01 Jan 2023

Disagreement Matters: Exploring Internal Diversification For Redundant Attention In Generic Facial Action Analysis

Xiaotian Li

Zheng Zhang

Xiang Zhang

Taoyue Wang

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/comsci_facwork/1366

Recommended Citation

X. Li and Z. Zhang and X. Zhang and T. Wang and Z. Li and H. Yang and U. Ciftci and Q. Ji and J. Cohn and L. Yin, "Disagreement Matters: Exploring Internal Diversification For Redundant Attention In Generic Facial Action Analysis," *IEEE Transactions on Affective Computing*, Institute of Electrical and Electronics Engineers, Jan 2023.
The definitive version is available at https://doi.org/10.1109/TAFFC.2023.3286838

Disagreement Matters: Exploring Internal Diversification for Redundant Attention in Generic Facial Action Analysis

Xiaotian Li, *Member, IEEE,* Zheng Zhang, *Member, IEEE,* Xiang Zhang, *Member, IEEE,* Taoyue Wang, *Member, IEEE,* Zhihua Li, *Member, IEEE,* Huiyuan Yang, *Member, IEEE,* Umur Ciftci, *Member, IEEE,* Qiang Ji, *Fellow, IEEE,* Jeffrey Cohn, *Senior Member, IEEE,* and Lijun Yin, *Senior Member, IEEE*

Abstract—This paper demonstrates the effectiveness of a diversification mechanism for building a more robust multi-attention system in generic facial action analysis. While previous multi-attention (e.g., visual attention and self-attention) research on facial expression recognition (FER) and Action Unit (AU) detection have been thoroughly studied to focus on "external attention diversification", where attention branches localize different facial areas, we delve into the realm of "internal attention diversification" and explore the impact of diverse attention patterns within the same Region of Interest (RoI). Our experiments reveal that variability in attention patterns significantly impacts model performance, indicating that unconstrained multi-attention plagued by redundancy and over-parameterization, leading to sub-optimal results. To tackle this issue, we propose a compact module that guides the model to achieve self-diversified multi-attention. Our method is applied to both CNN-based and Transformer-based models, benchmarked on popular databases such as BP4D and DISFA for AU detection, as well as CK+, MMI, BU-3DFE, and BP4D+ for facial expression recognition. We also evaluate the mechanism on Self-attention and Channel-wise attention designs for improving their adaptive capabilities in multi-modal feature fusion tasks. The multi-modal evaluation is conducted on BP4D, BP4D+, and our newly developed large-scale comprehensive emotion database BP4D++, which contains well-synchronized and aligned sensor modalities, addressing the scarcity of annotations and identities in human affective computing. We plan to release the new database to the research community, fostering further advancements in this field.

Index Terms—Facial action unit detection, facial expression recognition, multi-modal feature fusion, attention, transformer, multi-channel, multi-head, diversity, disagreement.

1 Introduction

Both automatic facial expression recognition (FER) and facial action unit (AU) detection are critical for revealing human affective states. Over the past few decades, deep learning has achieved remarkable performance in automatic facial action analysis, leveraging its capacity to capture intricate high-level abstractions through hierarchical architectures with multiple layers of non-linear transformations and representations [21]. Attention, a fundamental behavioral and cognitive process that involves selectively focusing on specific information while disregarding other perceptible information, plays a crucial role in human cognition. Originally introduced in the field of natural language processing (NLP) for machine translation, the concept of Neural Attention Network (NAN) has been extended to the realm of computer vision [56].

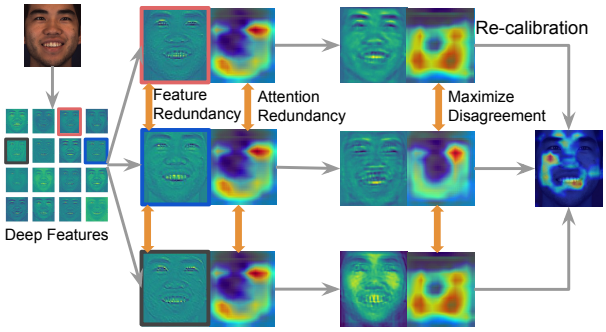


Fig. 1: Overview of the "Feature-to-Attention" dependency. Redundant attention maps are often generated by similar features. We leverage a diversification mechanism to maximize attention distance and perform re-calibration, constructing enhanced attention with diverse representations.

In recent facial analysis works [31], [37], [42], [45], [46], [71], multiple attention modules are commonly used to localize different regions of interest (RoIs) and learn discriminative facial features by disentangling a global face into smaller parts. However, this approach has limitations.

- Xiaotian Li, Zheng Zhang, Xiang Zhang, Taoyue Wang, Zhihua Li, Umur Ciftci, and Lijun Yin are with the Department of Computer Science, Binghamton University, 4400 Vestal Pkwy E, Binghamton, NY 13902. E-mail:{xli210, zzhang27, zxiang4, twang61, zli191, uciftci, lyin}@binghamton.edu.
- Huiyuan Yang is with the Department of with Electrical and Computer Engineering, Rice University, 6100 Main St, Houston, TX 77005. E-mail: hy48@rice.edu.
- Qiang Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180. E-mail: jiq@rpi.edu.
- Jeffrey Cohn is with the Department of Psychology, Psychiatry, and Intelligent Systems, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260. E-mail: jeffcohn@pitt.edu.

First of all, the strong dependencies and mutual exclusive relation among local facial parts are ignored. For instance, the AU1 and AU2 often co-exist due to the constraints of facial muscles, but decoupling the global face into isolated parts undermines their relationship. Second, the lack of a global view weakens the learning performance and compromise the representational power of the network. More importantly, paying attention to specific facial parts might capture irrelevant information. For instance, both AU9 and AU10 have the same muscle activation on levator labii superioris which is visually indivisible. Our investigation of related works [29], [44], [45], [46], reveals that attention maps between (AU4 and AU7), (AU12, AU14, and AU15), and (AU23 and AU24) exhibit obvious regional overlap, leading to ambiguous localization of facial actions and less discriminative features. We categorize these approaches as as "external attention diversification".

Unlike attention design, general CNN-based features learn to present different patterns in local receptive fields, indicating that distinct representational quality can yield varied performance when learning local feature. Therefore, designing a more robust attention framework with diversified patterns remains a challenging task due to the lack of focus diversity. This has prompted us to ponder several intriguing questions: (1) Is there any redundancy for unconstrained multi-attention? (2) What is the origin of redundant attention maps and why are they present in the model? (3) Are there more sophisticated approaches, beyond network pruning techniques, to effectively handle redundant information in the model? (4) Can the integration of diversifying attention modules assist in learning refined features?

Both CNN attention and Transformer self-attention typically utilize linear transformations, reshaping, or feature pooling techniques to derive multiple channels from the same input feature. However, due to the high dependency between input features and attention maps, simple linear transformations cannot fundamentally avoid redundant information. In Fig. 1, we selected some similar deep features, and the resulting attention maps reveal significant overlapping areas. While conventional network pruning can remove this redundancy, they primarily focus on reducing network size and improving training efficiency. In contrast, we disentangle a collection of feature maps extracted by a backbone network, and map them to several sub-spaces for attention initialization. By maximizing the distance between attention maps and re-calibrating their contribution weights, an optimized attention system is constructed to encode more distinctive focus information. The diversity function encourages multiple attentions to focus the same RoIs while reserve a certain level of disagreement between them. Our proposed method, referred to as "internal attention diversification," focuses on the patterns of attention features rather than their physical positions. Besides, attention mechanisms have been widely used in multi-modal deep learning address various (e.g., misalignment, and heterogeneity). Attention allows the models to dynamically assign different weights for multi-modalities, based on their importance for the task at hand. If we view the multi-attention module as a voting system, when all votes are similar, the functionality of the system are greatly compromised and fragile. In this study, we discovered that a diversified attention mechanism for feature fusion, can not only selectively attend to relevant features, but also ensure a robust feature fusion decision system. Unlike "external attention diversification," which requires fine-grained supervision for specific locations, our proposed "internal attention diversification" provides more generality and flexibility for multi-modal tasks due to its self-diversified feature. Our contribution lies in four-fold:

- We highlighted the robustness and redundancy issue of multi-attention, which are often overlooked in conventional facial action analysis.
- We provide a new understanding of internal attention diversification mechanism, which explicitly enhances the representational and focusing abilities of the attention system without introducing many trainable parameters, external knowledge guidance, or regularizers.
- We have demonstrated the flexibility and generality of the proposed mechanism, by applying to CNN and Transformer attention designs, and extending it to multi-modal feature fusion tasks.
- We constructed a new spontaneous multi-modal emotion database BP4D++ by capturing 3D geometric facial sequences, 2D facial videos, thermal videos, and physiological data sequences from 233 participants across two years period. The new database will be released to the research community along with the paper being published.

2 RELATED WORK

2.1 Facial action analysis

Facial expression and action unit features can be broadly classified into two categories: shallow features and deep features. In this section, our main focus is on models based on deep features. Mollahosseini et al. [36] presented a new deep neural network (DNN) architecture to deal with the FER task. Prior work [9], [59] demonstrate the deep features perform an impressive function on both AU multi-label classification and intensity estimation. Predefined attention was a popular method to learn the ROIs regarding both AU detection and FER. [22] proposed the EAC-Net by using predefined attention, whose location is derived by face landmarks, to enhance and crop the RoIs of AUs. Sanchez et al. [53] adopted the Gaussian distribution to predefine the central location of specific AU according to the fixed landmark for AU intensity estimation. However, due to lacking the accurate definition of the contour and location bias cross identity, ethnicity, gender, age, and facial expressions, the predefined ROIs fall into sub-optimal and lose the adaptation during the recognition process. To tackle this issue, self-attention mechanism have been applied to discover the meaningful facial locations with more adaptations for FER, AU detection, AU intensity estimation, and face alignment tasks [27], [37], [42], [45], [46], [71]. In addition, self-attention is also widely used for capturing the correlations among AUs [45], spatio-temporal relation [24], [47], and modality-level relation [63]. For instance, Li et al. [24] proposed a dynamic knowledge distillation to expand the spatial AU correlation to structured temporal AU correlation knowledge by combing attention mechanism, smooth

regularization, and pseudo labeling with limited discrete labels in a semi-supervised way.

2.2 Attention mechanism

Attention emerged in CNN to filter information and allocate weights to the neural network efficiently. Attention mechanisms direct the operational focus of deep neural networks to areas where there is more saliency information. This mechanism has been successfully applied to deep CNN-based image enhancement methods. For example, Zhang et al. [69] proposed a residual channel attention network (RCAN) in which residual channel attention blocks (RCAB) allow the network to focus on the more informative channels. Huet al. [11] proposed a Squeeze-and Excitation (SE) block to improve the quality of representations produced by a network by explicitly modeling the inter-dependencies between the channels of its convolutional features. Woo et al. [58] proposed channel attention (CA) and spatial attention (SA) modules to exploit both inter-channel and inter-spatial relationships of feature maps.

Multi-head attention [55] is a mechanism that can boost the performance of the vanilla self-attention layer in NLP. A related work [48] extended the single relevance score to multi dimensional attention weights, demonstrating the effectiveness of modeling multiple features for attention networks. Researchers [2], [23] proposed a weighted mechanism to combine multi-head self-attention rather than adopting simple concatenation operation. In order to address the Transformer's limitation on computer vision, ViT [7] splits an image into fixed-sized patches to linearly embed them with position embedding and feeds the whole sequence to a standard multi-head attention Transformer. Recent work [24] and [26] utilize the multi-head transformer to learn the latent information of AU temporal dependencies and Spatial correlations. These research have demonstrated that the multi-head design is beneficial for learning enhanced features maps. However, there lacks a work that explains how to build a robust attention framework. Our method differs from existing works in that it utilizes attention at two different levels (Inter-AU and Inter-attention), capturing high-dimensional information for facial analysis, and does not require disentangling a face for focusing on each facial part independently, reducing design complexity. The proposed mechanism showcases flexibility and generality by applying to CNN and Transformer attention designs, extending to multi-modal feature fusion tasks, and being versatile for a plethora of applications.

2.3 Multi-modal learning

Unstructured real-world data can inherently take many forms, also known as modalities, often including diverse representations of content. Multi-modal learning refers to an embodied learning situation that engages multiple sensory systems and action systems of the model. In the past decade, multi-modal learning has been studied on both facial expression recognition and action unit detection tasks. The extraction and synthesis of rich information from a multi-dimensional data space require the use of an intermediate mechanism to facilitate decision making in intelligent systems. Li et al. [19] proposed a deep fusion network to learn the optimal combination weights of 2D and 3D facial representations for multi-modal 2D+3D FER. Researchers [13] propose to utilize RGB-Thermal-Depth images for pain estimation. Li et al. [25] explained that EEG signals show a strong correlation with facial actions and eye blinking of both posed and spontaneous expressions. They utilize the early fusion of EEG feature and RGB feature to boost both posed and authentic facial actions detection. Liu et al. [30] designed a cross-modal translation network encoding the RGB images to reconstruct thermal images. Recently, the transformer-based [55] attention mechanism has been widely applied in multi-modal features fusion. For example, MulT [54] first attempted to fuse multi-modalities by transformer, which was applied to image, audio, and text. TransFuser [39] incorporates the global context of the 3D scene into the feature extraction layers of different modalities. Researchers [67] approach RGB-Depth features fusion by utilizing multi-head fusion attention to fuse AU features of two modalities in a transformer. In this paper, we will focus on how to utilize the internal diversification mechanism to boost the traditional MLP-based and Transformer-based attention systems.

3 METHODOLOGY

In this section, we elaborate on variants of the proposed module, including (1) A CNN-based model called "Self-Diversified Multi-Channel Attention Network" (SMA-Net) as shown in Fig.3; (2) A Transformer-based model called "Self-Diversified Multi-Channel Attention Transformer" (SMA-ViT) as shown in Fig.4; and (3) Extended modules for multi-modal feature fusion, which combine vanila Tranformer and MLP with the proposed internal attention diversification mechanism as shown in Fig.5.

3.1 Backbone networks and location

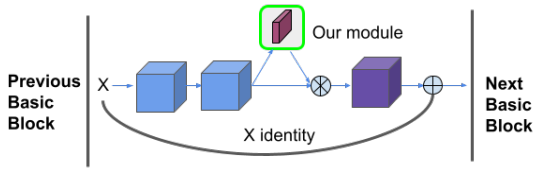


Fig. 2: The location of the multi-attention module in the basic-block of ResNet.

Our method, SMA-Net, employs ResNet-18 as the backbone network, while SMA-ViT utilizes ViT base as the backbone network. Fig. 2 and Fig. 5 shows the exact location of our module. The modules are integrated into every basic-block of ResNet-18 and ViT (with a standard multi-head Transformer) to ensure consistent and comprehensive inference. In each basic-block, we first obtain a feature map. This feature map serves as the input I for the SMA modules. After being processed through the SMA module, each feature map can obtain a feature O that has been optimized with self-diversified multi-attentions, which is then returned to the backbone network for further forward propagation. In the following sections, we will discuss the implementation details of the proposed SMA module, which comes into play after the features are passed through the SMA modules.

3.2 CNN-based model

SMA-Net is powered by three key components that work in harmony: (1) Multiple "Feature-to-Attention channel" (F2A) are meticulously initialized and trained with deep features

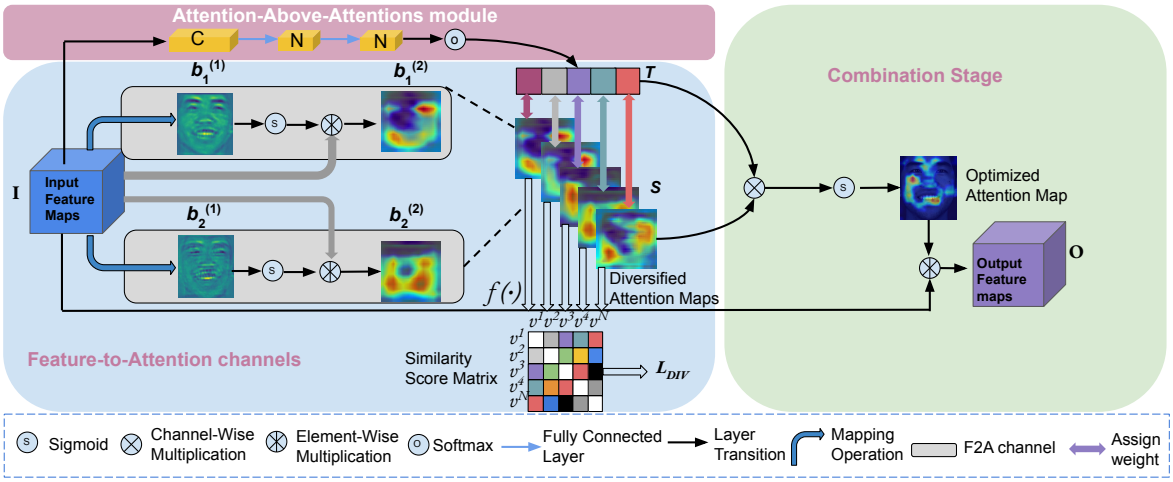


Fig. 3: Overview of the basic multi-attention block in our Self-Diversified Multi-Channel Attention Network (SMA-Net). The model takes the extracted feature block as input. After mapping the feature block into N attention channels (here N = 2), the diversity losses l_{div} encourage the model to learn the same active parts yet present diverse representational patterns. The diversified attention maps are evaluated and aggregated into the next stage for getting the optimized attention map. Noting that "Self-Diversified" means the representational diversity of attention maps is generated with an unsupervised loss function l_{div} without using any ground-truth information.

mapped from the local receptive field. In each channel, we deploy a spatial attention to attend the facial RoIs, and learn the pixel-level relations for facial actions in a holistic way; (2) To foster diversity and robustness for attention maps, we introduce the "Internal Diversification", which contains an attention similarity matrix and a diversity loss function. It encourages attentions from different channels to focus on identical locations while simultaneously learning diverse representations of attention maps, serving as a regularization technique; and (3) "Attention-Above-Attentions" (AAAtention), akin to channel-wise attention, is designed to selectively emphasize optimal attention channels while suppressing less useful ones. It simulates the inter-relation among F2A channels, which we term the "inter-attention" correlation.

Feature-to-Attention channel (F2A) The input feature I is initially obtained by the basic-block of ResNet18 as the input of this module. Fig.3 shows the overview of the proposed multi-attention block in SMA-Net. Inspired by the multi-head attention design [7], [55], the input feature is split into several sub-channels. Our method does not follow previous principles, such as using global pooling or average pooling, for the initialization of each channel. Instead, we expect the multiple attention maps to show diversity from the initial stage. In addition, considering the common symptoms of over parameterization for conventional multi-head attention works [5], this module reduces the channel dimension from C to N by applying a simple convolutional operation, where N is the number of attention channels. After channel mapping operation, we get the feature maps b_n^1, where n denotes the nth (from 1 to N) channel and 1 means the first stage of the F2A channel. The channel number N can be set as a hyper-parameter of the model.

Zhu et al. [45] adopt spatial attention to capture AU-related local features and the pixel-level relation within independent AUs. In contrast, our holistic approach avoids disentangling the face into multiple parts, preventing ambiguity from invalid segmentation. The spatial attention in each F2A channel learns AU-related features and inter-AU correlations. Specifically, we further extract b_n^1 using another convolutional layer F_n^1, and compute the spatial attention map S_n in each branch by:

$$S_n = I \circledast \sigma(F_n^1(I)) \tag{1}$$

where F_n^1 indicates 2D convolutional operation that takes 1/1 as the input/output channel. Here σ denotes the sigmoid function, and $\circledast$ denotes the element-wise multiplication, where $I \in \mathbb{R}^{C\times H\times W}$ and $S_n \in \mathbb{R}^{1\times H\times W}$. The sigmoid function is used for attention activation in F2A channels.

Internal Diversification (ID) To reduce the computational cost, we incorporate a linear projection head, denoted as $f(\cdot)$ in Fig. 3, that maps attention representations S_n to a space where the diversity loss is applied. These down-sampled vectors v are then used to construct an attention similarity score matrix. The diversity loss $\mathcal{L}_{DIV}$ is designed to encourage the model to learn diverse attention patterns by maximizing the cosine distance between pairs of attention maps from the F2A channels in an unsupervised manner. To achieve this, we calculate the cosine similarity $\mathcal{D}^{m,n}$ between the down-sampled attention vectors v^n and v^m, and minimize their cosine similarity. Note that the cosine similarity is normalized from 0 to 1 for ease of training. The equations are formulated as:

$$\mathcal{D}^{m,n} = \frac{v^n \cdot v^m}{\|v^n\| \, \|v^m\|} \tag{2}$$

$$\mathcal{L}_{DIV} = \frac{1}{(N-1)^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \frac{\mathcal{D}^{m,n}+1}{2} \tag{3}$$

where N indicates the total number of attention channels, and $n \neq m$. To further reduce computational burden, the the upper or lower half of the symmetric attention similarity matrix can be omitted.

Attention-Above-Attentions (AAAtention) The Attention-Above-Attentions (AAAtention) module, shown in Fig. 3, is constructed as a global attention system to learn to weigh

the performance of F2A channels and re-calibrate them accordingly. Inspired by Domain Attention in [57], which uses the squeeze-and-excitation mechanism [11] for domain-sensitive weighting in universal object detection, our module selectively emphasizes optimal attention channels while suppressing less useful ones, and captures the inter-relation among F2A channels. T in Fig. 3 stands for the correlation map of F2A channels. This module applies average-pooling to remove spatial information from I, denoted as a descriptor F_{avg}. The channel-wise attention weight $T \in \mathbb{R}^{1\times1\times N}$ can then be computed as:

$$T = \varphi(L^1(L^2(F_{avg}(I)))) \tag{4}$$

where L^1 and L^2 denote two full-connected layers for affine transformation, and φ denotes the softmax function.

The combination module performs element-wise multiplication with the channel-wise attention map T and the aggregated multi-channel attention maps S. We use $*$ to denote the channel-wise multiplication operation. The optimized attention map $A \in \mathbb{R}^{1\times H\times W}$ is denoted as:

$$A = \sigma(F_{avg}(T * S)) \tag{5}$$

where F_{avg} indicates the average-pooling operation. The refined output feature $O \in \mathbb{R}^{C\times H\times W}$ can be expressed as:

$$O = A \circledast I \tag{6}$$

where $\circledast$ denotes the element-wise multiplication. The output feature O is then returned to the backbone network for further forward propagation.

3.3 Transformer-based model

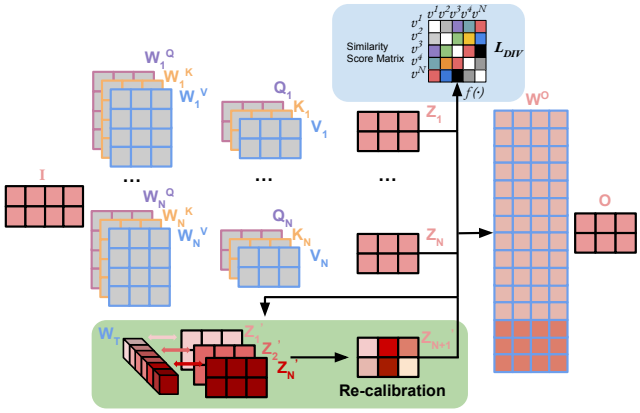


Fig. 4: Illustration of the multi-attention block in SMA-ViT.

The standard multi-head attention mechanism in Transformer models is designed to attend to information from different representation subspaces at different positions. However, existing approaches often reshape and transpose input vectors I into different subspaces holistically, treating a sub-attention path as a positional head. Therefore, treating the branch of a standard multi-head attention as a channel, rather than a positional head, is more appropriate. This approach results in redundant attention due to the lack of explicit guidance for localizing specific regions of interest (RoIs). Instead of using unconstrained multi-head attention with keys, values, and queries, we propose diversifying the patterns of self-attention.

In Fig.4, we compute the attention function on a set of queries packed together into matrices $Q_n \in \{Q_1, Q_2, ..., Q_N\}$, with keys and values packed into matrices K_n and V_n, where n indicates the nth head of multi-attention. The matrix of each output z_n is formulated as $\text{Attention}(Q_n, K_n, V_n) = softmax(\frac{Q_n K_n^\top}{\sqrt{d_k}})V_n$ where $Q_n = IW_n^Q, K_n = IW_n^K, V_n = IW_n^V$, and d_k indicates the scaling factor. I indicates the input embedding for the basic multi-attention block in SMA-ViT. It is obtained from each layer of standard multi-head Transformer in ViT base. The layer is parameterized by a query matrix W_n^Q, a key matrix W_n^K and a value matrix W_n^V. However, adjusting the similarity via queries, keys, or values separately in a large number of independent matrices is impractical and inefficient. Instead, we approach internal diversification by enlarging the distance between z_n values following Eq.3. Afterward, we re-weight the diversified matrices $z_n^{'}$ using a trainable parameter W_T, as no significant performance improvement is observed when applying the Attention-Above-Attention module (i.e., squeeze-and-excitation mechanism) in SMA-Net.

In contrast to SMA-Net, which reduces multiple attention maps into one to obtain aligned outputs, SMA-ViT incorporates the re-calibrated attention matrix $z_{N+1}^{'}$ as an extra head concatenated with other heads $z_n^{'}$. To implement this module, we cautiously expand the dimension of W^O without adding substantial parameters, considering that reducing multiple heads z_n to a single one may lead to information loss. Setting an expanded attention ensures that the model retains the learned correlation knowledge. Formally, the output feature O of the multi-head attention is formulated as:

$$O = \text{Concat}(z_1^{'}, ..., z_N^{'}, z_{N+1}^{'})W^O \tag{7}$$

where $z_n^{'} = \text{Diver}(z_n)$, $z_n = \text{Attention}(IW_n^Q, IW_n^K, IW_n^V)$, and $z_{N+1}^{'} = \text{Recalib}(z_1^{'}, ..., z_N^{'})$. "Diver" indicates the self-diversifying process that guided by Eq. 3, and "Recalib" indicates the Re-calibration operation for diversified attention maps. The parameter dimension of W_n^Q, W_n^K, and W_n^V is exactly aligned with the original multi-head self-attention, while the head dimension N in W^O are expanded to $N+1$.

3.4 Multi-modal feature fusion

We integrated the internal attention diversification with a Transformer-based self-attention and MLP-based Channel-wise attention (SE-Net) for multi-modal feature fusion tasks. The Transformer attention mechanism is highly effective in capturing long-range dependencies and modeling global contextual information, making it well-suited for simulating complex interactions between features that may span across different regions or modalities. On the other hand, SE-Net is particularly advantageous for feature fusion as it offers a flexible and efficient way to re-calibrate feature channels based on global context. Therefore, combining SE-Net and Transformer with the proposed module may amplify their respective strengths. In this paper, we employed a deep fusion strategy to investigate whether internal attention diversification can enhance the adaptation ability of feature fusion.

Initially, multiple streams of pre-trained ResNet-18 are used to encode and extract higher-level feature abstractions from different modalities. As shown in Fig. 5, embedding I_1 indicates the deep feature of 2D texture face, while I_2 is the deep feature obtaining from data of other modalities (e.g., a subject's 3d depth face, thermal face, or physiological signal). The input is then created by concatenating deep embeddings I_1 and I_2, and then fed into the Transformer-based (self-attention) module and SE-based (channel attention) module in Fig. 5. We select one-layer of SMA-ViT for Transformer-based feature fusion. This module does not use the patches splitting strategy of ViT in the early stage of feature extraction, as using a multi-patch ViT backbone would significantly increase the network complexity. The distance of multi-head attention is controlled and regularized with the diversity loss in Eq. 3. Similarly, the same internal diversification mechanism is applied to the SE-based module. In contrast to the pixel-level spatial attention in the original SMA-Net, these non-neural attention modules lack efficient methods, such as convolution operations, to map a single input into sub-spaces with diverse initializations. One alternative approach is to use distinct pooling operations to simulate the diversity. However, using different pooling strategies in different F2A channels shows only marginal improvement in performance. We speculate that the diversity loss, which provides sufficient control for diversifying the high-level abstraction of attention maps, diminishes the contribution of the initialization process.

We conducted multi-modal evaluations that involves the fusion of 2D RGB visual images with 3D depth maps from BP4D, thermal images from BP4D+, and physiological data from our newly developed BP4D++. For BP4D++, we went beyond traditional visual facial activity capture using a photoelectric sensor, and also employed multiple contact biosensors to record the subject's physiological status. These comprehensive measurements, including continuous blood pressure, electrocardiogram (ECG), pulse rate, and respiration rate, provide a wealth of authentic emotional information and cues that have been overlooked in previous studies. While existing works have utilized physiological signals for sequence-level multi-modal emotion recognition [8], [15], [49], where datasets provide coarse-grained video-level annotations, image-level facial action analysis using physiological measurements remains an untapped frontier in the field of affective computing. Fortunately, our newly developed BP4D++ is frame-wise annotated with AU labels, which brings the more challenging image-level multi-modal fusion task into the research community. However, unlike other visual modalities, physiological modality contains relatively less emotion-related information that can be effectively utilized for facial behavior detection. Consequently, regular attention fusion methods run the risk of overfitting to the dominant image modality and fail to fully exploit the complementary information from physiology. In contrast, our self-diversified attention generates varied attention masks to decrease the likelihood of multiple overlapping attention focusing solely on the dominant modality, thereby mitigating the overfitting issue. Specifically, the process begins with synchronizing the image and physiological signals, followed by down-sampling the physiological signals to match the video frame rate, and aligning each

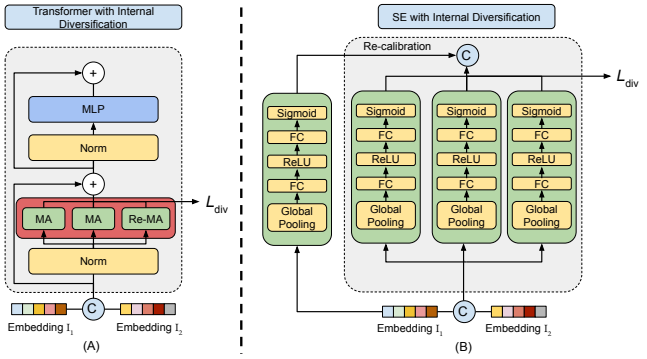


Fig. 5: Two extended modules for multi-modal feature fusion with the combination of Vanila Transformer, vanila channel-wise attention with the internal diversification mechanism. Figure (A) shows one layer in the proposed SMA-ViT, with the red portion indicating the location of the basic multi-attention block of SMA-ViT.

image to its corresponding physiological values. To detect the AUs of one frame in a video, we locate the physiological signals at the same timestamp and select a segment of M frames. The paired P-dimensional physiological signals are then compacted to a matrix $I_2 \in \mathbb{R}^{M \times P}$. Next, we feed it to GRUs (gated recurrent units) to obtain the temporal evolution features. Finally, the physiological features are concatenated with the image features I_1 and fed into the multi-modal fusion module. Note that only physiological data adopts GRUs as the backbone network.

3.5 Overall objective and loss function

Considering the issue of data imbalance that can skew the training process, we choose weighted BCE logits loss for AU detection. The loss function can be described as:

$$\mathcal{L}_{AU} = -\sum_{n=1}^{N} w_n [y_n log\sigma\left(x_n\right) + \left(1 - y_n\right) log\left(1 - \sigma\left(x_n\right)\right)] \tag{8}$$

where w_n is calculated by the nth AU's occurrence ratio, and less likely occurred AUs have higher weights. $\sigma(x)$ is the corresponding predicted probability of input x. y_n is the ground truth.

For the facial expression recognition task, the cross-entropy loss is selected as the loss function which can be described as:

$$\mathcal{L}_{FE} = -\sum_{c=1}^{C} y_c log(\frac{exp(x_c)}{\sum_{i=1}^{C} exp(x_i)}) \tag{9}$$

where x is the input, y is the ground truth, C is the number of classes, and c indicates the current class of input x. The overall function can be can be described as:

$$\mathcal{L}_{ALL} = \mathcal{L}_{AU} + \alpha\mathcal{L}_{DIV} \text{ and } \mathcal{L}_{ALL} = \mathcal{L}_{FE} + \alpha\mathcal{L}_{DIV} \tag{10}$$

where α is the balance factors serving as a hyperparameter.

4 Experiments

4.1 Datasets and Settings

The proposed framework is evaluated on two benchmark datasets: BP4D [68] and DISFA [34] for action unit detection. In addition, it is evaluated on four FER databases: Extended Cohn-Kanade (CK+) [33], MMI [38], BU-3DFE [66], and BP4D+ [70].

BP4D and **DISFA**: BP4D contains 41 subjects and around 140,000 frames captured under laboratory environments. There are 27 subjects and around 130, 000 frames in the DISFA database. Following the previous settings of DISFA, the frames with intensities equal to or larger than 2 are regarded as AU occurrences while the rest are absent. We divide both datasets into subject-independent 3 folds and report the performance through cross-validation for a fair comparison with other algorithms. We observed a severe data imbalance issue in DISFA which affects the evaluation significantly for the less representative AUs. To tackle such an issue, we apply a selective data re-sampling technique that enhances the less occurred AUs. The selective oversampling process duplicates the minority classes in the training set. Meanwhile, setting the threshold p can control unnecessary oversampling from other majority classes which may cause overfitting.

CK+, **MMI**, **BU-3DFE** and **BP4D+**: CK+ consists of 593 video sequences from 123 subjects, and we use the last three frames of each sequence, resulting in a set of 981 images. MMI comprises 236 image sequences from 31 subjects, and we select three frames from the middle of each sequence in frontal view, resulting in a dataset of 624 images. BU-3DFE contains 2,500 pairs of static 3D face models and texture images from 100 subjects with diverse ages and races. Only the 2D texture images with high-intensity expressions are used in our experiments. BP4D+ is a spontaneous emotion corpus with 140 subjects, and we select 2468 frames from 72 subjects based on the FACS codes, and use 2D texture images for four expressions (happiness, surprise, pain, neutral) as the ground truth. The images are split into 10 folds with mutually exclusive subjects.

Extended database BP4D++[1]: The existing facial action datasets are limited in terms of subjects number, diversity, and metadata. Thanks to the existing available multi-modal datasets BP4D [68] and BP4D+ [70], we extend to develop a larger-scale multi-modal emotion database BP4D++, which consists of **233** participants (132 females and 101 males). The data is significantly expanded in terms of participants number as compared to the existing emotion databases: DISFA (27 subjects), MMI (44 subjects) [38], BP4D (41 subjects) [68], BP4D+ (140 subjects) [70]. Following ethical principles, our data collection was approved by the institutional review board (IRB). Each subject signed an informed consent form. A professional performer/interviewer applied a procedure containing 10 seamlessly-integrated tasks as [68] [70] that resulted in effective elicitation of spontaneous emotions.

233 participants were recruited from our University. There are 132 females and 101 males, with ages ranging from 18 to 70 years old. Ethnic/Racial Ancestries include Asian, Black, Hispanic/Latino, White, and others (e.g., Native American). Our data collection system consists of a 3D dynamic imaging camera system, a thermal sensor, a physiological signal sensor system, and a studio-quality audio recorder. Ten tasks were performed to elicit a wide range of spontaneous emotion expression (from positive, to neutral, and to negative) and inter-personal facial action behavior by a professional interviewer. The system setup and synchronization method are basically consistent with

1. Contact Lijun Yin at lijun@cs.binghamton.edu for dataset access.

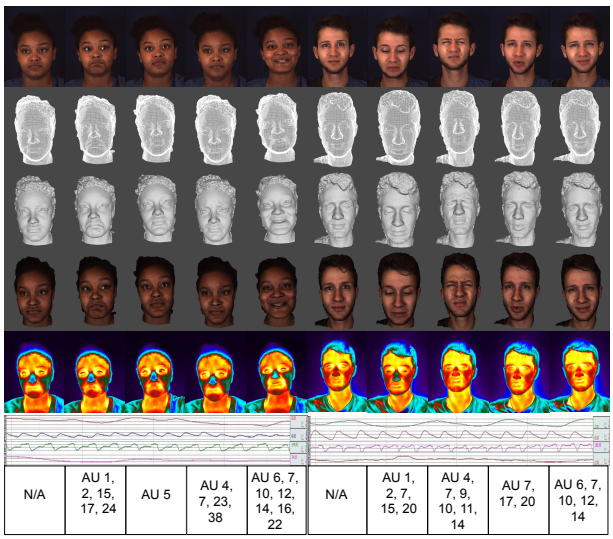


Fig. 6: A sample sequence from our BP4D++. 2D texture image, 3D mesh model, 3D shaded model, 3D texture model, thermal image, and physiological signal (respiration rate, blood pressure, EDA, heart rate) and corresponding AU occurrence are shown from top to bottom.

BP4D+. Each subject is associated with 10 different emotions and multi-modal data including the 3D sequence, 2D RGB sequence, thermal sequence, and the sequences of physiological data (i.e., blood pressure, EDA, heart rate, and respiration rate). The sample sequences of different modalities from two subjects are shown in Fig.6. Besides, the metadata including manually labeled action units occurrence and intensity, 3D/IR facial landmarks, and 3D head poses are also generated for better analysis of automatic human facial action. Around 94,000 frames were well-annotated by three expert FACS coders for AU coding. The new database is ready for public and will be released to the research community by the time of the paper being published.

4.2 Implementation details

For the AU detection, we first process the image by the cropping operation to cut off redundant area which is not relevant to the face recognition. Then the images are resized to 256*256*3 (H*W*C) to fit the model. Each of the training images is randomly rotated (-45 to 45 degrees), flipped horizontally (50% possibility), and with color jitters (saturation, contrast, and brightness). The backbone network is pre-trained on ImageNet [40]. We choose SGD as the optimizer with an initial learning rate of 0.01. It is decreased to 0.001 after the first 2 epoch training. The weight decay and momentum are set as 0.0001 and 0.9 respectively. The number of attention branches, denoted as N, is set as 7, 7, 7, and 5 on BP4D, BP4D+, BP4D++, and DISFA datasets.

For the FER, we cropped the image and resized it to 64*64*3 (H*W*C). Before training, the images are randomly horizontal flipped (50% possibility), rotated (-15 to 15 degrees), and with color jitters (saturation, contrast, hue, and brightness). We choose SGD as the optimizer with an initial learning rate of 0.01. The learning decay rate is 0.99 and it decays every 10 epoch until the 100 epoch is finished. The number N of attention branches is set as 7.

TABLE 1: F1 score of the state-of-the-art models on BP4D. Numbers with underline indicate the best performance.

AU	DSIN	JAA	ARL	LP	SRERL	HMP-PS	UGN	SMA-Net	SMA-ViT
1	51.7	47.2	54.8	43.4	46.9	53.1	54.2	56.5	52.7
2	40.4	44.0	39.8	38.0	45.3	46.1	46.4	45.1	45.6
4	56.0	54.9	55.1	54.2	55.6	56.0	56.8	57.0	59.8
6	76.1	77.5	75.7	77.1	77.1	76.5	76.2	79.5	83.8
7	73.5	74.6	77.2	76.7	78.4	76.9	76.7	79.5	79.2
10	79.9	84.0	82.3	83.8	83.5	82.1	82.4	84.5	83.5
12	85.4	86.9	86.6	87.2	87.6	86.4	86.1	86.4	87.2
14	62.7	61.9	58.8	63.3	63.9	64.8	64.7	66.1	64.0
15	37.3	43.6	47.6	45.3	52.2	51.5	51.2	55.8	54.1
17	62.9	60.3	62.1	60.5	63.9	63.0	63.1	64.2	61.2
23	38.8	42.7	47.4	48.1	47.1	48.5	50.8	48.7	52.6
24	41.6	41.9	55.4	54.2	52.3	54.5	53.6	56.8	58.3
Avg.	58.9	60	61.1	61.0	62.9	63.4	63.3	65.0	65.2

TABLE 2: F1 score of the state-of-the-art models on DISFA. Numbers with underline indicate the best performance.

AU	DSIN	JAA	ARL	LP	SRERL	HMP-PS	UGN	SMA-Net	SMA-ViT
1	42.4	43.7	43.7	29.9	45.7	38.0	43.3	53.4	51.2
2	39.0	46.2	42.1	24.7	47.8	45.9	48.1	54.2	49.3
4	68.4	56.0	63.6	54.2	72.7	65.2	63.4	64.0	64.7
6	28.6	41.4	41.8	46.8	47.1	50.9	49.5	57.0	48.3
9	46.8	44.7	40.0	76.7	49.6	50.8	48.2	47.0	50.6
12	70.8	69.6	76.2	87.2	72.9	76.0	72.9	76.6	87.6
25	90.4	88.3	95.2	63.3	93.8	93.3	90.8	92.0	85.1
26	42.2	58.4	66.8	45.3	65.0	67.6	59.0	55.2	61.2
Avg.	53.6	56.0	58.7	61.0	56.9	61.0	60.0	62.4	62.2

4.3 Result and Discussion of Single-modal Tasks

In this section, we will evaluate the the proposed method on both facial action unit detection and expression recognition tasks.

Action unit detection: we compared the proposed network with other state-of-the-art algorithms. Table 1 and Table 2 show the F_1 score matrices, which can be described as $F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recal}$, is applied to estimate the performance in AU detection. Our method outperforms other benchmark algorithms (DSIN [6], JAA [44], ARL [45], LP [37], SRERL [17], HMP-PS [51], and UGN [50]), including the attention-based models, on both BP4D and DISFA. It shows our model achieves the superior performance for AU detection, particularly achieving the best F_1 in nine individual AUs(AU1, 4, 6, 7, 10, 14, 15, 17, 23, and 24) on BP4D.

Facial expression recognition: we conducted further evaluation on the facial expression recognition task. Table 3 and Table 4 demonstrate that both SMA-Net and SMA-ViT achieves competitive performance in terms of the accuracy against the state-of-the-art models, including HOG 3D [16], STM-Explet [28], IACNN [35], DTAGN [14], MSFLBP [65], FMPN [4], DeRL [10], Berretti's model [41], Yang's work [60], Attention CNN [20], Lo's work [18], Lopes's work [32], FERAtt [62], SEnet18 [11], and CBAM [58] on four datasets.

TABLE 3: Accuracy of the state-of-the-art models on CK+ and MMI. Bold numbers indicate the best performance. Numbers with underline means sub-optimal performance.

CK+	Data	Accuracy	**MMI**	Data	Accuracy
HOG 3D	sequence	91.44	HOG 3D	sequence	60.89
STM-Explet	sequence	94.19	STM-Explet	sequence	75.12
IACNN	image	95.37	IACNN	image	70.24
DTAGN	sequence	97.25	DTAGN-Joint	sequence	71.55
MSFLBP	image	99.12	-	-	-
FMPN	image	98.06	FMPN	image	82.74
Attention CNN	image	98.68	-	-	-
DeRL	image	97.30	DeRL	image	73.23
ResNet18	image	97.67	ResNet18	image	75.62
ViT base	image	96.82	ViT base	image	76.20
Our SMA-Net	image	**99.17**	**Our SMA-Net**	image	82.75
Our SMA-ViT	image	**99.17**	**Our SMA-ViT**	image	**83.26**

TABLE 4: Accuracy of the state-of-the-art models on BU3DFE and BP4D+. Bold numbers indicate the best performance. Numbers with underline means sub-optimal.

BU3DFE	Data	Accuracy	**BP4D+**	Data	Accuracy
Berretti et al.	3D	77.54	-	-	-
Yang et al.	3D	84.80	-	-	-
Lo et al.	2D + 3D	**86.32**	-	-	-
Lopes	2D	72.89	-	-	-
FERAtt	2D	85.15	-	-	-
DeRL	2D	84.17	DeRL	2D	81.39
-	-	-	SEnet18	2D	94.25
-	-	-	CBAM	2D	94.20
ResNet18	2D	84.16	ResNet18	2D	93.74
ViT base	2D	83.67	ViT base	2D	93.19
Our SMA-Net	2D	85.42	**Ours SMA-Net**	2D	95.41
Our SMA-ViT	2D	85.15	**Ours SMA-ViT**	2D	**96.63**

4.4 Ablation study

In this section, we investigate the effectiveness of each module in the proposed model. Table 5 presents the results of ablation study in terms of the F_1 score on BP4D and DISFA. The baseline model is ResNet-18.

Effect of Key modules: As shown in Table 5, the average F_1 score increases from 60.8 to 62.7 by adding the F2A module to the baseline ResNet-18 model. The "multi-channel" design in Table 5 refers to applying multiple spatial attention with average-pooling based initialization without diversifying the pattern of attention branches. This result indicates that without diversifying the attention initialization, the performance of the multi-head design may be weakened to some extent. The Attention-Above-Attentions (AAA) module refines the model by re-calibrating the entire multi-attention system. However, we argue that the AAA module may miss out on spatial information that is more sensitive to facial action analysis. Nevertheless, by combining the F2A and AAA modules, our proposed model achieves a remarkable performance improvement over the baseline, with an increase of 5.2 F_1 score (3.6 on BP4D and 6.7 on DISFA). For a fair comparison, We also applied the design that splitting 12 AU-related attention (F2A+AAA+local-AU) to each corresponding AU respectively for finer local region learning. According to [46]. We split 12 attention branches from deep features I. Each independent attention branch is employed to attend the local AU features and supervised by a single AU ground-truth. However, the results show that the AU-related local attention design does not surpass our proposed model under the same circumstance. We believe this is due to the lack of a robust attention learning mechanism. Specifically, although the conventional multi-head structure allows the model to pay attention to different parts, it cannot guarantee that each branch has strong executive power. In contrast, our proposed model avoids this issue by querying AU-specified attention from multiple patterns.

Effect of Diversification Mechanism: The lower section of Table 5 demonstrates the impact of the proposed Internal Diversification (ID) on our model. In comparison to External Diversification (ED), which enhances the distance between AU-specific multi-attention methods, Internal Diversification leads to improved performance, with an average increase of 1.5 F_1 score when compared to ED.

4.5 Investigation of attention heads number

To investigate the impact of branch number on recognition performance, we visualize the performance curve for varying values of N values as shown in Fig. 8. The results show

TABLE 5: Ablation study on BP4D and DISFA. Bold numbers indicate the best performance. ED indicates "External Diversification" which is equivalent to AU-specific local attention methods, while ID indicates the proposed "Internal Diversification" method.

Method	BP4D	DISFA	Avg
Baseline	60.8	53.6	57.2
multi-channel	62.3	56.5	59.4
F2A	62.7	57.9	60.3
AAA	61.2	54.1	57.7
multi-channel+AAA	62.5	57.6	60.1
F2A+AAA	**64.4**	**60.3**	**62.4**
F2A+AAA+local-AU	63.3	59.7	61.5
Ours+ED	63.7	60.9	62.2
Ours+ID (SMA-Net)	65.0	**62.4**	**63.7**
Ours+ID (SMA-ViT)	**65.2**	62.2	**63.7**

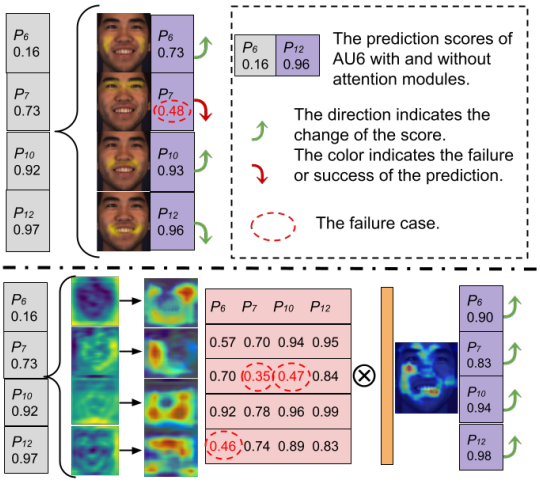


Fig. 7: Prediction score and visualization using different methods. The upper part is the conventional multi-attention for AU detection. In the lower part, we focus on AU6, AU7, AU10, and AU12 with positive ground truth, and show the results for 4 out of the 7 attention channels used. By re-calibrating the prediction scores from multi-channels, our method effectively suppresses failure cases that are challenging for previous multi-attention models.

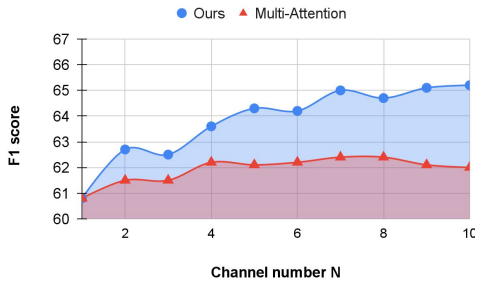


Fig. 8: Comparison of the AU detection performance between SMA-Net and Vanila Multi-Attention networks by increasing the channel number N on BP4D.

that performance generally improves with an increase in N until it reaches a certain threshold. However, we observed that a high value of N may induce saturation of attention diversity, resulting in overlapping attention maps that could dominate the model training and omit valuable information from other branches. Nevertheless, our proposed module effectively mitigates this side effect of attention saturation. As illustrated in Fig. 8, the network with embedded SMA exhibits delayed in the onset of saturation compared to vanilla multi-attention, manifesting at a higher value of N. In addition, SMA-Net surpasses regular multi-attention in achieving a higher upper-bound in terms of recognition performance.

4.6 Result and Discussion of Multi-modal Tasks

In this section, we evaluate if the proposed internal attention diversification mechanism can empower the adaptation ability of conventional attention in multi-modal feature fusion tasks. We compare our approach with the baseline attention modules from standard ViT transformer and vanilla SE-Net attention module. We also compare with some state-of-the-art multi-modal methods: MTUT [1], TEMT-Net [30], and AMFT [64]. The results are reported in Table 6 and Table 7. The multi-head attention design with ID (Internal Diversification) mechanism achieves the best performance of 65.3% F1-score on BP4D, 64.7% on BP4D+, and 53.5% on BP4D++, which is 2.2%, 2.2% and 1.6% higher than the baseline model without using ID. It proves that the proposed mechanism can be generically and effectively applied to different deep fusion works of multiple modalities including 2D texture images, 3D depth maps, thermal images, and physiological data. The two attention frameworks do not show significant performance gaps in the process of deep feature fusion tasks, which indicates that compared with the MLP-based SE attention framework, the residual design and positional embedding of transformer-based method shows subtle advantage for learning and fusing high-abstract features. Note that the uni-modal result using physiological data is not reported, for it is incapable of detecting facial AUs independently due to missing certain valid information and containing some inaccurate signal for calibration during the data collection process. Thus, the physiological signal, as the peripheral information, can only be used as complementary features for benefiting the multi-modal spontaneous facial action analysis.

4.7 Visualization and Failure Cases

Fig.7 illustrates how our method improves the prediction ability of attention learning in AU detection. The conventional multi-attention method, in Fig.7 (up), disentangles a face into several AU-related branches and use partial attention to learn discriminative features. However, the prediction of each AU highly relies on the one-dimensional confidence table due to one-to-one mapping from attention to specific AU. This figure shows it achieves high confidence for AU6, AU10, and AU12 but gets a failure prediction for AU7. This is caused by the uncertainty of the attention learning mechanism. There is no additional measures to restore or cure a single attention branch with low quality. Fig.7 (down) shows our method. We expand the one-dimensional prediction confidence table to a two dimensional one (the pink matrix). With the diverse representations of attention

TABLE 6: Comparison with state-of-the-art methods using F1 score in terms of individual AUs. The upper part is the F1 score on BP4D; The bottom part is the F1 score on BP4D+. Bold numbers indicate the best performance. "V" indicates 2D visual RGB images. "D" indicates 3d depth maps generated from 3D data. "T" indicates thermal (heat-map) images.

Model	Modalities	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.
ResNet18	V	48.3	45.7	57.6	77.7	74.4	81.5	85.9	63.8	48.3	58.2	44.7	43.8	60.8
ResNet18	D	44.5	47.3	54.7	77.2	74.4	83.6	87.8	59.1	53.1	60.6	42.9	36.1	60.1
MTUT	V, D	51.2	50.2	62.2	77.2	71.7	83.8	88.2	61.4	54.4	57.9	45.8	42.2	62.2
TEMT-Net	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	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
Multi-head ViT	V, D	48.4	44.2	56.9	78.3	78.0	82.0	88.0	60.5	54.6	62.3	52.5	50.4	63.0
Multi-head SE	V, D	46.3	43.3	59.5	76.5	**80.4**	80.3	88.5	61.7	55.1	63.3	51.1	50.9	63.1
Multi-head ViT + ID	V, D	**55.5**	47.1	59.1	**79.4**	76.2	84.2	**88.6**	60.5	56.6	64.0	**53.1**	54.8	64.9
Multi-head SE +ID	V, D	51.3	46.8	60.7	79.3	79.2	84.3	88.3	64.9	55.1	**65.2**	51.3	**56.7**	**65.3**

Model	Modalities	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	Avg.
ResNet18	V	45.3	47.1	24.7	83.7	87.3	87.8	86.1	79.2	46.5	36.4	53.1	61.5
ResNet18	T	35.2	33.5	37.7	83.1	86.9	88.6	87.9	77.2	41.4	45.5	54.3	60.0
MTUT	V, T	42.6	41.7	36.8	83.3	88.3	89.7	87.6	80.1	48.5	44.8	53.2	63.3
AMF	V, T	43.2	**42.3**	34.8	85.2	87.4	89.1	89.3	**82.4**	47.3	45.1	53.9	63.6
Multi-head ViT	V, T	44.4	37.9	29.9	85.1	89.0	90.4	88.2	81.5	44.3	44.8	51.8	62.5
Multi-head SE	V, T	43.6	37.2	31.7	85.4	**89.6**	90.7	88.9	81.2	48.7	43.6	53.0	63.1
Multi-head ViT + ID	V, T	**51.4**	**42.3**	**36.9**	85.4	89.1	**91.2**	88.8	81.7	44.6	**45.4**	54.8	**64.7**
Multi-head SE +ID	V, T	47.9	42.2	32.9	**85.7**	89.4	86.4	**89.8**	82.1	**48.9**	44.7	**58.2**	64.6

TABLE 7: Comparison with other methods using F1 score in terms of individual AUs on BP4D++. Bold numbers indicate the best performance. "V" indicates 2D visual RGB images. "P" indicates physiological signal.

Model	Modalities	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU23	Avg.
ResNet18	V	37.1	27.1	33.2	84.5	22.6	85.8	31.8	75.7	47.6	56.7	50.2
ResNet18 + RNN	V, P	28.2	26.0	35.4	84.2	21.6	86.4	36.4	76.2	47.0	63.3	50.5
SE-Net18 + RNN	V, P	27.9	31.7	34.1	83.0	24.9	82.4	35.7	77.9	46.3	62.3	50.6
Multi-head ViT	V, P	34.9	29.4	37.4	**85.0**	31.4	86.9	36.5	**79.8**	53.4	45.1	52.0
Multi-head SE	V, P	34.6	**33.6**	35.3	84.1	28.0	85.0	**39.1**	78.1	48.3	53.3	51.9
Multi-head ViT + ID	V, P	35.4	33.4	38.6	84.4	23.9	86.3	36.0	76.8	**54.9**	**63.6**	53.3
Multi-head SE +ID	V, P	**44.9**	33.1	**39.7**	84.4	**32.7**	**87.1**	30.5	78.6	50.4	53.8	**53.5**

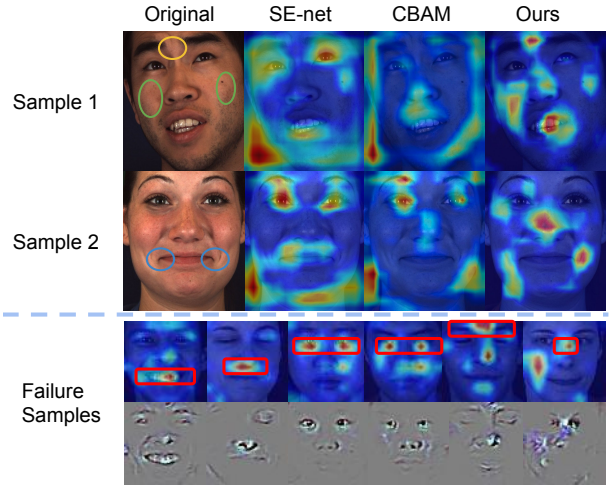


Fig. 9: Visualization of attention maps on BP4D. Each column shows the results using different models. The first row shows the original images. Yellow: AU4; Green: AU6; Blue: AU12, AU14, AU15

maps, our model can selectively choose which channel is more convincing and which one is not. Thus, the failure cases in the pink matrix has been effectively suppressed.

To further elaborate on how the proposed approach affects the recognition task, we provide a qualitative analysis. We apply the Grad-GAM [43] to visualize the attention maps of two subjects in BP4D. We compare it with several attention models on BP4D for AU detection in Fig.9. Through observation, we found our method can concentrate more on AU-related regions. For AU4 (Brow Lowerer) and AU6 (Cheek Raiser) in sample 1 and AU12 (Lip Corner Puller), AU14 (Dimpler), and AU15 (Lip Corner Depressor) in sample 2, our model performs better in focusing on the target area accurately than the channel-wise attention model (SE-Net) and Mixture attention model (CBAM). Besides, unlike the baseline attention models, our model alleviates the irrelevant regions(e,g. background area and pupil area) of the face in column two and column three. We argue that irrelevant areas can cause negative interference to facial parts localization and image detection.

In Fig. 9, we also visualize some interesting failure cases of our attention model. The attention model focus on some unrelated areas or actions (e.g., mouth or eyes closing and opening). This is caused by the lack of human-defined guidance for attention learning. However, it is worth noting that the inability to focus on the target region does not necessarily indicate an incorrect prediction. These seemingly unrelated areas may still be informative for inferring the targets in an indirect manner. For example, teeth are not directly related to AU6 or 12, but smiling is often accompanied by showing teeth. This inspires us to explore the impact of these "non-correlated" regions on face behavior prediction in future work. Despite not being human-understandable, these regions could still offer crucial contextual cues for comprehending facial actions and expressions. Further investigation into the role of these "non-correlated" regions in face behavior prediction can lead to new insights and potentially improve the performance of attention models.

4.8 Complexity Evaluation

In this section, we study to quantize the complexity of the proposed module in terms of parameter scale, training

TABLE 8: Complexity evaluation for different SMA variants versus baseline models. M indicates million, SPS indicates samples per seconds, and MB means megabyte. MM indicates multi-modal.

Model	Parameter(M)	Training speed(SPS)	Memory cost(MB)
A1. SMA-Net	11.9	50.3-63.0	1,161
A2. ResNet18	11.2	213.2-322.7	1,105
B1. SMA-ViT	89.5	11.0-16.5	7,984
B2. ViT base	86.9	47.9-53.6	7,653
C1. MM SMA-ViT	24.5	23.5-34.1	1296
C2. MM ViT base	23.7	57.2-66.0	1253
D1. MM SMA-SE	22.9	29.9-37.4	1259
D2. MM SE-Net	22.4	63.8-79.1	1235

speed, and memory cost. The backbone models in Tab. 8 includes: (A2) ResNet18; (B2) Base version ViT; (C2) Two-stream ResNet18 with Transformer feature fusion; (D2) Two-stream ResNet18 with channel-wise attention (SE-Net) fusion. All other SMA-based models from A1 to D1 are variants that are derived from the backbone networks by merging Internal Diversification features. To ensure fair comparison, we train all the models on a GTX 2080Ti GPU using samples from the BP4D dataset. The input image size is set to 224x224. We use a mini-batch size of 8 for uni-modal AU detection, and a mini-batch size of 4 for multi-modal learning (combining RGB texture images and thermal images).

The results in Tab. 8 indicate that the proposed module does not suffer from a bottleneck in terms of parameter size and memory cost. The parameters added by the SMA variants are only 3% to 6% compared to the backbone networks, and the marginal increase in memory cost ranges from 2% to 5%. This shows SMA variants have a clear advantage in terms of parameter-efficiency and memory-efficiency over popular backbone models such as ResNet-50, Inception-v4 [52], ResNeXt-101 [12], EfficientNet-B4 [61], ViT-H/14 [7] and NFNet-F0 [3].

The training speed, as shown in Tab. 8, refers to the number of images that the machine can process per second during the training stage. There is a variation in the training speed reduction across different SMA variants, which can be attributed to the dense calculation involved in estimating attention distances in Eq. 3. The results in Tab. 8 are obtained by calculating the diversity loss without dimension reduction of diversified attention. To address this issue, a simple linear layer $f(\cdot)$ is introduced in Fig. 3 to reduce the dimension of attention vectors v. This helps to reduce the speed gap to less than 30% without sacrificing predictive performance. However, the dimension reduction should be carefully chosen, as there exists a trade-off between computational cost and accuracy. It is important to note that Eq. 3 is not used during the inference stage, so there is no significant disparity in the test speed between the models.

5 CONCLUSIONS

In conclusion, this article addresses the robustness issue of attention-based learning through a self-diversified multi-channel attention approach that introduces an internal diversification mechanism. We integrate this approach with Transformer-based and CNN-based attention designs for various facial recognition tasks, showcasing its versatility and applicability. Besides, our approach demonstrates effectiveness and flexibility in improving the adaptive power of Transformer-based and Channel-wise attention designs in multi-modal feature fusion tasks, contributing to the advancement of facial action analysis research.

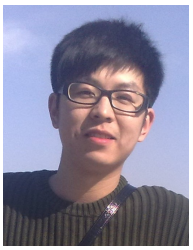
6 ACKNOWLEDGMENTS

The material is based on the work supported in part by the NSF under Grant CNS-1629898, Grant CNS-1629716, Grant CNS-1629856 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

REFERENCES

[1] M. Abavisani et al. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *CVPR*, 2019.
[2] K. Ahmed et al. Weighted transformer network for machine translation. *CoRR*, 2017.
[3] A. Brock, S. De, S. L. Smith, and K. Simonyan. High-performance large-scale image recognition without normalization. *arXiv*, 2021.
[4] Y. Chen et al. Facial motion prior networks for facial expression recognition. In *VCIP*, 2019.
[5] J.-B. Cordonnier, A. Loukas, and M. Jaggi. Multi-head attention: Collaborate instead of concatenate. *arXiv*, 2021.
[6] C. Corneanu et al. Deep structure inference network for facial action unit recognition. In *ECCV*, September 2018.
[7] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint:2010.11929*, 2020.
[8] M. Egger, M. Ley, and S. Hanke. Emotion recognition from physiological signal analysis: A review. *ENTCS*, 343:35–55, 2019.
[9] O. Ertugrul et al. D-pattnet: Dynamic patch-attentive deep network for action unit detection. *Frontiers in Computer Science*, 2019.
[10] M. Fernandez et al. Feratt: Facial expression recognition with attention net. In *CVPR Workshops*, 2019.
[11] J. Hu et al. Squeeze-and-excitation networks. In *CVPR*, 2018.
[12] Y. Huang et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv preprint:1811.06965*, 2019.
[13] R. Irani et al. Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition. In *CVPR Workshops*, pages 88–95, 2015.
[14] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *ICCV*, pages 2983–2991, 2015.
[15] K. H. Kim, S. W. Bang, and S. R. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427, 2004.
[16] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.
[17] G. Li et al. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*, 2019.
[18] H. Li et al. An efficient multimodal 2d + 3d feature-based approach to automatic facial expression recognition. *Computer Vision and Image Understanding*, 140:83–92, 2015.
[19] H. Li et al. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, 19(12):2816–2831, 2017.
[20] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju. Attention mechanism-based cnn for facial expression recognition. *Neurocomputing*, 411:340–350, 2020.
[21] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.
[22] W. Li et al. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. 2017.
[23] X. Li, Z. Li, H. Yang, G. Zhao, and L. Yin. Your "attention" deserves attention: A self-diversified multi-channel attention for facial action analysis. In *FG*, 2021.
[24] X. Li, X. Zhang, T. Wang, and L. Yin. Knowledge-spreader: Learning facial action unit dynamics with extremely limited labels. *arXiv*, 2022.
[25] X. Li, X. Zhang, H. Yang, W. Duan, W. Dai, and L. Yin. An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis. In *FG*, 2020.
[26] Z. Li, X. Deng, X. Li, and L. Yin. Integrating semantic and temporal relationships in facial action unit detection. In *ACM MM*, 2021.
[27] X. Liang et al. Patch attention layer of embedding handcrafted features in cnn for facial expression recognition. *Sensors*, 2021.
[28] M. Liu et al. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, 2014.

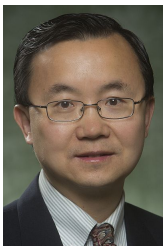
[29] M. Liu et al. Deeply learning deformable facial action parts model for dynamic expression analysis. In *ACCV*, 2015.
[30] P. Liu, Z. Zhang, H. Yang, and L. Yin. Multi-modality empowered network for facial action unit detection. In *WACV*, 2019.
[31] Z. Liu et al. Relation modeling with graph convolutional networks for facial action unit detection. In *MultiMedia Modeling*, 2020.
[32] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 2017.
[33] P. Lucey et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010.
[34] S. M. Mavadati et al. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 2013.
[35] Z. Meng et al. Identity-aware convolutional neural network for facial expression recognition. In *FG*, 2017.
[36] A. Mollahosseini et al. Going deeper in facial expression recognition using deep neural networks. In *WACV*, 2016.
[37] X. Niu et al. Local relationship learning with person-specific shape regularization for facial action unit detection. In *CVPR*, 2019.
[38] M. Pantic et al. In *Multimedia and Expo*, 2005.
[39] A. Prakash et al. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021.
[40] O. Russakovsky et al. Imagenet large scale visual recognition challenge. *IJCV*, 2014.
[41] S, Berretti et al. A set of selected sift features for 3d facial expression recognition. In *CVPR*, pages 4125–4128, 2010.
[42] S, Happy et al. Automatic facial expression recognition using features of salient facial patches. *IEEE TAC*, 2015.
[43] R. R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.
[44] Z. Shao et al. Deep adaptive attention for joint facial action unit detection and face alignment. In *ECCV*, September 2018.
[45] Z. Shao et al. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing*, 2019.
[46] Z. Shao et al. JÂa-net: Joint facial action unit detection and face alignment via adaptive attention. *IJCV*, 2020.
[47] Z. Shao et al. Spatio-temporal relation and attention learning for facial action unit detection. 2020.
[48] T. Shen et al. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*, 2018.
[49] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.
[50] T. Song, L. Chen, W. Zheng, and Q. Ji. Uncertain graph neural networks for facial action unit detection. *AAAI*, pages 5993–6001, 2021.
[51] T. Song et al. Hybrid message passing with performance-driven structures for facial action unit detection. In *CVPR*, 2021.
[52] C. Szegedy et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.
[53] E. Sánchez-Lozano et al. Joint action unit localisation and intensity estimation through heatmap regression. In *BMVA*, 2018.
[54] Y.-H. H. Tsai et al. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019.
[55] A. Vaswani et al. Attention is all you need. *CoRR*, 2017.
[56] F. Wang and D. M. J. Tax. Survey on the attention based RNN model and its applications in computer vision. *arXiv preprint:1601.06823*, 2016.
[57] X. Wang et al. Towards universal object detection by domain attention. In *CVPR*, 2019.
[58] S. Woo et al. Cbam: Convolutional block attention module. In *ECCV*, 2018.
[59] S. Wu et al. In *Deep Facial Action Unit Recognition from Partially Labeled Data*, 2017.
[60] X, Yang et al. Automatic 3d facial expression recognition using geometric scattering representation. In *FG*, pages 1–6, 2015.
[61] S. Xie et al. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.
[62] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *CVPR*, pages 2168–2177, 2018.
[63] H. Yang et al. Exploiting semantic embedding and visual feature for facial action unit detection. In *CVPR*, 2021.
[64] H. o. Yang. Adaptive multimodal fusion for facial action units recognition. In *ACM Multimedia*, pages 2982–2990, 2020.
[65] S. Yasmin et al. Development of a robust multi-scale featured local binary pattern for improved facial expression recognition. *Sensors*, 20(18), 2020.
[66] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3d facial expression database for facial behavior research. In *International Conference on FGR*, 2006.
[67] X. Zhang and L. Yin. Multi-modal learning for AU detection based on multi-head fused transformers. In *FG*, 2021.
[68] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 2014.
[69] Y. Zhang et al. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.
[70] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, 2016.
[71] K. Zhao et al. Deep region and multi-label learning for facial action unit detection. In *CVPR*, pages 3391–3399, 2016.



Xiaotian Li received his MS degree in Computer Science department from Binghamton University, New York State, USA, in 2019. He is currently working in the Graphics and Image Computing (GAIC) Laboratory of SUNY Binghamton as a PhD candidate. His research interests include computer vision, machine learning, pattern recognition, multi-modal signal processing, affective computing, relational learning, self-supervised learning, and video understanding.

Qiang Ji received his Ph.D degree in electrical engineering from the University of Washington. He is currently a Professor with the Department of Electrical, Computer, and Systems engineering at RPI. From January, 2009 to August, 2010, he served as a program director at the National Science Foundation, managing NSF's machine learning and computer vision programs.

Jeffrey Cohn is Professor of Psychology, Psychiatry, and Intelligent Systems at the University of Pittsburgh, Adjunct Professor of Computer Science at the Robotics Institute at Carnegie Mellon University, and Chief Scientist and Co-Founder of Deliberate.ai. He leads interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis and synthesis of facial expression and prosody and applies those tools to research in human emotion, social development, nonverbal communication, psychopathology, and biomedicine.



Lijun Yin received the Ph.D. degree in computer science from the University of Alberta, Canada, and the master's degree in electrical engineering from Shanghai Jiao Tong University, China. He is a Professor of computer science, the Director of the Center for Imaging, Acoustics, and a Perception Science with Binghamton University, the Director of Graphics and Image Computing Laboratory, and the Co-Director of Seymour Kunis Media Core, T. J. Watson School of Engineering and Applied Science, State University of New York at Binghamton. His research has been funded by the NSF, AFRL/AFOSR, NYSTAR, and the SUNY Health Network of Excellence.