Adaptive Multimodal Fusion for Facial Action Units Recognition

Huiyuan Yang, Taoyue Wang and Lijun Yin
{hyang51,twang61}@binghamton.edu,lijun@cs.binghamton.edu
Department of Computer Science, State University of New York at Binghamton
Binghamton, NY, USA

ABSTRACT

Multimodal facial action units (AU) recognition aims to build models that are capable of processing, correlating, and integrating information from multiple modalities (*i.e., 2D images from a visual sensor, 3D geometry from 3D imaging, and thermal images from an infrared sensor*). Although the multimodel data can provide rich information, there are two challenges that have to be addressed when learning from multimodal data: 1) the model must capture the complex cross-modal interactions in order to utilize the additional and mutual information effectively; 2) the model must be robust enough in the circumstance of unexpected data corruptions during testing, in case of a certain modality missing or being noisy. In this paper, we propose a novel **A**daptive **M**ultimodal **F**usion method (**AMF**) for AU detection, which learns to select the most relevant feature representations from different modalities by a re-sampling procedure conditioned on a feature scoring module. The feature scoring module is designed to allow for evaluating the quality of features learned from multiple modalities. As a result, AMF is able to adaptively select more discriminative features, thus increasing the robustness to missing or corrupted modalities. In addition, to alleviate the over-fitting problem and make the model generalize better on the testing data, a cut-switch multimodal data augmentation method is designed, by which a random block is cut and switched across multiple modalities. We have conducted a thorough investigation on two public multimodal AU datasets, BP4D and BP4D+, and the results demonstrate the effectiveness of the proposed method. Ablation studies on various circumstances also show that our method remains robust to missing or noisy modalities during tests.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Biometrics*; *Image representations.*

KEYWORDS

AU; Facial Action Units; Multi-modalities; Multimodal fusion.

ACM Reference Format:

Huiyuan Yang, Taoyue Wang and Lijun Yin. 2020. Adaptive Multimodal Fusion for Facial Action Units Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413538

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413538

1 INTRODUCTION

Facial action unit (AU) detection has been an essential task for human emotion analysis. Conventionally, most state-of-the-art AU detection methods exploit images collected from the visible-spectrum based RGB cameras [50][24][31][29][32][39][3]. However, as AU analysis relies on the detection of subtle facial muscle movement, the visual-only detection methods have found to be insufficient for detecting subtle changes from the single modality. Recent advancements in multimodal sensor development present a promise in study of AU detection through multiple modalities. For example, the public database BP4D+ provides a set of synchronized data with multiple modalities, i.e., *2D visual*, *3D depth* and *thermal* modalities [49], allowing us to investigate various features from different modalities for AU detection, *i.e.*,, AU6 (*Cheek Raiser*), involves the deformation of *Orbicularis oculi* and *pars orbitalis* muscles in the cheek area, which only show subtle changes in visual images, while a better geometric changes can usually be observed in depth images. Similarly, *microcirculation* and *blood flow* may vary along the contraction or relaxation of certain muscles, which results in the change of skin surface temperature.

Recently, there has been an advancement by extending machine learning methods to learn additional information presented in the data from multiple modalities. For example, Li et al. [21][22] combined the 2D and 3D feature for facial expression recognition. Irani et al. [16] utilized the visual, depth and thermal modalities for pain study. Lakshminarayana et al. [19] explored physiological signals in combination with visual images to predict action units. Although the presence of multiple modalities provides additional valuable information, challenges still remain when learning features from multiple modalities [35][28][44], which requires 1) the models must capture the complex cross-modal interactions in order to utilize the additional and mutual information effectively; 2) the models must be robust to unexpected data corruption, such as in the presence of missing and noisy modalities during testing. In this paper, we propose a novel **A**daptive **M**ultimodal **F**usion method (**AMF**) for AU detection. First, a feature scoring module is designed for evaluation of the features learned from multiple modalities, and then a sampling based feature selection process is conditioned on the feature scores. As a result, our model learns to select the most relevant feature representations from different modalities, while avoiding useless or misleading information. More importantly, our model is able to learn to rely on the most discriminative features from individual modality adaptively, making it robust to various imaging conditions, especially in the case of missing or corrupted modalities during testing. Built upon the selective and adaptive feature

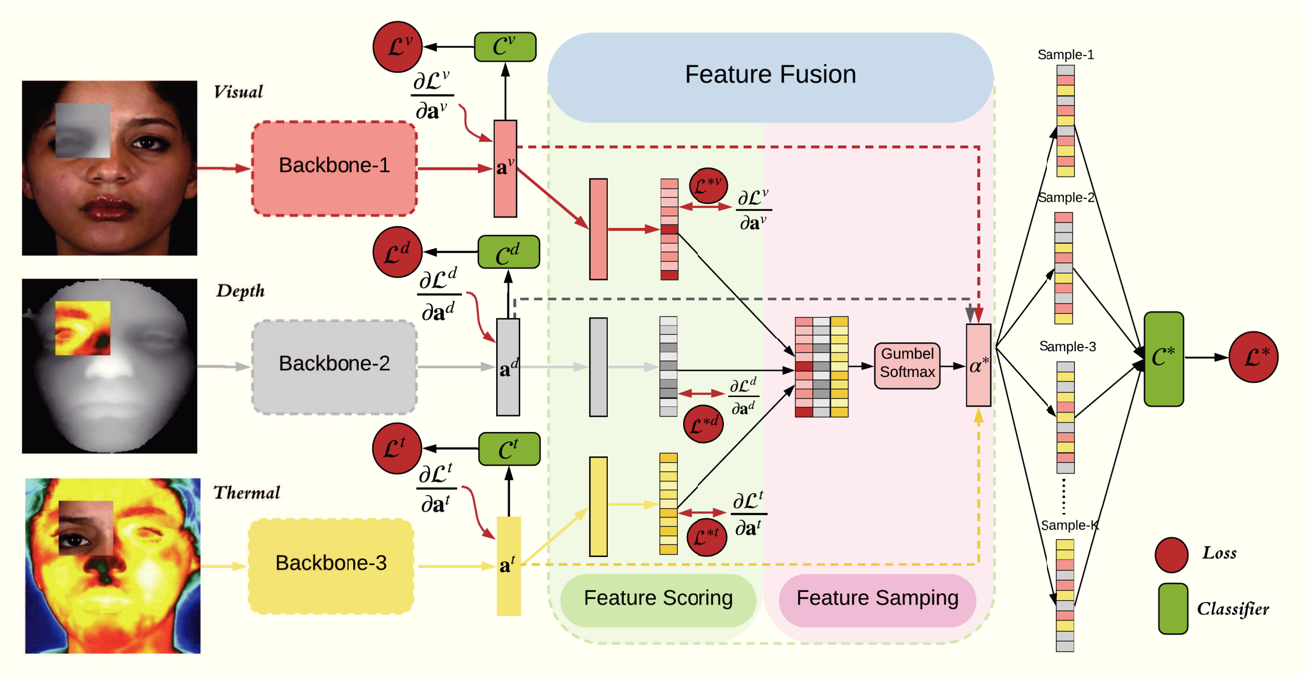

Figure 1: Framework of the proposed adaptive multimodal fusion model (AMF) with three modalities (i.e., visual, depth, and thermal). AMF learns to select the most relevant feature representations from different modalities by the Gumbel-Softmax resampling trick conditioned on a feature scoring module. The feature scoring module is designed to allow for evaluating the quality of features learned from multiple modalities. As a result, AMF is able to automatically select the more discriminative features, and robust to missing or corrupted modalities. To alleviate the over-fitting issue and let the model generalize better on the testing data, a cut-switch multimodal data augmentation method is also applied, in which a random block is cut and switched across modalities.

fusion strategy, we further propose a cut-switch multimodal data augmentation method by randomly cutting and switching a block across modalities. By doing so, we can alleviate the over-fitting problem to a certain degree, and make the model generalize better on the testing data. We have conducted a thorough evaluation on two public datasets (BP4D and BP4D+), and scrutinized the performances with respect to various combinations of multiple modalities, cut-switch strategy, and different levels of noises, demonstrating the effectiveness and robustness of the proposed AMF method.

The contributions of this work are listed in the following threefold:

- We present an adaptive multimodal fusion method for facial action unit detection, which is able to effectively select features from multiple modalities, enabling more accurate and robust AU detection.
- We propose a cut-switch multimodal data augmentation method, which has been proved to be an effective way of improving performance.
- Extensive experiments are conducted to evaluate the performance of multimodal based AU detection, showing the advantage of the proposed AMF method and its robustness to missing or corrupted data.

2 RELATED WORKS

2.1 Action Unit Detection

In recent years, deep features of 2D visible images have been widely used for AU detection. Deep learning approach developed by Gudi et al. [11] is one of the pioneer works in AU detection, which demonstrated impressive performance on both AU occurrence detection and intensity classification tasks. Zhao et al. [50] proposed a network called DRML which applied a region layer to capture local structural information on different facial regions. Li [23–25] defined several regions of interest (RoI) around AU-related facial landmarks to enhance the feature map intensities at different levels. Furthermore, those works [23–25] cropped the trained features maps into 3×3 and learned a separated set of features through fully connected layers. In order to leverage the temporal information, Chu et al. [6] and and Li et al. [23] aggregated CNN output into Long Short-Term Memory (LSTM) for AU predictions, while Yang et al. [42] proposed to learn the temporal information from static image. Shao et al. [31, 32] gave insight into the spatial attention mechanism which applied the multi-scale region learning to extract the AU related local features. Most recently, Niu et al. [29] tried to capture the local information and the relationship of individual local face regions, aiming to improve the AU detection robustness.

2.2 Multimodal Machine Learning

Multimodal machine learning aims to build models that can process, correlate, and integrate information from multiple modalities [2]. The success of multimodal machine learning has been demonstrated in a wide range of applications, e,g, human action analysis [1, 4, 37, 38] , person/object localization and tracking [15, 34, 47] and image segmentation [14, 51].

In the field of emotion related tasks such as action unit detection and facial expression recognition, we have seen a trend of extending machine learning methods to learn additional information presented in the multiple modalities. Li et al. [21][22] applied 2D + 3D feature-based approaches for facial expression recognition. Zhang et al. [46] combined 2D texture images with facial landmarks for expression recognition. Wu et al. [40] proposed a novel deep two-view approach to learn features from both texture and thermal images and adopted the commonality in between for expression recognition. Irani et al. [16] applied RGB-Thermal-Depth images for pain estimation. Lakshminarayana et al. [19] conducted a exploratory work by combining physiological signals with color images to predict action units. Liu et al. [26] proposed a thermal empowered multi-task network for facial action unit detection, which made a good use of the strength and correlation of visual and thermal modalities and achieved a good performance in AU detection.

One of the key steps in multimodal machine learning is the multimodal fusion, with the aim at integrating features of multiple modalities for enabling more accurate and robust performance. Three types of fusion strategies (i.e., early, late, and hybrid fusion) are the commonly used techniques for multimodal feature fusion [2][12][5]. Our proposed adaptive feature fusion strategy is particularly related to late fusion with a focus on the selection mechanism in order to choose the most relevant feature representations from different modalities, meanwhile it can avoid useless or misleading information. Consequently, our model remains fully robust to missing or corrupted modalities during testing.

3 PROPOSED METHOD

In this section, we describe our approach of the selective feature fusion across different modalities.

3.1 Problem Formulation and Notation

A multimodal dataset consists of N labeled frames defined as $\mathbf{X} = (\mathbf{X}^v, \mathbf{X}^d, \mathbf{X}^t)$ for *visual*, *depth* and *thermal* modalities respectively. The dataset is indexed by N such that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N)$ where $\mathbf{X}_i = (\mathbf{X}_i^v, \mathbf{X}_i^d, \mathbf{X}_i^t)$, $1 \leq i \leq N$. The corresponding labels for these N frames are denoted as $y = (y_1, y_2, ..., y_N)$, $y_i \in \{0, 1\}^C$, where C is the number of AUs.

3.2 Multimodal Fusion

High-level features are extracted by an individual backbone network (i.e. ResNet-18 [13]) $\mathbf{f}$ with parameter θ, represented as $\mathbf{a} = (\mathbf{a}^v, \mathbf{a}^d, \mathbf{a}^t)$, $\mathbf{a}^v, \mathbf{a}^d, \mathbf{a}^t \in \mathbb{R}^D$, where D is the dimension of feature:

$$\mathbf{a}^k = \mathbf{f}(\mathbf{X}^k; \theta_1^k); k \in \{v, d, t\} \tag{1}$$

The straight-forward approach to fuse multimodal data is to combine them at the input or feature level, namely early-fusion or late-fusion respectively. However, they are not optimal for multimodal data. First, the model cannot capture the complex interaction among modalities. Second, the model is sensitive to missing or noisy input by considering different modalities equally.

Intuitively, the features from individual modality offer different strengths for the task of AU recognition; more importantly, collecting data from multiple sensors inevitably increases the chance of having missing or corrupted modalities. Therefore, it is desirable to design a mechanism to adaptively fuse the features based on the condition of modalities.

Feature scoring is designed to evaluate the discriminability of the extracted features. Similar to the widely applied attention mechanism [36][41], this function learns to evaluate each feature conditioned on the extracted features, thus allowing the feature scoring function to be jointly trained with other modules.

$$\alpha^k = \mathbf{g}(\mathbf{a}^k; \theta_2^k); k \in \{v, d, t\} \tag{2}$$

where $\alpha^k = [\pi_1^k, \pi_2^k, ..., \pi_D^k], k \in \{v, d, t\}$, and $\pi_i^k \in [0, 1]$ representing the score for individual feature extracted from different modalities. Instead of re-weighting each feature by the corresponding score, we apply a stochastic fusion method[5] to select the feature from different modalities.

Feature Sampling aims to re-sampling a feature index α^* based on the scores across modalities:

$$\begin{aligned} \alpha^* &= Sampling(\alpha^v, \alpha^d, \alpha^t) \\ where, \alpha^* &= [\pi_1^*, \pi_2^*, ..., \pi_D^*], \pi_i^* \in \{0, 1\}^M \end{aligned} \tag{3}$$

where D is the dimension of feature, M is the number of modalities, in our case, $M = 3$ for the *visual*, *depth* and *thermal* modalities. However, the sampling step with discrete variables are difficult to train because the back-propagation algorithm cannot be applied directly to non-differential layers. The reparameterization trick is proposed in VAE [18] to construct a differential unbiased estimator of the lower bound in a model with continuous latent variables, but fails on discrete variables. The Gumbel-Softmax trick [17][27] is a variation of the reparameterization trick, but capable of handing discrete variables. The Gumbel-Softmax trick allows us to draw samples α^* from a categorical distribution efficiently, given the class probabilities α^k and a random variable ϵ^k via:

$$\alpha^* = one\_hot\Big(arg\max_k \big(log(\alpha^k) + \epsilon^k\big)\Big) \tag{4}$$

where $k \in \{v, t, d\}$ is the index of modality. In practice, the random variable ϵ is sampled from a gumbel distribution, which is a continuous distribution on the simplex that can approximate categorical samples:

$$\epsilon = -log(-log(u)), u \sim Uniform(0, 1) \tag{5}$$

However, the *argmax* operation is not differential in Eq.4, hence a softmax function is used as a continuous, differentiable approximation to *argmax*:

$$h^k = \frac{\exp\left(\left(\log\left(\alpha^k\right) + \epsilon^k\right)/\tau\right)}{\sum_{i \in \{v,d,t\}} \exp\left(\left(\log\left(\alpha^i\right) + \epsilon^i\right)/\tau\right)} \tag{6}$$

where $\tau > 0$ is the temperature that modulates the re-sampling process: when the temperature τ approaches 0, samples from the Gumbel-Softmax distribution become one-hot and Gumbel-Softmax distribution becomes identical to the categorical distribution; but when τ approaches to $+\infty$, samples will become uniform distribution [33][17]. Finally, h^k is transformed into index α^* through the one-hot function, which is further used to select features $\mathbf{a}^*$ from different modalities.

3.3 AU Recognition

First of all, we define a cross-entropy loss function for the ground truth y and the prediction $\bar{y}$:

$$\mathcal{L}_{C\mathcal{E}} = -\left[y^T \times log(\bar{y}) + (1-y)^T \times log(1-\bar{y})\right] \tag{7}$$

For the labeled training data $(\mathbf{X}_i^v, \mathbf{X}_i^d, \mathbf{X}_i^t, y_i)$, we have three classifiers C^v, C^d and C^t to map individual feature into predictions, thus the supervised loss for each modality is represented as $\mathcal{L}^v, \mathcal{L}^d, \mathcal{L}^t$:

$$\mathcal{L}^k = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{C\mathcal{E}}\left(C^k(\mathbf{X}_i^k), y_i\right), k \in \{v, d, t\} \tag{8}$$

K features are constructed by running the Gumbel-Softmax resampling procedure K times, and those **K** features are mapped into prediction by classifier C^*. An average voting strategy is applied for the final prediction, and the loss function is defined as follow:

$$\mathcal{L}^* = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{C\mathcal{E}}\left(\frac{1}{\mathbf{K}}\sum_{j=1}^{\mathbf{K}}C^*(\mathbf{a}_{i,j}^*), y_i\right) \tag{9}$$

3.4 Reverse gradient guided feature scoring

Ideally, the feature scoring function $g(.)$ should be able to automatically learn to evaluate the quality of features $(\mathbf{a}^v, \mathbf{a}^d, \mathbf{a}^t)$ through training. However, it is not guaranteed to realize it in practice. Therefore, it is desirable to design an extra constraint that encourages the feature scoring function $g(.)$ to fulfil its goal. The idea of gradient reversal layer was first proposed by Ganin et al.[9] for unsupervised domain adaptation through adversarial training. In our work, we follow a similar idea but simplify it as a reverse gradient guidance, which is defined as:

$$\mathcal{L}^{*k} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - dist\left(\alpha^k,\ \left\|\frac{\partial\mathcal{L}^k}{\partial\mathbf{a}^k}\right\|\right)\right), k \in \{v, d, t\} \tag{10}$$

Where $\frac{\partial\mathcal{L}^k}{\partial\mathbf{a}^k}$ is the gradient of loss function $\mathcal{L}^k$ regarding to the latent feature $\mathbf{a}^k$, *dist* is a distance function. We use *Cosine* as the *dist* function in our experiments.

3.5 Cut-Switch for data augmentation

With limited training data, deep model is prone to overfitting, especially in our case of AU recognition, where the frames were collected from a small number of subjects with limited variations. Therefore, an effective data augmentation method is desirable to alleviate the overfitting issue.

Data augmentation methods, such as CutOut [8], MixUp [45] and the recent CutMix [43], have been proposed and demonstrated an effective way of alleviating the overfitting issue. However, a patch is removed in CutOut method, which leads to information loss and inefficiency during training. Both MixUp and CutMix rely on the proportionally mixed ground truth labels and areas, which make them inapplicable to the multi-label AU recognition task.

We propose a simple but effective cut-switch multimodal data augmentation method, as shown in Fig.2, where three patches are cropped and randomly switched among three modalities based on a randomly sampled box. The benefits of cut-switch is two-fold: First, as compared to CutOut [8], our method can augment training data without information loss by cutting and switching blocks at the aligned face area. Second, without relying on mixed labels, our cut-switch data augmentation method can be applied to the multi-label task, while still maintains the benefits of mixing area as used in MixUp [45] and CutMix [43]. To our knowledge, this is the first work for multimodal data augmentation, and experimental results have shown the effectiveness of such a cut-switch strategy.

3.6 Full objective of the networks

Combining the aforementioned objectives, our overall full objective for training the network corresponding to the *visual, depth and thermal* modalities is defined as follows:

$$\mathcal{L} = \lambda_*\mathcal{L}^* + \sum_{k\in\{v,d,t\}}\left(\lambda_k\mathcal{L}^k + \lambda_k^*\mathcal{L}^{*k}\right) \tag{11}$$

where λ are positive regularization parameters.

4 EXPERIMENTS

In this section, we evaluate the proposed method in terms of its capability to improve multi-modal fusion as well as its robustness for missing or noisy inputs.

4.1 Datasets

BP4D [48] is a widely used dataset for evaluating AU detection performance. The dataset contains 328 2D and 3D videos collected from 41 subjects (23 females and 18 males) under eight different tasks. As mentioned in the dataset, the most expressive 500 frames (around 20 seconds) are manually selected and labeled for AU occurrence from each one-minute long sequence, resulting in a dataset of around 140,000 AU-coded frames. For a fair comparison with the state-of-the-art methods, a three-fold subject-exclusive cross validation is performed on 12 AUs.

BP4D+ [49] is a multimodal spontaneous emotion dataset, where high-resolution 3D dynamic models, high-resolution 2D videos, thermal (infrared) images, and physiological data were acquired from 140 subjects. There are 58 males and 82 females, with ages ranging from 18 to 66 years old. Each subject experienced 10 tasks corresponding to 10 different emotion categories, and the most facially-expressive 20 seconds from four tasks were AU-coded from all 140 subjects, resulting in 192,000 AU-coded frames. Following a similar setting in BP4D dataset, 12 AUs are selected and the performance of three-fold cross-validation is reported.

4.2 Implementation details and evaluation metrics

In our experiments, we use two modalities in the BP4D dataset: *2D visual image* and *3D face model*; and three modalities in the BP4D+ dataset: *2D visual image*, *3D face model* and *thermal image*. Face areas are cropped from the *visual* and *thermal* modalities using a

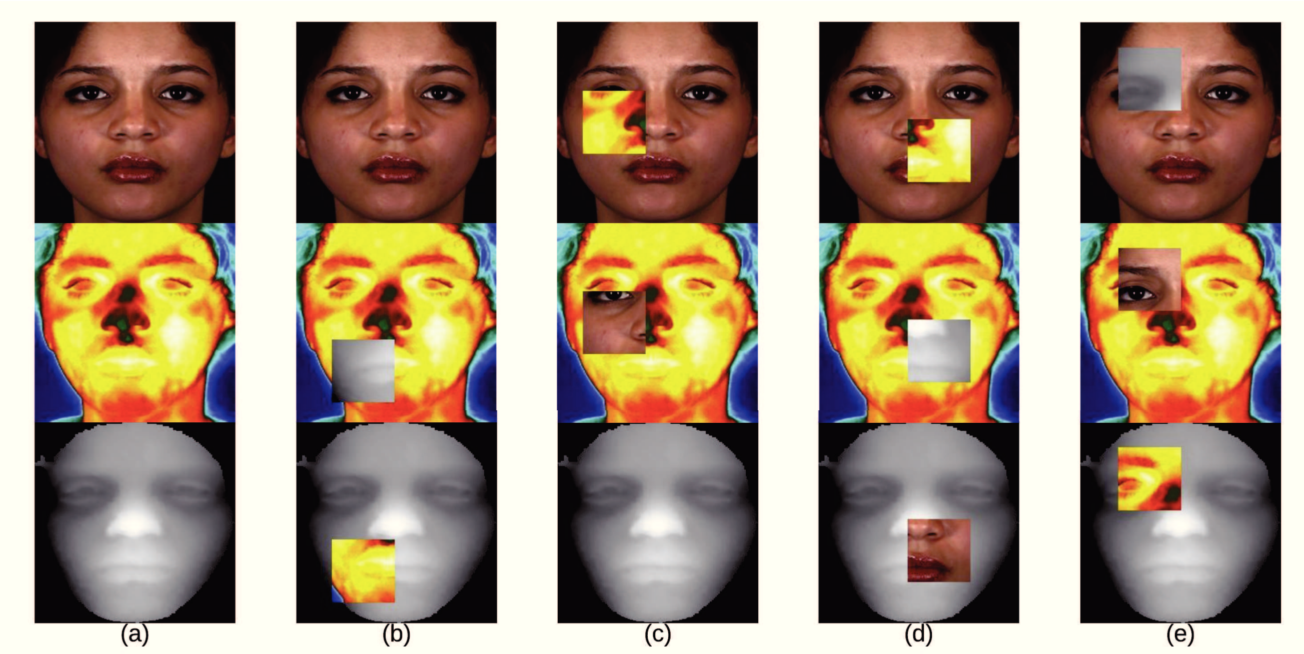

Figure 2: Examples of the Cut-Switch data augmentation method. (a) is the original input multi-modal pairs (X^v, X^t, X^d); (b~e) are the potential examples after data augmentation.

tracking algorithm [10] provided in OpenCV. For 3D face model, we first crop the ROIs of 3D meshes and then project the meshes into depth maps. All the face images are further aligned and cropped to the size of 256×256, and then randomly cropped to 224×224 for training, center-cropping for testing. Random horizontal flip is also applied during training.

The hyper-parameter λ_* is set to 2, and λ_k, λ_k^* are set to 1. The block size for cut-switch is set to 50, and the number of samples **K** is set to 100. The temperature τ is set to 1 at begining, and gradually decreased towards 0.5 over each epoch of the training process. We use an Adam optimizer with learning rate of 0.0001 and mini-batch size 50 with early stopping. Cross-validation is applied to find the best parameters. We implement our method with the Pytorch[30] framework and perform training and testing on the NVIDIA GeForce 2080Ti GPU.

For the AU recognition task, we use the F1-score for comparison study with the state of the arts. F1-score is defined as the harmonic mean of the precision and recall. As the distribution of AU labels are unbalanced, F1-score is a preferable metric for performance evaluation.

4.3 Experimental results

4.3.1 ***Comparison with single-modal based methods*** *.* To prove that multimodal provides additional valuable information for AU detection, we first compare our method to the single modality based methods, including Deep Structure Inference Network (DSIN) [7], Joint AU Detection and Face Alignment (JAA) [31], Optical Flow network (OF-Net) [42], Local relationship learning with Person-specific shape regularization (LP-Net) [29], Semantic Relationships Embedded Representation Learning (SRERL) [20], and ResNet18.

The upper part of Table.1 shows the results of different methods on the BP4D database using visual-only modality, where SRERL achieves the highest performance, around 3.3% higher than the corresponding ResNet-18. However, by using both *visual* and *depth* modalities, our method outperforms all the single modality (*visual*) based state-of-the-art methods, achieving around **3%** improvement in F1-score than the SRERL method, and **5.5%** higher than the ResNet-18-Depth. As no related results have been reported on the BP4D+ dataset, we compare our method with the ResNet-18 in Table 2. A similar finding is also observed that our multimodal based method outperforms the single-modal based ResNet-18, improving the F1-score by **4.5%**.

In short, *the experiments show the superiority of our multimodal fusion approach over the single-modal based approaches for AU detection on the both datasets.*

4.3.2 ***Comparison with multimodal based methods****.* As previously discussed, our method is designed to combine information from multiple modalities for improving AU detection performance. In this section, we examine if the proposed method can improve the performance when using multiple modalities. The early fusion and late fusion are currently the most common fusion techniques when facing multimodal data, so we use early and late fusion strategy with ResNet-18 backbone as baseline. We also compare with the ResNet-18 with channel attention mechanism (CAM), and the state-of-the-art multimodal methods: MTUT [1] and TEMT-Net [26].

Table 1: F1 scores in terms of 12 AUs are reported for the proposed method and the state-of-the-art methods on the BP4D database. V and D represent visual and depth modality. Bold numbers indicate the best performance; bracketed numbers indicate the second best. * *indicts the result from our own implementation.*

Method	Modal	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	Avg.
DSIN [7]	Visual	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
JAA [31]	Visual	47.2	44.0	54.9	77.5	74.6.	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
OF-Net [42]	Visual	50.8	45.3	56.6	75.9	75.9	80.9	88.4	63.4	41.6	60.6	39.1	37.8	59.7
LP-Net [29]	Visual	43.4	38.0	54.2	77.1	[76.7]	83.8	87.2	63.3	45.3	60.5	**48.1**	**54.2**	61.0
SRERL [20]	Visual	46.9	45.3	55.6	77.1	**78.4**	83.5	87.6	63.9	52.2	**63.9**	[47.1]	[53.3]	62.9
ResNet-18	Visual	48.0	46.7	57.0	77.5	71.6	83.5	85.0	63.8	47.1	58.2	39.4	37.3	59.6
ResNet-18	Depth	44.6	49.3	54.4	77.5	74.8	83.7	88.4	59.0	53.3	60.6	41.9	36.2	60.3
Early fusion	{V, D}	44.1	50.0	50.6	75.7	63.8	84.8	[89.3]	[65.0]	39.0	62.6	35.7	29.8	57.5
Late fusion	{V, D}	51.2	46.8	61.1	80.5	73.8	**87.7**	88.9	62.4	47.7	61.1	41.2	31.4	61.1
ResNet-18+CAM*	{V, D}	**55.4**	[50.3]	**62.9**	[81.5]	72.1	[87.6]	88.2	63.1	49.9	**65.3**	44.5	43.8	[63.7]
MTUT[1]*	{V, D}	51.3	50.2	[62.2]	77.2	71.7	83.8	88.2	61.4	54.3	57.9	45.8	42.2	62.2
TEMT-Net[26]*	{V, D}	53.7	47.1	60.5	77.6	75.6	84.8	87.4	**67.0**	[57.2]	61.3	44.7	41.6	63.2
AMF	{V, D}	[55.1]	**58.3**	62.0	**82.5**	75.6	87.2	**89.6**	60.9	**59.1**	62.4	45.0	52.0	**65.8**

Table 2: F1 scores in terms of 12 AUs are reported for the proposed method and the state-of-the-art methods on the BP4D+ database. V, D and T represent the corresponding visual, depth and thermal modality. Bold numbers indicate the best performance; bracketed numbers indicate the second best. * *indicts the result from our own implementation.*

Method	Modal	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	Avg.
ResNet-18	Visual	47.8	[47.0]	24.5	84.3	[88.0]	89.8	87.2	80.6	47.5	36.7	[54.7]	27.4	59.6
ResNet-18	Depth	40.9	39.2	30.4	83.8	86.7	**90.9**	[90.2]	79.6	38.2	44.0	52.5	**39.4**	59.6
ResNet-18	Thermal	39.0	34.0	25.0	82.2	84.0	87.6	87.2	79.2	32.1	36.5	43.9	7.9	53.2
Early fusion	{V, D, T}	39.0	34.6	26.2	80.1	86.1	89.5	87.7	74.0	41.0	33.5	44.9	15.8	54.4
Late fusion	{V, D, T}	38.5	38.9	[38.8]	82.8	84.0	89.5	89.2	78.4	42.6	32.3	52.2	22.1	57.4
MTUT[1]*	{V, D, T}	[49.9]	**49.5**	36.8	[85.4]	**88.6**	90.5	88.0	[81.0]	[49.4]	[44.6]	54.0	[35.4]	[62.7]
TEMT-Net[26]*	{V, D, T}	-	-	-	-	-	-	-	-	-	-	-	-	-
AMF	{V, D, T}	**50.1**	46.3	**44.4**	**85.8**	87.7	[90.6]	**90.8**	**83.8**	**51.0**	**47.6**	**57.5**	33.9	**64.1**

MTUT is designed to improve the testing performance in hand gesture recognition task by encouraging the networks to learn a common understanding across different modalities while avoiding negative transfer. TEMT-Net is a thermal empowered multi-task deep model which learns the latent representative by transferring the visual modality to the thermal modality. Since the source code for both MTUT and TEMT-Net are not released, we implement the corresponding methods, and report the results in Table.1 and Table.2. For the BP4D dataset, our model outperforms all the related methods, and achieves the highest F1-score **65.8%**, which is around **8.3%**, and **4.7%** higher than the early and late fusion methods,**2.1%** higher than the ResNet-18 + CAM, and **3.6%** and **2.6%** higher than the MTUT and TEMT-Net. The improved performance is also observed in BP4D+ dataset, as shown in Table.2, our model achieves the highest performance **64.1%**, showing 9.7%, 6.7% improvement over the early and late fusion methods, and 1.4% improvement over the MTUT. Note that the structure of TEMT-Net is incapable of being extended to three modalities, so no result reported in the BP4D+ dataset.

4.4 Ablation study

4.4.1 ***Results on BP4D+ with fusion of different modalities.*** We conduct experiments to examine the effects of fusion of different modalities, and report the results in Table.3. There are some interesting findings: 1) *different modalities are not contributing equally for AU detection, and they may have their own strength and weakness.* The fusion of {*depth*, *thermal*} is almost always achieving the worst performance than fusion of other modalities in all three methods; 2) *adding more modalities to the model does not always help for increasing the performance unless the model is able to capture the complex cross-modal interactions.* As we can see, the worst performance on late fusion is observed when using the *visual*, *depth* and *thermal* modalities. On the contrary, our model achieves the highest performance when using all the three modalities than using any two of them.

4.4.2 ***Effectiveness of individual part for AU detection.*** To answer the question of impact of individual part of proposed method, we conduct experiments on the BP4D dataset under different settings, and report the results in Table.4. A late fusion based ResNet-18 is trained *with* and *without* the cut-switch data augmentation

Table 3: Ablation study on BP4D+ dataset with fusion of different modalities. Bold numbers indicate the best performance for individual method.

Method	Modal	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	Avg.
Early fusion	{V, D}	35.6	32.7	26.9	80.2	85.8	89.8	88.0	77.0	37.3	34.1	46.9	15.4	54.1
	{V, T}	**39.1**	33.8	**30.0**	**83.7**	85.0	**90.5**	**89.2**	75.3	**43.4**	35.8	**50.0**	17.5	**56.1**
	{D, T}	24.2	24.3	25.0	83.1	82.3	89.0	88.2	**81.4**	36.4	**40.0**	49.0	**19.9**	53.5
	{V, D, T}	39.0	**34.6**	26.2	80.1	**86.1**	89.5	87.7	74.0	41.0	33.5	44.9	15.8	54.4
Late fusion	{V, D}	43.9	**46.1**	**38.9**	83.4	**89.0**	89.1	88.4	79.3	**47.6**	42.9	53.0	23.3	**60.4**
	{V, T}	**44.4**	42.5	34.0	83.0	86.5	89.5	89.3	78.8	46.9	35.7	55.6	15.3	58.5
	{D, T}	31.0	34.7	38.8	**85.4**	87.3	**90.1**	**89.5**	**81.0**	43.2	**45.6**	**55.7**	**24.3**	58.9
	{V, D, T}	38.5	38.9	38.8	82.8	84.0	89.5	89.2	78.4	42.6	32.3	52.2	22.1	57.4
AMF	{V, D}	45.3	42.5	34.8	**85.9**	**87.9**	89.5	90.4	82.6	50.1	45.5	55.7	**42.1**	62.7
	{V, T}	**53.2**	**50.4**	36.0	84.3	86.7	90.4	90.1	82.6	45.7	47.4	56.5	39.4	63.5
	{D, T}	39.6	40.7	32.8	84.3	85.3	89.2	89.3	77.6	45.4	44.3	56.3	37.6	60.2
	{V, D, T}	50.1	46.3	**44.4**	85.8	87.7	**90.6**	**90.8**	**83.8**	**51.0**	**47.6**	**57.5**	33.9	**64.1**

Table 4: Ablation study of effectiveness of individual part of our model on BP4D dataset. Bold numbers indicate the best.

Method	Modal	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	Avg.
Resnet-18 **w/o** cut-switch	{V, D}	51.2	46.8	61.1	80.5	73.8	**87.7**	88.9	**62.4**	47.7	61.1	41.2	31.4	61.1
Resnet-18 + cut-switch	{V, D}	53.8	51.5	58.6	79.4	73.5	86.2	89.1	59.6	44.8	**64.8**	45.3	46.6	62.8
AMF w/o cut-switch	{V, D}	52.1	51.0	**64.5**	79.2	73.9	86.4	88.3	60.5	55.3	64.2	**47.7**	49.2	64.4
AMF + cut-switch	{V, D}	**55.1**	**58.3**	62.0	**82.5**	**75.6**	87.2	**89.6**	60.9	**59.1**	62.4	45.0	**52.0**	**65.8**

method using the *visual* and *depth* modalities. As shown in the table.4, the performance is improved from 61.1% to 62.8% by training with the cut-switch method, which proves the effectiveness of our proposed cut-switch data augmentation method. **1.4%** performance improvement is also achieved by training **AMF** *with* and *without* the cut-switch.

Without cut-switch, we compare our method with late fusion based ResNet-18, as such, any performance improvement can be attributed to our feature fusion module. As shown in Table.4, around **3.3%** higher F1-score is achieved by comparing our proposed feature fusion method (*third row*) with the directly late fusion method (*first row*), which shows the effectiveness of our proposed feature fusion method.

4.4.3 *Robustness for noisy input.*

To show the performance when unexpected data corruption occur during testing, for example in the scenario of missing modality or noisy input, we conduct further experiments to evaluate the robustness of our model.

To emulate the scenario of missing modality, we replace one of the designated missing modality with all zero, and report the results in Fig.4. We can find that the performance of RestNet-18 (*late fusion*) *w/o* CAM decrease dramatically at the absence of one modality. It is especially true when visual modality is missing, the performance decreased from **61.1%** and **63.7%** to **23.6%** and 29.8% for ResNet-18 and ResNet-18+CAM respectively. However, another interesting fact is that the performance of ResNet-18 only decreased from 61.1% to 48.8%, which indicates the late fusion based ResNet-18 learns to put more weight on the visual modality than the depth modality through a biased classifier, even under the condition of missing visual modality. On the other hand, our proposed method remains robust to missing modality, achieving **60.7%** and **60.6%** F1-score for missing *visual* and *depth* modality respective, which is about **37.1%** and **11.8%** higher than the corresponding ResNet-18 model. *It is worth noting that, even with missing modality, our model still outperforms the single modality based ResNet-18 (ResNet-18-Visual and ResNet-18-Depth).*

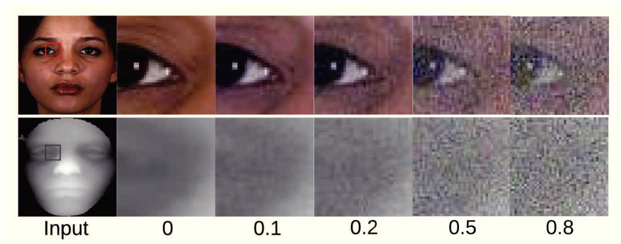

Figure 3: Example images for noisy modality corresponding to Fig.5 . Gaussian noise $\sigma = 0.1, 0.2, 0.5, 0.8$ are added to the normalized visual and depth images (range from -1 to 1). Images from a small area labeled as red box are used to show the difference.

We further evaluate the performance of our method and the ResNet-18 under the setting of corrupted modality, and report the results in Fig.5. As we can see, both our model and ResNet-18 model are robust to Gaussian noise with variance less than 0.2, and the performance changes as increasing the variance. The **red** and **blue** line in Fig.5 represent our model with Gaussian noise added to the *visual* or *depth* modality respectively, which shows comparable performance even with the variance increased from 0.2 to 0.8. The example images of corrupted modality is shown in

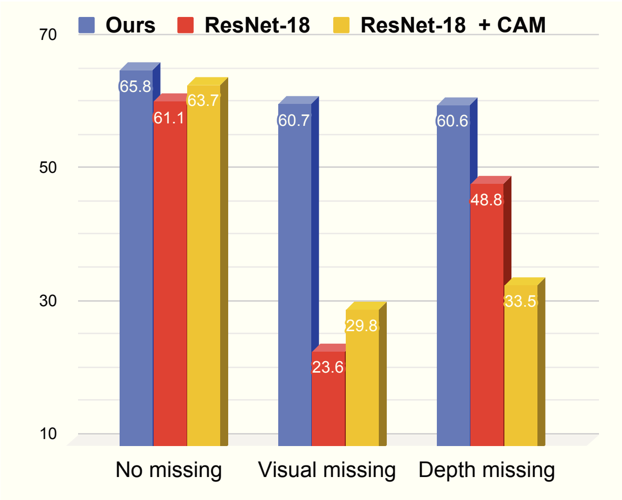

Figure 4: Ablation study of model robustness respect to missing modality on BP4D.

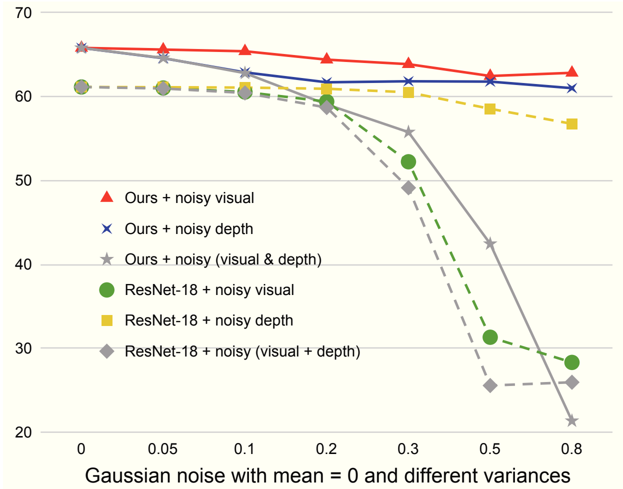

Figure 5: Ablation study of model robustness respect to noisy input on BP4D.

Fig.3. The worst performance is observed at the point *(variance=0.8, F1-score= 61%)*, which is close to the performance of ResNet-18 with clean inputs. We attribute the improved robustness to the feature scoring and sampling steps in our proposed method, which is able to evaluate the quality of features learned from individual modality and sample the feature based on their corresponding scores. On the other hand, the performance of ResNet-18 decreases dramatically when the variances (i.e., noise level) increase in the visual modality, as shown in the **green** line of Fig.5. The yellow line shows a certain robustness to the corrupted depth modality, which is consistent to our finding that the late-fusion based ResNet-18 model relies heavily on the visual modality (as shown the depth missing in Fig.4). Such a performance is due to the ResNet-18 being as a biased classifier.

When Gaussian noise is added to both modalities, as shown in grey lines, the performances of both our model and ResNet-18 decrease dramatically when the variance increases, as both modalities are corrupted and not enough information available. Note that such an extremely worst case rarely occurs in real applications though.

5 CONCLUSION

In this paper, we proposed a novel adaptive multimodal fusion (AMF) framework for AU detection. A feature scoring module is designed to evaluate the features learned from multiple modalities. The adaptive feature fusion process is conditioned on the feature scores with the Gumbel-Softmax resampling tricks to select the most relevant features from different modalities, while avoiding useless or misleading information. To alleviate the over-fitting issue, and make the model generalize better on the testing data, a cut-switch multimodal data augmentation strategy is also proposed. Extensive experiments demonstrate that our proposed model outperforms the single modality and both early and late fusion based multimodal models, as well it shows a better performance than the state-of-the-art peer approaches. In order to investigate the performance in various data degradation conditions, we conduct experiments to study the influence of missing or corrupted modalities, and the results show that our models are robust to various imaging conditions in terms of missing modality and noisy input.

It is worth noting that our proposed AMF framework is expandable to any number of modalities. Our future work will investigate feature fusion schemes from more modalities including audio and physiological signals, as well as more efficient data augmentation scheme across multi-dimension and multi-modal data.

6 ACKNOWLEDGEMENT

The material is based on the work supported in part by the NSF under grant CNS-1629898 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

REFERENCES

[1] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M Patel. 2019. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1165–1174.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[3] Mina Bishay and Ioannis Patras. 2017. Fusing multilabel deep networks for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 681–688.

[4] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 375–388.

[5] Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni. 2019. Selective sensor fusion for neural visual-inertial odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10542–10551.

[6] Wen-Sheng Chu, Fernando De la Torre Frade, and Jeffrey Cohn. 2017. Learning Spatial and Temporal Cues for Multi-label Facial Action Unit Detection. In *Automatic Face and Gesture Recognition (FG)*.

[7] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. 2018. Deep structure inference network for facial action unit recognition. In *Proceedings of the European*

Conference on Computer Vision (ECCV). 298–313.

[8] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[9] Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* (2014).

[10] Helmut Grabner, Michael Grabner, and Horst Bischof. 2006. Real-time tracking via on-line boosting.. In *Bmvc*, Vol. 1. 6.

[11] Amogh Gudi, H Emrah Tasli, Tim M den Uyl, and Andreas Maroulis. 2015. Deep learning based FACS action unit occurrence and intensity estimation. In *Automatic Face and Gesture Recognition Workshops*.

[12] William Grant Hatcher and Wei Yu. 2018. A survey of deep learning: platforms, applications and emerging research trends. *IEEE Access* 6 (2018), 24411–24432.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. 2018. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis* 49 (2018), 1–13.

[15] Yongtao Hu, Jimmy SJ Ren, Jingwen Dai, Chang Yuan, Li Xu, and Wenping Wang. 2015. Deep multimodal speaker naming. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1107–1110.

[16] Ramin Irani, Kamal Nasrollahi, Marc O Simon, Ciprian A Corneanu, Sergio Escalera, Chris Bahnsen, Dennis H Lundtoft, Thomas B Moeslund, Tanja L Pedersen, Maria-Louise Klitgaard, et al. 2015. Spatiotemporal analysis of RGB-DT facial images for multimodal pain level recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 88–95.

[17] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[18] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[19] Nagashri N Lakshminarayana, Nishant Sankaran, Srirangaraj Setlur, and Venu Govindaraju. 2019. Multimodal Deep Feature Aggregation for Facial Action Unit Recognition using Visible Images and Physiological Signals. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–4.

[20] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. 2019. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8594–8601.

[21] Huibin Li, Huaxiong Ding, Di Huang, Yunhong Wang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. 2015. An efficient multimodal 2D+ 3D feature-based approach to automatic facial expression recognition. *Computer Vision and Image Understanding* 140 (2015), 83–92.

[22] Huibin Li, Jian Sun, Zongben Xu, and Liming Chen. 2017. Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia* 19, 12 (2017), 2816–2831.

[23] Wei Li, Farnaz Abtahi, and Zhigang Zhu. 2017. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 6766–6775.

[24] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. 2017. EAC-Net: A Region-based Deep Enhancing and Cropping Approach for Facial Action Unit Detection. In *Automatic Face and Gesture Recognition (FG)*.

[25] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. 2018. EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2018).

[26] Peng Liu, Zheng Zhang, Huiyuan Yang, and Lijun Yin. 2019. Multi-modality empowered network for facial action unit detection. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2175–2184.

[27] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).

[28] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. *ICML* (2011).

[29] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. 2019. Local Relationship Learning with Person-specific Shape Regularization for Facial Action Unit Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11917–11926.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 8024–8035.

[31] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. 2018. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 705–720.

[32] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. 2019. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing* (2019).

[33] Chen Shen, Guo-Jun Qi, Rongxin Jiang, Zhongming Jin, Hongwei Yong, Yaowu Chen, and Xian-Sheng Hua. 2018. Sharp attention network via adaptive sampling for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (2018), 3016–3027.

[34] Luciano Spinello, Rudolph Triebel, and Roland Siegwart. 2010. Multiclass multimodal detection and tracking in urban environments. *The International Journal of Robotics Research* 29, 12 (2010), 1498–1515.

[35] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. *ICLR* (2019).

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[37] Gyanendra K Verma and Uma Shanker Tiwary. 2014. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage* 102 (2014), 162–172.

[38] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 569–576.

[39] Can Wang and Shangfei Wang. 2018. Personalized multiple facial action unit recognition through generative adversarial recognition network. In *Proceedings of the 26th ACM international conference on Multimedia*. 302–310.

[40] Chongliang Wu, Shangfei Wang, Bowen Pan, and Huaping Chen. 2016. Facial expression recognition with deep two-view support vector machine. In *Proceedings of the 24th ACM international conference on Multimedia*. 616–620.

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[42] Huiyuan Yang and Lijun Yin. 2019. Learning Temporal Information From A Single Image For AU Detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.

[43] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*. 6023–6032.

[44] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *EMNLP* (2017).

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[46] Wei Zhang, Youmei Zhang, Lin Ma, Jingwei Guan, and Shijie Gong. 2015. Multimodal learning for facial expression recognition. *Pattern Recognition* 48, 10 (2015), 3191–3202.

[47] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. 2019. Robust Multi-Modality Multi-Object Tracking. In *Proceedings of the IEEE International Conference on Computer Vision*. 2365–2374.

[48] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706.

[49] Zheng Zhang, Jeff M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. 2016. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[50] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. 2015. Joint patch and multi-label learning for facial action unit detection. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[51] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*. 465–476.