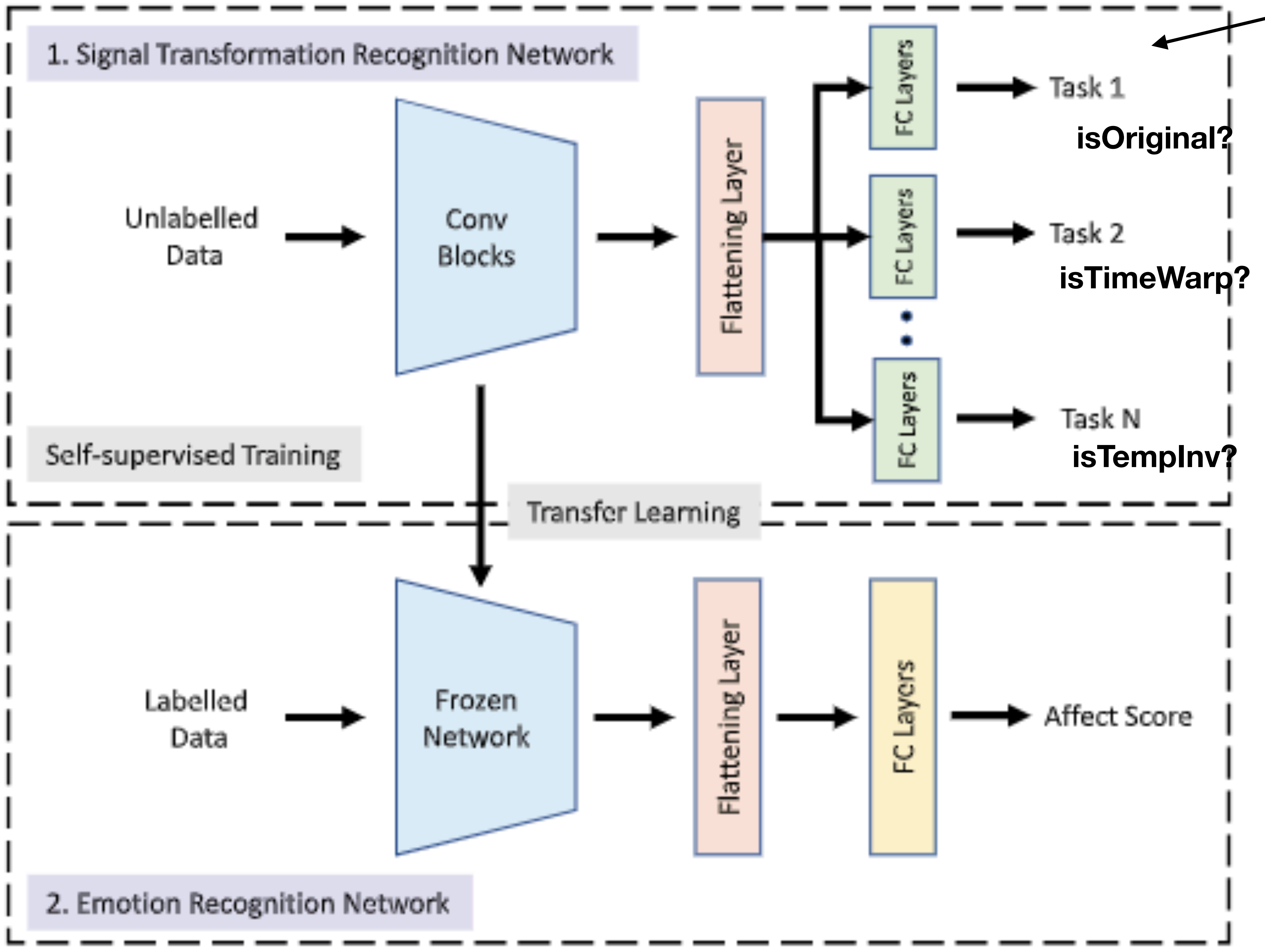


Cross Entropy Binary Loss

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\Big(p\left(y_i\right)\Big) + \left(1-y_i\right) \cdot \log\Big(1-p\left(y_i\right)\Big)$$

Self-supervised ECG Representation Learning for Emotion Recognition

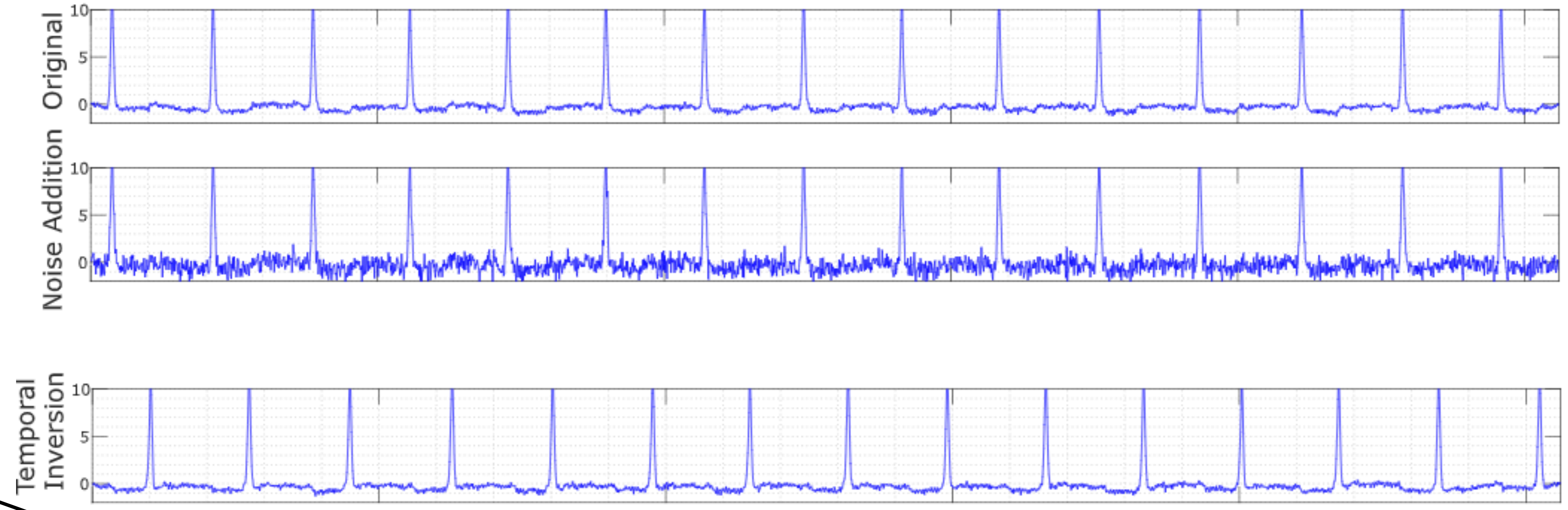
What are these Tasks?



True

False

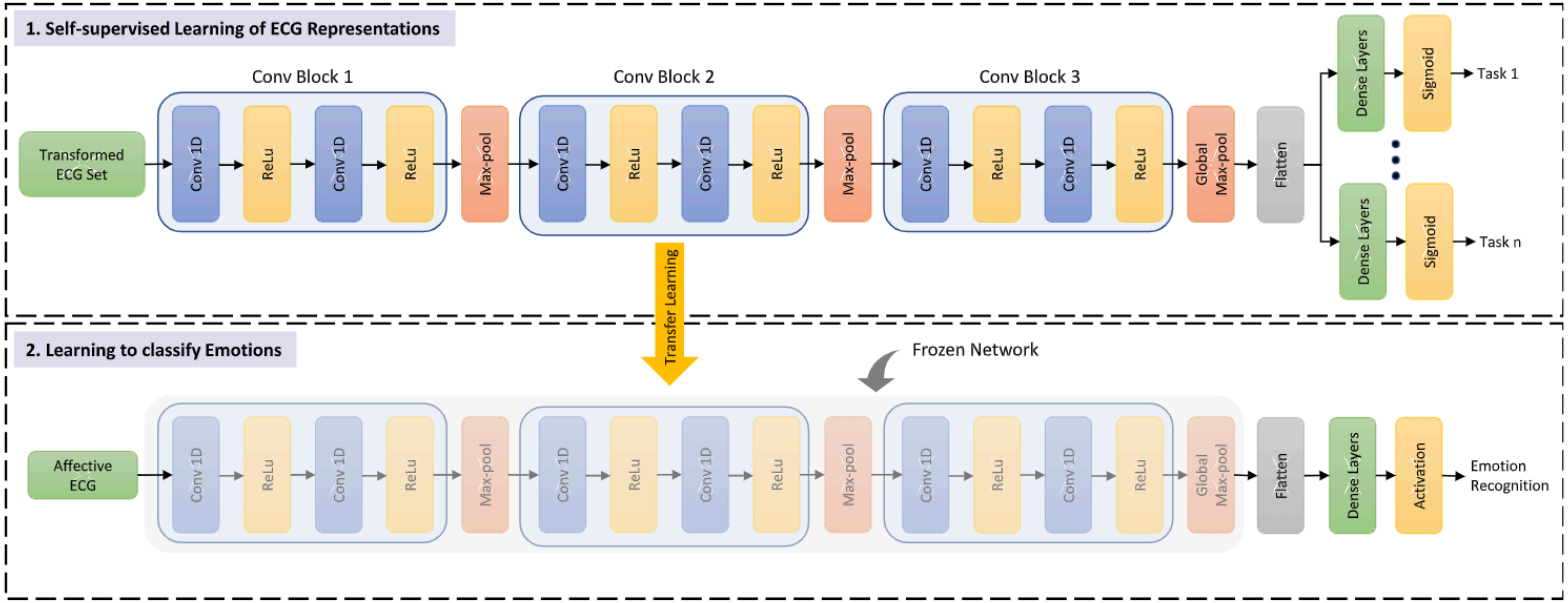
True



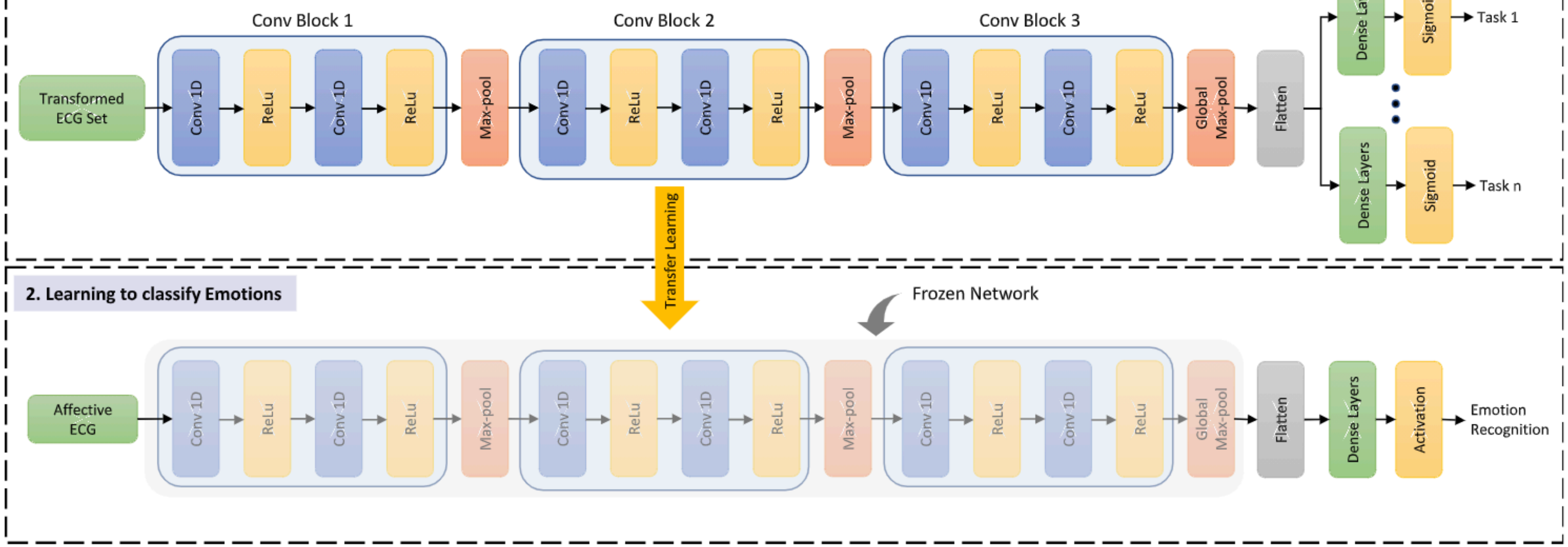
Cross Entropy Binary Loss

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right)$$

https://github.com/grazai/SSL-ECG-Paper-Reimplementaton/blob/main/src/pretext_training.py



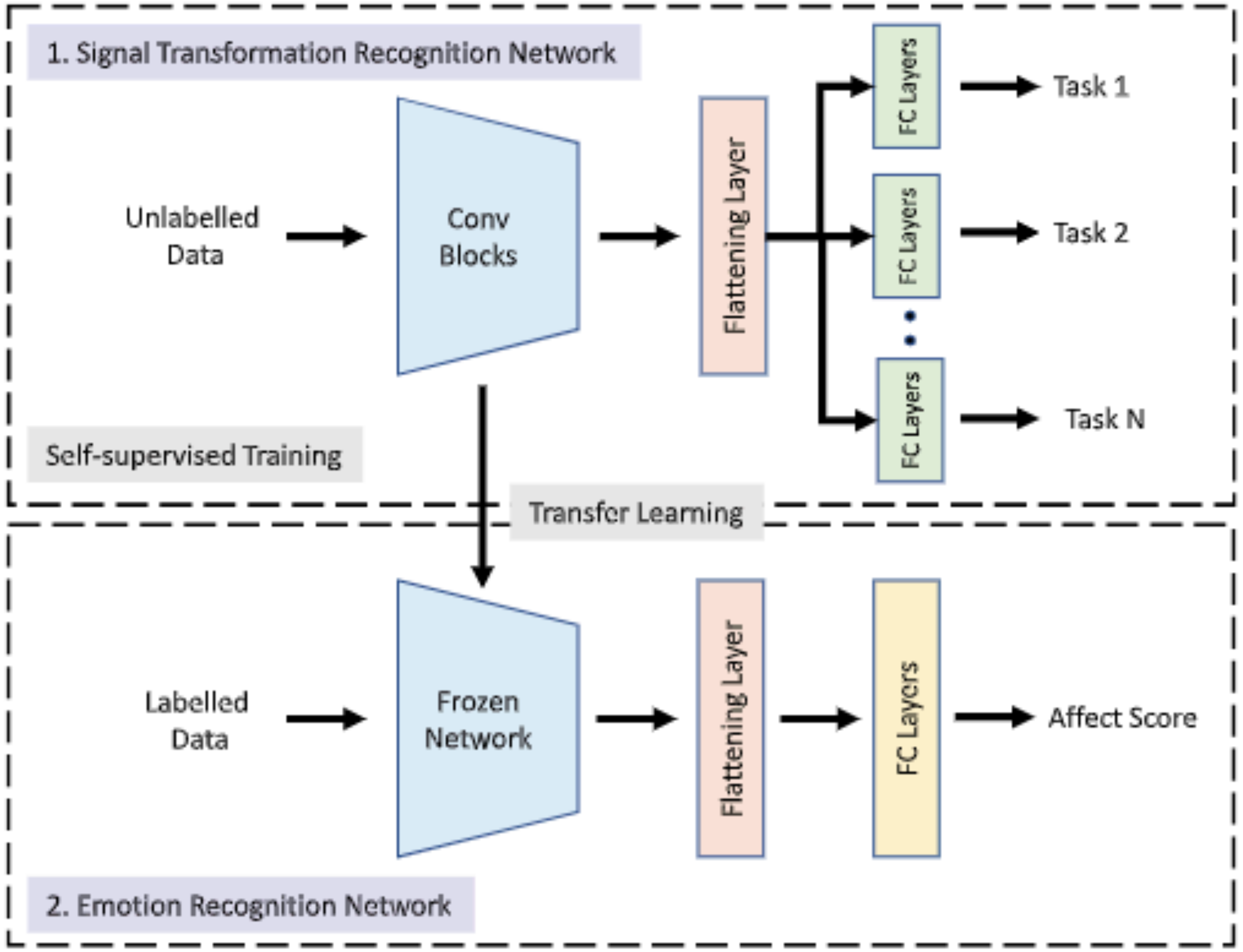
Module	Layer Details	Feature Shape
Input	–	2560×1
Shared Layers	$[conv, 1 \times 32, 32] \times 2$	2560×32
	$[maxpool, 1 \times 8, stride = 2]$	1277×32
	$[conv, 1 \times 16, 64] \times 2$	1277×64
	$[maxpool, 1 \times 8, stride = 2]$	635×64
	$[conv, 1 \times 8, 128] \times 2$	635×128
	global max pooling	1×128
Task-Specific Layers	$[dense] \times 2$ $\times$ 7 *parallel tasks*	128
Output	–	2



Module	**Layer Details**	**Feature Shape**
Input	–	2560×1
Shared Layers	$[conv, 1 \times 32, 32] \times 2$	2560×32
	$[maxpool, 1 \times 8, stride = 2]$	1277×32
	$[conv, 1 \times 16, 64] \times 2$	1277×64
	$[maxpool, 1 \times 8, stride = 2]$	635×64
	$[conv, 1 \times 8, 128] \times 2$	635×128
	global max pooling	1×128
Task-Specific Layers	$[dense] \times 2$ $\times$ *7 parallel tasks*	128
Output	–	2

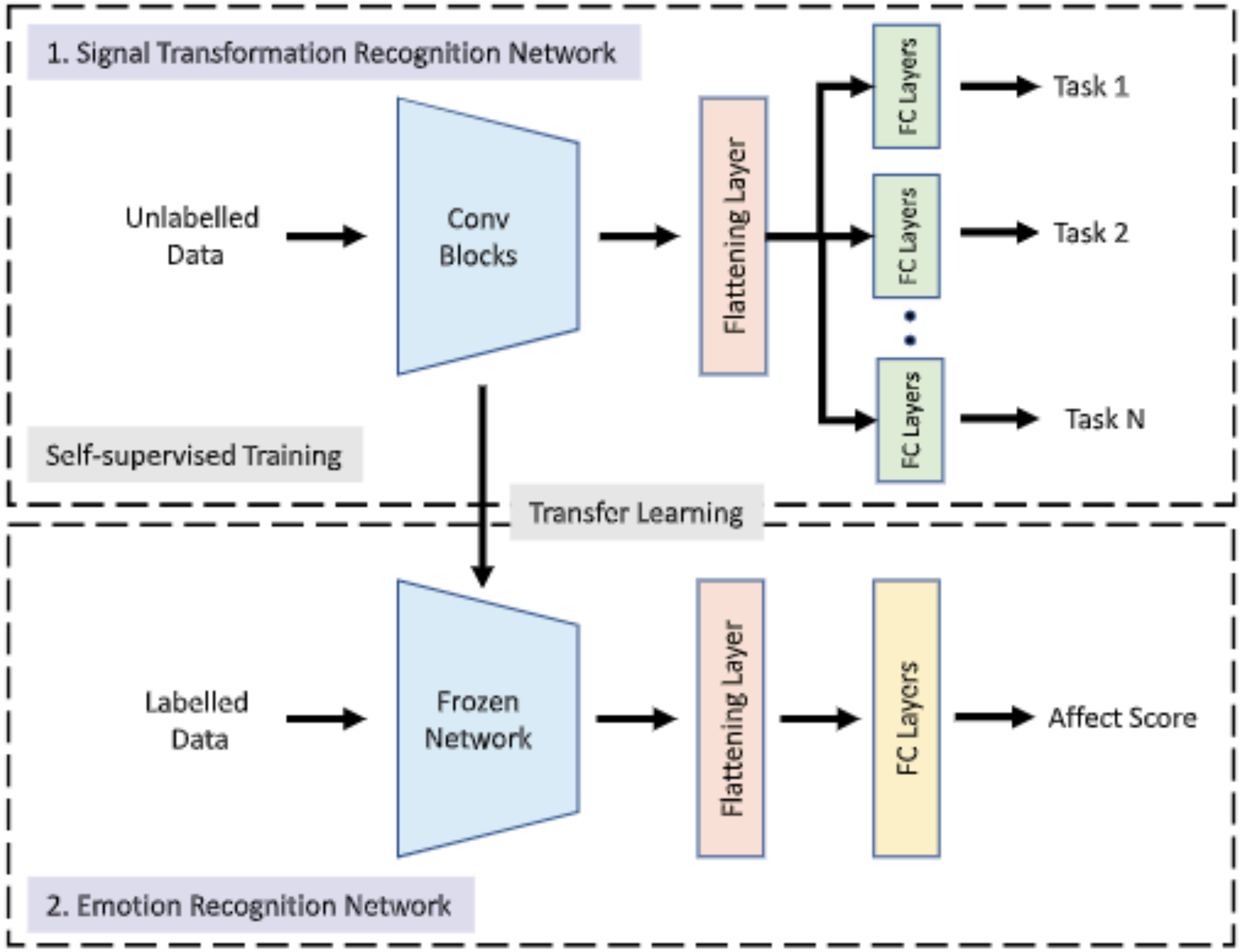
https://github.com/grazai/SSL-ECG-Paper-Reimplementaton/blob/main/src/model.py

Self-supervised ECG Representation Learning for Emotion Recognition



First Stage:
Learn to discriminate "valid" ECG towards corrupted
Learn "what" makes an signal an ECG signal

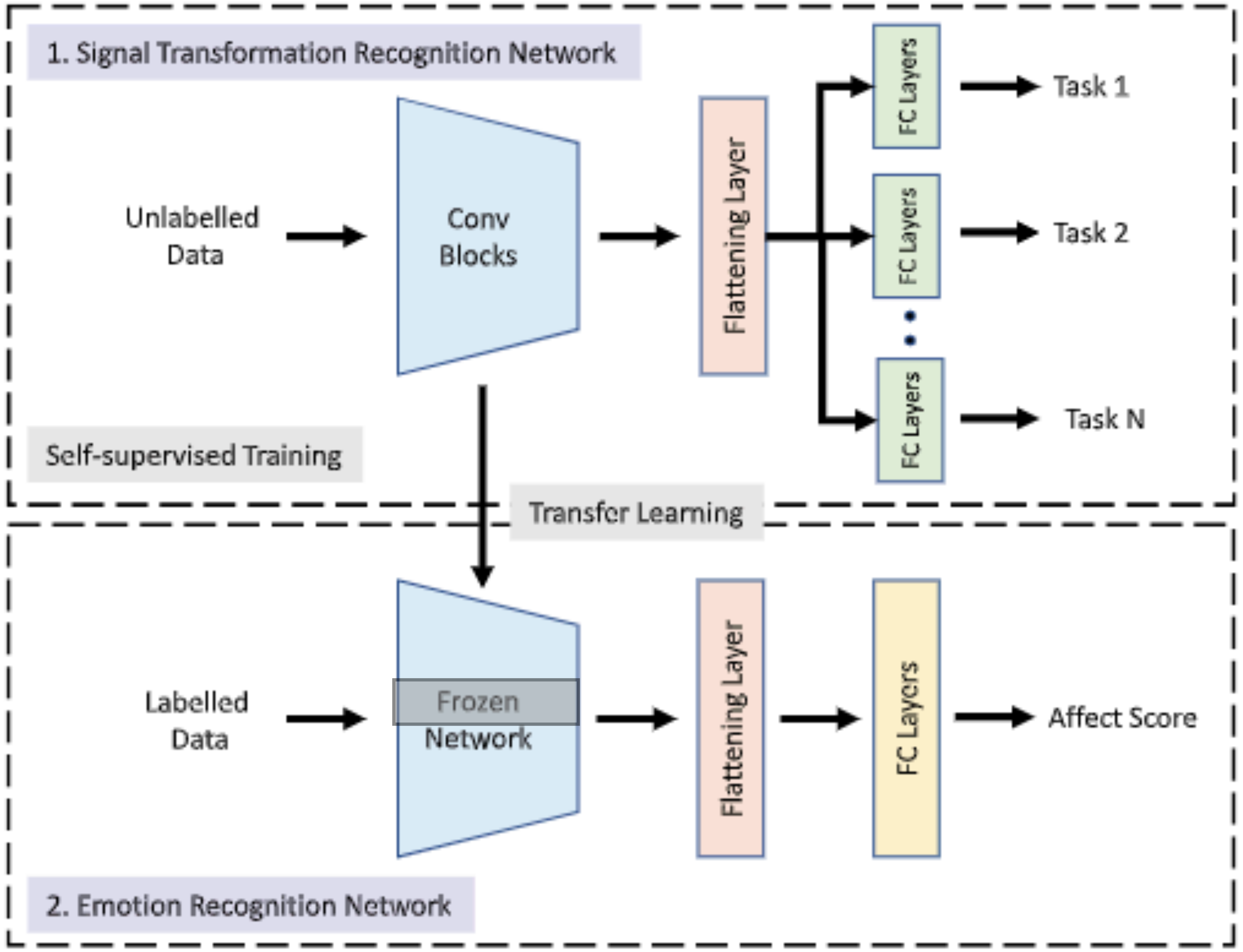
Self-supervised ECG Representation Learning for Emotion Recognition



First Stage:
Learn to discriminate “valid” ECG towards corrupted
Learn “what” makes an signal an ECG signal

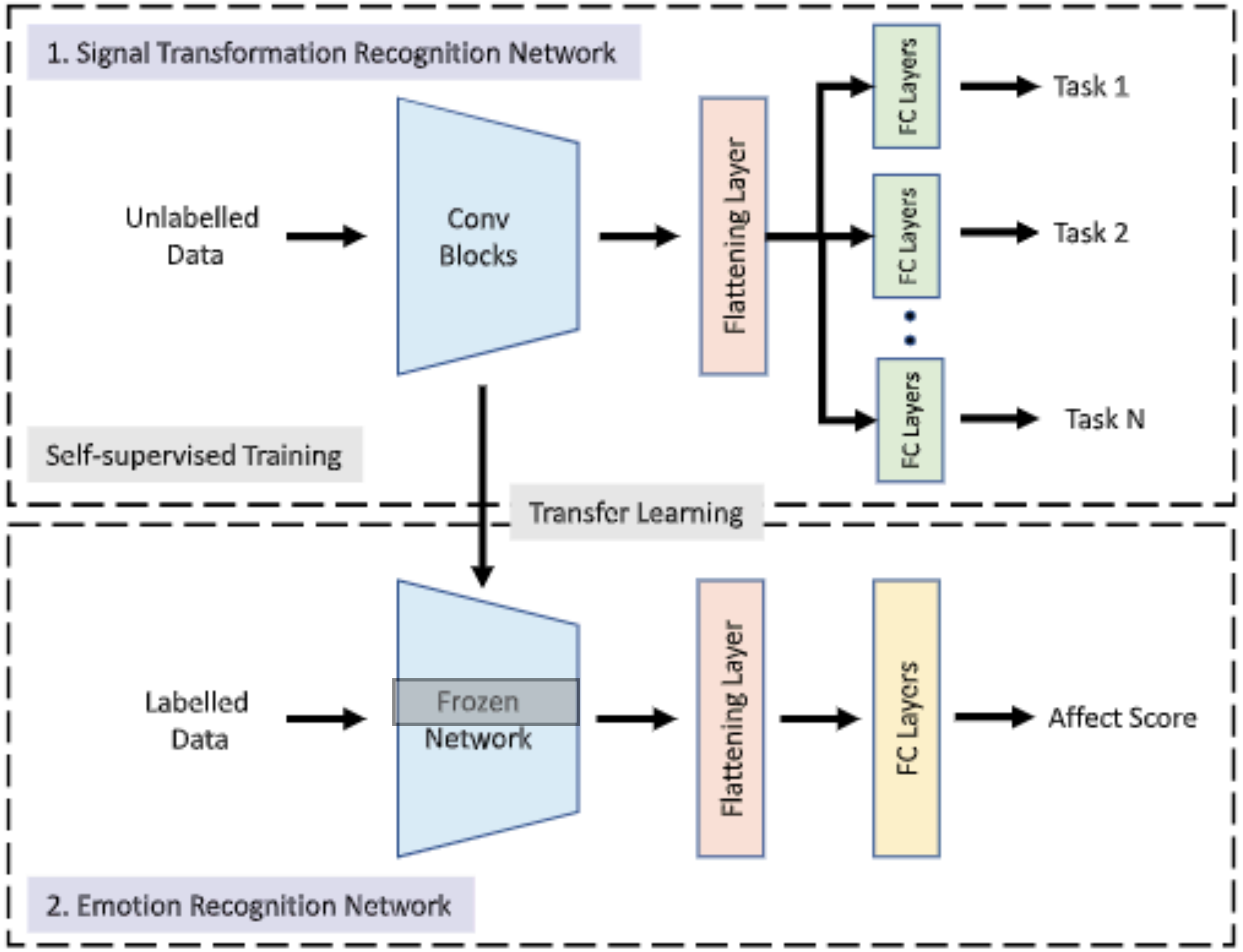
Second Stage:
Learn new Head for Dataset specific labels

Self-supervised ECG Representation Learning for Emotion Recognition



Second Stage:
Learn new Head for Dataset specific labels
Hard Feature approach (instead of Fine-tuning)

Self-supervised ECG Representation Learning for Emotion Recognition



Second Stage:
Learn new Head for Dataset specific labels
Hard Feature approach (instead of Fine-tuning)

https://github.com/grazai/SSL-ECG-Paper-Reimplementaton/blob/main/src/finetune_to_target.py

A: AMIGOS

Ref.	Method	Arousal		Valence	
		Acc.	F1	Acc.	F1
[5]	GNB	–	0.545	–	0.551
[29]	CNN	0.81	0.76	0.71	0.68
Ours	Fully-Supervised CNN	0.844	0.835	0.811	0.809
	Self-Supervised CNN	**0.889**	**0.884**	**0.875**	**0.874**

B: DREAMER

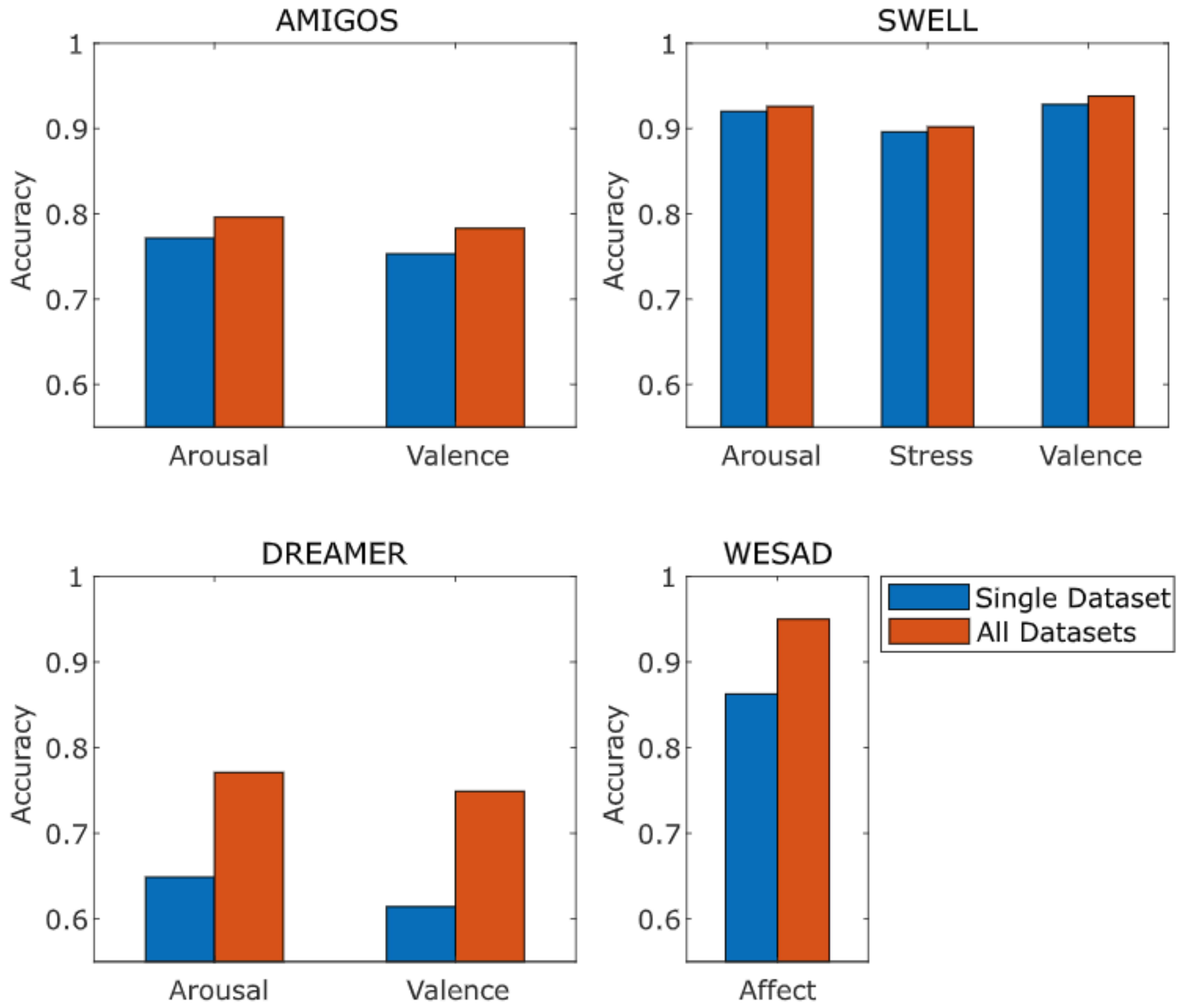
Ref.	Method	Arousal		Valence	
		Acc.	F1	Acc.	F1
[23]	SVM	0.624	0.580	0.624	0.531
Ours	Fully-Supervised CNN	0.707	0.708	0.666	0.658
	Self-Supervised CNN	**0.859**	**0.859**	**0.850**	**0.845**

C: WESAD

Ref.	Method	Affect State	
		Acc.	F1
[24]	*k*NN	0.548	0.478
	DT	0.578	0.517
	RF	0.604	0.522
	AB	0.617	0.525
	LDA	0.663	0.560
[31]	CNN	0.83	0.81
Ours	Fully-Supervised CNN	0.932	0.912
	Self-Supervised CNN	**0.969**	**0.963**

D: SWELL

Ref.	Method	Stress		Arousal		Valence	
		Acc.	F1	Acc.	F1	Acc.	F1
[32]	*k*NN	0.769	–	–	–	–	–
	SVM	0.864	–	–	–	–	–
Our	Fully-Supervised CNN	0.894	0.874	0.956	0.962	0.961	0.956
	Self-Supervised CNN	**0.933**	**0.924**	**0.967**	**0.964**	**0.973**	**0.969**



Repositories

- https://code.engineering.queensu.ca/17ps21/SSL-ECG
  - Tensorflow 1.14
  - Problem: They assume a folder with preprocessed ECG, and did not provide that. You would still have to code that by yourself.
- https://github.com/grazai/SSL-ECG-Paper-Reimplementaton
  - I could not parse SWELL, so I replaced that with DEAP

References

- P. Sarkar and A. Etemad, "Self-supervised ECG Representation Learning for Emotion Recognition," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2020.3014842.

- Jionghao Lin, Shirui Pan, Cheng Siong Lee, and Sharon Oviatt. 2019. “An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals,” In Proceedings of the 28th ACM Int. Conf. on Information and Knowledge Management (CIKM '19) doi :https://doi.org/10.1145/3357384.3358160

- Y. Zhang, Y. Liu, F. Weninger and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 4990-4994, doi: 10.1109/ICASSP.2017.7953106.

Thanks for Attention

Any questions?