ReactioNet: Learning High-order Facial Behavior from Universal Stimulus-Reaction by Dyadic Relation Reasoning

Xiaotian Li Taoyue Wang Geran Zhao Xiang Zhang Xi Kang Lijun Yin
State University of New York at Binghamton
{xli210, twang61, gzhao10, zxiang4, xkang3, lyin}@Binghamton.edu

Abstract

Diverse visual stimuli can evoke various human affective states, which are usually manifested in an individual's muscular actions and facial expressions. In lab-controlled emotion datasets, such a critical component (i.e., stimulus) was commonly designed in a limited way, making researchers incapable of generalizing the universal correlation and causation of stimulus-reaction as well as predicting possible emotions from context, timing, and relation. In this paper, we collected a large-scale spontaneous facial behavior database ***ReactioNet****, which contains 1.1 million coupled stimulus-reaction tuples (visual/audio/caption from both stimuli and subjects). We introduce a new facial behavior detection scenario,* ***Dyadic Relation Reasoning*** *(DRR), which aims to detect facial actions by reasoning their relations with stimuli. By aggregating the dyadic information, our method essentially forms a relation prototype* ***Universal Stimulus Reaction*** *(U-SR), which encodes the low-order and high-order relationships between stimulus agents and facial reactions. A framework with both non-graph and graph modules is further developed to evaluate DRR-based facial action unit detection, facial expression recognition, and scene classification. Specifically, to learn "what" arouses a facial reaction, the non-graph module associates and projects the fine-grained stimulus-reaction features into common subspaces using cross-domain contrastive learning. To learn "how" stimulus-reaction pairs are mutually related, the graph module adopts Graph Convolution Network to represent, converge, and infer the dyadic U-SR relation under two relation assumptions (i.e., homophily and heterophily [68]). Extensive experiments demonstrate the effectiveness of the proposed work. The new dataset will be available for the research community.*

1. Introduction

Human perception [10] is the transformation of external stimuli into an accessible, subjective, and reportable experience [15], which in consequence arouses human emotions

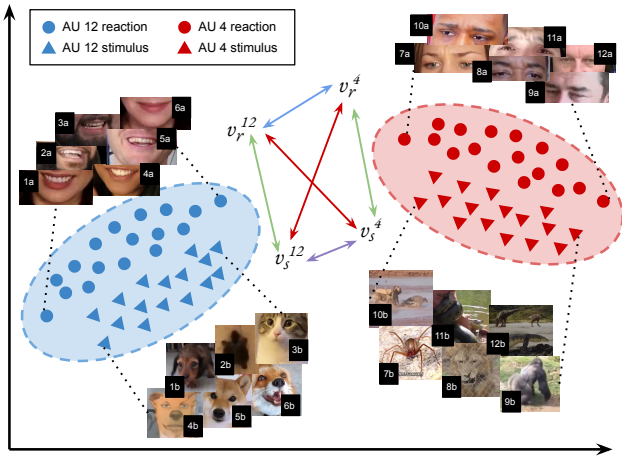

Figure 1. **A hypothetical graph of the common embedding space**. AU 12 reactions v_r^{12} are triggered by the corresponding stimuli v_s^{12} (e.g., harmless and attractive animals). AU 4 reactions v_r^4 are triggered by the stimuli v_s^4 (e.g., fierce and aggressive creatures). Arrows with distinct colors represent "four types" relations between the nodes. The samples are one-to-one corresponded.

that are reflected in behavioral responses e.g., face, gesture, and voice [23, 24]. Various emotions can be elicited by distinct stimuli, making the selection of appropriate stimuli crucial for creating affective databases. However, most current spontaneous emotion databases are restricted to lab-controlled environments without real-life experiences, such as, BP4D [77] with only 10 stimulus activities moderated by an interviewer, MPED [62] with 28 video stimuli, DEAP [26] with 40 music videos, BUEEG [37] with mixed stimuli of 6 videos, 11 still pictures, 2 physical experience, and RML [8] with 10 different emotion sentences. With limited variety of stimulus scenes, it is difficult to scrutinize the connection between external stimuli and human reactions (e.g., via face and voice), neither to generalize the connection statistically.

To overcome this limitation, we created a facial behavior database called "ReactioNet" which includes 2,486 reaction video clips and 1.1 million reaction images in the wild, along with corresponding stimuli. Each reaction video con-

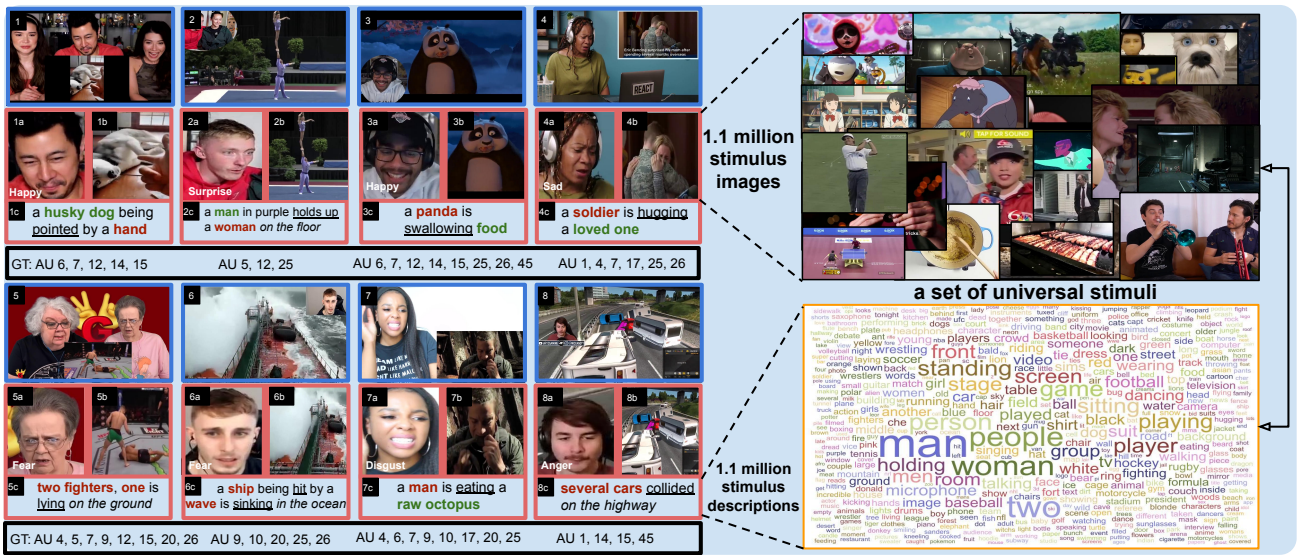

Figure 2. **A snapshot of ReactioNet.** For each sample on the left, the **top** image is the untrimmed reaction scene. The **bottom** data triplet contains: a) the subject's face; b) the stimulus scene; and c) a textual summary of the stimulus scene. GT indicates the facial action unit (AU) labels of FACS [52]. The WordCloud is generated based on the 8049 different words in textual descriptions by frequency.

tains well-synchronized dyadic data of both stimulus and subject. A "Reaction" video captures the subject's spontaneous expressions and vocalizations in response to contents on video platforms [64]. It is a unique user-generated text that reflects the overlapping complexities of the emotional, embodied, virtual, and offline performances of authenticity and social engagement [63]. We densely populate synsets in WordNet [2] according to categories of popular reaction videos for collecting a corpus with multiple scenes, including **animation**, **film**, **game**, **object**, **show**, **sports**, **interview/public speech**, and **self-made video**. Fig. 2 displays a snapshot of ReactioNet.

Three aspects distinguish ReactioNet from existing databases: **First**, it enables the learning of facial behavior associated with the explicit stimuli an individual perceives. As shown in Sample 3 of Fig. 2, the man sees a cartoon panda swallowing food. By forming a consensus between his cognition and the stimulus scene's semantics, his internal feeling "funny" is activated. This, in turn, triggers a smile reaction combined with a series of facial action units (AUs). However, without stimulus scenes as context, a smile on a man's face becomes ambiguous and less meaningful. For instance, a smile in the real world is a multi-faceted dynamic expression that can signal much more than "happy", e.g., rapport, polite disagreement, sarcasm, frustration, pain, and more [21]. Some recent context-aware work suggests to use spatial surroundings of a subject as an auxiliary clue for emotion recognition [30, 76]. However, the facial expressions of subjects in context-aware datasets are mostly non-spontaneous (e.g., posed facial expressions from actors) [48], and lack adequate quantity and diversity in terms of stimuli types, reaction types, reactions per stimuli, and synchronized annotations of both sides across multiple modalities. These issues are well addressed by the dyadic views of subjects and stimuli in our ReactioNet. **Second**, ReactioNet allows for new exploration on cross-domain tasks, such as stimulus and reaction domain adaptation [5] and generalization [81]. **Third**, it derives a new facial behavior detection scenario **Dyadic Relation Reasoning** (DRR). DRR aims to detect high-level facial behaviors by reasoning their relations with stimuli and vice versa. In low-level facial behavior detection tasks, an action on a face can be recognized, yet what causes the action and how it triggers the behavior remain unknown. The lack of such a reasoning ability makes the current studies less informative in terms of interpretability and logicality. DRR, in contrast, is intended to utilize the methods (e.g., relation reasoning [54], causality inference [74]) to parse what and how the stimulus influences the corresponding reaction. In this work, we apply DRR for facial action unit detection (AUD), facial expression recognition (FER), and scene classification (SC). Note that all diagrams presented in this paper use AUD as an example.

ReactioNet is well-grounded in multi-modal data, such as visual faces, visual stimuli, and textual descriptions of stimuli. We bring these crucial information into a DRR-based framework consisting of non-graph and graph modules. (1) To learn "what" elicits a specific reaction, the non-graph module leverages cross-domain contrastive learning to associate and map stimulus-reaction features into different AU-related embedding subspaces. As shown in Fig. 1, the AU 12 (lip corner puller) and AU 4 (brow lowerer) related stimulus/reaction are divided into two clusters separately. The stimulus photos of cute animals (e.g., pup-

pies, kittens, turtles) arouse people's desire to protect, care, and smile (AU 12). Thus, the dependency between cuteness perception [80] and AU 12 can be built. On the contrary, the fierce and aggressive creatures trigger more serious facial expressions that are often accompanied by AU 4. In this way, the non-graph module identifies the coupled regions of interest (ROIs) for stimulus-reaction pairs, and groups their feature nodes for the graph module with knowledge from only one domain (e.g., reaction domain). (2) To learn "how" stimuli and reactions are related, we present a relation prototype "**Universal Stimulus Reaction**" (U-SR). As shown in Fig. 1, it consists of one first-order Reaction-Reaction (i.e., both have the AU ground-truth) (blue), two second-order direct Stimulus-Reaction$_1$ (i.e., one may deviate from the AU ground-truth, and stimulus-reaction are directly related) (green), two second-order indirect Stimulus-Reaction$_2$ (i.e., one may deviate from the AU ground-truth, and stimulus-reaction are indirectly related) (red), and one high-order Stimulus-Stimulus (i.e., both may deviate from the ground-truth) (purple). U-SR differs from prior relation models [57, 40, 31] that only encode low-order/inter-class relations (e.g., AU co-occurred or mutually exclusive relation); instead, it builds a more complex yet robust graph by encoding both low-order and high-order relations. More details are elaborated in Sec. 3 and Fig. 3.

The contribution of this work lies in four-fold: **(1)** To our best knowledge, ReactioNet is the first spontaneous facial behavior database with well-synchronized stimulus-reaction data, diverse stimulus-reaction types, and annotations across multi-modalities. It can serve as a benchmark for broader visual/linguistic understanding applications. **(2)** We systematically investigate the relationships that exist between facial reactions and stimuli, and introduce a high-level facial behavior detection scenario that reasons about their dyadic relation. **(3)** We devise a unified framework that utilizes cross-domain contrastive learning, and a hybrid GCN with low-order/high-order relation encoded under homophily/heterophily assumptions to simultaneously addressing two questions in DRR-based tasks: a) learning which stimulus triggers a specific facial behavior; and b) reasoning how stimulus-reaction are related. **(4)** We provide three benchmarks for DRR-based facial action unit detection, expression recognition, and scene classification on ReactioNet. Extensive experiments demonstrate the generalization ability and flexibility of the proposed framework under different settings across ten existing affective datasets.

2. New database - ReactioNet

The content-wise diversity is an essential factor for generalizing the universal relation between stimulus and human reaction. Apart from inheriting some advantages of previous emotion datasets (e.g., spontaneous facial behavior [46], multi-modal data sources [79], temporal dynamics [78], large-scale data pool, diverse meta-data), ReactioNet offers novel and inspiring features including coupled data from dyadic domains, diversity and hierarchy in multi-scene [70], linguistic meta-data, and cross-domain learning supportability.

ReactioNet comprises 1168 long reaction videos (approx. 61.3 million frames) from YouTube. We carefully selected and edited 2486 short clips (around 1.1 million frames) with highlighted facial responses and unique stimulus scenes. It provides 8 types of stimulus scenes (including animation, film, game, object, show, sports, self-made video, interview/public speech) and 59 types of finer-grained sub-scenes. Multi-modal data from different domains in ReactioNet includes visual/audio/caption from stimulus, subject, and the global view. Deepface [55] is employed to roughly analyze the demographics of ReactioNet by predicting subjects' facial attributes (e.g., age, ethnicity). The proposed dataset contains around 1566 subjects with ages ranging from 20 to 70 years old. Ethnic ancestries include Asian, Black, Hispanic/Latino, Indian, Middle-Eastern, and White. A large set of metadata is created, including type tags of stimulus videos, facial landmarks, head pose tracking, gaze tracking, FACS coding, facial expression coding, and textual descriptions of stimuli. The rich meta-data in ReactioNet can potentially benefit a wide range of visual/linguistic understanding tasks, such as image captioning [29, 20], scene graph generation [14], text-video retrieval [16], and visual/textual question answering [18], etc. 50,000 key frames are sparsely selected to generate a compact data collection for getting high-quality annotations. Both AU occurrence and intensity (including AU 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, and 45), and seven primary facial expressions are manually encoded by three expert FACS coder. Deepface [55] is employed to roughly analyze the demographics of ReactioNet by predicting their facial attribution (e.g., age, ethnicity). The proposed dataset contains around 1566 subjects, with ages ranging from 20 to 70 years old. Ethnic ancestries include Asian, Black, Hispanic/Latino, Indian, Middle-Eastern, White, and others (e.g., Native American).

Three highlighted features of ReactioNet include: (1) **Coupled data from dyadic domains**. As shown in Fig. 2, the global views of samples are all based on split-screen display, which allows multiple screens to be projected onto the same one. Each reaction image usually contains one view of the stimulus scene and one (or multiple) views of the subject's faces. For each video clip, we randomly choose one major subject. For subjects' faces, we employ a semi-automatic cropping method. Initially, we manually crop the target face in the first frame and then use an Boosting-based face tracking model to extract the remaining frames. Since the location of stimulus scenes is generally fixed, no tracking method is employed after setting the first frame. By

semi-automatic cropping the subject's face and the stimulus scene, we extract one-to-one corresponding data pairs from dyadic domains. These cropped and aligned data pairs serve as crucial metadata in our dataset, as fully automatic segmentation of such complex images is impractical. (2) **Diversity and hierarchy in multi-scene**. By referring to the WordNet, Wikipedia, and other online resources, we densely populate synsets of the common video scenes to form a hierarchical keyword pool (e.g., ImageNet[7]). About 1000 keywords are collected and fed to the search engine for finding the best-matched videos. Thus, these videos can cover a large variety of reaction scenes, providing both impressive stimulus diversity and reaction diversity. Please refer to the supplementary materials for all of the fine-grained scenes and searching keywords. (3) **Textual metadata**. We generated the textual description of stimulus scenes using the caption generator from BLIP [32] with both beam search and nucleus sampling. However, the models cannot generalize well to every fine-grained scene, such as animation and game. Thus, we manually checked and corrected the generated textual descriptions to get more credible textual metadata.

The protocol of data collection, processing, and release has been approved by the Institutional Review Board (IRB). All collected videos were provided under Creative Commons license, granting us permissions and defining the terms of use, sharing, and modification. Our data collection and dissemination efforts abide by platform guidelines. Adequate caution was taken to not store any user information, videos, or metadata on permanent storage outside the computing infrastructure of the social media platform. We aim to disseminate the data upon request and log all access to the dataset, which will only be available for research purposes. Please refer to the supplementary material for more details of ReactioNet.

3. Relation prototype

In this section, we elaborate on the design philosophy of Universal Stimulus-Reaction (U-SR).

Non-graph module: U-SR builds upon insights from Contrastive Domain Adaptation (CDA) [59, 67] and Contrastive Language-Image Pretraining (CLIP) [51]. CLIP's contrastive learning on (image, text) pairs yields powerful semantic representations, but it lacks fine-grained understanding and contextual comprehension. Domain adaptation addresses this by leveraging source domain annotations for adaptation but struggles to bridge the text-image gap. To overcome these limitations, we propose triplet pairings of samples (including stimuli image, stimuli text, reaction image) under the fine-grained supervision (e.g., individual AUs) from the source domain. This bridges facial reactions with visual stimuli while capturing the rich linguistic semantics derived from the real world. In addition, unlike

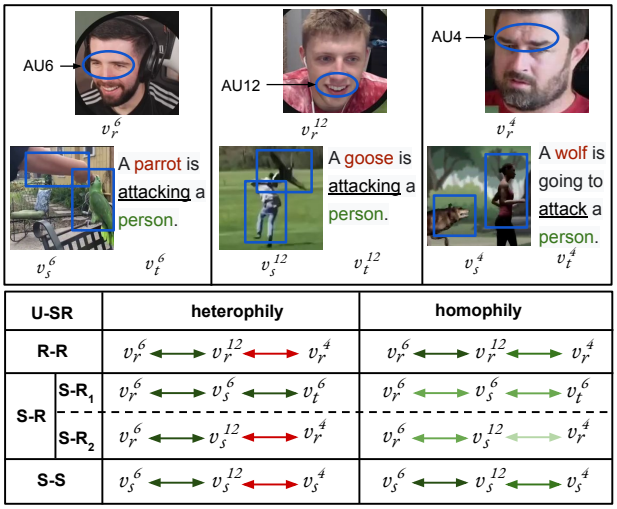

Figure 3. **Four types in U-SR relation prototype under two assumptions.** Green arrows are connected and red arrows are not. s indicates stimulus, t indicates the textual description of stimulus, and r indicates the visual reaction.

aforementioned cross-domain and multi-modal tasks where samples from source domain and target domain belong to the same or similar categories, the stimuli and facial reaction domains in U-SR exhibit lesser category-wise resemblance. The samples in target domain serve as the empirical context for inferring the predictions of source domain. Thus, we harness the potential of these triplet sample pairings to identify fine-grained ROIs that serve as reflective markers of causal relations across both the source and target domains.

Graph module: As the Non-graph module lacks ability in casual relation reasoning, U-SR incorporates graph convolutions networks with domain adaptation to derive high-order cross-domain relationships from low-order uni-domain relations. (1) **Node definition**. In Fig. 2, complicated stimulus scenes contain various entities (e.g., object, object's action, interaction between objects, and the context), making it impractical to identify specific entities as independent nodes. To simplify the graph structure and reduce the annotation cost, we treat all stimuli that elicit a behavioral response as the same node. For example, ROIs that trigger AU6 are AU6-specific stimulus nodes, and ROIs of AU6 on subjects' faces are AU6-specific reaction nodes. This allows extracting dyadic node features with labels only from one domain. (2) **Relation definition**. Stimulus-reaction data share the same supervision information but belong to different domains, resulting in features that exhibit both homogeneity and heterogeneity in the representational space. Thus, we define U-SR under two assumptions (including *homophily* and *heterophily* [68]). We explain the "four types" relation with three samples in Fig. 3: (1) "a parrot is attacking a person" triggers a smile and AU 6; (2) "a goose is attacking a person" arouses

the subject's smile and AU 12; and (3) "a wolf is going to attack a person" elicits the subject's concern and AU 4. Under the **heterophily** assumption [68], nodes with homologous semantic expressions tend to form edges, regardless of whether they are similar or not in feature-wise. The "four types" relations in U-SR are: (1) First-order Reaction-Reaction (R-R). In this case, all reaction features v_r^6, v_r^{12}, v_r^4 have the ground-truth. According to the inter-AU relation, AU 6 and AU 12 are usually co-occurred in a smiling face, whereas AU 4 is frequently found in negative expressions (e.g., concern, fear, sad). Thus, v_r^6 and v_r^{12} are connected, while v_r^{12} and v_r^4 are unconnected. (2) Second-order direct Stimulus-Reaction$_1$ (S-R$_1$). As the stimulus (i.e., a parrot is attacking a person) is the direct trigger of AU 6, we assume the visual stimulus v_s^6, textual stimulus v_t^6 and face reaction v_r^6 are all positively connected. (3) Second-order indirect Stimulus-Reaction$_2$ (S-R$_2$). The stimulus of AU 12 v_s^{12} (i.e., a goose is attacking a person) is an indirect trigger of AU 6 v_r^6. However, considering that AU6 and AU12 always appear together, they can be assumed to be related indirectly. Likewise, the visual stimulus AU 12 v_s^{12} and the reaction AU 4 v_r^4 are unrelated because AU 12 and AU 4 are mutually excluded. (4) High-order Stimulus-Stimulus (S-S). Even if all of the stimulus features may deviate the AU ground truth, their relationship can still be reasoned. For instance, both two subjects feel it funny to see the non-lethal and small-sized animals (e.g., parrot, goose) attack a people. Thus, the stimulus nodes v_s^6 and v_s^{12} that elicit AU 6 and AU 12 should be positively connected. Yet, if the attacker is a wolf, it may lead to the death of the woman, eliciting the subject's the concern expression and AU 4. In this way, the AU 4 stimulus v_s^4 and AU 12 stimulus v_s^{12} are defined to be unrelated, even if their visual semantics tend to be similar, i.e., animals are attacking people. Note that, under this assumption, all three higher-order relations are derived from the basic pre-define first-order AU relation by following Eq. (4). Under the **homophily** assumption, nodes with similar feature expressions prone to connect to each other. The homophily degree [68] describes the extent to which two nodes belong to the same class. The nodes in a homophily graph are fully connected, and takes their representational similarity as the edge weights. In Fig. 3, the intensity of green indicates the similarity of two nodes, and the examples are for reference only. In this way, the two assumptions are compatible with both semantic-wise homogeneity and feature-wise heterogeneity in order to achieve better balance in this special cross-domain task.

4. Framework

In this section, we present the details of DRR-based framework. Let $\mathcal{Q} = (\mathcal{R}, \mathcal{S}, \mathcal{T})$ denote the dataset, where $\mathcal{R}, \mathcal{S}, \mathcal{T}$ is the set of face reaction, visual stimulus, and textual descriptions of stimuli respectively. The input triplet consists of a face image $x_r \in \mathcal{R}$, a stimuli image $x_s \in \mathcal{S}$, and a stimulus text $x_t \in \mathcal{T}$. The textual description serves as a valuable source of coarse-grained prompt for effectively locating ROIs within the stimulus scenes. Our proposed framework, shown in Fig. 4, includes two modules: a non-graph module and a graph module. The non-graph module projects stimulus-reaction data into multiple AU-related embedding subspaces, each containing unique features for learning which stimulus associates with a specific AU. The graph module aims to converge the U-SR under homophily and heterophily assumptions to explore how stimuli and reactions are related.

4.1. Non-graph module

This section outlines the framework's process for encoding visual and textual features, addressing domain shifts in visual stimuli, activating local AU areas, associating them with relevant stimuli, and projecting and clustering them into shared embedding subspaces.

Visual and textual encoder The visual features of face and stimulus are extracted by the base models with sharing weights. We adopt ResNet-18 [19] pre-trained on ImageNet [53] as the visual encoder $f_{\mathcal{SR}}(\cdot)$, obtaining two feature maps from the last convolution layer. Let $z_s = f_{\mathcal{SR}}(x_s) \in \mathbb{R}^{H\times W\times C}$, and $z_r = f_{\mathcal{SR}}(x_r) \in \mathbb{R}^{H\times W\times C}$ where $H \times W \times C$ is $7 \times 7 \times 512$ in this work. We adopt a standard transformer pre-trained on CLIP [51] as the textual encoder $f_{\mathcal{T}}(\cdot)$, obtaining a feature map $z_t = f_{\mathcal{T}}(x_t) \in \mathbb{R}^C$, where C is 512.

Universal domain adaptor To adapt domain shifts across various stimuli (e.g., animation, film, game, self-made video) and reactions, we insert a domain-sensitive attention component into the visual backbone network $f_{\mathcal{SR}}(\cdot)$. Note that even stimuli from the same scene may exhibit multiple domain style. For instance, animation scenes may feature distinctive visual styles (e.g., 2D, 3D, motion graphics, cut-out). To handle this, we adopt a light-weighted attention module known as universal domain adapter (UDA) [69]. UDA employs a channel-wise domain attention that learns to assign weights to SE adapters (i.e., spatial attention modules), each of which corresponds to an underlying unknown domain pattern. The attention module dynamically increases the weight of relevant adapters based on the domain pattern of a feature, while suppressing the influence of unrelated adapters, thus inferring domain information without requiring specific domain labels. A self-diversified attention design [35, 38] is applied to increase the pattern diversity of domain adapters and avoid attention redundancy.

S-R associating To locate, associate, and align the stimulus features with local AUs, we apply the conventional multi-head design [36]. The global average pooling (GAP) is employed to flatten the feature maps z_s, z_r, and z_t. For each feature, it is duplicated and fed into N small MLP projection heads, where N represents the number of AUs. Let v_s^n,

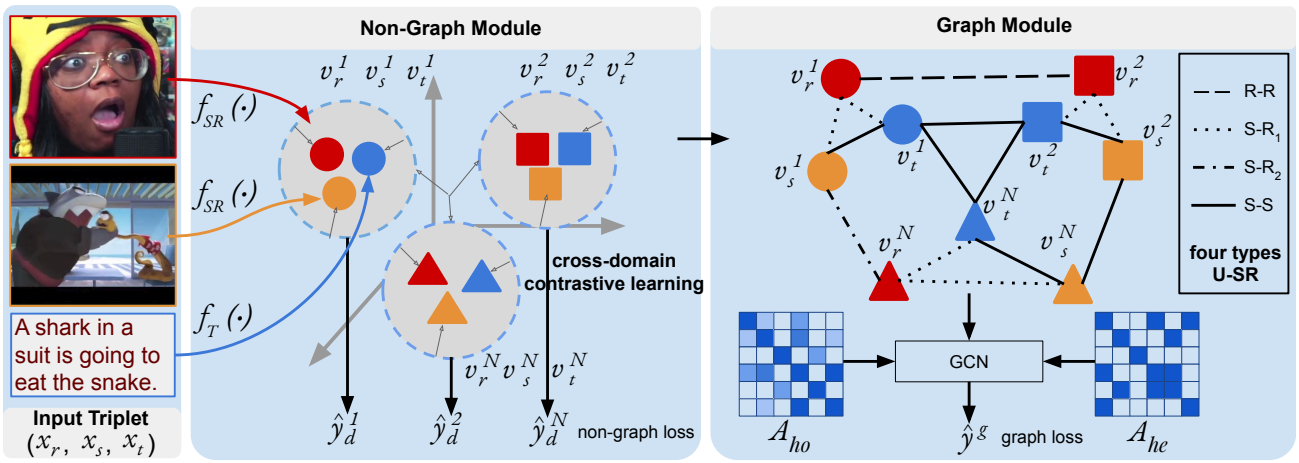

Figure 4. **Overview of the proposed framework.** The color indicates the data domain and shape means the target AU class. Four dashed lines represent four types relation encoded by U-SR. In FER and SC tasks, we replace the AU types in the framework with the corresponding label types.

v_r^n, and $v_t^n \in \mathbb{R}^C$ to be the nth AU-related feature triplet in Fig. 4, where the feature dimension C is 12. These features are forwarded to independent fully-connected layers, getting the estimated AU occurrence probability $\hat{y}_d^n$, where n is the nth AU, and d is the dth domain in $\mathcal{D} = (\mathcal{R}, \mathcal{S}, \mathcal{T})$. Each output is supervised with a binary AU label for activating the most related region of interest (ROI) in different domains. Note that each AU-related stimulus shares the same supervision information with corresponding AU-related reaction. We chose weighted BCE with logits as the multi-label classification loss. The non-graph AU loss function is defined as:

$$\mathcal{L}_{AU} = \sum_{d=1}^{D}\sum_{n=1}^{N} w_d^n [y_d^n log \hat{y}_d^n + (1 - y_d^n) log (1 - \hat{y}_d^n)] \tag{1}$$

where w_d^n is calculated by the nth AU's occurrence ratio [56] in the dth domain, and less likely occurred AU is assigned with higher weights to address imbalance issue for the multi-label classification. Note that the nth AU's occurrence ratios in different domains are the same. In this way, the stimulus-reaction features are well associated according to different AUs.

S-R common embedding space For the AU-related features triplet (v_s^n, v_r^n, v_t^n), we build cross-domain contrastive learning to further summarize and represent the common features of AU-related stimuli and reactions. To be specific, we encourage AU-related embeddings from domains $\mathcal{R}$, $\mathcal{S}$, and $\mathcal{T}$ to be close to each other, while ensuring that embeddings of different AU classes are far apart, regardless of the domain they belong to. Here, bringing the AU-related features in $\mathcal{S}$ and $\mathcal{T}$ closer is aimed at activating the corresponding stimulus ROIs through textual prompt, while encouraging the proximity of features in $\mathcal{S}$ and $\mathcal{R}$ serves to further associate mutually related regions for the stimulus-reaction pairs. Formally, we consider the mth AU related feature v_i^m from the ith domain as an anchor, and it forms a positive pair with the features v_j^m in the same AU area from other domains. For each sample in a mini-batch, the cross-domain contrastive loss is formulated as:

$$\mathcal{L}_{CDC} = -\log \frac{\exp(\text{csim}(v_i^m, v_j^m)/\tau)}{\sum_{n=1}^{N}\sum_{k=1}^{D} \mathbb{1}_{[n \neq m \forall \mathcal{A}_{n,m}=0]} \exp(\text{csim}(v_i^m, v_k^n)/\tau)} \tag{2}$$

where τ is a temperature parameter, $i, j, k \in D$ is the domain number, $m, n \in N$ is the AU number, and $\mathbb{1}_{[n \neq m, \mathcal{A}_{n,m}=0]}$ is an indicator function evaluating to 1 iff $n \neq m$ and $\mathcal{A}_{n,m} = 0$. $\text{csim}(u, p) = \frac{u^\top p}{\|u\|_2 \|p\|_2}$ denotes the cosine similarity between u and p. It is worth noting that, unlike the original NEC loss [3], we explicitly select the negative samples using an AU relation matrix $\mathcal{A}_{n,m} \in \mathbb{R}^{N \times N}$. Considering the overlapping regions among some Action Units (AUs), it's not appropriate to maximize the distance between all negative pairs without any constrains. For instance, there exists a shared area (e.g., orbicularis oculi) between AU6 cheek raiser and AU7 lid tightener. We utilize the adjacency matrix $\mathcal{A}$, established by [40], as a prior to identify and exclude related negative pairs. $\mathcal{A}_{n,m} = 1$ indicates that the two AUs are related.

4.2. Graph module

In this section, the graph module constructs the representational features in a non-Euclidean space. We encode low/high-order relations of the "four types" in U-SR. This strategic approach ensures our model versatile enough to compatibilize both homophily and heterophily.

Overview The cross-domain graph is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1^1, v_1^2, \cdots, v_D^N\}$ represents the set of $N \times D$ node features across different domains, and $\mathcal{E}$ is the

edge set representing the relationship between the nodes. The inputs of the graph module v are the AU-related features from non-graph module.

Homophily U-SR relation In this section, we utilize the feature similarity to construct a dynamic adjacency matrix that represents the homophily relation between the nodes. The dynamic graph [61] [12] is designed to capture the continuously changing of relations between individual's facial appearance and stimuli entities. The homophily adjacency matrix $\mathcal{A}_{ho} \in \mathbb{R}^{3N\times 3N}$ is formulated as:

$$\mathcal{A}_{ho} = \mathcal{A}_{i,j}^{m,n} = \frac{\text{csim}(v_i^m, v_j^n) + 1}{2} \tag{3}$$

where $i, j \in D$, $m, n \in N$. The homophily adjacency matrix is dynamically updated in each mini-batch, and each edge is assigned with a weight (i.e., the homophily degree) that indicates the strength of the node connection.

Heterophily U-SR relation defines the static method of extending low-order pre-defined relations to high-order relations in U-SR. The heterophily adjacency matrix $\mathcal{A}_{he} \in \mathbb{R}^{3N\times 3N}$ is formulated as:

$$\mathcal{A}_{he} = \mathcal{A}_{i,j}^{m,n} = \begin{cases} 1 & \text{if } \mathcal{A}_{m,n} = 1 \\ 0 & \text{if } \mathcal{A}_{m,n} = 0 \end{cases} \tag{4}$$

where the basic adjacent matrix $\mathcal{A}$ is pre-defined using the widely adopted probability statistics [40] and FACS definition. It serves as the initial inter-AU relation for formulating the extended matrix $\mathcal{A}_{he}$. $\mathcal{A}_{m,n}$ indicates the relation between AU m and AU n, with 1 meaning related and 0 denoting unrelated.

GCN encoder The homophily and heterophily relation are encoded with the two-layer GCN $f_{\mathcal{G}}$ with sharing weights. The lth layer of the two branch GCN is denoted as:

$$Z_{ho}^{l} = \text{ReLU}(A_{ho}^{l-1} Z_{ho}^{l-1} W^{l-1}) \tag{5}$$

$$Z_{he}^{l} = \text{ReLU}(A_{he}^{l-1} Z_{he}^{l-1} W^{l-1}) \tag{6}$$

where W^{l-1} denotes the learnable weight matrix, the inputs of the first layer Z^0 are the features v_s^n, v_r^n, and v_t^n getting from non-graph module. $\text{ReLU}(\cdot)$ is the activation function. Finally, the two output features of GCN are fused to one with their mean values, and forwarded to a fully connected layer to obtain the estimated AU occurrence probability $\hat{y^g}$. The loss function of the graph module is denoted as:

$$\mathcal{L}_{\mathcal{G}} = \sum_{n=1}^{N} w_n [y_n log \hat{y}_n^g + (1 - y_n) \, log \, (1 - \hat{y}_n^g)] \tag{7}$$

The overall loss function is expressed as follow:

$$\mathcal{L} = \mathcal{L}_{\mathcal{AU}} + \lambda_1 \mathcal{L}_{\mathcal{G}} + \lambda_2 \mathcal{L}_{\mathcal{CDC}} \tag{8}$$

where λ_1 and λ_2 are the trading-off hyper-parameters. In this paper, all experiments are conducted with $\lambda_1 = \lambda_2 = 1$.

Table 1. Comparison with baselines and benchmarks in three detection tasks on ReactioNet. Bold numbers indicate best performance, and underlined numbers indicate sub-optimal results.

Model	Domain	AUD(F1)	FER(Acc)	SC(Acc)
ResNet18 [19]	$\mathcal{R}$ or $\mathcal{S}$	54.8	69.6	51.6
ResNet50 [19]	$\mathcal{R}$ or $\mathcal{S}$	55.3	70.3	52.8
ViT [9]	$\mathcal{R}$ or $\mathcal{S}$	53.7	68.5	50.5
GCN [25]	$\mathcal{R}$ or $\mathcal{S}$	55.9	71.7	52.1
Feature fusion [6]	$\mathcal{RST}$	54.6	67.8	50.7
UDA [69]	$\mathcal{RST}$	56.5	70.6	52.9
DA-GCN [17]	$\mathcal{RST}$	57.1	71.2	53.6
SEV-Net [73]	$\mathcal{RST}$	57.7	74.1	53.0
without CDC	$\mathcal{RST}$	58.9	73.9	54.5
without HO	$\mathcal{RST}$	59.1	74.3	53.2
without HE	$\mathcal{RST}$	59.3	73.5	56.1
DRR static	$\mathcal{RST}$	60.6	76.2	58.7
DRR dynamic	$\mathcal{RST}$	**61.3**	**78.5**	<u>60.2</u>
DRR dyadic	$\mathcal{RST}$	<u>61.1</u>	<u>77.9</u>	**60.9**

5. Experiment

5.1. Quantitative evaluation

-Evaluation on ReactioNet-

Our method was compared with substantial baselines and state-of-the-art (SOTA) algorithms on ReactioNet for three DRR-based tasks, including **action unit detection** (AUD), **facial expression recognition** (FER), and **scene classification** (SC). Scene classification categorizes scenes from photos based on object layout and ambient context. In our case, we use stimulus video types as labels for scene classification. Among them, AUD and FER leverage data from the reaction domain, while SC relies on data from the stimulus domain. In this study, we evaluated our approach across 12 AUs, 7 primary expressions, and 24 selected fine-grained sub-scene categories. To facilitate a comprehensive evaluation, we introduced three variants of the DRR-based model: (1) **DRR static**, a basic model for assessing a single task; (2) **DRR dynamic**, a temporal model using sequential frames; and (3) **DRR dyadic**, a bidirectional reasoning model based on multi-task learning, which includes all three tasks. DRR dyadic is achieved by simultaneously inferring AU and FE reactions induced by the stimulus scene, and inferring the stimulus type that can cause relevant reactions from the other domain. To evaluate the effectiveness and contribution of the key components in our proposed method, we conducted an **ablation study** in the lower part of Tab. 1, where we removed the following variants: (1) cross-domain contrastive learning (CDC); (2) the homophily relation module (Ho); and (3) the heterophily relation module (He). Refer to the supplementary material for more implementation details.

In Tab. 1, our model and its variants demonstrate superior quantitative performance across all three tasks. Notably, for baselines such as ResNet18, ViT, and GCN, which lack dyadic knowledge of stimulus-reaction domain, our proposed base model, DRR dynamic, outperforms these models by 6.5%, 7.6%, and 5.4% in AUD and 8.9%, 10%,

and 6.8% in FER, and achieves 8.6%, 9.7%, and 8.1% improvements in SC. While cross-domain methods such as Feature Fusion, UDA, DA-GCN, and SEV-Net can integrate Stimuli-Reaction knowledge, our model maintains an outstanding advantage in performance. These results suggest that stimulus-reaction based tasks does not rely solely on learning low-level features or representations, but also benefits from learning high-level semantic relation and effective cross-domain adaptation. The experiments also reveal that DRR dynamic and DDR dyadic improve the model's performance by enhancing the ability to learn temporal context, and the interdependent relationships among multiple tasks. Note that the performance outcomes presented for ReactioNet have been based on partial annotations. It is essential to acknowledge that the official evaluation results may undergo fluctuations in tandem with the revisions and updates to our annotation procedures.

-Evaluation on Benchmark Datasets-

Table 2. Comparison with SOTAs on ten affective datasets.

AUD (F1)	EmotioNet[11]	CK+ [42]	DISFA [46]	BP4D [77]	BP4D+ [79]
Ran[50]	-	60.7	-	-	-
EAC[34]	-	-	48.5	55.9	-
ViT[9]	43.2	58.9	58.7	60.3	59.6
Swin[41]	43.8	60.4	59.3	62.6	59.3
JAA[56]	-	-	56.0	60.0	-
HMP-PS[60]	-	-	61.0	63.4	-
SEV-Net[73]	-	-	58.8	63.9	61.5
FAUDT[22]	47.3		61.5	64.2	-
AMF[72]	-	-	-	64.4	62.7
MEF-RF[43]	-	-	**63.1**	64.7	-
Baseline GCN	42.7	57.0	56.8	59.2	58.5
DRR UM	**48.9**	64.4	62.3	64.5	62.9
DRR MM	-	-	-	**66.5**	**64.7**
DRR PR	48.5	**64.7**	62.9	64.9	63.1
FER (Acc)	FER+ [1]	RAFDB [33]	AffectNet [47]	MMI [49]	BU3D [75]
VGG-FACE[27]	-	77.5	60.0	-	-
RAN[66]	88.5	86.9	59.5	-	-
SCN[65]	88.0	87.3	63.4	-	-
ViT[9]	87.6	87.2	57.9	-	-
FMPN[4]	-	-	61.5	82.7	-
DeRL[71]	-	-	-	73.2	84.1
FERatt[45]	-	-	-	**83.2**	77.9
SMA-Net[35]	-	-	-	82.7	85.4
DMSRL[39]	-	-	52.3	72.1	61.7
DMUE[58]	88.6	88.7	-	-	-
VTFF[44]	**88.8**	88.1	64.8	-	-
Ad-Corre[13]	-	86.9	63.3	-	-
Baseline GCN	85.7	79.1	56.6	78.5	78.4
DRR UM	87.5	88.4	65.7	83.1	85.3
DRR MM	-	-	-	-	**87.4**
DRR PR	88.1	**89.6**	**65.9**	82.9	85.8

Unlike ReactioNet, existing affective datasets fall short in establishing frame-wise correspondence between stimulus-reaction, making it difficult to validate the proposed method on a comparable database. To address this limitation, we extend DRR with three settings to assess our method using only the reaction domain: (1) **DRR UM**, a basic relation model for assessing uni-modal (i.e., reaction) tasks; (2) **DRR MM**, a high-order relation model for assessing multi-modal tasks; and (3) **DRR PR**, a pre-trained model based on DRR dyadic. For multi-modal learning, we replace the stimulus data with face images from other modalities (e.g., thermal face and 3D depth). For DRR PR, we initialized only the visual encoder with pre-trained weights and fine-tuned it for downstream tasks, including AUD and FER. The pre-trained model is based on the DRR Dyadic trained on the entire database for AUD and SC.

We compared our model with state-of-the-art methods on five AUD and five FER datasets, and our DRR variants achieved the best performance on seven of ten, as shown in Tab. 2. The results of DRR PR demonstrate that transferring the high-order stimulus-reaction relation can benefit general low-level facial behavior detection tasks. The preeminent performance of DRR MM illustrates the flexibility of our proposed framework. It demonstrates that addressing issues of feature-wise homogeneity and heterogeneity, as well as learning high-order action relation can benefit not only cross-domain problems, but also multi-modal tasks. We also performed experiments on the official Aff-Wild2 [28] validation set, and achieved impressive F1 scores for both AU detection and FER respectively. Note that we leveraged both temporal and multi-modal (including visual and audio) information in the downstream task, adapting the setup of Aff-Wild2 accordingly. Since the labels for the test set are not available, only training and validation sets were used in our experiments.

-Evaluation on Fine-grained Scenes-

Table 3. Evaluate the performance of DRR based AUD in terms of scene types. S1-8 is eight basic scenes: animation, film, game, object, show, sports, interview/public speech, and self-made video.

Scenes/Models	S1	S2	S3	S4	S5	S6	S7	S8	AVG
ResNet-18	59.8	**53.9**	47.2	53.3	57.4	52.2	56.5	57.2	54.7
ViT	59.1	44.0	**50.5**	53.8	57.0	50.6	56.7	57.7	53.7
DRR Static	**61.6**	50.4	48.6	**56.6**	**58.8**	**52.4**	**59.4**	**59.4**	**55.9**
Change	↗	↘	↘	↗	↗	↗	↗	↗	↗

We evaluated our model's performance for AU detection in eight basic scenes to assess the difficulty for dyadic relation reasoning. Models were trained with 66% data samples and tested on the rest. In Tab. 3, the proposed model outperformed baselines in most cases, but faced challenges with film and sport scenes due to the lack of long-term temporal context. Unlike other scenes, where users get instant feedback, it often takes a longer time for users to react when watching a movie, which makes it difficult to establish direct correlations between stimuli and reactions in the current frame. This highlights the need to investigate the role of long-term dependencies and reaction delay in stimulus-reaction tasks in the future.

5.2. Qualitative evaluation

In this section, we explore two advantages of our model, including context-aware ability and perception of high-order stimulus-reaction relation.

Context-aware detection As depicted by sample 46 in Fig. 5 (a), our DRR model is capable of inferring a positive emotion by learning from stimulus events, such as "a man lying in bed with a cat", even if subject's face is invisible due to self-occlusion. Moreover, the model effectively

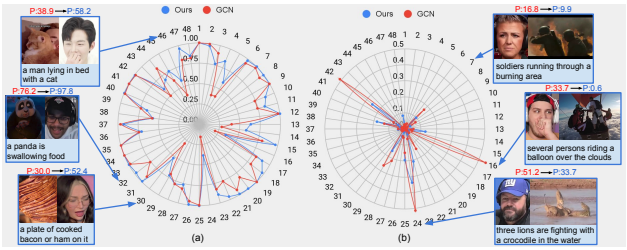

Figure 5. **Comparisons of predictions for samples in ReactioNet.** We sampled 48 test data with active AU 12 (a) and inactive AU 12 (b), and represented their estimated probabilities on radar charts. For active AU12 in (a), correct predictions are points on the periphery with values above 0.5. For inactive AU12 in (b), correct predictions are points in the center with values below 0.5.

mitigates the mis-classification of negative samples as positives, as evident in Fig. 5 (b), by inferring reactions when perceiving serious stimuli such as "soldiers running through a burning area". Overall, the DRR based method offers enhanced detection capability by interpreting reactions within the context of external knowledge.

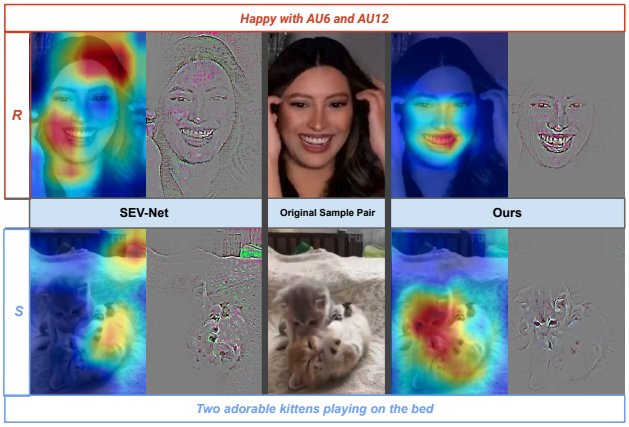

Figure 6. **Comparison of the paired attention heatmaps**.

High-order relation In Fig. 6, we visualize dyadic attention maps to compare DRR with a SOTA cross-modality attention model, SEV-Net. We observe that our method can concentrate more on semantic regions. For instance, it focuses on the activity of two kittens playing in the stimulus domain, and discerns the triggered smiling face with AU6 (i.e., the orbicularis oculi muscle that raises the cheeks) and AU 12 (i.e., the zygomaticus muscle that pulls the corners of the mouth upward and outward) on the reaction side. In contrast, SEV-Net appears to primarily encode low-order AU relations and tends to fixate on limited components within the stimulus scene (e.g., only one cat is highlighted). On the other hand, our model excels in capturing multiple meaningful regions and effectively linking relevant stimuli

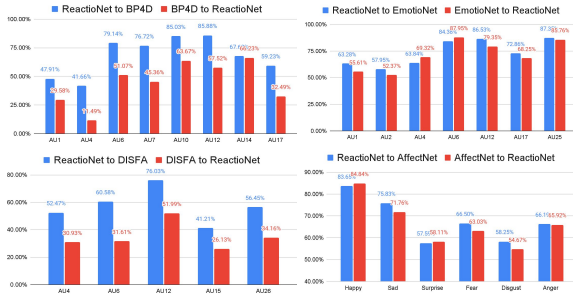

Figure 7. **Cross-database validation with benchmarks.**

with their corresponding reactions. This disparity in performance can be attributed to the high-order semantic relationships encoded by our model.

5.3. Cross-dataset evaluation

We conducted cross-database validation to assess the annotation quality of ReactioNet and its generalizability across multiple domains. Four benchmark databases, BP4D, DISFA, EmotioNet, and AffectNet, were selected for comparison using Vanilla ResNet-18. The results in Fig. 7 show that ReactioNet-to-others outperformed others-to-ReactioNet by up to 23% in F1 score and accuracy for AUD and FER. This is attributed to ReactioNet's diverse data in context, identity, individual variations (e.g., gender, ethnicity, and age), head movements, and lighting conditions. This rich variety mirrors the distribution of facial expressions in real-world scenarios and underscores the high-quality annotations.

6. Conclusion and Future work

This work introduces a large-scale human reaction database (ReactioNet) with the presence of synchronized stimuli and human reactions for the research community. Our study delves into the understanding of facial behaviors through the systematic exploration of dyadic relationships between stimuli and reactions, and forms a relational prototype encoded with low-order/high-order connections. We develop a new framework to learn "what" arouses specific facial reactions, and understand "how" the stimulus-reaction are related, and demonstrate the advancement in improving the performance over the traditional facial behavior detection approaches. We plan to perform more intricate object annotations within the stimuli domain, aiming to broaden the applicability of dyadic relation reasoning in future research.

7. 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] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016. 8

[2] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *ECCV*, September 2018. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 6

[4] Yuedong Chen, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. Facial motion prior networks for facial expression recognition. In *2019 IEEE Visual Communications and Image Processing*, 2019. 8

[5] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *CVPR*, June 2019. 2

[6] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[8] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. Rml: a generic language for integrated rdf mappings of heterogeneous data. In *Ldow*, 2014. 1

[9] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint:2010.11929*, 2020. 7, 8

[10] Eric Eich, John F Kihlstrom, Gordon H Bower, Paula M Niedenthal, Joseph P Forgas, et al. *Cognition and emotion*. Oxford University Press on Demand, 2000. 1

[11] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016. 8

[12] Yingruo Fan et al. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12701–12708, 2020. 7

[13] Ali Pourramezan Fard and Mohammad H. Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, pages 26756–26768, 2022. 8

[14] Kaifeng Gao, Long Chen, Yulei Niu, Jian Shao, and Jun Xiao. Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In *CVPR*, pages 19497–19506, June 2022. 3

[15] Howard Gardner. *Disciplined mind: What all students should understand*. Simon & Schuster, 2021. 1

[16] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, pages 5006–5015, June 2022. 3

[17] Lei Guo, Li Tang, Tong Chen, Lei Zhu, Quoc Viet Hung Nguyen, and Hongzhi Yin. Da-gcn: A domain-aware attentive graph convolution network for shared-account cross-domain sequential recommendation. In *International Joint Conference on Artificial Intelligence*, 2021. 7

[18] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *CVPR*, pages 5078–5088, June 2022. 3

[19] Kaiming He et al. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 7

[20] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, pages 13450–13459, June 2022. 3

[21] Mohammed Hoque, Louis-Philippe Morency, and Rosalind W Picard. Are you friendly or just polite?–analysis of smiles in spontaneous face-to-face interactions. In *International Conference on Affective Computing and Intelligent Interaction*, pages 135–144. Springer, 2011. 2

[22] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 8

[23] Jacob Jacoby. Stimulus-organism-response reconsidered: an evolutionary step in modeling (consumer) behavior. *Journal of consumer psychology*, 2002. 1

[24] Myung Ja Kim, Choong-Ki Lee, and Timothy Jung. Exploring consumer behavior in virtual reality tourism using an extended stimulus-organism-response model. *Journal of travel research*, 59(1):69–89, 2020. 1

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 7

[26] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011. 1

[27] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 2020. 8

[28] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 8

[29] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *CVPR*, pages 17969–17979, June 2022. 3

[30] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10143–10152, 2019. 2

[31] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 2019. 3

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 4

[33] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[34] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 8

[35] Xiaotian Li, Zhihua Li, Huiyuan Yang, Geran Zhao, and Lijun Yin. Your "attention" deserves attention: A self-diversified multi-channel attention for facial action analysis. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 5, 8

[36] Xiaotian Li, Xiang Zhang, Taoyue Wang, and Lijun Yin. Knowledge-spreader: Learning facial action unit dynamics with extremely limited labels, 2022. 5

[37] Xiaotian Li, Xiang Zhang, Huiyuan Yang, Wenna Duan, Weiying Dai, and Lijun Yin. An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 336–343. IEEE, 2020. 1

[38] Xiaotian Li, Zheng Zhang, Xiang Zhang, Taoyue Wang, Zhihua Li, Huiyuan Yang, Umur Ciftci, Qiang Ji, Jeffrey Cohn, and Lijun Yin. Disagreement matters: Exploring internal diversification for redundant attention in generic facial action analysis. *IEEE Transactions on Affective Computing*, pages 1–12, 2023. 5

[39] Yingjian Li, Zheng Zhang, Bingzhi Chen, Guangming Lu, and David Zhang. Deep margin-sensitive representation learning for cross-domain facial expression recognition. *IEEE Transactions on Multimedia*, 2022. 8

[40] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. Relation modeling with graph convolutional networks for facial action unit detection. In *International Conference on Multimedia Modeling*, pages 489–501. Springer, 2020. 3, 6, 7

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 8

[42] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 8

[43] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *IJCAI*, 2022. 8

[44] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021. 8

[45] Pedro D Marrero Fernandez, Fidel A Guerrero Pena, Tsang Ren, and Alexandre Cunha. Feratt: Facial expression recognition with attention net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 8

[46] S. M. Mavadati et al. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 3, 8

[47] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017. 8

[48] Michael T Motley and Carl T Camden. Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Communication (includes Communication Reports)*, 52(1):1–22, 1988. 2

[49] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005. 8

[50] Guozhu Peng et al. Weakly supervised facial action unit recognition through adversarial training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4, 5

[52] Erika L Rosenberg and Paul Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 2020. 2

[53] Olga Russakovsky et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2014. 5

[54] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, 2017. 2

[55] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 3

[56] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018. 6, 8

[57] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention

and relation learning. *IEEE Transactions on Affective Computing*, 2022. 3

[58] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[59] Ankit Singh. Clda: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:5089–5101, 2021. 4

[60] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8

[61] Tengfei Song et al. Dynamic probabilistic graph convolution for facial action unit intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2021. 7

[62] Tengfei Song, Wenming Zheng, Cheng Lu, Yuan Zong, Xilei Zhang, and Zhen Cui. Mped: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, 7:12177–12191, 2019. 1

[63] Anna Lee Swan. Transnational identities and feeling in fandom: Place and embodiment in k-pop fan reaction videos. *Communication Culture & Critique*, 11(4):548–565, 2018. 2

[64] Jessica Vogele. Where's the fair use: The takedown of let's play and reaction videos on youtube and the need for comprehensive dmca reform. *Touro L. Rev.*, 33:589, 2017. 2

[65] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 8

[66] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 2020. 8

[67] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022. 4

[68] Tao Wang, Di Jin, Rui Wang, Dongxiao He, and Yuxiao Huang. Powerful graph convolutional networks with adaptive propagation mechanism for homophily and heterophily. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4210–4218, 2022. 1, 4, 5

[69] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7289–7298, 2019. 5, 7

[70] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *CVPR*, pages 20922–20931, 2022. 3

[71] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[72] Huiyuan Yang, Taoyue Wang, and Lijun Yin. Adaptive multimodal fusion for facial action units recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 8

[73] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2021. 7, 8

[74] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021. 2

[75] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 8

[76] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156, 2019. 2

[77] Xing Zhang et al. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 1, 8

[78] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013. 3

[79] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, 2016. 3, 8

[80] Zhicong Zhang and Jiaxian Zhou. Cognitive and neurological mechanisms of cuteness perception: A new perspective on moral education. *Mind, Brain, and Education*, 14(3):209–219, 2020. 3

[81] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2