MuSe-Personalization 2023: Feature Engineering, Hyperparameter Optimization, and Transformer-encoder Re-discovery

Ho-min Park
Ghent University Global Campus
Incheon, South Korea
IDLab, ELIS, Ghent University
Ghent, Belgium
homin.park@ghent.ac.kr

Ganghyun Kim
Ghent University Global Campus
Incheon, South Korea
ganghyun.kim@ghent.ac.kr

Arnout Van Messem
Department of Mathematics, University of Liège
Liège, Belgium
arnout.vanmessem@uliege.be

Wesley De Neve
Ghent University Global Campus
Incheon, South Korea
IDLab, ELIS, Ghent University
Ghent, Belgium
wesley.deneve@ghent.ac.kr

ABSTRACT

This paper presents our approach for the MuSe-Personalization sub-challenge of the fourth Multimodal Sentiment Analysis Challenge (MuSe 2023), with the goal of detecting human stress levels through multimodal sentiment analysis. We leverage and enhance a Transformer-encoder model, integrating improvements that mitigate issues related to memory leakage and segmentation faults. We propose novel feature extraction techniques, including a pose feature based on joint pair distance and self-supervised learning-based feature extraction for audio using Wav2Vec2.0 and Data2Vec. To optimize effectiveness, we conduct extensive hyperparameter tuning. Furthermore, we employ interpretable meta-learning to understand the importance of each hyperparameter. The outcomes obtained demonstrate that our approach excels in personalization tasks, with particular effectiveness in Valence prediction. Specifically, our approach significantly outperforms the baseline results, achieving an Arousal CCC score of 0.8262 (baseline: 0.7450), a Valence CCC score of 0.8844 (baseline: 0.7827), and a combined CCC score of 0.8553 (baseline: 0.7639) on the test set. These results secured us the second place in MuSe-Personalization.

CCS CONCEPTS

• **Computing methodologies** → *Knowledge representation and reasoning*; **Neural networks**; *Object recognition*; • **Applied computing** → *Health informatics*.

KEYWORDS

Emotion detection, Human pose, Multimodal fusion, Multimodal sentiment analysis

MuSe' 23, October 29, 2023, Ottawa, Canada

ACM ISBN 979-8-4007-0270-9/23/10.
https://doi.org/10.1145/3606039.3613104

ACM Reference Format:
Ho-min Park, Ganghyun Kim, Arnout Van Messem, and Wesley De Neve. 2023. MuSe-Personalization 2023: Feature Engineering, Hyperparameter Optimization, and Transformer-encoder Re-discovery. In *Proceedings of the 4th Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation (MuSe '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3606039.3613104

1 INTRODUCTION

Emotions, the subtle shifts in our mental state, can significantly influence our autonomic nervous system, which governs involuntary bodily functions such as heart rate and breathing [12]. These emotional alterations can manifest themselves in voice intonation, facial expressions, and body gestures—elements intuitively perceived by humans. Sentiment analysis plays a critical role in interpreting these manifestations, combining human emotional states with the physical responses they elicit, quantified via video, audio, and sensor data. By fusing these modalities, we can delve deeper into the complexities of emotional responses.

In this research, we develop a multimodal machine learning model for detecting human stress levels by harnessing sentiment analysis techniques, including facial expression and tone of voice analysis, along with physiological response evaluation. Our model is applied to a unique dataset, Ulm-TSST [25], consisting of videos simulating stressful job interview scenarios. In more detail, this paper describes our participation in the MuSe-Personalization sub-challenge, which is part of the fourth Multimodal Sentiment Analysis Challenge (MuSe 2023) [7]. Building on the MuSe-Stress sub-challenge that took place last year, this sub-challenge aims to enhance and personalize stress prediction models by accounting for individual characteristics. In MuSe-Stress 2022 [8], we proposed a Transformer-encoder model that took advantage of Pose features [22], with this model obtaining the third place in the competition. For this year's MuSe-Personalization sub-challenge, we

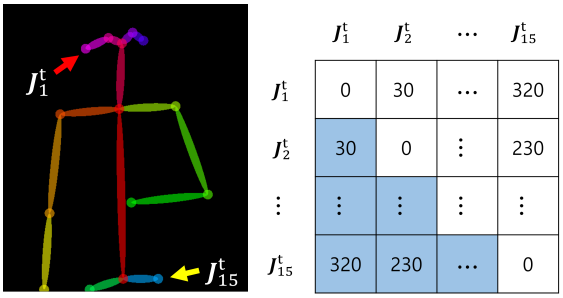
decided to extend this approach, focusing on hyperparameter tuning and feature engineering.

Our paper presents four key contributions:

- **Revised Transformer-encoder:** We utilize and refine the Transformer-encoder model that was previously proposed in [22], addressing issues related to memory leakage and segmentation faults under specific hyperparameter settings.
- **Novel Pose Feature Extraction:** We further explore the potential of Pose features, proposing a novel extraction method based on joint pair distance. This process incorporates the application of feature and time normalization.
- **Self-Supervised Learning-Based Feature Extraction:** In contrast to the baseline study, we apply feature extractors based on self-supervised learning. In particular, we made use of Wav2Vec2.0 [3] and Data2Vec [2] to extract features from the available raw audio.
- **Interpretable Meta Learning on Hyperparameters:** After extensive hyperparameter tuning, we build a meta-learning model to assess the performance impact of each hyperparameter, leading to improved understanding and optimization.

In essence, we enhanced and integrated already existing models and feature extraction techniques, resulting in an improved approach towards the problem of stress detection.

The remainder of this paper is organized as follows. Section 2 explains our methodology. Next, our experimental setup is detailed in Section 3, whereas our experimental results are summarized in Section 4. These results are then further interpreted in Section 5, discussing their implications and also paying attention to the limitations of the adopted methodology. Finally, Section 6 provides conclusions and potential directions for future research.



(a) Pose and joint **(b) Distance matrix**

Figure 1: Our proposed pose feature extraction process: (a) a human pose skeleton of the frame at time t, highlighting two example joints J_1^t and J_{15}^t and (b) a table representation of the Euclidean distance between each pair of joints, with the blue-colored section being utilized as a new pose feature. Our approach aims to capture more precise and comprehensive movement information, enhancing the ability of models to accurately associate movements with stress levels.

2 METHODS

2.1 Transformer-encoder

In this study, we adopt the Transformer-encoder model previously utilized in [22] as our primary model. The decision to augment the baseline model, which was composed of Gated Recurrent Units (GRU) [6], with a Transformer approach was based on the Transformer's capability to effectively learn long-range dependencies inherent in sequential data. This efficacy is due to the model's self-attention mechanism [28], which allows for the consideration of interactions amongst all elements within a sequence simultaneously. The original Transformer architecture includes an encoder and a decoder. However, since our task—a regression task—computes the output directly from the input, there is no need for a decoder that generates sequences of values in output. Therefore, we only utilize the encoder. Additionally, we addressed the issue of forced training termination due to segmentation faults by optimizing the hyperparameter grid. Our improved code can be found at the following GitHub URL: https://github.com/kyleok/MUSE2023_clean.

2.2 Novel Pose Feature Extraction

In this study, we introduce a more effective method of utilizing human posture as an input for stress detection, compared to our initial use of a pose feature for MuSe 2022 in [22]. In particular, we devised a method to extract a new type of feature using the two-dimensional joint positions (x, y) of a skeleton generated by OpenPose [5]. This method involved applying OpenPose to video frames sampled at 0.5-second intervals (2Hz) and calculating the Euclidean distance of the difference in body part positions before and after each time interval. This technique of feature extraction characterizes the movement of a subject over a defined time interval and assists the model in correlating the movement of this subject with their stress level.

Our pose approach achieved the second highest Concordance Correlation Coefficient (CCC) score in predicting emotional dimensions in the development (validation) set compared to other features. However, this approach encountered multiple challenges. For instance, a combined CCC of 0.1921, discovered during an evaluation using the test set, indicated susceptibility to a lack of generalization compared to other features. A substantial disparity between parts with large movements (like hands) and parts with smaller movements (like hips, neck, and shoulders) was noted, and the approach failed to register complex movements. Upon re-examining the feature extraction process, we found that the movement of the hands and arms was substantial compared to the movement of all other body parts. This means that some feature values were having very large variances, which can cause them to be dominantly used in predictions and obscure the importance of other features. This is likely due to the fact that the data were collected in a job interview environment.

To mitigate these issues, we drew inspiration from the Joint Collection Distances (JCD) concept, as introduced by Yang et al. [30], to calculate the Euclidean distance between all pairs of joints of a skeleton in each frame, thereby proposing this as a new pose feature for stress detection. In doing so, we aim to capture spatially and temporally invariant information in skeleton-based motion recognition. An intuitive explanation for this can also be found in Figure 1.

While our contribution to advancing pose features is grounded in the ideas of Yang et al. [30], it diverges in two main aspects. First,

we consider the limitation imposed by the height of the camera and the desk it is placed on, which often precludes the appearance of the lower body, primarily below the knees. As a result, we exclude the body joints below the hip and compute JCD for the remaining 15 joints. Secondly, we strive to accommodate the diversity of body sizes and variable movements over time. To achieve this, we normalize the JCD over time and space, which results in three types of features: (1) JCD-feature, (2) JCD-time, and (3) JCD-both. The JCD-feature normalizes the distance between all joint pairs, creating a feature invariant to the gender and body size of a subject. JCD-time, on the other hand, normalizes over time, assisting the model in identifying joint pairs active at specific times. Lastly, JCD-both, which applies both JCD-feature and JCD-time, was developed to produce features with invariance across both feature and time.

2.3 Self-Supervised Learning-Based Feature Extraction

We employ Wav2Vec2.0 [3] and Data2Vec [2], which are both self-supervised learning (SSL) [19]-based feature extractors, to derive unique features from raw audio. This differs from the baseline method [7] in terms of the datasets used for pre-training and the types of features extracted. While the baseline method uses the MSP-Podcast [15] dataset and Wav2Vec2.0 to extract a single 1,024-dimensional feature, we use LibriSpeech [20] and both Wav2Vec2.0 and Data2Vec to extract two types of features: audio (512-dimensional) and context (768-dimensional). We designated those features as W2V-audio, W2V-context, D2V-audio, and D2V-context, thereby allowing us to dissect and examine their efficiency. For the sake of distinction, the Wav2Vec2.0 feature provided in the baseline paper is separately named W2V-original.

Our feature extractors were pre-trained and fine-tuned on 960 hours of LibriSpeech [20] at a sampling rate of 16kHz. We re-sampled the raw audio to 16kHz and subsequently extracted the features. This input goes into both Data2Vec and Wav2Vec2.0, and is then converted into audio and context feature vectors, with a length equal to the input audio (seconds) $\times$ 16,000 (Hz). Since both the General Model and the Personal Model take input features at a frequency of 2Hz, we sampled one feature vector over every 8,000 vectors.

Unlike the emotion-labeled MSP-Podcast [15], LibriSpeech [20] is based on speech input and text script output, and it offers at least twice as much training data. As a result, we expect our features to provide a broader representation and act as intermediaries between text and audio.

Wav2Vec2.0 utilizes contrastive learning [19] to pre-train on unlabeled data. This training approach comprises two primary components: masking and quantization. Masking involves randomly concealing parts of the input data, which the model then attempts to predict, thus enhancing its ability to extract features and predict missing elements from the input speech signal. Furthermore, quantization simplifies the input data, facilitating improved predictability for the model. Indeed, this stage is recognized for its efficacy in handling noisy speech [9].

Data2Vec is also utilized as an audio feature extractor, similar to Wav2Vec2.0. The distinction between them lies in their pre-training methods. While Wav2Vec2.0 uses both masking and quantization, Data2Vec's approach for pre-training excludes the quantization step. By skipping quantization, Data2Vec retains more information from the raw audio, potentially leading to a richer representation of the input data.

2.4 Interpretable Meta Learning on Hyperparameters

Meta learning provides valuable insights into the effectiveness of various machine learning methodologies across a range of tasks [27]. In our study, we employ a Random Forest regressor to predict the development CCC of a model trained with a given combination of hyperparameters. This methodology allows us to estimate the most optimal hyperparameters for our model. Additionally, we use SHapley Additive exPlanations (SHAP) [16], a method for interpretability that helps us better understand the importance of different hyperparameters. Specifically, with this approach, we can achieve better insight into how these hyperparameters influence CCC scores, aiding in further optimization.

3 EXPERIMENTAL SETUP

3.1 Multimodal Training Procedure and Terms

Due to their size and complexity, the video data are not directly used as model inputs. Instead, we employ a feature extractor to convert them into a more manageable format in a process known as feature extraction. This process samples video, audio, text, and sensor values from each interview video at 0.5-second intervals, transforms the sampled data via a feature extractor to generate output, and condenses this output into a single matrix representing a video. These generated features are referred to as unimodal features. Models trained with these unimodal features extracted from training videos and their corresponding emotional dimensions are termed unimodal models. Finally, for enhanced emotional dimension prediction, we can combine different types of features through a process known as fusion. In our study, we apply late fusion, where predictions from unimodal models are averaged.

Our overall approach towards MuSe-Personalization is visualized in Figure 2. The GRU and Transformer-encoder in each box represent the models utilized in our study. These models are trained and evaluated using a dataset that comprises recordings from 69 subjects. Specifically, the training set (red) and the development set (blue) are used to train the General Model. For each subject, a copy of this General Model serves as a pre-trained model, after which fine-tuning is applied based on subject-specific test data. This process involves the division of the test set (green) into a sub-training set (light green), a sub-development set (green), and a sub-test set (dark green).

Through the completion of this personalization or fine-tuning step, the General Model is transformed into several Personal Models tailored to individual characteristics. The final evaluation of the efficacy of these Personal Models is then carried out through the sub-test set.

In summary, our approach encompasses initial model training on a comprehensive scale of data without considering individual traits. Following this, a personalization step is performed that considers individual characteristics, thereby converting the General Model into different Personal Models.

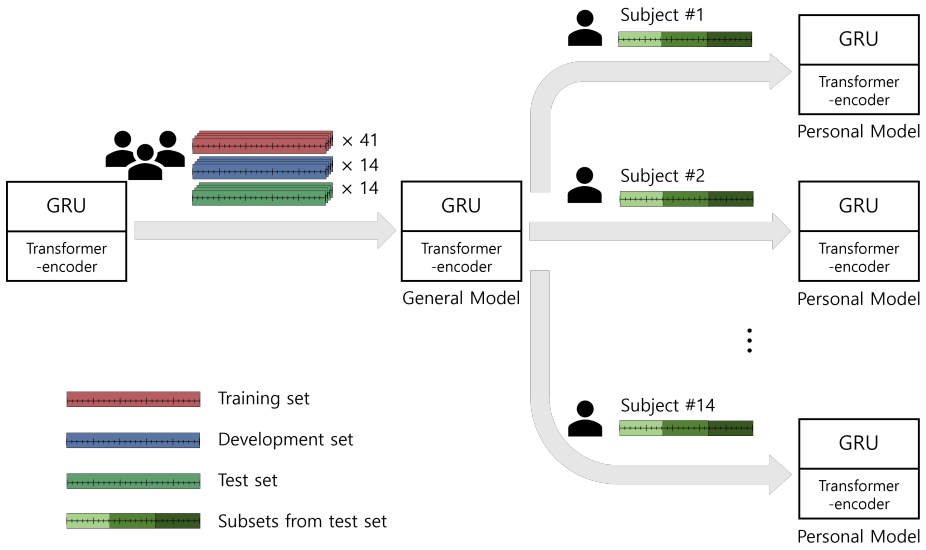


Figure 2: Overview of our approach towards MuSe-Personalization. The GRU and Transformer-encoder represent the models utilized. Training and evaluation are conducted using a dataset comprising recordings from 69 subjects. The training set (red) and development set (blue) are utilized for General Model training, followed by a fine-tuning step based on test data belonging to an individual subject, partitioned into a sub-training set (light green), a sub-development set (green), and a sub-test set (dark green). This results in the creation of Personal Models, adapted to individual characteristics. The final model efficacy is assessed through the sub-test set.

3.2 Hardware and Software Settings

In this study, we performed training using a distributed network of six workstations. We conducted experiments on both the General Models and the Personal Models, publishing our best models and prediction results [17, 18]. For efficient experiment execution, we used the logging and sweep functions of WandB [4] to train individual models on six workstations, with results reviewable within WandB. We also enhanced the baseline code by adding Transformer-encoder models [22], implemented in PyTorch [23] for our experiments.

For feature extraction, we used various software tools. We employed ffmpeg [26] and OpenPose [5] for pose extraction. Furthermore, we leveraged Wav2Vec2.0 and Data2Vec through the Hugging Face feature extractor [29]. Lastly, we used Pandas [21] and NumPy [14].

3.3 CCC Loss

We trained all our models utilizing a loss function based on CCC. The CCC loss is given by:

$$\mathcal{L} = 1 - \text{CCC} \tag{1}$$

The CCC itself is defined as follows:

$$\text{CCC} = \frac{2\rho\sigma_{\hat{Y}}\sigma_{Y}}{\sigma_{\hat{Y}}^2 + \sigma_{Y}^2 + (\mu_{\hat{Y}} - \mu_{Y})^2} \tag{2}$$

In these equations, $\mathcal{L}$ represents the CCC-based loss function, while ρ is the Pearson correlation coefficient between $\hat{Y}$ (predicted values) and Y (ground truth values). The mean and standard deviation are denoted by μ and σ, respectively. We also use the CCC value as a performance metric for our predictive models.

3.4 Interpretable Meta Learning on Hyperparameters

Table 1 summarizes the hyperparameter settings utilized during the training phase of both the General and Personal Models. The table lists the name of each hyperparameter, the minimum and maximum values explored during tuning, and the total number of configurations tested. We determined the minimum and maximum values for all hyperparameters by referring to the parameters provided in the baseline code and subsequently expanded the search range incrementally. This approach ensured a balance between doing a comprehensive search and computational efficiency.

We conducted as many experiments as possible by combining hyperparameters from Table 1, tailoring them to the model and feature characteristics. We designated the highest CCC score out of 20 seeds as the representative value of each configuration. After training both the General and Personal Models, we derived 2,990 combinations for the General Models and tested 2,051 hyperparameter combinations for the Personal Models. We split these logs into logs for the GRU and the Transformer-encoder and trained a Random Forest Regressor using these hyperparameter combinations as inputs and the corresponding development CCC as outputs. We allocated 80% of the hyperparameter combinations and development CCC pairs for model training and the remaining 20% for model testing. Training included hyperparameter tuning through cross-validation and grid search. We analyzed the obtained meta-learning

Table 1: Hyperparameter configuration for training General and Personal Models

Model type	Hyperparameter name	Min value	Max value	Number of configurations
General	Window length	200	400	3
	Learning rate	0.0001	0.01	4
	Hop length	50	300	3
	Model complexity	2	128	7
	Number of layers	2	16	4
Personal	Window length	2	60	10
	Learning rate	0.0001	0.05	14
	Hop length	2	25	7

model using SHAP, but as this meta-learning model was designed for hyperparameter analysis, we did not conduct a performance evaluation that for instance targeted the prediction of errors.

4 RESULTS

4.1 Development CCC

4.1.1 Baseline Features. Table 2 displays the results of predicting personalized emotional dimensions—arousal and valence—using various feature combinations. Each cell corresponds to the development CCC achieved by a model trained on the feature and emotional dimension denoted by its respective row and column. For example, the development CCC of 0.8999 is achieved by a Personal Model trained on the ViT feature for Arousal. We denote a Transformer-encoder model result in blue and a GRU [6] model result in black. For ease of comparison, we present each result alongside the results from the baseline paper [7], denoting any absence of a baseline result with a hyphen.

Given our outcomes, all features, except for eGeMAPS, surpassed the CCC of the baseline paper for Arousal in the unimodal setting. Moreover, for Valence, all features consistently outperformed the baseline CCC, exhibiting improvements between 0.24 to 0.66. Consequently, the combined CCC exceeded the baseline across all features. For Arousal predictions, the GRU model outperformed the Transformer-encoder model for 17 out of 24 unimodal features. However, for Valence, barring FAU [11], the Transformer-encoder achieved the highest CCC across all features. In comparison to the effectiveness that was obtained in [22] for MuSe 2022 [8], where the Transformer-encoder did not show significant strength in unimodal prediction, our study shows that the Transformer-encoder appears to be particularly advantageous for personalization tasks.

The final three rows of Table 2 show late fusion results. We utilized the late fusion technique from the baseline paper and calculated the fusion results by averaging multiple unimodal predictions, paralleling the method from the baseline paper. For MuSe-Personalization, only the A+V fusion results were disclosed, representing the mean value of the unimodal prediction results for Audio (A) and Video (V), and specifically, the average of DeepSpectrum [1], eGeMAPS [13], Wav2Vec2.0 [3], ViT [10], FaceNet [24], and FAU [11]. The development CCC for Arousal in the baseline paper was given by 0.9145, and the development CCC for Valence by 0.8559. We generated results by separately fusing A and V, in addition to A+V, and these results are displayed in the lower rows of Table 2. Ultimately, our A+V fusion outcome registered a combined CCC that was 0.0779 higher than the combined CCC of the baseline paper. Furthermore, both our Audio fusion (A) and Video fusion (V) results outperformed the baseline late fusion (A+V).

4.1.2 Additional Features. Based on the detailed results that can be found in Table 2, Figure 3 presents a visual comparison between the combined development CCC for (a) the newly introduced Pose features and (b) the SSL-based features.

In terms of Pose features, "Original" refers to the feature that was extracted using the method that was proposed earlier in [22], calculating temporal changes in joint movements. Conversely, the rest embody features extracted through JCD. Our experimental results show that JCD-both outperforms Original, with an increase of approximately 0.032 in CCC. Despite exhibiting a lower CCC relative to the other video features, JCD-both demonstrates considerable enhancement over Original. This implies the potential of these novel features in improving emotional dimension predictions.

An interesting pattern emerges within the SSL-based features. We employed Wav2Vec2.0 [3] and Data2Vec [2] to extract audio features and context features, respectively, and utilized these to train our model for emotional dimension predictions. All the features, excluding D2V-audio, exhibited superior CCC scores compared to W2V-original. Interestingly, context features consistently produced higher combined development CCC scores than their audio counterparts. On average, Wav2Vec2.0, designed specifically for audio, produced higher CCC scores than Data2Vec. Among all, W2V-context achieved a CCC of 0.927, doing better than all other unimodal features in terms of development CCC.

4.2 Test CCC

The competition permitted us to make a total of five submissions, thereby enabling us to experiment with different late fusion feature combinations. Guided by the insights derived from the findings delineated in Section 4.1, we opted for Personal Models that exhibited superior performance in terms of development CCC to undertake testing. The selected features and their respective development CCC scores are listed below, with the model used (TF denotes Transformer-encoder) and the corresponding development CCC indicated within parentheses.

For arousal:

- Audio (A): DeepSpectrum (TF, 0.9376), eGeMAPS (GRU, 0.8783), W2V-context (TF, 0.9287)
- Video (V): ViT (GRU, 0.8999), FaceNet (GRU, 0.8766), FAU (GRU, 0.9378)

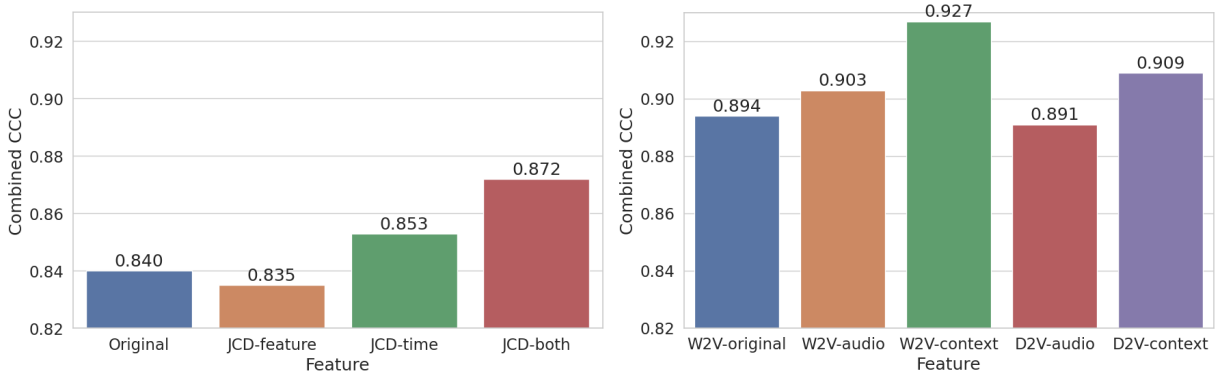
For valence:

- Audio (A): DeepSpectrum (TF, 0.9059), eGeMAPS (TF, 0.9296), W2V-context (TF, 0.9258)
- Video (V): ViT (TF, 0.8947), FaceNet (TF, 0.8939), FAU (GRU, 0.8124)

A comprehensive summary of the test CCC scores can be found in Table 3. Here, 'A' represents audio (acoustic) features, 'V' signifies video features, 'Base' denotes the baseline CCC scores, and 'Ours' refers to the CCC scores obtained for our optimal model. The table also presents those feature combinations for which results could not

Table 2: Summary of the development CCC scores obtained by Personal Models. In the "Type" column, "S" stands for Biosignal (electrocardiogram, respiration, and heart rate), "A" stands for audio (acoustic), and "V" stands for video. "Dim" represents the dimension of an input feature. "Baseline CCC" refers to baseline paper results, while "Our best CCC" refers to the best results we obtained for both GRU and Transformer-encoder. We used blue to indicate when the best result is obtained for the Transformer-encoder. A hyphen ("-") is inserted if there are no baseline results. If both "Baseline CCC" and "Our best CCC" have a numerical value, "Our best CCC" is the result of testing more hyperparameter combinations with the same features in our environment. If a feature has a higher value than the baseline, it is highlighted by putting it in bold. W2V, D2V, and Pose refer to newly added features; W2V represents Wav2Vec2.0 and D2V represents Data2Vec. To provide a fair comparison with the baseline, Fusion does not include W2V, D2V, and Pose.

Feature info			Arousal		Valence		Combined	
Feature name	Type	Dim	Baseline CCC	Our best CCC	Baseline CCC	Our best CCC	Baseline CCC	Our best CCC
Biosignal	S	3	-	0.8716	-	0.6651	-	0.7684
DeepSpectrum	A	1024	0.8064	**0.9376**	0.3536	**0.9059**	0.5800	**0.9218**
eGeMAPS		78	0.9073	0.8783	0.5892	**0.9296**	0.7483	**0.9040**
Wav2Vec2.0		1024	0.7421	**0.8775**	0.5142	**0.9096**	0.6282	**0.8936**
ViT	V	384	0.2691	**0.8999**	0.6050	**0.8947**	0.4371	**0.8973**
FaceNet		512	0.8260	**0.8766**	0.6491	**0.8936**	0.7376	**0.8851**
FAU		20	0.6382	**0.9378**	0.1468	**0.8124**	0.3925	**0.8751**
W2V-audio	A	512	-	0.9336	-	0.8718	-	0.9027
W2V-context		768	-	0.9287	-	0.9258	-	0.9273
D2V-audio		512	-	0.9110	-	0.8713	-	0.8912
D2V-context		768	-	0.9186	-	0.9001	-	0.9093
Pose, original	V	26	-	0.8022	-	0.8775	-	0.8399
Pose, JCD-feature		105	-	0.8684	-	0.8017	-	0.8351
Pose, JCD-time		105	-	0.8884	-	0.8180	-	0.8532
Pose, JCD-both		105	-	0.8649	-	0.8649	-	0.8718
A	Fusion	3	-	0.9577	-	0.9590	-	0.9584
V		3	-	0.9478	-	0.9373	-	0.9426
A+V		6	0.9145	**0.9625**	0.8559	**0.9636**	0.8852	**0.9631**



(a) Combined CCC on Pose Features

(b) Combined CCC on SSL-based Features

Figure 3: Performance comparison between the newly added features.

be procured, symbolized by a hyphen ("-"). Results that surpassed the baseline numbers are highlighted in bold.

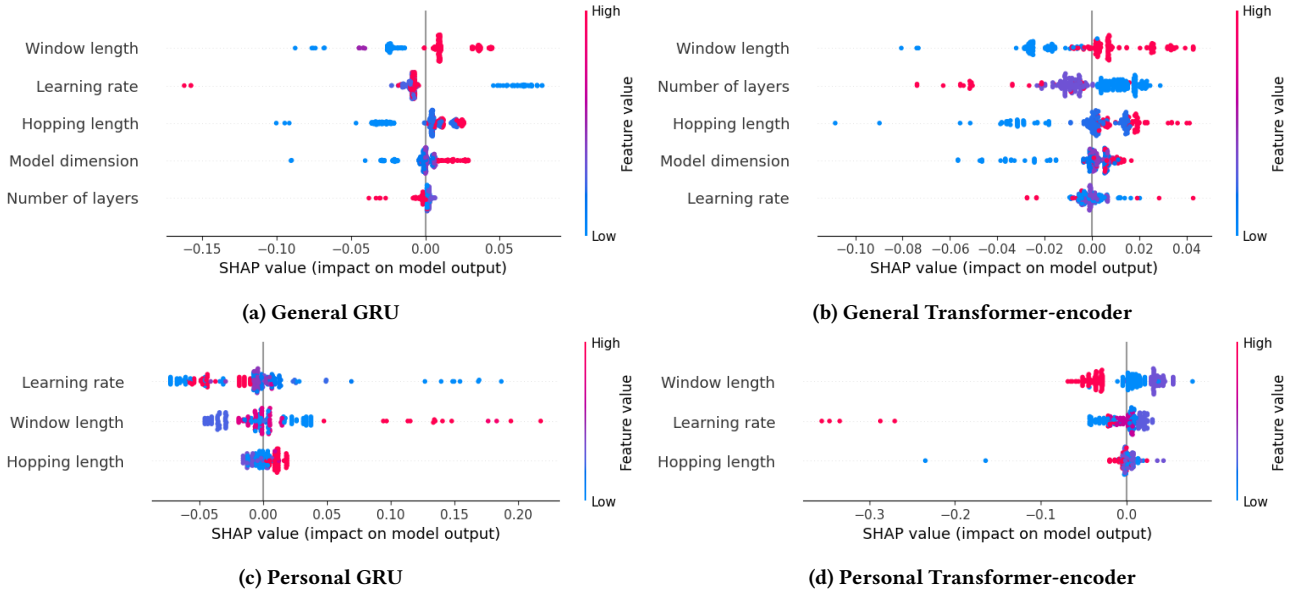
Within the constraints of the five submissions allowed, we aimed at finding the optimal test CCC scores by adding or removing the necessary features for fusion based on the development CCC scores obtained. Ultimately, we achieved the following results, leading to a second place in MuSe-Personalization: Arousal: 0.8262, Valence: 0.8892, and Combined: 0.8577.

4.3 Interpretable Meta Learning on Hyperparameters

The impact of the hyperparameters, as discerned by our meta-learning models, is visualized in the four beeswarm plots shown in Figure 4. Each plot organizes the parameters in a descending order of importance, from top to bottom. For instance, in the GRU-based General Model depicted in Figure 4(a), the Window length ranks as the most critical hyperparameter, while the Number of layers

Table 3: Summary of test CCC scores. The "Features" column enumerates the late fusion combinations we submitted, organized in the order of Arousal and Valence. In this context, the "+" sign denotes the features that were added, while "-" indicates the ones removed. The label "normalize" refers to an option that can be found in the baseline code. We made a total of five submissions, and our final results are as follows, securing us a second place in the MuSe-Personalization 2023 challenge: Arousal: 0.8262, Valence: 0.8892, and Combined: 0.8577.

Features	Arousal [CCC]		Valence [CCC]		Combined [CCC]	
	Base	Ours	Base	Ours	Base	Ours
A+V, A+V	0.7450	**0.8262**	0.7827	0.8844	0.7639	0.8553
A+V normalize, A+V-FAU	-	0.7875	-	**0.8892**	-	0.8384
A+V-FaceNet-eGeMAPS, A	-	0.8046	-	0.8434	-	0.8240
A+V+FAU+DeepSpectrum+W2V-context, A+V-FAU+eGeMAP+W2V-context	-	0.8258	-	0.8847	-	0.8553
A+V, A+V-FAU	-	**0.8262**	-	**0.8892**	-	**0.8577**



(a) General GRU

(b) General Transformer-encoder

(c) Personal GRU

(d) Personal Transformer-encoder

Figure 4: Beeswarm plots depicting the influence of various hyperparameters on the four meta-learning models, as interpreted by SHAP. For the General Models, the Window length is the most influential hyperparameter, with an evident trend of larger Window lengths correlating positively with the development CCC. For the Personal Models, the GRU and Transformer-encoder show differing crucial factors: the Learning rate and the Window length, respectively. Notably, the Transformer-encoder demonstrates a reverse trend, with increased values leading to a decrease in the development CCC.

holds the least importance. The color-coded points on the plot, with darker reds and blues signifying higher and lower hyperparameter values, respectively, provide further insight. Points plotted left of 0.00 on the x-axis indicate a decrease in the development CCC, while those plotted on the right suggest an improvement in outcomes.

In the case of the GRU-based General Model (see Figure 4(a)), it is evident that a larger Window length tends to yield better results. The influence of the Learning rate, however, is less clear-cut. A similar pattern regarding Window length is noticeable in the Transformer-encoder-based General Model (see Figure 4(b)).

In contrast, the Personal Models display a divergence in the importance of hyperparameters for the GRU and Transformer-encoder. The Learning rate assumes critical importance for the GRU, while, akin to the General Model, the Window length emerges as the key determinant for the Transformer-encoder. A noteworthy trend is observable in the Transformer-encoder-based Personal Models: an increase in the Window length value inversely affects the development CCC.

5 DISCUSSION

We can categorize the findings derived from our experimental results, and their implications, into three major items:

- **Re-discovery of the Transformer-encoder:** Our experimental results show that the Transformer-encoder architecture excels in personalization tasks, particularly in Valence predictions. With the exception of the FAU feature, the model achieved the highest development CCC scores across all features. We attribute the Transformer-encoder's success in personalization to its ability to capture long-range

dependencies using its attention mechanism. However, long-range dependencies may not fully explain this success. Our analysis highlights the importance of the window length hyperparameter for the Transformer-encoder's performance in Personal Models. Notably, for the Personal Transformer-encoder, longer window lengths reduced effectiveness, as measured by development CCC. This suggests an optimal window length for capturing personalization features, beyond which extraneous information may degrade effectiveness.

- **Benefits of More Hyperparameter Tuning:** We undertook efforts towards optimizing a wider selection of hyperparameters compared to the baseline paper. As a result, we managed to surpass the baseline development CCC in all unimodal predictions, except for the Arousal-eGeMAPS pair. The significance of hyperparameters, as revealed through meta-learning, pointed to learning rate and window length as particularly crucial factors. Moreover, an increase in window length in the personalized Transformer-encoder was found to negatively impact the development CCC. We were unable to directly use the results of meta-learning for our hyperparameter optimization. Nonetheless, we hope that sharing the knowledge gained from this study will benefit future researchers working in this field.
- **Potential of Newly Crafted Features:** Our newly introduced features, such as the Pose features extracted through JCD and the different SSL-based features (Wav2Vec2.0 [3] and Data2Vec [2]), showed considerable promise in improving emotional dimension predictions. For instance, our JCD-based features demonstrated a notable enhancement over the original Pose feature [22], whereas the SSL-based features, particularly context-based ones, consistently scored higher CCC scores compared to their audio counterparts.

Furthermore, our study encountered the following three limitations:

- **Absence of a Novel Emotional Detection Model:** We focused on refining and effectively using the Transformer-encoder model that we previously adopted for MuSe-Stress 2022 [22], rather than introducing an entirely new model structure. Future research may explore the potential for a novel model to further improve personalized stress detection.
- **Questions on the Usability of Pose Features:** Although JCD-both achieved a combined CCC of 0.872, exceeding the combined CCC of 0.840 obtained by the original Pose feature [22], it fell slightly short of the combined CCC obtained by the other six audio and video features. Consequently, during testing, we excluded Pose from the fusion step due to a limitation on the number of test submissions.
- **Interpretability at the Model Level:** While we explored a variety of hyperparameters and searched for the combination that best suited our data, including conducting a thorough analysis of the hyperparameters, this can be regarded as an indirect interpretation compared to an interpretation of the model. A method for a more concentrated model interpretation seems necessary.

6 CONCLUSIONS

The research efforts presented in this paper underscore the importance of hyperparameter tuning, the potential of new feature engineering, and the effectiveness of the Transformer-encoder model in emotion prediction tasks. This is reflected by our test set evaluation results, leading to a provisional second place in the MuSe-Personalization 2023 challenge with an Arousal score of 0.8262, a Valence score of 0.8844, and a combined score of 0.8553. However, our research efforts also demonstrate the need for a greater focus on model interpretability and the development of entirely new model structures. In future research, we therefore plan to keep working on more comprehensive and robust solutions for emotion prediction tasks, further building on our current achievements. We also plan to investigate the generalizability of our newly engineered pose features by testing them across different use cases that involve stress detection (e.g., driver behavior monitoring).

ACKNOWLEDGMENTS

This research effort was supported by the National Research Foundation (NRF) Korea (NRF-2020K1A3A1A68093469), funded by the Ministry of Science and ICT (MSIT) Korea, and by the Department of Biotechnology (India) (DBT/IC-12031(22)-ICD-DBT). This research effort was also supported by Ghent University Global Campus (GUGC) in Korea.

REFERENCES

[1] Shahin Amiriparian, Nicholas Cummins, Sandra Ottl, Maurice Gerczuk, and Björn Schuller. 2017. Sentiment analysis using image-based deep spectrum features. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 26–29. https://doi.org/10.1109/ACIIW.2017.8272618

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 1298–1312. https://proceedings.mlr.press/v162/baevski22a.html

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[4] Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. https://www.wandb.com/ Software available from wandb.com.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[6] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR* abs/1406.1078 (2014). arXiv:1406.1078 http://arxiv.org/abs/1406.1078

[7] Lukas Christ, Shahin Amiriparian, Alice Baird, Alexander Kathan, Niklas Müller, Steffen Klug, Chris Gagne, Panagiotis Tzirakis, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2023. The MuSe 2023 Multimodal Sentiment Analysis Challenge: Mimicked Emotions, Cross-Cultural Humour, and Personalisation. arXiv:2305.03369 [cs.LG]

[8] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2022. The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge* (Lisboa, Portugal) *(MuSe' 22)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3551876.3554817

[9] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 244–250. https://doi.org/10.1109/ASRU51503.2021.9688253

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[11] Paul Ekman and Wallace V Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

[12] Paul Ekman, Robert W Levenson, and Wallace V. Friesen. 1983. Autonomic Nervous System Activity Distinguishes Among Emotions. *Science* 221, 4616 (1983), 1208–1210. https://doi.org/10.1126/science.6612338 arXiv:https://www.science.org/doi/pdf/10.1126/science.6612338

[13] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[14] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[15] Reza Lotfian and Carlos Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Transactions on Affective Computing* 10, 4 (2019), 471–483. https://doi.org/10.1109/TAFFC.2017.2736999

[16] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[17] Ho min Park, Ganghyun Kim, Arnout Van Messem, and Wesley De Neve. 2023. best general models. (8 2023). https://doi.org/10.6084/m9.figshare.23798262.v2

[18] Ho min Park, Ganghyun Kim, Arnout Van Messem, and Wesley De Neve. 2023. prediction results for fusion. (8 2023). https://doi.org/10.6084/m9.figshare.23798256.v1

[19] Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Ho min Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. 2023. Know Your Self-supervised Learning: A Survey on Image-based Generative and Discriminative Training. arXiv:2305.13689 [cs.CV]

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[21] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.3509134

[22] Ho-min Park, Ilho Yun, Ajit Kumar, Ankit Kumar Singh, Bong Jun Choi, Dhananjay Singh, and Wesley De Neve. 2022. Towards Multimodal Prediction of Time-Continuous Emotion Using Pose Feature Engineering and a Transformer Encoder. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge* (Lisboa, Portugal) *(MuSe' 22)*. Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/3551876.3554807

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W Schuller. 2021. The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. 5–14.

[26] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux Journal* 2006, 146 (2006), 10.

[27] Joaquin Vanschoren. 2018. Meta-Learning: A Survey. *CoRR* abs/1810.03548 (2018). arXiv:1810.03548 http://arxiv.org/abs/1810.03548

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[30] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. 2020. Make Skeleton-Based Action Recognition Model Smaller, Faster and Better. In *Proceedings of the ACM Multimedia Asia* (Beijing, China) *(MMAsia '19)*. Association for Computing Machinery, New York, NY, USA, Article 31, 6 pages. https://doi.org/10.1145/3338533.3366569