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

1. New database-ReactioNet

1.1. File hierarchy

ReactioNet contains 2486 short clips (around 1.1 million frames) with highlighted facial responses and unique stimulus scenes. The files of ReactioNet are organized in a hierarchical way. The root file contains 8 scenes; each scene contains several sub-scenes; each sub-scene contains several tasks; each task contains several sequences from one untrimmed original video; and each sequence contains a sequence of consecutive frames. The statistics of the files in ReactioNet are shown in Fig. 1. Note that the total number of frames is slightly higher than 1.1 million due to some invalid tasks included. The fine-grained scenes are listed in Sec. 1.8.

1.2. Multi-modal data

Fig. 2 shows the sample sequence in three modalities (including visual, audio, and caption) and two dyadic domains (including stimulus and reaction) in ReactioNet.

1.3. Metadata

Fig. 3 shows the data samples from the reaction domain. Note that the face landmarks, head poses, and eye gazes are generated by OpenFace 2.0 [1], and are further manually inspected.

1.4. Demographics

Deepface [10] 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. Tab. 1 shows the detailed ethnicity and age distribution of subjects. Note that "unknown" indicates that the results predicted by Deepface are invalid. One subject (i.e., participant) may emerge in multiple videos, and one video may contain multiple subjects.

Table 1. **Ethnicity and age distribution.**

Ethnicity	Number	Age	Number
Asian	128	20-24	115
Black	206	25-29	352
Hispanic/Latino	80	30-34	320
Indian	21	35-39	126
Middle-Eastern	95	40-44	39
White	434	45-49	9
-		50+	3
Unknown	602	Unknown	602

1.5. Annotation

Except for the manually annotated 50,000 key frames, we adopt a semi-automatic way to annotate the remaining frames of 1.1 million reaction images. Both AU occurrence and AU intensity are encoded. We first use OpenFace 2.0 [1] to extract the perdition results of 17 AUs' intensity (including AU 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, and 45) and occurrence. As human facial muscles move smoothly, facial appearance also changes smoothly overtime. As a dynamic database, the AU labels of ReactioNet also follow the temporal smoothness assumption [6, 12]. By following the principle, we checked if a frame was incorrectly annotated. The intensity difference $D_{i,j}$ between two neighbor frames should be similar:

$$D_{i,j} = \sum_{m=1}^{M} |y_{i,m} - y_{j,m}| \quad (1)$$

where y indicates the AU intensity drived by OpenFace, i and j denote the ith and jth neighbor frames, and M indicates the total number of AUs. In this paper, we set the $D_{i,j} >= 2.0$ as the threshold for detecting the suspicious frames. Similarly, we assume that the difference in AU occurrence $T_{i,j} = |C_i - C_j|$ for two neighbour frames i and j is small. Here C denotes the number of occurred AUs. In this paper, we set the $T_{i,j} >= 5$ as the threshold for picking out suspicious frames. Any suspicious frames will be further checked by the expert FACS coder.

major scenes	number of sub-scenes	number of tasks	number of sequences	number of frames
sport	14	170	426	325571
show	10	177	489	240949
animation	7	94	268	209058
film	7	89	296	127073
game	7	119	314	268120
object	6	163	473	346484
self-made	6	88	246	131246
interview/public speech	2	69	197	139747

Figure 1. **Statistics of the files' number in ReactioNet.**

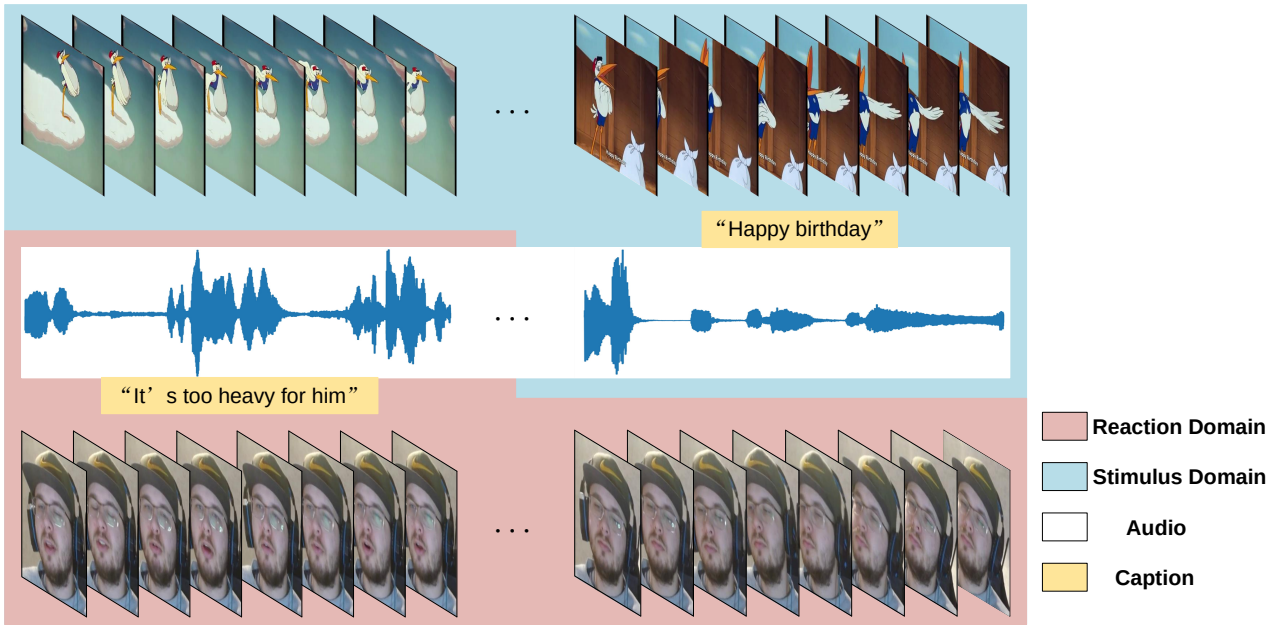


Figure 2. **The sampled multi-modal sequence from a subject in ReactioNet.** The data are from three modalities (including visual, audio, and caption) and two dyadic domains (including stimulus and reaction). Of note, the reaction data and stimulus data share one audio track.

1.6. Multi-label balance

Label imbalance is a major challenge for training multi-label classification models such as facial action unit detection. The performance bottleneck is commonly caused by the less occured AUs such as AU 1, AU 2, AU 4, AU 5, AU 9 ,AU 17, AU 20. Fig. 4 shows the proportion of frames with less occured AUs to the total frames. In general, the number of less occured AUs in ReactioNet are more sufficient than most existing benchmark datasets, providing a more balanced AU occurrence annotation. Fig. 5 shows the statistics of AUs occurrence for 50,000 key frames in ReactioNet.

1.7. Inter annotator agreement

In this work, three experienced FACS coders coded the seven facial expression, seventeen facial action units occurrence with 50,000 key frames. We recruited three graduate students from our school's Department of Psychology who have many years of work experience in fields related to face analysis to perform this task. To quantify the inter-annotator agreement, all sequences of ReactioNet coded by three coders. We report the kappa reliability of all the tasks in the following TABLE Tab. 2 as a supplement.

1.8. Searching keywords

The following is a list of the indexed entries that are manually organized for searching matched Reaction videos. The hierarchical keyword pool is built to cover the most com-

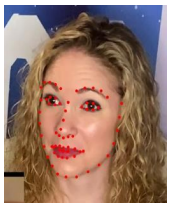
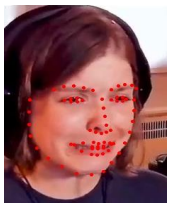
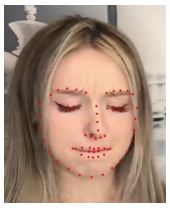
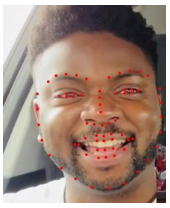
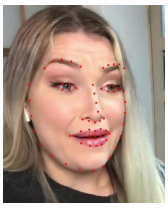
Faces with Landmarks					
Head Poses	(0.232, 0.411, 0.044)	(0.283, -0.373, 0.089)	(0.203, -0.124, 0.038)	(0.118, -0.282, -0.011)	(0.098, -0.006, -0.08)
Eye Gazes	Left (-0.188, 0.163, -0.969) Right (-0.165, 0.194, -0.967)	Left (0.246, 0.25, -0.937) Right (0.048, 0.235, -0.971)	Left (0.247, 0.345, -0.905) Right (-0.16, 0.271, -0.949)	Left ([0.287, 0.195, -0.938) Right (0.074, 0.149, -0.986)	Left (0.123, 0.173, -0.977) Right (-0.051, 0.199, -0.979)
Action Units	AU1,2,5,7,14,15	AU6,7,12,14,20,26,28	AU1,4,14,45	AU6,7,9,10,12, 14,20,23,25	AU1,2,5,10,12, 14,15,25,26
Facial Expression	Surprise	Sad	Fear	Happy	Fear

Figure 3. **The data samples with multiple metadata from the reaction domain in ReactioNet.** The metadata includes, faces with landmarks (first row), head pose estimation (second row), eye gaze estimation (third row), AU occurrence (fourth row), facial expression (fifth row).

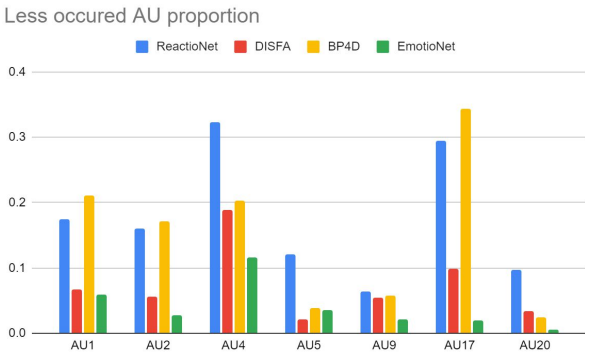


Figure 4. **The proportion of less occured AUs in the total number of frames**. Our ReactioNet provided a more balanced AU label distribution when compared to benchmarks such as BP4D [15], DISFA[9], and EmotioNet [2].

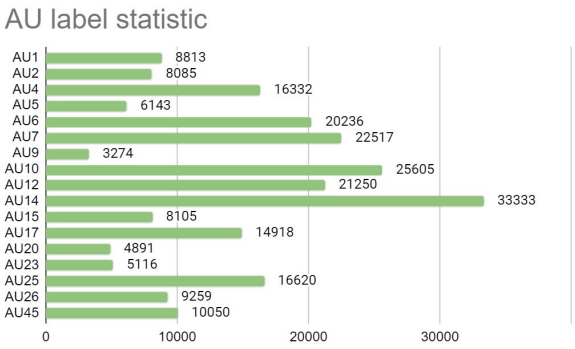


Figure 5. **Statistics of AUs occurrence in ReactioNet.**

mon reaction video categories. When searching for reaction videos, "Reaction to " was concatenated with each keyword. For some keywords (e.g., NBA), only one search can match and achieve multiple videos. Thus, limited keywords can still yield sufficient search results. Each major scene contains several fine-grained sub-scenes. Note that the number of listed sub-scenes is slightly higher than 59, as some invalid search keywords are included.

- Sport
  - Soccerball: Man City, Bayern Munich, Liverpool, Barcelona, Chelsea, Messi, C Ronaldo, Neymar, world cup, Maradona, Pele, English Premier League, Bundesliga, La Liga, UEFA.
  - Cricket: cricket, ICC, cricket world cup.
  - Hockey: hockey, NHL.
  - Volleyball: volleyball, volleyball game, beach volleyball
  - table tennis: table tennis, best table tennis player, ping pang, Ma Long
  - Basketball: NBA.
  - Baseball: Baseball.
  - MLB: MLB.
  - Golf: golf, PGA, Tiger Woods.
  - MMA: UFC, MMA, knockout, Mcgregor, Khabib.
  - Wrestling: Wrestling.
  - Boxing: Boxing.

Table 2. Descriptive statistics of kappa reliability for all tasks.

Session name	Task number	Task name	Kappa reliability
Facial expressions	1	Anger	0.90
	2	Disgust	0.84
	3	Fear	0.87
	4	Happy	0.99
	5	Sad	0.91
	6	Surprise	0.92
	7	Neutral	0.96
Facial action units	1	AU1	0.92
	2	AU2	0.84
	3	AU4	0.92
	4	AU5	0.93
	5	AU6	0.97
	6	AU7	0.93
	7	AU9	0.82
	8	AU10	0.93
	9	AU12	0.96
	10	AU14	0.95
	11	AU15	0.81
	12	AU17	0.86
	13	AU20	0.92
	14	AU23	0.79
	15	AU25	0.93
	16	AU26	0.87
	17	AU45	0.94

  - Badminton: badminton, Lin Dan, Lee Cong Wei.
  - rugby(football): NFL, rugby, Aaron Donald, super bowl game.
  - Skiing: skiing, skier.
  - Racing: racing, F1, nascar, indycar, off-road racing.
  - Cycling: cycling.
  - Race: 100m, 4X100, long-distance race.
  - Skydiving: skydiving.

- Film
  - Action: action movies, Jackie Chan, Bruce Willis, Tom Cruise, Arnold Schwarzenegger, Sylvester Stallone, Dwayne Johnson.
  - Comedy: Jim Carrey, Mr. Bean, Jordan Peele, stephen chow, comedy.
  - Drama: drama movies, drama master, drama club, K drama.
  - Fantasy: fantasy movie, marvel movies, Crouching Tiger, Hidden Dragon, DC movies, Harry Potter, Lord of the Ring, walking dead, monster movie.
  - Horror: horror movie, comedy horror, scary movie, gothic horror, Japanese horror, netflix horror.
  - Mystery: mystery movie, suspense, crime movie, murder.
  - Romance: romance movies, romantic movies, love movies.
  - Tragedy: tragedy movies, sad movies, heartbroken, moved.
  - Western: western, cowboy.

- Show
  - Talk show: chat show, talk show, Jimmy show.
  - Reality show: reality show, Saturday Night Live, Carpool Karaoke, Master chef, Voice of China, Man vs. Wild, American Idol, The Real Housewives
  - Standup comedy: Stand-Up Comedy
  - Acrobatic show: acrobatic, Acro dance, Aerial.
  - Magic show: magic show, magician.
  - Dance show: dance show, dancer, hip hop dance, jazz dance, poppin dance, Michael Jackson, Kpop dance.
  - Concert show: concert show, music concert, Lady Gaga concert, Taylor Swift, Justin Bieber, concert, Bruno Mars, super bowl halftime show, DJ live, band.
  - music show: singing show, singer, billie eilish, best singer, music, dua lipa, mv.
  - Play: stage play, theater, musical theater, Broadway, Broadway performance.
  - Fashion show: fashion show, Victoria's Secret, model catwalk, fashion style, trendy outfit, hairstyle.

- Object
  - Beauty-Related Content: beauty hacks, beauty tips, natural beauty, beauty of nature, beauty challenge.
  - Health-Related Content: best for health, health tips, fitness goals, fitness, weight loss, gym, workout.
  - Food-Related Content: my recipe, Chinese food, Indian food, French food, US food, weird food, delicious food, Asian food, hamster, otter.
  - Craft-Related Content: diy craft, craft challenge, 5 min craft, art and craft, crafting, craft time.
  - Living-organism-Related Content: animal, animal fight, pets, cat, tiger, lion, dog, fox, elephant, hippo, zebra, horse, monkey, koala bear, bear, panda, frog, lizard, turtle, centipede, tick, ant, whale.

Figure 6. **The WordCloud maps of textual description for the stimulus scenes on different AUs.** The larger the word, the higher its frequency.

  - Instrument-Related Content: musical instrument, violin, guitar, trumpet, french horn, trombone, piano, drum, banjo, harmonica, flute, oboe, saxophone.

- Self-made video
  - Funny related: prank, tik tok comedy, funny videos, laugh challenge.
  - Angry related: angry video, road rage, rage, gamer rage, wrath, bad temper, frustrated.
  - Fear related: fear video, fright, creeps, horror, try not scream, scary video, apprehension, foreboding, cold feet.
  - Sadness related: sad tik tok, try not to cry, sad moment in life.
  - Disgust related: disgust, hate.
  - Surprise related: amazing, surprise, amazing video, amazing skill, crazy, unbelievable moments, insane moments, incredible moments, Surprising moment, shocking video, astonishing video.

- Animation
  - Family-friendly: ratatouille, the incredibles, zootopia, coco, cars, despicable me, up, wall-e, Moana.
  - Comedy: Kung Fu Panda, Baby Boss, Meet the robinsons, Rio2.
  - Fantasy: Soul, Brave, Frozen, Sonic, Turning Red, Demon Slayer, Belle, Your Name, Pikachu , Lego movie, Akira, fantasy movie, EVA.
  - Action: Bad Guys, into the spider verse, dragon ball, big hero 6, megamind, Lilo and Stitch.
  - Music: Sing 2, Encanto, Vivo, Over the Moon, Lion King.
  - Romance: the book of life, isle of dogs, bae, the peanuts, the red turtle, paperman, feast, the croods, wonder park, the lorax.
  - Drama: Mulan, Spirited Away, hundred and one dalmatians, dumbo, pinocchio, perfect blue, grave of the fireflies, howl's moving castle, the iron giant, paprika, lady and the tramp.

- Game
  - Action: HALO INFINITE, CALL OF DUTY, BLACK MESA, DOOM ETERNAL, PUBG, Apex Legends, Destiny 2, fortnite, Overwatch, Super Mario, Mortal Kombat, The King of Fighters, Super Smash Bros, Ball FighterZ, God of War, The Legend of Zelda, Batman: Arkham Knight, Monster Hunter World.
  - RPG: final fantasy, THE WITCHER 3, ELDEN RING, CYBERPUNK 2077, DIABLO, Monster Hunter, Nier Automata, Demon's Souls, Assassin's Creed Odyssey, Legend of Zelda.
  - Adventure: A Way Out, The Wolf Among Us, Gone Home, Until Dawn, Shadow of the Tomb

Raider, The Last of Us, Marvel's Spider-Man, Red Dead Redemption 2.

  - Simulation: MICROSOFT FLIGHT SIMULATOR, , WORLD OF WARSHIPS, FARMING SIMULATOR 19, F1 2020, EURO TRUCK SIMULATOR 2, Cooking Simulator, Planet Coaster.

  - Strategy: OFFWORLD TRADING COMPANY, CIVILIZATION, COMMAND and CONQUER, ENDLESS LEGEND, STARCRAFT.

  - Sports: FOOTBALL MANAGER, FIFA, MADDEN NFL, GOLF WITH YOUR FRIENDS, NBA 2K, RIDERS REPUBLIC, ROCKET LEAGUE, Fight Night Champion, UFC4.

  - Puzzle: Portal 2, Baba is You, The Talos Principle, Little Nightmares, Hitman, Human: Fall Flat, Keep Talking and Nobody Explodes, Return of the Obra Dinn, HIDDEN FOLKS, THE ROOM THREE, KRYSTOPIA: NOVA'S JOURNEY, Poly Bridge, Escape Simulator, Braid.

- Interview/Public speech

  - Interview: horrible, worst, Disturbing, funny, hilarious, awkward, uncomfortable, embarrassing, best, stupid, dumb, street, live TV.

  - Speech: great, iconic, best, Tedx, best, Trump, public, court, Biden, press conference, apology, graduation , debate, political, apple event, lecture, Jobs.

1.9. Linguistic meta-data

A 20-words-length text description is generated for each stimulus image. To better understand the content of the stimulus scenes, we count the frequency of non-repetitive words in the textual descriptions. There are still 8049 words left after filtering out meaningless words (e.g., "the," "a," "an," "in," "in") with a Python library named "wordstop." The WordCloud maps corresponding to specific AUs are shown in Fig. 6. We observed that the patterns of WordCloud vary according to different AUs. For instance, except common words (e.g., man), some of the most frequent stimuli words with AU 4 reaction are "fighting, middle, ball, street, red", while the words for AU 12 are "white, played, tie, baseball, singing". The differences suggest that a relational model can be able to uncover more underlying reasoning information between stimulus-reaction. Accordingly, we utilize cross-domain contrastive learning to activate and associate related ROIs of stimulus scenes that arouse related reactions in the common embedding spaces.

AU	AU 1	AU 2	AU 4	AU 6	AU 7	AU 10	AU 12	AU 14	AU 15	AU 17	AU 23	AU 26
AU 1	1	1	0	0	0	1	0	0	0	0	0	1
AU 2	1	1	0	0	0	1	1	0	0	0	0	1
AU 4	0	0	1	0	1	0	0	0	0	0	0	1
AU 6	0	0	0	1	1	1	1	1	1	1	0	0
AU 7	0	0	1	1	1	1	1	1	1	1	1	1
AU 10	1	1	0	1	1	1	1	1	1	1	1	0
AU 12	0	1	0	1	1	1	1	1	1	1	1	0
AU 14	0	0	0	1	1	1	1	1	1	1	1	0
AU 15	0	0	0	1	1	1	1	1	1	1	0	0
AU 17	0	0	0	1	1	1	1	1	1	1	1	0
AU 23	0	0	0	0	1	1	1	1	0	1	1	0
AU 26	1	1	1	0	1	0	0	0	0	0	0	1

Figure 7. **The AU relation matrix.** "1" indicates connected, while "0" indicates unconnected.

2. Framework

2.1. Prediction

The proposed model output multiple estimated predictions from graph and non-graph models. The first output is $\hat{y}_d^n$, where n is the nth AU, and d is the dth domain in $(\mathcal{R}, \mathcal{S}, \mathcal{T})$, and the second is $\hat{y^g}$. Although both are used for calculating the graph and non-graph loss functions, only the outputs of the non-graph module from the reaction domain are used for predicting the AU occurrence probabilities at inference. Our consideration is that the outputs from stimulus domain and relation learning may deviate from the AU ground-truth. The late fusion result in Table 1 of the original paper provides a strong evidence for our assumption. It demonstrates that when applying an unconstrained feature fusion for stimulus-reaction features, the outcome tends to be even worse than the result using only reaction data.

2.2. Relation prior matrix

The AU relation matrix, FE relation matrix, and scene relation matrix is employed as prior matrices for both cross-domain contrastive learning, and building the extended adjacent matrices under in the graph module. It is shown in Fig. 7. As [7] does not provide any information of AU 26, we derive it from the original FACS definition. In FACS, AU 26 is always occured in a surprise expression (accompanied with AU 1, 2, 5) and a fear expression (accompanied with AU 1, 2, 4, 5, 7, 20). Thus, we define AU 26 are connected with AU 1, 2, 4, and 7. The FE relation matrix is derived from the co-occurring statistic from [4]. For scene relation matrix, we treat sub-classes belonging to the same scene category are related.

3. Experiment

3.1. Affective Databases

AU detection is evaluated on five datasets, including EmotioNet, CK+, DISFA, BP4D, and BP4D+. **EmotioNet** is a in-the-wild database which contains nearly on million Internet images with large variations in illumination, pose

and occlusions. We evaluated on the 25K images in the validation set. **CK+** contains 593 sequences from 123 subjects performing posed expressions. Among them, we use 309 sequences of 106 subjects that are annotated with six basic expressions and AUs (AU1, 2, 4, 5, 6, 7, 9, 12, 17, 23, 24, and 25). **DISFA** is a benchmark dataset for AU detection, which contains video from left view and right view of 27 subjects. 8 of 12 AUs with intensity greater than 1 from the left camera are used. F1-score is reported based on subject-exclusive 3-fold cross-validation. **BP4D** and **BP4D+** are widely used datasets for evaluating AU detection in the lab-controlled condition. BP4D contains 328 2D and 3D videos collected from 41 subjects. In BP4D+, high-resolution 3D dynamic model, high-resolution 2D video, thermal image and physiological data were acquired from 140 subjects. For a fair comparison with the state-of-the-art methods, 12 AUs are selected and performance of 3-fold cross-validation is reported.

FER is evaluated on five datasets, including FER+, RAFDB, AffectNet, MMI, and BU3D. **FER+** is an extended dataset based on FER2013, where 8 emotions are annotated. The accuracy on the test set with 3,589 testing images is reported for a fair comparison. **RAFDB** contains 15,000 facial images with 7 expressions annotated. In this paper, we select 12,271 images for training and remaining samples for validation. AffectNet is a large facial expression dataset with around 0.4 million images manually labeled for the presence of eight (neutral, happy, angry, sad, fear, surprise, disgust, contempt) facial expressions. In this paper, we choose AffectNet-7 without contempt class, which contains 283,901 and 3,500 images, for training and validating. **MMI** consists of 236 image sequences from 31 subjects. Each sequence is labeled as one of the six prototypical facial expressions. We selected three frames in the middle of each sequence as peak frames in frontal video. **BU-3DFE** contains 2500 pairs of static 3D face modes and texture images from 100 subjects with a variety of ages and races. Both 3D depth map and 2D texture images are used in the DRR MM.

3.2. Experimental setup

Most of the experiments in this work follow the following experimental setup: 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). We chose SGD as the optimizer with an initial learning rate of 0.01 with 64 as the batch size. After the first two epochs of training, it is reduced to 0.001. Note that only ViT uses AdamW as the optimizer according to its original setup. The data from the visual stimulus domain and visual reaction domain were augmented using the same strategy. No data augmentation is applied to the textual description domain. The weight decay and momentum are set at 0.0001 and 0.9, respectively. All models are implemented in PyTorch and trained on an Nvidia 3090 RTX GPU. We apply three-fold cross validation on the full database. The three-fold annotation files will be released for reference. The adjacent matrix of GCN and cross-domain GCN is constructed under the heterophily assumption in Sec.5.5. Note that some of the implementation details varied slightly when training and applying fine-tuning on different datasets.

Scene Classification (SC) is a task in which scenes from photographs are categorically classified. Unlike object classification, which focuses on classifying prominent objects in the foreground, Scene Classification uses the layout of objects within the scene, in addition to the ambient context, for classification. In this work, 24 sub-classes are selected for evaluating SC on ReactioNet.

3.3. Results of AUD in terms of individual AUs

Tab. 3 illustrates the experimental results with respect to individual AUs, measured by F1 score on ReactioNet.

3.4. Results of AUD in terms of accuracy

This section is the extended work of Sec.5.1 in the original paper. In Tab. 4, we compare the proposed framework with the models in Tab 1 of the original paper on DRR task using accuracy. Note that the F1-score is still the most convincing criteria for evaluating multi-label classification tasks.

3.5. Additional qualitative evaluation

We further apply the proposed framework to the multi-modal AU detection task without using stimulus-reaction information. In this section, the non-graph module is designed to learn cross-modality relations instead. Different from the cross-domain setup, we fuse the outputs of the graph module and non-graph module as the final prediction, which is a common approach for multi-modal feature fusion operations. We test the proposed framework on BP4D with two modalities (RGB texture and depth map) and on BP4D+ with three modalities (RGB texture, depth map, and thermal map).

In the upper part of Tab. 5 and Tab. 6, we first compare our method to the single modality based benchmarks, including JAA [5], SEV-Net [14], and UGN-B [8] and single modality based baselines, including ResNet-18 (visual), and ResNet-18 (depth). On BP4D, our method outperforms the state-of-the-art method UGN-B by about 3.2%. Our method outperforms all the single modality methods, achieving around 3.6% improvement in F1-score over SEV-Net. In the lower part of Tab. 5 and Tab. 6, we compare our method to the multi-modality based benchmarks, including TEMT-Net [11], AMF [13], and MFT [3], and

Table 3. F1 score in terms of individual AUs.

Model	Domain	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU26	Avg.
ResNet18	$\mathcal{R}$	41.5	37.0	62.8	71.3	70.22	73.9	75.8	81.6	36.8	41.6	21.7	43.7	54.8
without CDC	$\mathcal{RST}$	44.8	39.0	68.1	81.4	72.4	80.3	81.4	84.3	37.8	46.5	27.7	43.3	58.9
without Ho	$\mathcal{RST}$	41.6	38.7	70.1	81.5	**73.9**	81.8	82.2	82.5	37.0	46.8	30.6	42.3	59.1
without He	$\mathcal{RST}$	43.8	37.4	69.8	80.6	72.7	81.4	83.2	83.1	38.6	43.3	33.6	43.7	59.3
DRR static	$\mathcal{RST}$	45.9	40.3	70.8	80.6	73.5	82.7	82.4	85.6	38.6	45.9	33.7	46.8	60.6
DRR dynamic	$\mathcal{RST}$	45.0	41.9	70.7	81.1	74.3	82.3	82.4	84.7	41.7	48.3	36.7	46.3	61.3
DRR dyadic	$\mathcal{RST}$	47.3	40.9	73.2	80.6	72.7	81.3	83.2	85.0	38.6	43.3	37.5	48.7	61.1

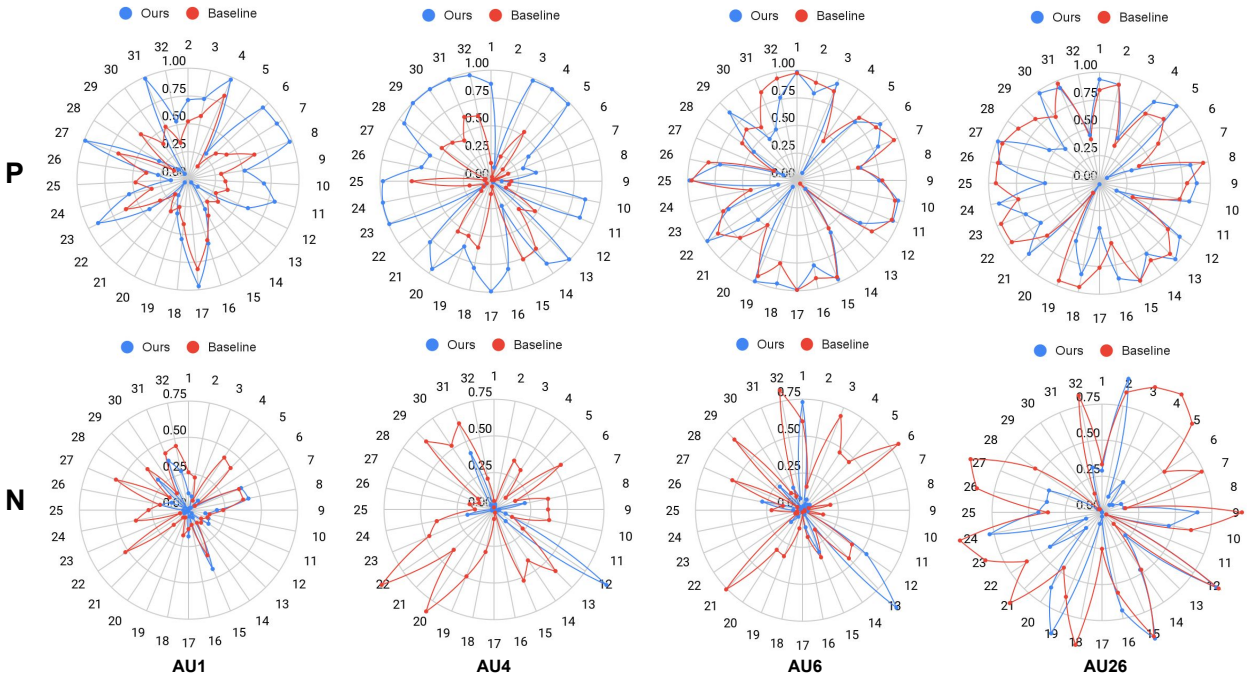


Figure 8. **Qualitative comparisons of DRR based AU detection on ReactioNet.** We randomly sampled 36 test data with positive AU 1, 4, 6, 26, and negative AU 1, 4, 6, 26. "P" indicates positive, and "N" indicates negative. The points on the radar charts represent the estimated probability of an AU occurrence. For positive samples, data points distributed on the periphery are better, and data values greater than 0.5 are correct predictions. For negative samples, data points distributed in the center are better, and data values less than 0.5 are correct predictions. The baseline algorithm is GCN using visual face data.

Table 4. **Comparison with baselines and benchmarks using accuracy on ReactioNet.**

Model	Domain	Accuracy (%)
ResNet-18	$\mathcal{R}$	76.6
ResNet-50	$\mathcal{R}$	76.9
ViT	$\mathcal{R}$	77.1
GCN	$\mathcal{R}$	77.1
Late fusion	$\mathcal{RST}$	77.3
UDA	$\mathcal{RST}$	76.9
Cross-domain GCN	$\mathcal{RST}$	77.3
SEV-Net	$\mathcal{RST}$	77.2
Ours	$\mathcal{RST}$	77.5

multi-modal baselines, including ResNet-18 with early feature fusion, ResNet-18 with late feature fusion, and multi-modal GCN. Our method outperforms all of the related algorithms, achieving the highest F1-score of 66.5% on BP4D and 64.7% on BP4D+.

In summary, the proposed framework is not only adapted to cross-domain learning but also shows remarkable performance improvement in multi-modal tasks.

This part is the extended work of qualitative comparison in Sec.5.1. In this section, we provide a global perspective with qualitative comparisons across different AUs. Note that the baseline method (single domain GCN) includes the majority of the key components proposed in this work (e.g., contrastive learning, relation learning under two assumptions, and combined graph/non-graph modules). The only difference is that our algorithm is trained with data from multiple domains (stimulus and reaction). As shown in Fig. 8, the proposed framework achieves better performance on different AUs. The conventional GCN-based baseline algorithm performs well on positive AU 6 and AU 26. However, when predicting the negative samples with high dif-

Table 5. Comparison with SOTAs using F1 score in terms of individual AU on BP4D for multi-modal AU detection. $\mathcal{V}$ represents RGB visual modality, and $\mathcal{D}$ represents depth visual modality.

Model	Modality	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.
JAA	$\mathcal{V}$	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
SEV-Net	$\mathcal{V}$	58.2	50.4	58.3	81.9	73.9	87.8	87.5	61.6	52.6	62.2	44.6	47.6	63.9
UGN-B	$\mathcal{V}$	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
ResNet-18	$\mathcal{V}$	48.0	46.7	57.0	77.5	71.6	83.5	85.0	63.8	47.1	58.2	39.4	37.3	59.6
ResNet-18	$\mathcal{D}$	44.6	49.3	54.4	77.5	74.8	83.7	88.4	59.0	53.3	60.6	41.9	53.3	60.3
Early fusion	$\mathcal{VD}$	44.1	50.0	50.6	75.7	63.8	84.8	89.3	65.0	39.0	62.6	35.7	29.8	57.5
Late fusion	$\mathcal{VD}$	51.2	46.8	61.1	80.5	73.8	87.7	88.9	62.4	47.7	61.1	41.2	31.4	61.1
Multi-modal GCN	$\mathcal{VD}$	48.7	40.8	55.8	79.0	76.7	84.3	87.0	64.0	54.0	62.7	49.4	53.3	63.0
TEMT-Net	$\mathcal{VD}$	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	$\mathcal{VD}$	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
MFT	$\mathcal{VD}$	51.6	49.2	57.6	78.8	77.5	84.4	87.9	65.0	56.5	64.3	49.8	55.1	64.8
DRR MM	$\mathcal{VD}$	52.7	48.3	62.4	81.9	77.6	85.2	88.7	66.1	57.2	65.5	56.2	55.9	66.5

Table 6. Comparison with SOTAs using F1 score in terms of individual AU on BP4D+ for multi-modal AU detection. $\mathcal{V}$ represents RGB visual modality, $\mathcal{D}$ represents depth visual modality, and $\mathcal{T}$ represents thermal visual modality.

Model	Modality	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.
JAA	$\mathcal{V}$	46.0	41.3	36.0	86.5	88.5	90.5	89.6	81.1	43.4	51.0	56.0	32.6	61.9
SEV-Net	$\mathcal{V}$	47.9	40.8	31.2	86.9	87.5	89.7	88.9	82.6	39.9	55.6	59.4	27.1	61.5
ResNet-18	$\mathcal{V}$	47.8	47.0	24.5	84.3	88.0	89.9	87.2	80.6	47.5	36.7	54.7	27.4	59.6
ResNet-18	$\mathcal{D}$	40.9	39.2	30.4	83.8	86.7	90.9	90.2	79.6	38.2	44.0	52.5	39.4	59.6
ResNet-18	$\mathcal{T}$	39.0	34.0	25.0	82.2	84.0	87.6	87.2	79.2	32.1	36.5	43.9	7.9	53.2
Early fusion	$\mathcal{VDT}$	39.0	34.6	26.2	80.1	86.1	89.5	87.7	74.0	41.0	33.5	44.9	15.8	54.4
Late fusion	$\mathcal{VDT}$	46.0	41.3	36.0	86.5	88.5	90.5	89.6	81.1	43.4	51.0	56.0	32.6	61.9
Multi-modal GCN	$\mathcal{VDT}$	43.8	40.9	36.9	85.0	88.9	91.1	90.5	83.0	40.6	48.3	54.6	44.0	62.3
AMF	$\mathcal{VDT}$	45.3	42.5	34.8	85.9	87.9	89.5	90.4	82.6	50.1	45.5	55.7	42.1	62.7
MFT	$\mathcal{VDT}$	49.6	42.0	43.5	85.8	88.6	90.6	89.7	80.8	49.8	52.2	59.1	38.4	64.2
DRR MM	$\mathcal{VDT}$	45.9	40.4	41.6	85.5	88.6	90.6	90.2	82.4	46.5	46.2	57.0	52.6	64.7

ficulties, such as AU 1, 4, 26, and the positive samples on AU 1, 2, the conventional GCN shows obvious misprediction. In addition, even correctly predicted values of GCN tend to be close to the uncertain median value of 0.5. It demonstrates the superiority of DRR-based AU detection.

4. Discussion

In this section, we discuss some extra findings based on our observations and experiments.

Noise We densely clean the data by following a reliable data processing flow. However, pursuing perfect data is impossible for high-level tasks (e.g., stimulus-reaction based DRR). In Fig. 9, three samples exhibit the special situations (including verbal communication, distraction, and face occlusion) from the visual reaction domain, which may diminish the performance of stimulus-reaction based method. Inspired by that, taking the subject's speaking content, belief (i.e., attention), and temporal context into account for learning facial behavior may improve our model. In Fig. 10, three samples illustrate the special situations (including blurred image, meaningless image, and verbal communication) from the visual stimulus domain, which may introduce noise to the proposed method. Considering that, integrating the stimulus's temporal context, audio, and caption to our model may solve the issues. Of note, the examples cannot cover all types of noise in the data given the task's high complexity.

Reaction delay Although the data from stimulus-reaction domains is well synchronized, humans' reaction delay may cause some side effects on the stimulus-reaction based DRR task. Fig. 11 shows an example of reaction delay in ReactioNet. The subject needs delayed time of more than 20 frames (25 fps video) to respond with a smile after seeing a funny cartoon character. This may cause an inconsistency between the stimuli presented in the current picture and the subject's reaction. Of note, different subjects' reaction times vary when perceiving distinct stimuli. The impact of the reaction delay is not assessed in this paper. We plan to do further analysis in future work using models that are more effective at learning long-term dependencies.

In summary, multi-modal data, long-term temporal dependencies, and the subject's beliefs can contribute to a more robust model in stimulus-reaction based tasks. On the other hand, integrating too many resources for performance improvement may come at the cost of model efficiency. How to trade off between them will be our another focus in the future.

References

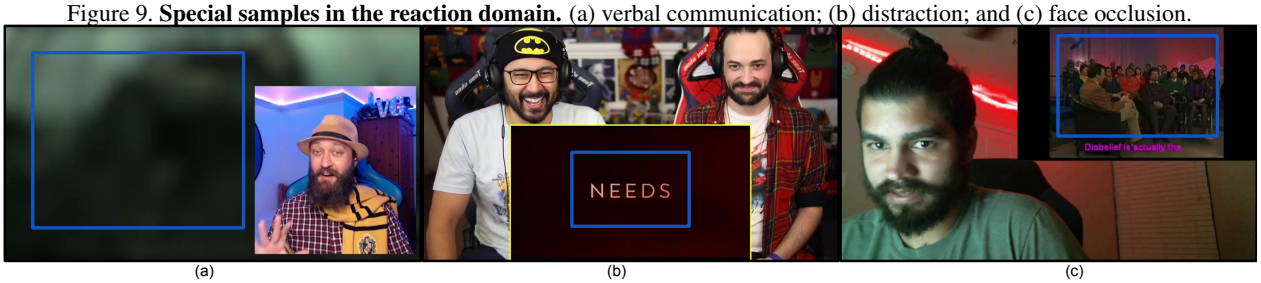
[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 1

[2] 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. 3

[3] 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. 7

(a) (b) (c)

Figure 9. **Special samples in the reaction domain.** (a) verbal communication; (b) distraction; and (c) face occlusion.



(a) (b) (c)

Figure 10. **Special samples in the stimulus domain.** (a) blurred image; (b) meaningless image; and (c) verbal communication.

[4] Shan Li and Weihong Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 2019. 6

[5] 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. 7

[6] Xiaotian Li, Xiang Zhang, Taoyue Wang, and Lijun Yin. Knowledge-spreader: Learning facial action unit dynamics with extremely limited labels, 2022. 1

[7] 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. 6

[8] 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. 7

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

[10] 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. 1

[11] 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. 7

[12] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, and Michel Valstar. Joint action unit localisation and intensity estimation through heatmap regression. In *BMVC*, 2018. 1

[13] 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. 7

[14] 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

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

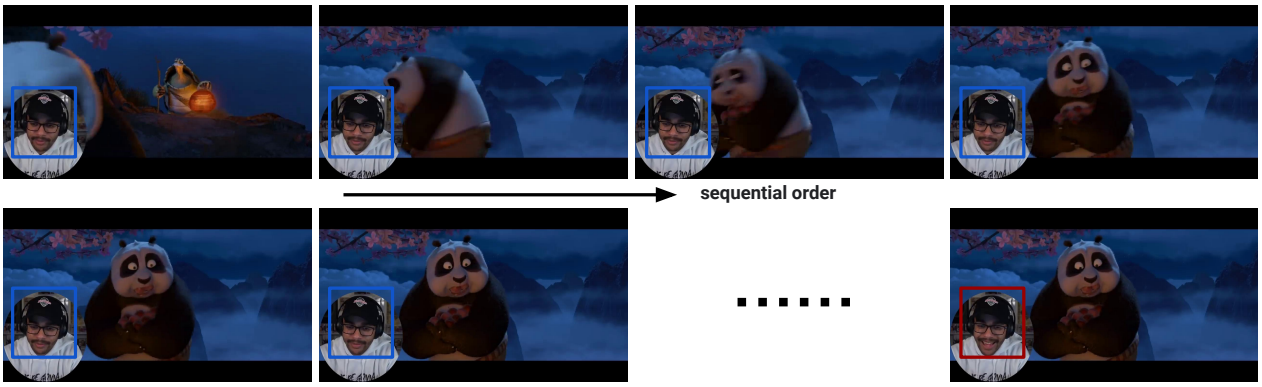


Figure 11. **An example of reaction delay.** The blue box indicates a face without a smile, and the red box indicates a face with a smile.