Supplementary Materials: Bridge then Begin anew: Generating Target-Relavent Intermediate Model for Source-Free Visual Emotion Adaptation

Anonymous submission

Overview

In the Supplementary Material of the Bridge then Begin Anew (BBA), we introduce more preliminary knowledge, detail the experimental setup, present the accuracy for each category, and provide an extended analysis of the ablation studies to demonstrate the effectiveness of the different components within our proposed BBA.

Preliminary Knowledge

Domain Gap for VER

Here, we thoroughly discuss the affective gap between the VER datasets and the inherent noise within the VER dataset. As illustrated in Figure 1, although all eight images depict the sky at sunset, they are categorized under positive sentiment in the FI dataset but negative sentiment in Emotion6. The differences in sentiment categorization across datasets lead to an affective gap. This discrepancy between datasets contributes to inaccurate pseudo-labeling. To address this, we propose the bridge model to mitigate domain affective differences and produce more accurate pseudo-labels.

In Figure 1 (a) of the main paper, the categories 'contentment' and 'anger' both depict a dog but convey different emotions, introducing confuse into the classifier. This noise reduces the separability of the categories, leading to overfitting of the source model. To reduce the impact of source domain noise, we train the target model from scratch, enabling it to learn from the target structure independently.

Source Model

The review paper (Li et al. 2024) states, 'the most significant difference between SFDA and UDA is that the UDA model can be trained using both the source and target domain data, while SFDA can only utilize the source model to initialize the target model and then update it with the unlabeled target data.' Like other SFDA methods, BBA uses the source data when training the source model. While during the adapting process, it only uses the source model and does not access source data.

We describe how to acquire the source model from the source domain. This learning process involves mapping inputs x_s from the source domain to the corresponding labels y_s to minimize cross-entropy loss. The network consists of a feature extractor g and a classifier f representing

Figure 1: Illustration of affective gap in the VER Datasets.

the source model as $\phi_s = f_s \circ g_s$. The cross-entropy loss function denoted L_{ce}, is computed as the expected negative log-likelihood of the correct labels:

$$L_{ce} = -\mathbb{E}_{x_s \in X_s} \left[\sum_{k=1}^{K} y_k^s \log \sigma(\phi_s(x_s)) \right], \qquad (1)$$

where $X_s = \{x_s^i\}_{i=1}^{N_S}$ and K indicates the number of categories. Following (Liang et al. 2022), we introduce label smoothing (LS) into the learning process. LS modifies the label distribution, which helps to improve model generalization. $y_k^s = (1-\alpha)y_k + \alpha/K$, y_k represents the label of the k^{th} category, α is the smoothing factor and σ is the softmax function.

Comprehensive Experimental Setup

Datasets

ArtPhoto (Machajdik and Hanbury 2010) dataset contains 806 artistic photographs, each of which is classified into one of Mikels' eight emotion categories (Zhao et al. 2016) by the original artists. These artists uploaded their work to an art-sharing platform using the emotion categories as keywords. They designed each image to evoke specific emotions by carefully manipulating composition, lighting, and color. The approach to photo creation and categorization allows us to examine how artists' intentional use of visual elements influences emotion classification in neural networks, providing insights into the interplay between artistic intent and automated emotion recognition.

FI (You et al. 2016) dataset is a collection of emotionally rich images from Flickr and Instagram. Similar to ArtPhoto,

Table 1: Classification accuracy comparison between BBA and state-of-the-art approaches on FI → Emotion6. SF means source-free. Pos and neg represent the positive and negative emotions, respectively.

Method	SF	pos	neg	avg
Source only	-	80.27	65.30	70.66
Oracle	-	71.91	89.18	82.99
SHOT	✓	84.28	58.77	67.90
SHOT++	✓	84.45	57.65	67.25
G-SFDA	✓	84.62	56.99	66.89
DaC	✓	86.96	54.20	65.93
AaD	✓	83.61	59.05	67.84
DINE	✓	79.60	66.70	71.32
TPDS	✓	**88.96**	48.88	63.23
BBA (ours)	✓	78.43	**70.34**	**73.23**
MCC	×	74.41	65.49	68.68
ELS	×	73.08	**70.34**	71.32
MIC	×	84.45	54.48	65.21

each image has one of the eight emotional categories defined by (Zhao et al. 2016), ensuring a nuanced range of human emotions. A group of 225 skilled Amazon Mechanical Turk workers labeled the images, with 23,308 images making the final cut based on receiving at least three matching ratings from these workers. The FI dataset contains a diverse range of human emotions, providing a valuable resource for advanced research in visual emotion recognition.

EmoSet (Yang et al. 2023) dataset comprises 3.3 million images, 118,102 carefully labeled by human annotators. EmoSet includes images from social networks and artistic images, which are well-balanced between different emotion categories. Motivated by psychological studies, in addition to the emotion category, each image is annotated with a set of describable emotion attributes: brightness, colorfulness, scene type, object class, facial expression, and human action, which help to understand visual emotions in a precise and interpretable way.

Emotion6 (Peng et al. 2015) dataset contains 1,980 images collected from Flickr. Each image is labeled by 15 annotators based on the Ekman model (Ekman 1992), covering six emotion categories: happiness, anger, disgust, fear, sadness, and surprise.

We adopt six settings of SFDA: EmoSet → FI, FI → EmoSet, Emotion6 → FI, FI → Emotion6, ArtPhoto → FI and FI → ArtPhoto. Given the categorical discrepancy between the FI and Emotion6 datasets, we consider Emotion6 → FI and FI → Emotion6 as binary classification tasks, focusing only on distinguishing between positive and negative emotional content in images. The dataset configuration is the same as WSCNet (She et al. 2019).

Evaluation Metrics

Building on the methodology outlined in (Zhao et al. 2022), we adopt **Accuracy** (Acc) as the primary metric to evaluate the performance of all methods. Accuracy is the ratio of correct predictions to the total number of samples. It provides a straightforward measure of methods across different

Table 2: Classification accuracy comparison between BBA and state-of-the-art approaches on Emotion6 → FI.

Method	SF	pos	neg	avg
Source only	-	60.65	86.44	68.18
Oracle	-	94.79	75.63	89.20
SHOT	✓	59.39	86.21	67.22
SHOT++	✓	60.49	**88.63**	68.71
G-SFDA	✓	55.97	88.24	65.39
DaC	✓	55.29	86.60	64.43
AaD	✓	62.53	88.24	70.04
DINE	✓	66.02	85.50	71.71
TPDS	✓	66.44	86.91	72.42
BBA(ours)	✓	**81.60**	69.67	**78.12**
MCC	×	74.52	72.88	74.04
ELS	×	71.39	77.35	73.12
MIC	×	66.28	85.65	71.94

emotion categories. In addition to overall performance, we analyze the accuracy of models on individual emotion categories. This detailed evaluation allows us to identify which emotions are recognized well and which may require further optimization.

To provide a comprehensive evaluation, we calculate both the **Macro Average** and **Weighted Average** metrics. The Macro Average metric treats each emotion category equally, providing an impartial view of model performance across all categories, regardless of their number of samples. The Weighted Average, on the other hand, considers the class distribution by weighting the contribution of each category according to its frequency in the dataset.

Baseline

We are the first to explore the source-free domain adaptation for visual emotion recognition tasks. The effectiveness of our proposed BBA is demonstrated through a comparative analysis with the following baselines:

Source only: It is the baseline in which the model is trained in the source domain and directly applied to the target domain.

SFDA methods: Our proposed BBA is mainly evaluated against state-of-the-art source-free domain adaptation techniques. It includes self-supervised approaches using pseudo-labels, like SHOT (Liang, Hu, and Feng 2020), which uses DeepCluster to compute class centroids. SHOT++ (Liang et al. 2021) extends SHOT by using MixMatch to increase the reliability of predictions from low-confidence samples by merging information with high-confidence matches. Furthermore, G-SFDA (Yang et al. 2021) applies local clustering to enforce consistency, while AaD (Yang et al. 2022) considers SFDA a clustering challenge and advocates prediction consistency based on feature space proximity. In addition, DaC (Zhang et al. 2022b) introduces an adaptive contrast learning framework designed for target-specific sample selection. DINE (Liang et al. 2022) and TPDS (Tang et al. 2024) propose adaptation strategies focusing on knowledge transfer and agent distribution streams to mitigate substantial source-target domain gaps.

UDA methods: Comparative evaluations also include

Table 3: Performance comparison between BBA and state-of-the-art approaches on FI ↔ ArtPhoto.

Method	SF	ArtPhoto → FI							FI → ArtPhoto						
		Acc	macro avg			weighted avg			Acc	macro avg			weighted avg		
			P	R	F1	P	R	F1		P	R	F1	P	R	F1
Source only	-	23.86	-	-	-	-	-	-	29.11	-	-	-	-	-	-
Oracel	-	66.11	-	-	-	-	-	-	43.67	-	-	-	-	-	-
SHOT	✓	26.73	26.21	30.10	26.87	27.65	26.73	26.06	33.95	40.41	37.41	35.45	40.81	33.95	34.21
SHOT++	✓	26.55	27.05	29.76	27.13	28.11	26.55	26.16	33.33	39.29	37.01	34.80	39.43	33.33	33.34
G-SFDA	✓	26.57	26.94	30.17	27.27	28.10	26.57	26.30	35.80	40.41	35.49	35.97	41.68	35.80	36.91
DaC	✓	25.54	25.06	26.09	23.79	26.09	25.54	24.32	35.19	36.94	36.19	35.01	37.62	35.19	34.83
AaD	✓	25.96	24.97	28.07	25.00	27.18	25.96	24.98	32.09	38.25	36.18	33.16	38.09	32.09	31.38
DINE	✓	27.03	24.83	27.99	24.61	26.36	27.03	25.34	33.33	40.56	37.73	33.56	40.21	33.33	31.36
TPDS	✓	28.20	29.62	31.96	28.94	31.85	28.20	28.19	32.72	38.93	34.69	33.73	39.40	32.72	33.07
BBA(ours)	✓	29.63	**35.65**	**32.51**	**30.02**	**35.77**	**29.63**	**28.92**	37.65	41.05	**39.88**	**38.57**	41.42	**37.65**	37.90
CycleEmotionGAN	×	25.20	-	-	-	-	-	-	37.18	-	-	-	-	-	-
CycleEmotionGAN++	×	**32.01**	-	-	-	-	-	-	**40.51**	-	-	-	-	-	-
MCC	×	25.75	27.10	26.68	25.03	30.85	25.75	26.44	36.42	**41.70**	39.01	37.01	**42.48**	36.42	36.25
ELS	×	25.29	26.12	27.56	25.12	28.48	25.29	25.19	37.04	41.20	39.00	38.10	42.19	37.04	**37.94**
MIC	×	25.04	24.53	27.80	24.19	28.27	25.04	24.79	34.57	36.72	36.28	34.15	37.61	34.57	33.75

UDA techniques that use source domain data during adaptation. In particular, our proposed BBA frequently outperforms UDA methods, including MCC (Jin et al. 2020), which introduces minimum class confusion loss, ELS (Zhang et al. 2022a), which advocates label smoothing, and MIC (Hoyer et al. 2023), which emphasizes context-inferred predictions for masked target images.

Emotion UDA methods: We also conduct comparisons with emotion-specific UDA methods, including CycleEmotionGAN (Zhao et al. 2019) and CycleEmotionGAN++ (Zhao et al. 2022), which use cyclic coherent adversarial networks for emotion domain adaptation.

Oracle: It assumes that classifier training and testing are performed exclusively within the target domain and serves as the upper bound of domain adaptation.

Implementation Details

Our experiments are conducted using PyTorch (Paszke et al. 2017) with a fixed random seed set to 2021 and conducted on a single NVIDIA A100 GPU. For a fair comparison, all methods implemented in this paper use ResNet-101 (He et al. 2016) as the backbone, which is pre-trained on ImageNet. We use SGD as the optimizer. Learning rates are set to 1×10^{-3} for pre-trained layers and 1×10^{-2} for other layers. Other settings include a momentum of 0.9, a weight decay of 1×10^{-3}, a batch size of 64, and a bottleneck size 256. In the TMA step, considering the model is unfamiliar with the target domain data initially, we use an exponentially decaying update scheme for α_t, formulated as $\alpha_t = \alpha_0 \cdot e^{-kt}$, where α_0 is the initial momentum, k is the updating rate, and the subscript t represents the training iter.

For inference, our model maintains the same computational complexity as SHOT, with 7.865G FLOPs and 43.026M parameters, which is the size of the source model (resnet101). Moreover, target and source models are not required to be the same in BBA, allowing smaller target model (*e.g.* resnet18) for faster inference.

Comprehensive Experimental Results

Overall Experimental Results on FI ↔ ArtPhoto

Due to space limitations in the main text, we present the experimental results on FI ↔ ArtPhoto here.

In Table 3, we adpot the same dataset settings as CycleEmotionGAN and CycleEmotionGAN++ for a direct comparison. These two methods perform unsupervised domain adaptation in VER. Macro average and weighted average values for Source only, Oracle, and the two methods are marked with a '-' as they are unavailable in their papers.

From the results, it is clear that our method shows commendable performance on both ArtPhoto → FI and FI → ArtPhoto. On ArtPhoto → FI, the overall accuracy of our method surpasses that of previous state-of-the-art SFDA methods and even surpasses the general UDA methods in some cases. BBA also outperforms CycleEmotionGAN, a UDA method tailored for emotion datasets. Although the performance of BBA does not surpass that of CycleEmotionGAN++, it is essential to acknowledge that UDA methods with access to the source domain data during adaptation generally achieve superior results. BBA also performs best in both macro and weighted average evaluation metrics.

For FI → ArtPhoto, BBA outperforms all the other methods except CycleEmotionGAN++. It also outperforms UDA methods that benefit from access to the source domain data during adaptation. All of the above demonstrate the effectiveness of our BBA.

Category-level Accuracy across Datasets

This section focuses on BBA's category-level accuracy on six benchmarks.

In Table 1, BBA excels at accurately identifying negative (neg) emotional states on FI → Emotion6, achieving the highest accuracy of 70.34%. It demonstrates BBA's nuanced understanding and effective classification of negative emotions. In addition, BBA achieves the best overall average performance with a score of 73.23%, demonstrating its superior ability to maintain high performance across different

Table 4: Classification accuracy comparison between BBA and state-of-the-art approaches on FI → EmoSet .

Method	SF	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
Source only	-	41.59	32.90	58.92	71.82	74.77	56.30	21.73	45.42	50.87
Oracle	-	70.47	77.93	77.57	57.33	80.79	76.41	64.59	74.15	71.96
SHOT	✓	39.31	37.3	66.19	60.39	79.83	59.52	30.60	54.09	53.31
SHOT++	✓	40.97	43.85	68.58	55.92	81.16	56.44	39.46	51.06	54.15
G-SFDA	✓	42.02	36.71	63.5	60.87	80.07	63.04	34.65	53.44	54.40
DaC	✓	**45.96**	40.11	63.81	55.76	68.51	50.27	38.70	50.41	51.55
AaD	✓	39.89	41.57	63.85	53.06	80.13	53.09	40.48	**55.61**	52.68
DINE	✓	43.97	36.71	64.47	**74.44**	**84.95**	62.87	30.50	50.14	56.36
TPDS	✓	43.32	37.12	64.73	73.51	83.80	60.78	33.03	48.78	55.88
BBA(ours)	✓	41.34	**44.61**	66.73	57.33	84.29	**72.93**	**56.82**	53.44	**58.76**
MCC	×	42.89	31.73	64.34	73.34	75.26	61.51	46.81	50.57	56.26
ELS	×	43.25	43.74	62.96	71.58	79.35	58.66	43.67	48.78	56.41
MIC	×	39.35	44.15	**69.25**	56.8	81.58	58.20	40.98	47.86	54.31

Table 5: Classification accuracy comparison between BBA and state-of-the-art approaches on FI → EmoSet.

Method	SF	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
Source only	-	65.08	35.59	64.47	38.60	56.74	39.27	55.67	53.70	51.57
Oracle	-	82.54	37.29	71.06	76.61	66.77	49.54	26.80	64.90	67.32
SHOT	✓	60.53	36.86	62.91	37.52	60.82	42.02	55.67	49.91	50.38
SHOT++	✓	53.33	47.88	64.64	39.28	61.44	45.87	56.7	48.20	50.42
G-SFDA	✓	53.97	**48.31**	64.82	39.77	60.82	45.32	56.70	47.06	50.47
DaC	✓	59.79	37.29	61.53	36.94	59.25	42.39	51.55	**54.65**	50.24
AaD	✓	51.53	38.98	63.26	38.30	63.32	49.91	**63.92**	46.49	49.90
DINE	✓	67.51	38.56	66.03	43.18	64.58	47.71	55.15	**54.65**	55.25
TPDS	✓	63.49	38.56	68.46	38.21	58.93	43.85	56.19	44.40	51.45
BBA(ours)	✓	**74.92**	39.41	63.08	47.08	**69.59**	47.71	51.03	52.56	**57.36**
MCC	×	67.83	32.20	**70.54**	48.05	62.70	**57.43**	51.03	46.11	56.58
ELS	×	68.25	34.75	70.36	49.32	54.55	49.36	51.03	53.89	56.42
MIC	×	66.24	31.36	67.07	**49.42**	65.51	55.05	55.15	47.06	56.26

emotional states. It highlights the robustness and adaptability of BBA, making it a valuable contribution to the emotion recognition community.

In Table 2, BBA gains outstanding performance, particularly in the positive (pos) emotion category, achieving the highest score of 81.60%. It indicates the ability of BBA to accurately identify positive emotional states on FI → Emotion6. Furthermore, BBA stands out in terms of average performance, achieving the highest average score of 78.12% across both emotional states. This superior average performance highlights the effectiveness of BBA in providing high-quality performance across different emotional contexts, demonstrating its robustness in emotion recognition.

Table 4 clearly shows that BBA stands out in several EmoSet → FI aspects. BBA achieves the highest performance in the Amusement (Amu) category with a score of 41.34%, demonstrating its effectiveness in identifying moments of joy and entertainment. Most impressively, BBA excels in the Anger (Ang) category, achieving the highest score of 44.61 %, indicating a solid ability to identify instances of anger accurately. In addition, BBA leads in the Excitement (Exc) and Fear (Fea) categories, with exceptional scores of 72.93% and 56.82%, respectively, beating the second best by margins of 10.01% and 9.89%. It demonstrates BBA's ability to recognize intense emotional states ranging from positive excitement to negative fear, underlining its comprehensive understanding of emotional dynamics. Besides the category-specific performance, BBA achieves the best overall average performance across all emotional states with a score of 58.76%. The superior average performance highlights BBA's robustness and adaptability, especially in diverse and complex emotional contexts.

In Table 5, BBA also shows its exceptional ability to recognize different emotional states on FI → EmoSet, in particular, excelling in the Amusement (Amu) and Disgust (Dis) with the highest scores of 74.92% and 69.59% respectively. The performance in identifying both a positive emotional state, such as amusement, and a negative one, such as disgust, highlights BBA's nuanced understanding of emotional recognition tasks. Furthermore, BBA achieves the highest overall average performance with a score of 57.36%, demonstrating its robustness and consistency across a wide range of emotional states. The average performance highlights the overall effectiveness of our method in emotion recognition.

As shown in Table 6, in the recognition of specific emotions, BBA scores 50.00% in the identification of contentment (Con), which shows that BBA is quite good at recognizing specific emotions, despite the challenges of not having direct access to the source data in SFDA. In terms of overall consistency, BBA has an average score of 37.65%,

Table 6: Classification accuracy comparison between BBA and state-of-the-art approaches on FI → ArtPhoto.

Method	SF	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
Source only	-	30.00	28.57	55.00	35.71	14.29	20.00	18.18	30.30	29.11
Oracle	-	55.00	35.71	30.00	14.29	42.86	55.00	59.09	45.45	43.67
SHOT	✓	14.29	42.86	50.00	30.00	83.33	17.39	22.73	38.71	33.95
SHOT++	✓	14.29	42.86	50.00	30.00	83.33	17.39	22.73	35.48	33.33
G-SFDA	✓	16.67	35.29	38.89	25.00	68.75	27.27	29.17	42.86	35.80
DaC	✓	23.53	**58.33**	33.33	35.29	60.0	21.74	15.79	41.46	35.19
AaD	✓	14.29	38.10	50.00	30.00	**91.67**	13.04	13.64	38.71	32.09
DINE	✓	23.80	38.10	58.33	45.00	83.33	8.70	9.09	35.48	33.33
TPDS	✓	19.05	42.86	41.67	40.00	58.33	17.39	22.72	35.48	32.72
BBA(ours)	✓	28.57	42.86	50.00	50.00	58.33	17.39	36.36	35.48	37.65
CycleEmotionGAN	×	20.00	28.57	**60.00**	42.86	28.57	**40.00**	27.27	43.75	37.18
CycleEmotionGAN++	×	30.00	36.00	40.00	21.43	14.29	40.00	**77.27**	**45.45**	**40.51**
MCC	×	28.57	38.10	50.00	45.00	66.67	8.70	36.36	38.71	36.42
ELS	×	28.57	42.86	50.00	35.00	58.33	13.04	45.45	38.71	37.04
MIC	×	**33.33**	42.86	33.33	**55.00**	58.33	26.09	9.09	32.25	34.57

Table 7: Classification accuracy comparison between BBA and state-of-the-art approaches on ArtPhoto → FI.

Method	SF	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
Source only	-	47.63	2.83	25.86	6.33	5.57	8.67	16.50	51.15	23.86
Oracle	-	77.24	44.88	72.03	65.59	60.12	61.40	48.00	65.37	66.11
SHOT	✓	34.71	31.36	33.62	9.06	55.80	12.29	30.93	33.02	26.73
SHOT++	✓	36.30	32.20	31.54	8.97	54.23	13.76	30.93	30.17	26.55
G-SFDA	✓	32.70	32.63	32.58	9.94	54.23	14.13	32.47	32.64	26.57
DaC	✓	44.02	36.02	3.29	18.71	52.98	9.36	14.95	29.41	25.54
AaD	✓	38.94	28.81	20.45	9.65	51.10	11.93	24.74	38.90	25.96
DINE	✓	**47.30**	41.10	4.16	17.93	57.99	8.26	14.95	32.26	27.03
TPDS	✓	28.78	26.27	29.29	13.55	58.31	**28.44**	37.63	33.40	28.20
BBA(ours)	✓	14.29	28.57	**50.00**	**35.00**	**66.66**	8.70	18.18	38.70	29.63
CycleEmotionGAN	×	35.39	16.26	33.44	6.81	35.40	23.05	20.20	35.23	25.20
CycleEmotionGAN++	×	44.86	40.49	18.33	32.99	30.96	17.88	**50.00**	27.53	**32.01**
MCC	×	26.67	12.29	24.09	16.86	34.80	28.44	31.44	38.90	25.75
ELS	×	33.75	21.18	19.76	10.43	45.45	13.94	31.96	**44.02**	25.29
MIC	×	27.83	**41.95**	18.37	11.31	55.17	22.94	8.25	36.62	25.04

comparing well with other SFDA and some UDA methods. Although CycleEmotionGAN++ has a higher average score, BBA still holds its own, considering it cannot use the source data directly.

It is clear from Table 7 that BBA achieves the highest scores in three categories: Amusement (Amu), Contentment (Con), and Disgust (Dis), with scores of 50.00%, 35.00%, and 66.66%, respectively. It indicates the ability of BBA to capture the nuances of these particular emotions.

BBA gets an average score of 29.63%, the highest among the SFDA approaches and the second only to CycleEmotionGAN++, which has the advantage of accessing source domain data. It shows the efficiency of BBA across different emotions in the target domain. While CycleEmotionGAN++ leads with an average score of 32.01%, it must remember that it uses source domain data, which may not always be feasible in real-world applications due to privacy or logistical constraints.

Ablation Study

The ablation studies presented in Tables 8, 9, 10 and 11, focus on evaluating the incremental impact of different components within the BBA. These ablation studies assess the model's ability to classify dominant emotions across datasets: EmoSet ↔ FI, Emotion6 ↔ FI, and FI ↔ ArtPhoto. The ablation components include the baseline, clustering, masking, $\mathcal{L}_{align}$, and $\mathcal{L}_{pol}$, with each addition intended to improve the model's ability to generalize and accurately classify emotions across domains. Furthermore, we analyze the benefits of fused distance metrics.

Baseline improvement: Starting from the baseline, each subsequent addition of components shows a consistent improvement in accuracy (Acc) and F1 scores across all three dataset pairs. It indicates that each component makes a positive contribution to the performance of the model.

Effect of clustering (+cluster): The inclusion of clustering shows a slight performance improvement, suggesting that grouping similar features or instances helps the model to understand the underlying emotional patterns better. This is because some pseudo labels are far from the centroid, which indicates that they may be outliers. By reassigning these outlier pseudo-labels to the nearest centroid, we can mitigate the problem of inaccurate pseudo-labels to some extent.

Effect of masking (+mask): Adding a masking component leads to a noticeable increase in accuracy and F1 scores, especially on EmoSet ↔ FI scenarios. It suggests that focus-

Table 8: Ablation study on different components of our proposed BBA on EmoSet ↔ FI datasets.

Method	step	FI → EmoSet							EmoSet → FI						
		Acc	macro avg			weighted avg			Acc	macro avg			weighted avg		
			P	R	F1	P	R	F1		P	R	F1	P	R	F1
baseline	A	52.63	54.33	52.52	51.98	54.41	52.63	0.52	51.59	48.36	51.45	48.73	54.04	51.59	51.70
+cluster	A	53.22	53.85	53.28	52.52	54.14	53.22	0.53	51.66	48.56	51.52	48.79	54.11	51.66	51.70
+mask	A	57.00	58.28	57.04	56.75	57.63	57.00	56.41	54.50	50.70	53.52	50.92	57.09	54.50	54.71
+$\mathcal{L}_{align}$	B	57.38	58.54	57.44	57.16	57.91	57.38	56.84	56.79	52.21	55.13	52.63	59.69	56.79	57.32
+$\mathcal{L}_{pol}$	B	**58.76**	**60.32**	**58.83**	**58.52**	**59.62**	**58.76**	**58.19**	**57.36**	**52.69**	**55.67**	**53.27**	**59.92**	**57.36**	**57.76**

Table 9: Ablation study on different components of our proposed BBA on Emotion6 ↔ FI datasets.

Method	step	FI → Emotion6							Emotion6 → FI						
		Acc	macro avg			weighted avg			Acc	macro avg			weighted avg		
			P	R	F1	P	R	F1		P	R	F1	P	R	F1
baseline	A	70.42	70.27	72.04	69.76	74.23	70.42	71.03	66.81	68.49	72.23	65.82	77.72	66.81	68.24
+cluster	A	72.28	71.70	73.56	71.51	75.47	72.28	72.84	73.86	72.23	76.68	72.10	80.08	73.86	75.02
+mask	A	73.05	72.35	74.24	72.26	76.04	73.05	73.59	76.65	73.63	77.62	74.35	80.48	76.65	77.55
+$\mathcal{L}_{align}$	B	**73.23**	**72.49**	**74.38**	**72.43**	**76.15**	**73.23**	**73.77**	**78.12**	**73.84**	**75.64**	**74.55**	**79.19**	**78.12**	**78.52**

Table 10: Ablation study on different components of our proposed BBA on ArtPhoto ↔ FI datasets.

Method	step	FI → ArtPhoto							ArtPhoto → FI						
		Acc	macro avg			weighted avg			Acc	macro avg			weighted avg		
			P	R	F1	P	R	F1		P	R	F1	P	R	F1
baseline	A	33.95	35.92	34.34	32.95	37.28	33.95	33.67	27.40	27.67	29.26	26.63	29.90	27.40	27.03
+cluster	A	34.57	35.74	34.86	33.66	37.29	34.57	34.45	28.08	26.45	29.36	27.09	27.58	28.08	27.20
+masking	A	36.42	40.32	39.05	37.58	41.38	36.42	37.13	28.73	27.20	30.62	28.34	27.86	28.73	27.93
+$\mathcal{L}_{align}$	B	37.04	41.19	39.64	38.29	42.34	37.04	37.88	28.91	30.92	31.67	29.53	33.51	28.91	29.23
+$\mathcal{L}_{pol}$	B	**37.65**	**41.05**	**39.88**	**38.57**	**41.42**	**37.65**	**37.90**	**29.63**	**35.65**	**32.51**	**30.02**	**35.77**	**29.63**	**28.92**

Table 11: Ablation study on different distance metrics for clustering.

Method	E→F	F→E	A→F	F→A	E6→F	F→E6
Cosine	56.22	58.33	29.26	37.58	77.99	72.70
Euclidean	56.42	58.29	29.56	36.41	77.76	72.52
Manhattan	56.45	58.21	28.55	37.03	77.16	72.74
BBA(ours)	**57.36**	**58.76**	**29.63**	**37.65**	**78.12**	**73.23**

ing on relevant features or regions within the data can improve the prediction capabilities of the model and leverage more distinctive features.

Effect of alignment (+$\mathcal{L}_{align}$): The alignment step further improves the performance of the model, particularly in the weighted average F1 scores, indicating improved consistency of performance across different emotional classes. It suggests that aligning the feature spaces of the source and target domains is critical for effective emotion classification. This is because the fact that the alignment approach allows the target model to benefit from the bridge model, which acquires domain-invariant knowledge across the source and target domains. At the same time, it learns the features of the target domain from scratch, thereby expanding the parameter space.

Effect of polarity (+$\mathcal{L}_{pol}$): The final addition of a polarity component leads to the highest observed accuracy and F1 scores in almost all scenarios, particularly on Emotion6 ↔ FI. It highlights the importance of understanding emotional polarities in achieving high classification performance. Polarity loss delves into the hierarchical features of emotional images. It introduces additional feature constraints to distinguish features based on their polarity, thereby addressing the lack of significant features within classes caused by label ambiguity.

Effect of fused distance metrics: We conduct experiments on six settings using different distance metrics. As the table 11 shows, no single distance measure consistently outperforms the others, and each setting favors a specific single measure. This is because different distance metrics focus on various parts, *e.g.*, Cosine distance cannot provide information on the magnitude; Euclidean distance is invariant to translation and rotation but sensitive to the range of input attributes; Manhattan distance remains unchanged with coordinate system reflections but sensitive to rotation. The table below shows that our fused distance metric improves robustness in diverse data types, thus giving the best results.

The ablation studies show that each component improves BBA's ability to classify emotions across different dataset pairs accurately. The consistent improvements across different metrics and datasets underscore the effectiveness of BBA in exploiting these components to achieve superior performance in emotion recognition tasks. By adding clustering, masking, alignment, and polarity components, we improve the model's accuracy and ability to work well in different emotional settings. It makes our approach solid and flexible for classifying emotions in different domains.

Future Work

We plan to extend our methodology to encompass datasets with class label distribution. This expansion aims to adapt our framework to regression tasks, addressing the complexities of label distribution issues, thereby broadening the applicability of our approach in the field of emotion recognition.

References

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hoyer, L.; Dai, D.; Wang, H.; and Van Gool, L. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11721–11732.

Jin, Y.; Wang, X.; Long, M.; and Wang, J. 2020. Minimum class confusion for versatile domain adaptation. In *Proceedings of the European Conference on Computer Vision*, 464–480. Springer.

Li, J.; Yu, Z.; Du, Z.; Zhu, L.; and Shen, H. T. 2024. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 6028–6039.

Liang, J.; Hu, D.; Feng, J.; and He, R. 2022. Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8003–8013.

Liang, J.; Hu, D.; Wang, Y.; He, R.; and Feng, J. 2021. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8602–8617.

Machajdik, J.; and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the ACM International Conference on Multimedia*, 83–92.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Peng, K.-C.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 860–868.

She, D.; Yang, J.; Cheng, M.-M.; Lai, Y.-K.; Rosin, P. L.; and Wang, L. 2019. WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia*, 22(5): 1358–1371.

Tang, S.; Chang, A.; Zhang, F.; Zhu, X.; Ye, M.; and Zhang, C. 2024. Source-free domain adaptation via target prediction distribution searching. *International Journal of Computer Vision*, 132(3): 654–672.

Yang, J.; Huang, Q.; Ding, T.; Lischinski, D.; Cohen-Or, D.; and Huang, H. 2023. EmoSet: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 20383–20394.

Yang, S.; Jui, S.; van de Weijer, J.; et al. 2022. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Proceedings of the Advances in Neural Information Processing Systems*, 5802–5815.

Yang, S.; van de Weijer, J.; Herranz, L.; Jui, S.; et al. 2021. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, 29393–29405.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, Y.; Liang, J.; Zhang, Z.; Wang, L.; Jin, R.; Tan, T.; et al. 2022a. Free Lunch for Domain Adversarial Training: Environment Label Smoothing. In *Proceedings of the International Conference on Learning Representations*.

Zhang, Z.; Chen, W.; Cheng, H.; Li, Z.; Li, S.; Lin, L.; and Li, G. 2022b. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In *Proceedings of the Advances in Neural Information Processing Systems*, 5137–5149.

Zhao, S.; Chen, X.; Yue, X.; Lin, C.; Xu, P.; Krishna, R.; Yang, J.; Ding, G.; Sangiovanni-Vincentelli, A. L.; and Keutzer, K. 2022. Emotional semantics-preserved and feature-aligned cyclegan for visual emotion adaptation. *IEEE Transactions on Cybernetics*, 52(10): 10000–10013.

Zhao, S.; Lin, C.; Xu, P.; Zhao, S.; Guo, Y.; Krishna, R.; Ding, G.; and Keutzer, K. 2019. Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2620–2627.

Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; Xie, W.; Jiang, X.; and Chua, T.-S. 2016. Predicting personalized emotion perceptions of social images. In *Proceedings of the ACM International Conference on Multimedia*, 1385–1394.