Weakly-Supervised Text-driven Contrastive Learning for Facial Behavior Understanding
Supplementary Material

1. Activity Descriptions

Table 1 and 2 show the activity descriptions of BP4D and BP4D+ respectively.

Table 1: 8 Activity descriptions the subjects participate in BP4D.

	Activity Description
A1	Talk to the experimenter and listen to a joke (Interview). The target emotion is happiness or amusement
A2	Watch and listen to a recorded documentary and discuss their reactions. The target emotion is sadness
A3	Experience sudden, unexpected burst of sound. The target emotion is surprise or startle
A4	Play a game in which they improvise a silly song. The target emotion is embarrassment
A5	Anticipate and experience physical threat. The target emotion is fear or nervous
A6	Submerge their hand in ice water for as long as possible. The target emotion is physical pain
A7	Experience harsh insults from the experimenter. The target emotion is anger or upset
A8	Experience an unpleasant smell. The target emotion is disgust

Table 2: 10 Activity descriptions the subjects participate in BP4D+.

	Activity Description
A1	Interview: Listen to a funny joke. The target emotion is happiness or amusement
A2	Graphic show: Watch 3D avatar of participant. The target emotion is surprise
A3	Video clip: 911 emergency phone call. The target emotion is sadness
A4	Experience a sudden burst of sound. The target emotion is startle or surprise
A5	Interview: True or false question. The target emotion is skeptical
A6	Improvise a silly song. The target emotion is embarrassment
A7	Experience physical threat in dart game. The target emotion is fear or nervous
A8	Cold pressor: Submerge hand into ice water. The target emotion is physical pain
A9	Interview: Complained for a poor performance. The target emotion is anger or upset
A10	Experience smelly odor. The target emotion is disgus

2. Label Semantic Descriptions

2.1. Facial Expression

Inspired by the work of SEV [4], we summarized 8 facial expression semantic descriptions based on the previous psychology study [1, 3].
Following descriptions are in **label name** : **label description** pattern.

Anger: The eyebrows are lowered and pulled closer together, and the eyelids become squinted or raised. The lips would tighten or curl inwards, the corners of the mouth would point downwards, and the Jaw is tense and might jut forward slightly.
Contempt: The eyes would be unengaged, one side of the mouth is pulled up and back. One eyebrow may pull upwards and the head may tilt back slightly, making the gaze follow down the nose.
Disgust: The eyebrows are pulled down, and the nose is wrinkled. The upper lip is pulled up and the lips are loose. The eyes are narrow, the teeth may be exposed, and the cheeks may be raised.
Fear: The eyebrows are pulled up and together, and the

upper eyelids are pulled up, and the lower eyelids are tense and drawn up. The mouth are stretched and drawn back, possibly exposing teeth. Vertical wrinkles may appear between the eyebrows.
Happiness: The eyes squint slightly, wrinkles appear at the corners of the eyes and the cheeks raise. The corners of the mouth move up at a diagonal, widening the mouth and the mouth may part, exposing teeth.
Neutral: The mouth is straight lined, the eyes are unfocused and the cheeks are slack. Not arch the eybrows, frown, smile or grimace.
Sadness: The eyebrows are lower and pulled closer together, and the inner corners of the eyebrows are angled up. The corners of the mouth are drawn downwards, and the lips may be either drawn in tightly or pouting outwards.
Surprise: The eyebrows are raised, and horizontal wrinkles would appear on the forehead. The jaw would go slack, the mouth would hang open loosely and the eyes would widen.

2.2. Facial Action Unit

The descriptions of AUs is written in SEV [4], which is based on the psychology study [2]. We then has slightly modified these descriptions, which are shown in a pattern:
AU id. label name : **label description**.

AU1. inner brow raiser: The inner corners of the eyebrows are lifted slightly, the skin of the glabella and forehead above it is lifted slightly and wrinkles deepen slightly and a trace of new ones form in the center of the forehead.
AU2. outer brow raiser: The outer part of the eyebrow raise is pronounced. The wrinkling above the right outer eyebrow has increased markedly, and the wrinkling on the left is pronounced. Increased exposure of the eye cover fold and skin is pronounced.
AU4. brow lowerer: The vertical wrinkles appear in the glabella and the eyebrows are pulled together. The inner parts of the eyebrows are pulled down a trace on the right and slightly on the left with traces of wrinkling at the corners.
AU6. cheek raiser: The cheeks are lifted without activately raising up the lip corners. The infraorbital furrow has deepened slightly and bags or wrinkles under the eyes must increase. The infraorbital triangle is raised slightly.
AU7. lid tightener: The lower eyelid is raised markedly and straightened slightly, causing slight bulging, and the narrowing of the eye aperture is marked to pronounced.
AU9. nose wrinkler: The nose is Wrinkled, the skin on bridge of the nose is drawn upwards, the nasal wings are lifted up, the infraorbital triangle is severely raised, and the upper part of the nasolabial fold is extremely deepened as the upper lip is drawn up slightly.
AU10. upper lip raiser: The upper lip is slightly raised and the nasolabial furrow is deepened.
AU12. lip corner puller: The corners of the lips are markedly raised and angled up obliquely. The nasolabial furrow has deepened slightly and is raised obliquely slightly. The infraorbital triangle is raised slightly.
AU14. dimpler: The lip corners are extremely tightened, and the wrinkling as skin is pulled inwards around the lip corners is severe. The skin on the chin and lower lip is stretched towards the lip corners, and the lips are stretched and flattened against the teeth.
AU15. lip corner depressor: The lip corners are pulled down slightly, with some lateral pulling and angling down of the corners, and slight bulges and wrinkles appear beyond the lip corners.
AU17. chin raiser: The chin boss shows severe to extreme wrinkling as it is pushed up severely, and the lower lip is pushed up and out markedly.
AU23. lip tightener: The lips are tightened maximally and the red parts are narrowed maximally, creating extreme wrinkling and bulging around the margins of the red parts of both lips.
AU24. lip pressor: The lips are severely pressed together, severely bulging skin above and below the red parts, with severe narrowing of the lips and wrinkling above the upper lip.
AU25. lips part: The teeth is clearly shown, and the lips are separated slightly. Nothing suggests that the jaw has dropped even though the upper teeth are not clearly visible.
AU26. jaw drop: The jaw is lowered about as much as it can drop from relaxing of the muscles. The lips are parted to about the extent that the jaw lowering can produce.

3. Pseudo-codes

We provide the pytorch-style pseudo-codes for both pretraining and finetuning in Algorithm 1 and 2.

4. Text Prompt templates

Let N denotes label name, D indicates label descriptions, and A represents activity descriptions. For label name prompting, only one template is used, i.e., "a photo of a person with {N}.". Label description prompting is randomly chose from one of the AU or expression templates.
AU Label Description Templates:

- "a photo of a person with {D}."
- "a photo shows a person that {D}."
- "a photo of one has {D}."
- "a photo of a person that {D}."
- "a photo of a face with {D}."
- "a photo of a person has {D}."

Algorithm 1: PyTorch-style pseudocode for CLEF in Pre-training

```
# encode_image: vision transformer
# encode_text: text transformer
# img1,img2: image inputs of two augmentation
# activity: activity text
# t1, t2: two learned temperature parameters
# targets: activity labels

# extract feature representations for image
i_f1 = encode_image(img1)
i_f1 = i_f1/i_f1.norm(dim=1, keepdim=True)
i_f2 = encode_image(img2)
i_f2 = i_f2/i_f2.norm(dim=1, keepdim=True)
# extract feature representations for
# activity description
a_f = encode_text(activity)
a_f = t_f/t_f.norm(dim=1, keepdim=True)
f_ii = torch.cat((i_f1, i_f2), 0)
f_ia = torch.cat((i_f1, a_f), 0)
# scaled cosine similarities
logit_ii = t1.exp()*i_f1 @ f_ii.t()
logit_it = t2.exp()*i_f1 @ f_ia.t()
# supervised contrastive loss function
loss_ii = sup_con_loss(logit_ii, targets)
loss_ia = sup_con_loss(logit_it, targets)
loss = (loss_ii + loss_ia)/2.0
```

- "a good photo of a person that {D}."
- "the photo of a face that {D}."
- "the photo of a person that {D}."
- "a photo of a face where {D}."
- "a photo shows facial action unit that {D}."
- "a cropped photo of face that {D}."
- "a clean photo of a person that {D}."
- "a facial action unit where {D}."

Expression Label Description Templates:

- "a photo of a person with {D}."
- "a photo shows a person with {D}."
- "a photo of one has {D}."
- "a photo of a face that {D}."
- "a photo of a person has {D}."
- "a good photo of a person in {D}."
- "the photo of a face in {D}."
- "a cropped photo of face that {D}."
- "a clean photo of a person with {D}."

Algorithm 2: PyTorch-style pseudocode for CLEF in Fine-tuning

```
# encode_image: Vision Transformer
# encode_text: Text Transformer
# img: image input
# n_text: label name text
# d_text: label description text
# t1: learned temperature parameter
# t2: learned temperature parameter
# lambda: fixed hyperparameter
# targets: facial expression or AU label

# extract feature representations for image
i_f = encode_image(img)
i_f = i_f/i_f.norm(dim=1, keepdim=True)
# extract feature representations for
# label name text
n_f = encode_text(n_text)
n_f = n_f/n_f.norm(dim=1, keepdim=True)
# extract feature representations for
# description text
d_f = encode_text(d_text)
d_f = d_f/d_f.norm(dim=1, keepdim=True)
# scaled cosine similarities
logit_in = t1.exp()*i_f @ n_f.t()
logit_dn = t2.exp()*d_f @ n_f.t()
# loss function
# if task is FER, task_loss: cross_entropy_loss
# if task is AUR, task_loss: bce_loss
loss_in = task_loss(logit_in, targets)
labels = torch.arange(n_text.shape[0])
loss_dn = cross_entropy_loss(logit_dn, labels)
loss = (lambda * loss_in + loss_dn)/2.0
```

- "a facial expression where {D}."
- "a photo of facial expression that {D}."

Activity description prompting is randomly chose from one of the following templates.

Activity Description Templates:

- "a photo of a person from an activity that {A}."
- "a photo shows a person in the activity that {A}."
- "a photo of an activity that {A}."
- "a photo of a person participated in an activity that {A}."
- "a photo of a face from the activity that {A}."
- "a photo of a person was in an activity that {A}."
- "a good photo of the activity where {A}."
- "a photo of a person joined in an activity that {A}."
- "a good photo of a person in an activity that {A}."
- "a cropped photo of face from an activity where {A}."

Table 3: Fine-tuning Settings

Database	epochs	lr	Warm-up epochs	lr schedule	weight decay
BP4D	3	0.0002	1	cosine decay: [1, 3]	0.01
BP4D+	3	0.0002	1	cosine decay: [1, 3]	0.01
DISFA	5	0.0001	0	steps: [2:0.1, 5:0.5]	0.01
Affect-Net	3	0.0002	1	cosine decay: [1, 3]	0.01
RAF-DB	5	0.0002	1	cosine decay: [1, 5]	0.01
FER+	7	0.0001	1	cosine decay: [3, 7]	0.05

- "a clean photo of a person in the activity that {A}."
- "an activity where {A}."

5. More Implementation Details

Table 3 and 4 show the detail implementation settings for fine-tuning and pre-training respectively. The settings not shown in Table 3 are the same as the pre-training settings. Note that only augmentation 1 is applied in the fine-tuning image augmentation.

Table 4: Pre-training Settings

config	value
Batch size	64
Vocabulary size	49408
Training epochs	5
Warm-up epochs	1
learning rate schedule	cosine decay
learning rate	10^{-5}
min learning rate	10^{-6}
weight decay	0.01
AdamW betas	(0.9, 0.999)
augmentation 1	HorizontalFlip
augmentation 2	ResizedCrop
	HorizontalFlip
	RandomRotation

6. More Ablation study

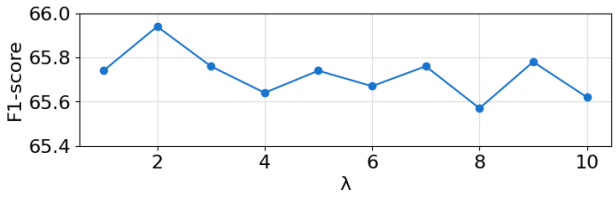
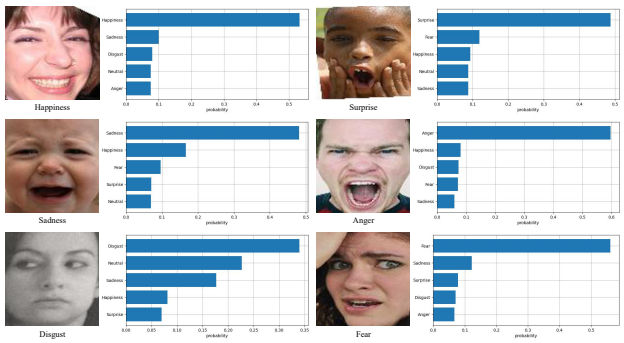


Figure 1: F1-score with different λ on BP4D

Evaluation of different λ. In this section, we evaluate the performance on BP4D by setting different hyperparameters λ, which can be seen in Firuge 1 The performance reaches its peak when λ is set to 2, which is attributed to the fact that loss from Image-Name pairs plays a major role in back propagation as Image-Name pairs are more diverse than Name-Description pairs.

7. More Visualization

Figure 2 shows more visualizations of prediction probability on RAF-DB. The query text is in "a photo of a person with {N}" format. Both success and failure examples are shown in it.



(a) successful prediction examples

(b) fail prediction examples

Figure 2: Visualization of image samples and the probabilities of their top 5 predictions on RAF-DB. The query texts are in the template of "a photo of a person with {N}"

References

[1] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 1

[2] Paul Ekman and Erika L Rosenberg, editors. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, 1997. 2

[3] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992. 1

[4] 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. 1, 2