

Knowledge-Aligned Counterfactual-Enhancement Diffusion Perception for Unsupervised Cross-Domain Visual Emotion Recognition

Supplementary Material

In this supplementary material, we add more method descriptions, implementation details, and experiment results. Specifically, the method section includes detailed explanations of the knowledge extractor in Sec. § A. The implementation details provide configurations of each module in our model in Sec. § B. In the part of experiments in Sec. § C, we present more main results, ablation studies, hyperparameter experiments, parameter sensitivity analysis, etc.

A. More Details of Method

A.1. Knowledge Parser

In the KADAP framework, a triplet extraction algorithm was developed to process the semantic role labels produced by the SRL model [2]. This algorithm generates a list of triplets for each input sample. The pseudo-code for the algorithm is presented as Algorithm 1. Specifically, B-ARG0 corresponds to the subject, B-V denotes the verb, B-ARG1 represents the direct object, B-ARG2 indicates the indirect object, and B-ARGM captures additional contextual information. Furthermore, I-* signifies the subsequent tokens associated with each component. Using the semantic labels L assigned to these tokens, the final triplet list T* is constructed.

B. More Implementation Details

The key hyperparameters for each module of the proposed model are detailed as follows.

For the BLIP caption generator, the hyperparameters were set to num_beams = 5, repetition_penalty = 5.0, max_new_tokens = 50, and min_new_tokens = 30. The input prompt for BLIP was defined as:

“What is the information of the subject in the picture? What is the facial expression and the body movement of the subject in the picture? What is the emotional meaning of the image? Please answer these questions with a descriptive sentence.”

For the UNet module, the input image size was set to 512×512 pixels, with F containing 4 feature maps. The i -th feature map F_i had spatial dimensions $H_i = W_i = 2^i + 2$, where $i = 1, 2, 3, 4$. The backbone architecture followed the configuration of Stable Diffusion [5].

In the MoE predictor, 8 experts were used, with $k = 2$ for the TopK operation. The number of attention heads was set to 8, while the MLP in CLIEA was implemented as a single linear layer. The text branch of CLIP [4] served as the text encoder \mathcal{T} .

Algorithm 1 Extract Triples from Semantic Role Labels

Input: T: initialized triple dict; d : initialized dicey subject; L: tags of sentence; W: words of sentence;

Output: T*: handled triple dict;

for each i, tag in L **do**

if tag is B-ARG0 **then** T[subject].append(w_i)
 add subsequent words w_i^{arg0} by I-ARG0

else if tag is B-V **then** T[verb].append(w_i)
 add subsequent words w_i^v by I-V

else if tag is B-ARG1 **then**

if T[subject] is empty **then**

 T[subject].append(w_i)

 add subsequent words w_i^{ARG1} by I-ARG1

else T[object].append(w_i)

 add subsequent words w_i^{ARG1} by I-ARG1

else if tag is B-ARG2 **then** T[object].append(w_i)

 add subsequent words w_i^v by I-ARG2

else if tag starts with B-ARGM **then**

if not d **then** Set d to w_i + subsequent words w_i^m

if T[object] is empty and d is not empty **then**

 Set T[object] to d

if T[subject], T[verb], T[object] are all not empty **then**

return T

For cross-domain dataset label alignment, all datasets were standardized to six emotion categories: *surprise, happiness, disgust, fear, sadness, and anger*.

The training process utilized the Emoset training set on a single A800 GPU. The KADAP framework was trained for 10 epochs, requiring approximately 40 hours. Additionally, training the CLIEA component required an extra 1 hour per epoch.

C. More Results of Experiments

C.1. Main Results

We also conducted a single-domain test for the Emoset [7] and SER30K [3] training sets: the training set and the test set belong to the same data set. As the Table 1 shows, our KADAP is 4 – 5% better than the previous SOTA model on both datasets. This proves that our method has a better ability to detect emotions.

C.2. Ablation Study

Variants of TIE. We performed ablation experiments on disturbing variables in TIE. This time, we will intervene

Methods	Accuracy(%)	
	Emoset	SER30K
MDAN [6]	75.75	59.38
LORA-V [3]	76.27	67.28
TGCA-PVT [1]	78.70	68.80
KADAP	83.38	72.97

Table 1. Experimental results of single domain setting on Emoset and SER30K dataset.

with the k variable, and we select samples with different labels under the same batch to operate. Specifically, we first fixed the P of each sample, which required us to choose the corresponding emotional label prompt according to the ground truth of the sample. Then, the knowledge variable k^* produced by other dissimilar samples in the same batch is used to form counterfactual alignment. Finally, counterfactual contrast learning is used to distinguish different similarity Y :

$$\mathcal{L}_{ccl}^k = - \sum \log \frac{\exp(\mathcal{S}(Y_{v_i}, Y_{k_i, p_i})/\tau)}{\sum_{j \neq y(i)}^B \exp(\mathcal{S}(Y_{v_i}, Y_{k_j, p_i})/\tau)}, \quad (1)$$

where τ is the temperature coefficient and B is the number of mini-batch. We use $CLIEA_k$ to indicate this variant, $CLIEA_p$ to indicate the method used in the body, and KADAP to indicate that CLIEA is not used and with no need for target domain samples. Our experimental results under multiple DA settings are shown in Table 2.

Task	KADAP	$CLIEA_k$	$CLIEA_p$
E→S	57.71	58.67(+0.96)	62.78(+5.07)
E→P	35.84	34.91(-0.93)	40.55(+4.71)
E→A	47.14	46.59(-0.55)	51.20(+4.06)
S→E	37.77	38.60(+0.83)	41.29(+3.52)
S→P	35.21	34.95(-0.26)	38.69(+3.48)
S→A	36.94	37.81(+0.87)	42.50(+5.56)

Table 2. Influence of selection of interference factors in TIE on model performance

It is clear that the results of this experiment prove that intervening with k to calculate TIE is indeed not an ideal choice.

Pseudo Labels We analyzed the pseudo-labels generated by CLIEA and compared the zero-shot alignment capabilities of the CLIP. Specifically, we use six-label emotions to form prompt words, like “A happy photo.” After input to the CLIP text encoder, the resulting text tag is embedded. The pseudo-label is obtained by direct comparison with the visual representation. We compared the pseudo-labels obtained by CLIEA with this native method to get the accuracy of the two pseudo-labels.

Methods	E→S		S→E	
	Emoset	SER30K	SER30K	Emoset
CLIP [4]	40.57	34.98	39.72	30.80
CLIEA	82.19	63.24	71.98	48.94

Table 3. Experimental results of different pseudo-labels generative method on E→S and S→E tasks.

More Cross-domain Results As shown in Figure 1, we report the correlated results of different datasets in the ablation experiment, that is, the DA tasks of E→P, E→A, S→P, S→A. Experimental results demonstrate that the proposed LoRA fine-tuning knowledge-guided cross-attention mechanism is highly effective compared to other fine-tuning strategies. Our proposed method provides a new lightweight and robust alternative of emotion recognition.

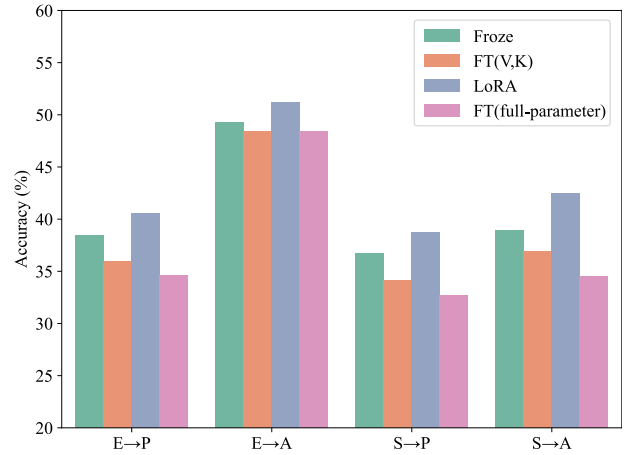


Figure 1. Effectiveness of different fine-tuning strategies on the more DA task. We only report the results on the target domain.

In addition to the ablation experiments of various modules of UC setting with SER30K as source domain data in the main paper, we present the experimental results of Emoset here, as shown in Table 4. It is observed that the introduction of MoE and triples t can improve the performance of the model, especially on the target domain data. For the input of MoE, processing both visual and knowledge representation can improve the generalization ability of the model and improve the performance in the target domain.

C.3. Hyper-parameter Study

We recorded the performance of the two variants of our model at the cross-domain training and analyzed the accuracy for each epoch. The first variant uses only KADAP, while the second one, KCDP, incorporates the CLIEA approach. As shown in Figure 2, we plotted the accuracy of

Modules			Datasets	
c	t	Classifier	E	S
✓	-	\mathbf{C}_{global}	82.72	55.62
-	✓	\mathbf{C}_{global}	83.10(+0.38)	56.33(+0.71)
✓	✓	\mathbf{C}_{global}	83.12(+0.40)	56.35(+0.73)
✓	-	MoE _v	83.24(+0.52)	56.78(+1.16)
-	✓	MoE _v	83.31(+0.59)	57.27(+1.65)
✓	✓	MoE _v	83.38 (+0.66)	57.16(+1.54)
✓	-	MoE _{v+k}	83.25(+0.53)	56.78(+1.16)
-	✓	MoE _{v+k}	83.28(+0.56)	57.28(+1.66)
✓	✓	MoE _{v+k}	83.32(+0.60)	57.71(+2.09)

Table 4. Ablation study on UC setting with training on Emoset. The symbols in the table are the same as those in the text part.

each method in both the source domain and the target test domain. We can observe that both variants exhibit relatively stable trends in the source domain, while the target domain shows more fluctuations. KCDP, due to the involvement of the CLIEA, performs better in the target domain, which can be attributed to its cross-domain generalization ability.

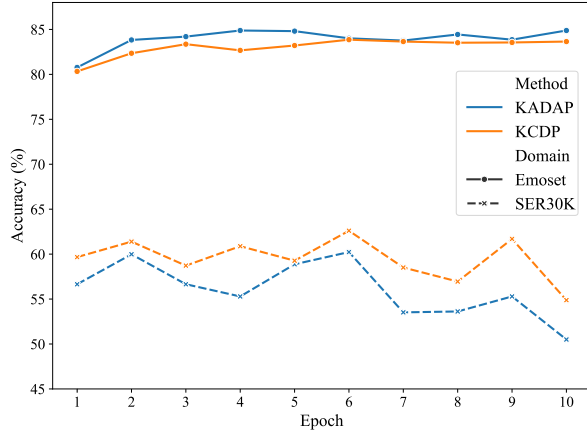


Figure 2. Effectiveness of different fine-tuning strategies on the more DA task. We only report the results on the target domain.

We also investigated the hyperparameters λ_1 and λ_2 in the loss function. The experiment was carried out on the E→S task based on DA setting. The experimental results are shown in the table below.

λ_1	0.2	0.4	0.6	0.8	1
Acc.	61.89	62.03	67.22	62.44	62.78

Table 5. Experimental results of different λ_1 , where $\lambda_2 = 0.7$

C.4. Influence of Module parameter

Samples of Target Domain We investigated the effect of the number of unlabeled target domain samples sampled

λ_2	0.2	0.4	0.6	0.8	1
Acc.	62.01	62.24	62.51	62.12	62.37

Table 6. Experimental results of different λ_2 , where $\lambda_2 = 0.7$

for training in CLIEA method on the cross-domain ability of the model. As can be seen from Figure 3, with the increase of the number of target domain samples k in the training process, the accuracy of the model in the target domain (SER30K) gradually increases, while the accuracy of the model in the source domain (Emoset) changes little. This shows that the increase of target domain data has a significant effect on improving the performance of UDA model in the target domain, especially in the case of small samples. However, as the number of samples increases to a certain threshold (for example, after $k = 3.2$), the accuracy of the target domain tends to increase gradually, indicating that the performance improvement of the model may be bottlenecked under a large number of target domain samples. This experimental result verifies the effectiveness of model tuning through incremental target domain data in cross-domain tasks and reveals the importance of reasonable selection of sample size in domain adaptation tasks.

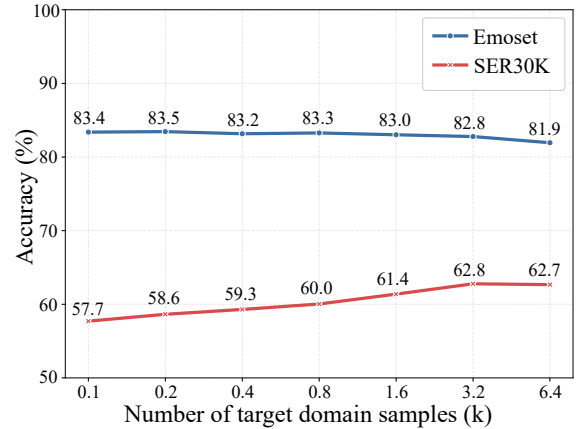


Figure 3. The influence of the number of unlabeled target domain samples k on the cross-domain capability of the model

Scale	E→S	S→E
$i \in \{1\}$	61.75	39.90
$i \in \{4\}$	61.02	39.27
$i \in \{1, 4\}$	61.83	40.18
$i \in \{1, 2, 4\}$	61.94	40.92
$i \in \{1, 2, 3, 4\}$	62.78	41.29

Table 7. Experimental results of different scales of attention maps.

Scale of Attention Maps We analyzed the impact of attention map scales in the denoising diffusion model. Different

scales, represented by i in A_i of the previous sections, were selected as the influencing variables. DA experiments were conducted on E→S and S→E, and the results are shown in Figure 7. Our model achieved the best performance by selecting the attention maps at the highest number of scales as visual features. This indicates that integrating multi-level feature maps captures more visual details, which benefits affective perception.

Key parameters of MoE Predictor We explored the effects of different key parameters in the MoE predictor on model performance, i.e. the number of experts N and k of TopK. The experiments were conducted on E→S and S→E tasks based on DA Settings. We fixed k to 2 and changed the number of experts, and the results were shown in Table 8. At the same time, we fixed the number of experts to 8 and changed the k variable. The results are shown in the Table 9. Sufficient empirical analysis shows that reasonable allocation of weights among experts can effectively deal with the visual and knowledge representation in the model.

N	2	4	6	8	16	20
E→S	62.04	62.13	61.95	62.78	62.41	62.24
S→E	40.31	40.26	40.98	41.29	41.07	41.17

Table 8. Experimental result with different numbers of experts N on DA setting.

k	1	2	3	4	8
E→S	62.34	62.78	62.75	62.27	61.94
S→E	41.20	41.29	41.34	40.58	40.67

Table 9. Experimental result with different k of TopK on DA setting.

References

- [1] Jian Chen, Wei Wang, Yuzhu Hu, Junxin Chen, Han Liu, and Xiping Hu. TGCA-PVT: Topic-guided context-aware pyramid vision transformer for sticker emotion recognition. In *ACM MM*, 2024. 2
- [2] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017. 1
- [3] Shengzhe Liu, Xin Zhang, and Jufeng Yang. SER30K: A large-scale dataset for sticker emotion recognition. In *ACM MM*, pages 33–41. ACM, 2022. 1, 2
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2
- [5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 1
- [6] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. MDAN: multi-level dependent attention network for visual emotion analysis. In *CVPR*, pages 9469–9478. IEEE, 2022. 2
- [7] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *ICCV*, pages 20326–20337. IEEE, 2023. 1