Cluster-level pseudo-labelling for source-free cross-domain facial expression recognition

Alessandro Conti[1]
alessandro.conti-1@unitn.it
Paolo Rota[1,2]
paolo.rota@unitn.it
Yiming Wang[3]
ywang@fbk.eu
Elisa Ricci[1,3]
e.ricci@unitn.it

[1] DISI - Department of Information Engineering and Computer Science
University of Trento

[2] CIMeC - Center for Mind and Brain Sciences
University of Trento

[3] Fondazione Bruno Kessler (FBK)

Abstract

Automatically understanding emotions from visual data is a fundamental task for human behaviour understanding. While models devised for Facial Expression Recognition (FER) have demonstrated excellent performances on many datasets, they often suffer from severe performance degradation when trained and tested on different datasets due to domain shift. In addition, as face images are considered highly sensitive data, the accessibility to large-scale datasets for model training is often denied. In this work, we tackle the above-mentioned problems by proposing the first Source-Free Unsupervised Domain Adaptation (SFUDA) method for FER. Our method exploits self-supervised pretraining to learn good feature representations from the target data and proposes a novel and robust cluster-level pseudo-labelling strategy that accounts for in-cluster statistics. We validate the effectiveness of our method in four adaptation setups, proving that it consistently outperforms existing SFUDA methods when applied to FER, and is on par with methods addressing FER in the UDA setting.
Code is available at https://github.com/altndrr/clup.

1 Introduction

Facial Expression Recognition (FER) [11, 31, 32, 33] refers to the task of automatically inferring the emotional state of a person from a facial image, which supports multiple application fields, such as assistive robotics and security monitoring. However, each individual shows their emotional state differently according to their personal traits or complicated cultural/ethical factors [3]. Such heterogeneity in the data space remains one of the main challenges for a generalisable model for FER. In the last twenty years, the efforts to improve such technologies have been mostly split between collecting larger and more diverse datasets [29, 37] and advancing learning algorithms for improving generalisation capability in the wild [11, 41]. Many recent techniques for FER exploit the attention mechanism [1, 9, 28, 30, 34], while

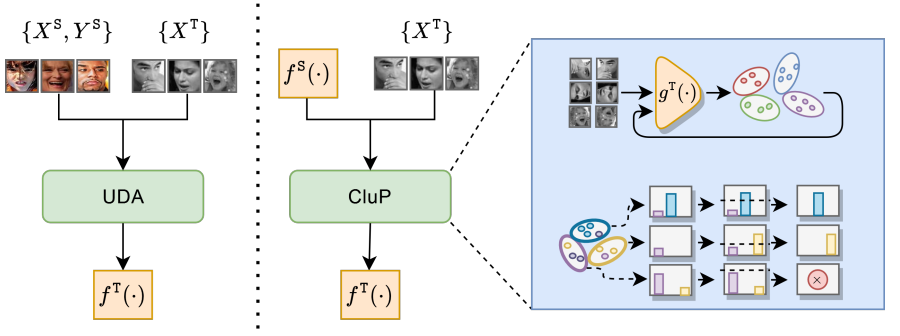


Figure 1: Comparison between previous works (the left part) and our CluP on cross-domain FER (the right part). Differently from past works, we aim to learn a target model $f^T(\cdot)$ with only source model $f^S(\cdot)$ and unlabelled target data $\{X^T\}$ without the source data $\{X^S, Y^S\}$, a very likely scenario due to privacy concerns. Our solution, CluP, is the first method on source-free domain adaptation for FER, exploiting self-supervised learning (SSL) to warm up the target feature extractor $g^T(\cdot)$ and a novel cluster-level pseudo-labelling technique.

some other works learn uncertainty via feature mixup [41], or improve feature representations by replacing the pooling layers to reduce padding erosion [32].

Recent works often frame the problem from an Unsupervised Domain Adaptation (UDA) perspective where labels of the target samples are not available [14, 20, 21]. For example, in [22], Li *et al.* introduce a novel loss function to preserve feature locality despite the domain shift. Such loss also organises facial expressions according to their intensity in the embedding space. A more recent method [6] exploits facial landmarks and holistic features to adapt to the target domain with adversarial learning applied on graphs.

While all these methods improve the adaptability of FER models across data distributions, the source data is required during adaptation. However, when dealing with facial images, the source data might not be available due to the increasingly stringent regulations concerning the privacy of citizens. Therefore, we are motivated to address the more challenging problem of Source-Free Unsupervised Domain Adaptation (SFUDA) for FER, given only the availability of the source pretrained model (see Fig. 1). To the best of our knowledge, we are the first to propose a domain adaptation solution for FER that works without the source facial data, embracing a privacy-preserving learning paradigm as the source data can remain private.

Our proposed method, CluP (**Clu**ster-level **P**seudo-labelling), exploits self-supervised learning (SSL) on the target data and proposes a novel cluster-level pseudo-labelling technique. Pseudo-labelling for UDA often extends the source model to the target domain using the source confidence to select the best target training inputs [26, 40]. However, the computation of confidence requires supervised training, which is only possible in UDA with the access to the source data. In the case of SFUDA, as the domain gap increases, one can expect a degrading representation capability of the source model on the target domain. Recent advances in SSL shows that a good data representation can be learnt without annotated labels [5, 7, 12]. In this work, we propose to exploit SSL techniques for a good starting feature representation for the target model, and further propose to improve the reliability of pseudo-labels with our newly introduced *cluster purity*, *i.e.* the local statistics of target samples that are clustered within the feature space expressed by the source model. We validate

CluP on a set of cross-domain FER benchmarks and prove its advantageous performance in terms of classification accuracy.

We summarise our contributions as follows:

- We present CluP, the first method addressing Source-free Unsupervised Domain Adaptation for FER, exploiting SSL to foundation our target model.
- CluP introduces a novel cluster-level pseudo-labelling scheme to improve the reliability of pseudo-labels based on in-cluster attributes that deviates from traditional confidence-based pseudo-labelling methods.
- We demonstrate that CluP surpasses competing methods for SFUDA and is comparable with UDA techniques on several FER adaptation benchmarks.

2 Related work

In the following, we present recent works on UDA methods for FER, and some general-purpose SFUDA solutions.

Unsupervised Domain Adaptation for FER. As a consequence of the domain bias, quite prominent among FER datasets, some works focus on domain adaptation [6, 14, 20, 21, 22, 38, 43]. In [14], Ji *et al.* apply late fusion on the outputs of two channels that learn intra-category and inter-category similarities of facial expressions. The authors of [22] introduce a locality preserving loss that draws samples of the same class closer. They also notice that neighbouring samples in the embedding space present similar emotional intensities. DETN [20] applies two variations of the Maximum Mean Discrepancy to assess the amount of divergence between the domains and to re-weight the class-wise source distribution to match the target. The authors extend the work in [21], where they additionally consider the differences in the conditional distributions. Differently from the above works, AGRA [6] focuses on the well-established approach of adversarial domain adaptation, leveraging facial landmarks alongside facial images. For the landmarks, they introduce two specialised graph neural networks while jointly considering the domain feature distributions, the local features (*i.e.* , landmarks), and the holistic features, achieving the best results on many benchmarks.

Compared to previous works, we consider a stricter setting where the source data is unavailable. We argue that, due to privacy issues, human behaviour understanding methods do not always have access to the source data. For this reason, we introduce a novel method for FER that adapt to a target domain in a source-free manner.

Source-Free Unsupervised Domain Adaptation. Recently, novel methods for source-free domain adaptation have been proposed [13, 17, 19, 23, 25, 39]. The setting represents a more complex but realistic scenario of UDA, where source data is unavailable. Some works resort to entropy-minimisation losses to adapt to the target domain without labels. For example, SHOT [23] employs an entropy loss alongside a classification loss on pseudo-labelled samples to adapt the network to the target domain. The work has been extended in [24] introducing an auxiliary head that solves relative rotation, leading to improved performance. Differently from the above, the authors of [13] frame the problem from an image translation perspective and translate the target images to the source style using only the source model. In [36], they perform self-training with a loss function that considers the intrinsic structure of the target domain via nearest neighbours. In the proposed work, we do not impose any constraint on the loss function, our refinement step works on source clusters, and we propose a novel score function to select the best samples to train on the target domain. Other works address open-set or universal domain adaptation [15, 16] without access to the source data.

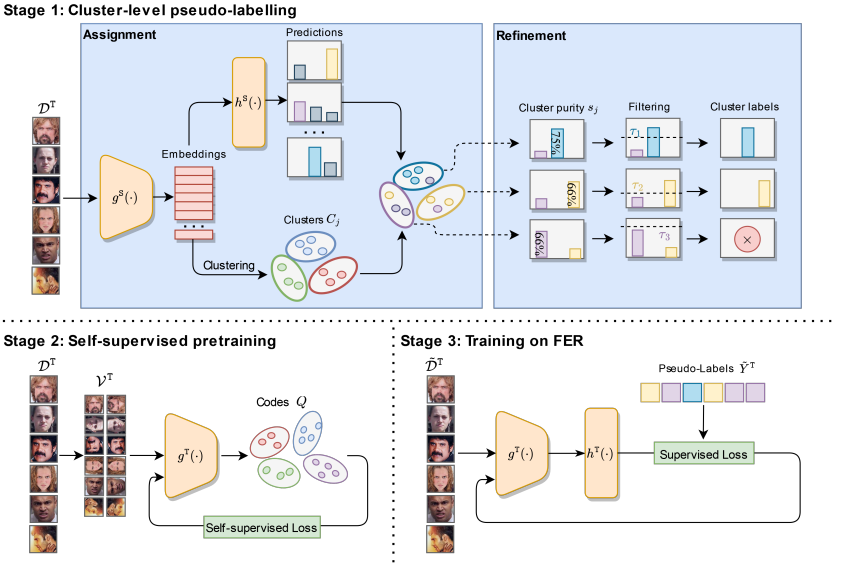


Figure 2: Our proposed CluP comprises of three-stage training, where the first stage produces trustworthy cluster-level pseudo-labels using the source model, the second stage warms up the target model in a self-supervised fashion, and finally the third stage performs the target model training with the refined pseudo-labels.

Unlike the previous works, our model does not rely only on the source model but is constructed based on independent self-supervised training on the target data. Moreover, we refine pseudo-labels by reducing unreliable samples using a novel decision metric at the cluster level based on cluster purity.

3 Method

The traditional closed-set UDA problem setting allows the access to the annotated source dataset $\mathcal{D}^{\mathrm{S}} = \{\mathbf{x}_i^{\mathrm{S}}, y_i^{\mathrm{S}}\}_{i=1}^{M^{\mathrm{S}}}$, and a target dataset $\mathcal{D}^{\mathrm{T}} = \{\mathbf{x}_i^{\mathrm{T}}\}_{i=1}^{M^{\mathrm{T}}}$ without annotations, where the target domain shares the same label space with the source, *i.e.* $\mathcal{Y}^S = \mathcal{Y}^{\mathrm{T}} = \{1, ..., N\}$.

Differently, the SFUDA protocol does not allow the access to the source dataset $\mathcal{D}^{\mathrm{S}}$, but only to a trained source model $f^{\mathrm{S}}(\cdot) : \mathcal{X}^{\mathrm{S}} \to \mathbb{R}^N$, which consists of a feature extractor $g^{\mathrm{S}}(\cdot) : \mathcal{X}^{\mathrm{S}} \to \mathbb{R}^Z$ and a classifier $h^{\mathrm{S}}(\cdot) : \mathbb{R}^Z \to \mathbb{R}^N$, where Z is the feature dimension.

Our proposed method CluP tackles the problem of SFUDA for FER. As illustrated in Fig. 2, CluP follows a three-stage training strategy where the first two can run in parallel: the first stage produces more trustworthy cluster-level pseudo-labels $\{\tilde{y}_i^{\mathrm{T}}\}_{i=1}^{\tilde{M}^{\mathrm{T}}}$ for a subset of $\tilde{M}^{\mathrm{T}}$ target samples $\tilde{\mathcal{D}}^{\mathrm{T}} = \{\mathbf{x}_i^{\mathrm{T}}\}_{i=1}^{\tilde{M}^{\mathrm{T}}}$ by exploiting the available $f^{\mathrm{S}}(\cdot)$ and our proposed cluster purity for pseudo-label refinement (described in Sec. 3.1), while in the second stage, a target feature extractor $g^{\mathrm{T}}(\cdot)$ is learned in a self-supervised fashion (described in Sec. 3.2). During the third stage, $g^{\mathrm{T}}(\cdot)$ is extended with a classifier $h^{\mathrm{T}}(\cdot)$ and the whole network is trained with the subset of target samples $\tilde{\mathcal{D}}^{\mathrm{T}}$ accompanied by their refined pseudo-labels (described in Sec. 3.3).

3.1 Cluster-level Pseudo-labelling

Pseudo-labels filtered by confidence that is produced by source model are often unreliable, particularly when the domain gap between the source and target is large. CluP exploits a clustering technique to group samples with similar characteristics (i.e *assignment*) and then uses a purity metric based on the source classifier to select the most reliable clusters (*i.e.* *refinement*).

Cluster pseudo-label assignment. First, the target features are extracted $\{\mathbf{z}_i^{\mathrm{T}}\}_{i=1}^{M^{\mathrm{T}}} \in \mathbb{R}^Z$ using the source feature extractor $g^{\mathrm{S}}(\cdot)$. Second, we cluster the features using K-means algorithm, resulting in a set of clusters $\{C_j\}_{j=1}^{K}$ Since FER often deals with highly unbalanced datasets, we perform over-clustering and consider $K \gg N$, to increase the chances that even minor classes can be expressed with some clusters. Leveraging the pseudo-labels predicted by the source model $\tilde{y}_i^{\mathrm{T}} = h^{\mathrm{S}}(\mathbf{z}_i^{\mathrm{T}})$ we assign to each cluster C_j a pseudo-label $\tilde{y}_j^{\mathrm{T}}$ that represents the majority-voted pseudo-label within each cluster.

Cluster pseudo-label refinement. As each cluster C_j should contain elements that are similar in the learned feature space, we might expect their pseudo-labels to expose an one-class distribution. Unfortunately, this is often not the case. However, a subset of clusters detaining a certain pseudo-label agreement can be defined using what we named as *cluster purity*.

Let us consider m_j^{T} as the cardinality of the j-th cluster, where $M^{\mathrm{T}} = \sum_{j=1}^{K} m_j^{\mathrm{T}}$. We define the *cluster purity* score s_j for each cluster C_j as the percentage of pseudo-labels $\{\tilde{y}_i^{\mathrm{T}}\}_{i=1}^{m_j^T}$ that agree with their cluster-level label $\tilde{y}_j^{\mathrm{T}}$:

$$s_j = \frac{\sum_{i=1}^{m_j^T} \mathbb{1}\left\{\mathbf{x}_i^{\mathrm{T}} \in C_j : \tilde{y}_i^{\mathrm{T}} = \tilde{y}_j^{\mathrm{T}}\right\}}{m_j^T}. \tag{1}$$

Given s_j per cluster, we can further refine the target dataset by only keeping clusters that have a *cluster purity* score higher than a threshold τ, *i.e.* the more reliable clusters, for training the target model. Considering that each category of the pseudo-labels might exhibit a different distribution, we design the *cluster purity* threshold τ to vary according to its category. Specifically, for the set of clusters that correspond to the same pseudo-label category $\{C_j\}_{\tilde{y}_j^{\mathrm{T}}=n}$ where $n \in \mathcal{Y}$, we select the Q-th percentile of their purity scores to serve as the threshold τ_n. Q is empirically set, and related experimental details are reported in Sec. 4.2.

After the cluster refinement, only clusters whose *cluster purity* score is higher than the threshold corresponding its category remain in the reduced target dataset $\tilde{\mathcal{D}}^{\mathrm{T}}$ and will be used for training the final target model $f^{\mathrm{T}}(\cdot)$.

3.2 Self-supervised pretraining

The pretraining of the target model is a delicate and important phase where the choice of the training method for warming up the backbone leads to relevant fluctuations in performance (see Sec. 4.2). In the specific, we noticed that a pretraining relying on self-supervision largely outperforms a model initialised with source weights. For this reason, inspired by SwAV [5], CluP exploits self-supervision on the target dataset to learn an initial feature extractor $g^{\mathrm{T}}(\cdot)$.

CluP performs clustering of the sample data while enforcing the consistency between cluster assignments produced for different augmentations of the same sample. First, target

features $\{g^{\mathrm{T}}(\mathbf{x}_i^{\mathrm{T}})\}_{i=1}^{M^{\mathrm{T}}}$ are grouped according to a similarity metric to retrieve N^P learnable prototypes $P = \{p_i\}_{i=1}^{N^P}$ and a set of codes $\{q_i^{\mathrm{T}}\}_{i=1}^{M^{\mathrm{T}}}$ where each sample is assigned to. Then, codes $\{q_i^{\mathrm{T}}\}_{i=1}^{M^{\mathrm{T}}}$ are used as targets to learn the optimal mapping to $\{g^{\mathrm{T}}(\mathbf{x}_i^{\mathrm{T}})\}_{i=1}^{M^{\mathrm{T}}}$ by minimising:

$$\mathcal{L}_c(\mathbf{x}_i, \mathbf{q}_i) = -\sum_{n=1}^{N^P} \mathbf{q}_i^{(n)} \log(\mathbf{p}_i^{(n)}) \tag{2}$$

where $\mathbf{q}$ is the one-hot vector of q and $\mathbf{p}$ is the softmax of the dot product of $g^{\mathrm{T}}(\mathbf{x}_i^{\mathrm{T}})$ and the cluster prototypes P.

By treating each sample as a class (*i.e.* $M^{\mathrm{T}} = N$), contrastive learning aims to learn a feature extractor $g^{\mathrm{T}}(\cdot)$ invariant to data augmentations. For each target image $\mathbf{x}_i^{\mathrm{T}}$, we generate an arbitrary number N^J of "views" by means of augmentation, *i.e.* $\{\mathbf{v}_{ij}^{\mathrm{T}} = t_j(\mathbf{x}_i^{\mathrm{T}})\}_{i=1\,j=1}^{M^{\mathrm{T}}\,N^J}$ with $t_j(\cdot) \sim \mathcal{T}$. Features extracted from views $\{g^{\mathrm{T}}(\mathbf{v}_{ij}^{\mathrm{T}})\}_{i=1\,j=1}^{M^{\mathrm{T}}\,N^J}$ instead of from inputs $\{g^{\mathrm{T}}(\mathbf{x}_i^{\mathrm{T}})\}_{i=1}^{M^{\mathrm{T}}}$ are then clustered. The feature extractor $g^{\mathrm{T}}(\cdot)$ aims to optimise for a "swapped" assignment problem between pairs of views $(j,k) \in \{1,...,N^J\}$ of the same input $i \in \{1,...,M^{\mathrm{T}}\}$:

$$\mathcal{L}_{swapped}((\mathbf{v}_{ij}, \mathbf{q}_{ij}), (\mathbf{v}_{ik}, \mathbf{q}_{ik})) = \mathcal{L}_c(\mathbf{v}_{ij}, \mathbf{q}_{ik}) + \mathcal{L}_c(\mathbf{v}_{ik}, \mathbf{q}_{ij}) \tag{3}$$

We minimise $\mathcal{L}_{swapped}$ for all the pairs generated from $\mathcal{D}^{\mathrm{T}}$ to get our pretrained $g^{\mathrm{T}}(\cdot)$.

Finally, the whole model is trained by alternating between clustering features and minimising Eq. (3). To work online, clustering is reformulated as an optimal transport problem (as in [2]) and is applied only on the features in a batch.

3.3 Training on FER

Finally, the target model (*i.e.* $f^{\mathrm{T}}(\cdot)$) is obtained training the pseudo-labelled subset $\tilde{\mathcal{D}}^{\mathrm{T}}$ (as detailed in Sec. 3.1) by using the self-supervised feature extractor $g^{\mathrm{T}}(\cdot)$ (as detailed in Sec. 3.2) and a new classifier $h^{\mathrm{T}}(\cdot)$. The model is trained with supervised cross-entropy loss between $\{\tilde{\mathbf{y}}_i^{\mathrm{T}}\}_{i=1}^{\tilde{M}^{\mathrm{T}}}$ and the prediction $\{f^{\mathrm{T}}(\mathbf{x}_i^{\mathrm{T}}\}_{i=1}^{\tilde{M}^{\mathrm{T}}}$ as in Eq. (4):

$$\mathcal{L}_{CE}^{T} = -\frac{1}{\tilde{M}^{\mathrm{T}}} \sum_{i=1}^{\tilde{M}^{\mathrm{T}}} \tilde{\mathbf{y}}_i^{\mathrm{T}} \log f^{\mathrm{T}}\left(\mathbf{x}_i^{\mathrm{T}}\right) \tag{4}$$

where $f(\cdot)$ already includes a softmax function for normalising the network logits into a probability distribution.

4 Experiments

We compare our method against the state-of-the-art methods for cross-domain FER with a set of benchmark datasets. We first introduce our experimental setup and then present the main comparison, followed by an extensive ablation study to justify our design choices.

Datasets. Following [6], we use AFE [6] and RAF-DB [22] as our source datasets, and ExpW [42] and FER2013 [10] as the target datasets.

- **AFE** [6] contains 54,901 images of thousands of Asian individuals, collected from Asian films. This dataset addresses cross-culture domain adaptation, as the other datasets in our experiment involve mainly European and American people.

Method	AFE → ExpW	AFE → FER2013	RAF-DB → ExpW	RAF-DB → FER2013
ICID [14]	54.85	46.44	68.52	53.00
DFA [43]	62.53	36.88	47.42	47.88
LPL [23]	54.51	49.82	68.35	53.61
DETN [20]	58.41	45.39	43.92	42.01
FTDNN [38]	55.29	48.58	68.08	53.28
ECAN [21]	62.52	46.15	48.73	50.76
CADA [27]	58.50	48.61	63.74	54.71
SAFN [35]	55.17	50.07	68.32	53.31
SWD [18]	56.56	51.84	65.85	53.70
AGRA [6]	*65.03*	51.95	*69.70*	*54.94*
SHOT-IM [24]	53.52	49.51	53.13	49.44
SHOT [24]	54.12	49.44	53.51	49.36
CluP (DeepClusterV2)	62.56	50.47	65.43	**53.83**
CluP (SwAV)	**65.00**	***52.51***	**66.60**	53.71

Table 1: Results of different methods in four domain adaptation settings, where the upper part lists methods for FER in the UDA setting with the source data accessible, while the lower part lists methods for FER in the SFUDA setting without accessing the source data. We highlight in *italic* the best result among all methods and in **bold** the best among SFUDA ones. Note that in "AFE → FER2013", CluP achieves the best result among all methods.

Method	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
SHOT-IM [24]	28.29	**45.05**	9.86	75.97	56.12	40.66	71.96
SHOT [24]	28.18	43.24	10.25	75.59	53.53	40.18	74.37
CluP (DeepClusterV2)	29.44	**45.05**	2.83	**83.15**	**77.70**	34.72	**76.65**
CluP (SwAV)	**37.89**	30.63	**13.57**	80.72	50.85	**44.51**	74.49

Table 2: Class-wise accuracy for RAF-DB → FER2013 in SFUDA setting.

- **RAF-DB** [22] contains 29,672 facial images from thousands of individuals that were collected from the Internet. We use RAF-DB as one of our source domain as it works as a counterpart of AFE.
- **ExpW** [42] contains 91,793 faces downloaded from Google Images, representing a large scale in-the-wild scenario with diverse ethic groups and facial poses.
- **FER2013** [10] is large-scale dataset collected with the Google Images Search API, containing 35,887 grey images of low resolution. We consider FER2013 as a target domain to demonstrate cross-colour domain adaptation.

Performance metric. To evaluate the performance of our method, we consider traditional top-1 classification accuracy. In addition, we also provide class-wise accuracy in our ablations to demonstrate how our method performs on different classes.

Implementation details. We implement our method using PyTorch and PyTorch Lightning, and run all the experiments on NVIDIA A100 GPUs. We pretrain ResNet18 for FER as our source model, while we perform the self-supervised learning on the target domain using the solo-learn library [8] for 1000 epochs with SGD and a cosine annealing scheduling policy. When performing cluster-level pseudo-labelling, we consider a large number of clusters for K-means to address imbalanced datasets. We consider $K = 1000$ for AFE → ExpW and RAF-DB → FER2013, and $K = 250$ for the others. We set the Q-th percentile per class to threshold the cluster purity, where Q is usually set to large values, depending on the adaptation setup. In detail, we use $Q = 0.9$ for AFE → ExpW and AFE → FER2013, while $Q = 0.7$ and $Q = 0.8$ for RAF-DB → ExpW and RAF-DB → FER2013. Our final target model is trained

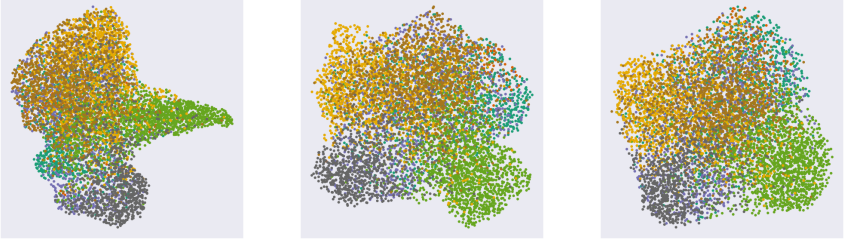
for 50 epochs using SGD, following a cosine annealing scheduling policy.

4.1 Comparisons

To the best of our knowledge, CluP is the first method to tackle SFUDA for FER, therefore we propose a comparison with state-of-the-art methods in the less restrictive UDA setting. To extend the comparison, we also report the results of a couple of general-purpose methods for SFUDA which we re-purposed for FER. CluP can be applied seamlessly to an arbitrary SSL method, to this end we report two versions where we apply different self-supervised pretraining on the target domain using SwAV [5] and DeepClusterV2 [4, 5].

Tab. 1 shows the classification accuracy of competing methods under different domain adaptation settings. Compared among SFUDA methods, our method with SwAV as self-supervised pretraining always performs better than SHOT by over ten points in most of the benchmarks. The same advantage holds when we adapt from the two source domains to FER2013, with a total improvement of +4%. More interestingly, CluP demonstrates comparable adaptation performance even when compared with UDA methods which have access to the source data. In particular, when we adapt from AFE to FER2013, our CluP scores the best performance among all methods.

For an in-depth investigation of how SFUDA methods performing on FER, we present the class-wise accuracy when adapting from RAF-DB to FER2013 in Tab. 2. Noticeably, CluP manages to consistently adapt better among all classes compared to SHOT and SHOT-IM, where for some classes, e.g. Surprise, Happiness, Sadness, the improvement is greater than +5%. We also notice that for minor classes under the SFUDA setting, e.g. Disgust, the classification accuracy is much lower compared to other major classes, mostly due to the limited samples for expressing the class in the target domain under a large domain gap.



(a) SHOT (b) DeepClusterV2 (Ours) (c) SwAV (Ours)

Figure 3: UMAP visualisations of the features spaces in the RAF-DB → FER2013 setting.

Qualitative Result. In Fig. 3, we present the UMAP visualisation of our methods with the target model pretrained with two self-supervised methods, i.e. DeepClusterV2 and SwAV, in the RAF-DB → FER2013 setting and compare them with SHOT. SHOT shows a more peculiar shape compared to our models, as it finetunes the source model to the new domain, thus having a tighter relationship with the source data. The inherited space indirectly constrains the target model to imitate the source domain when moulding the target domain. Therefore, starting from the source model can hinder the adaptability to the new domain in situations of extreme domain shift. On the contrary, the UMAPs of our models seem to fit better to the target domain. While similar, these two embedding spaces present subtle differences. Visually, the DeepClusterV2 space manages to better separate emotions.

Dataset	ImageNet	Source	DeepClusterV2	SwAV
ExpW	54.45	66.80	60.54	**69.13**
FER2013	36.38	56.99	58.10	**60.90**

Table 3: Top-1 accuracy with different pretrained target backbones: a model pretrained on ImageNet, the source model, and two self-supervised models, *i.e.* DeepClusterV2 and SwAV.

Backbone	Score	AFE → ExpW	AFE → FER2013	RAF-DB → ExpW	RAF-DB → FER2013
Source	Conf.	56.43	48.36	59.79	50.47
Source	Purity	56.54	47.34	61.18	**54.29**
SwAV	Conf.	62.88	51.27	63.22	50.68
SwAV	Purity	**65.00**	**52.51**	**66.60**	53.71

Table 4: Top-1 accuracy of various versions of CluP evaluated on four adaptation setups.

4.2 Ablation study

We present a thorough analysis of the main design choices of CluP. We first investigate different pretrained networks to validate the effectiveness of the self-supervised pretraining of the target model. We then compare our novel cluster purity score against the traditional confidence score to justify its advantages in providing more reliable pseudo-labels. Finally, we examine different combinations of our proposed elements and show how they impact the final adaptation performance on FER.

Does the self-supervised pretrained backbone work better? In order to understand how each pretrained backbone model serves as the target model, we experiment four options including (i) a model pretrained on ImageNet, (ii) the source model, *i.e.* "Source", and (iii) the DeepClusterV2 and (iv) the SwAV self-supervised models. For all models, we train a linear classifier applied on top of their frozen feature extractor with ground-truth target labels. As shown in Tab. 3, the self-supervised pretraining on the target domain using SwAV scores the *best* classification accuracy on the two target datasets. DeepClusterV2 demonstrates less consistent improvements over the source model on the two target domain, with $+1.1\%$ improvement on FER2013 dataset, but with -6.3% on ExpW. This might be due to the superiority of SwAV to learn discriminative feature representations over DeepClusterV2.

Does cluster purity perform better than confidence? We ablate our novel cluster purity score in comparison to the traditional confidence to prove its capability of providing more reliable pseudo-labels. We also show the impacts of different threshold on Q ranging from 0.5 and 0.9 on the adaptation performance. Fig. 4 shows the top-1 accuracy of CluP on FER2013, when adapting from AFE (the green plots) and RAF-DB (the orange plots), with varying thresholds on the confidence (the dashed line) and our cluster purity score (the solid line). We can observe a general increasing tendency of the accuracy as the threshold value increases, as more reliable pseudo-labels are selected due to a stricter criterion. Our cluster purity consistently outperforms the confidence at all threshold values. Specifically, when adapting from RAF-DB to FER2013, cluster purity outperforms confidence at the threshold of 80% by more than $+3\%$.

How do all the components interact with one another? We show how different pretrained target models and different pseudo-label criteria incrementally impact the performance of our proposed method. We consider two backbones, Source and SwAV, and two pseudo-label criteria, confidence and cluster purity score. We present the classification accuracy under different adaptation setups in Tab. 4. Regarding the target model, self-supervised

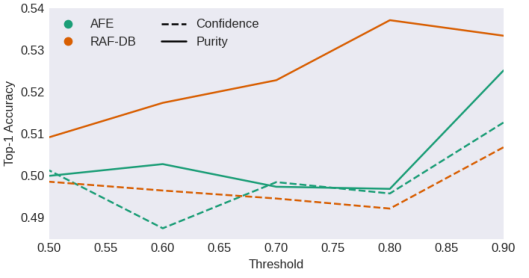


Figure 4: Top-1 accuracy of CluP on FER2013, when adapting from AFE (the green plots) and RAF-DB (the orange plots), with varying thresholds on the confidence (the dashed line) and our cluster purity score (the solid line). Best viewed in colour.

pretraining, *i.e.* SwAV, outperforms the source model under the majority of the adaptation setups, regardless the usage of either confidence or cluster purity score for pseudo-label refinement. When our proposed cluster purity is applied, we observe a consistent improvement of about $+3\%$ over all the adaptation setups with a cluster-based self-supervised feature extractor. When applied on the source model, on the other hand, its advantages are not stable.

5 Conclusions

In this work, we presented the first Source-Free Unsupervised Domain Adaptation solution for Facial Expression Recognition, motivated by the privacy-sensitive nature of facial images. Our method, CluP, employs self-supervised pretraining on the target domain for warming up the target model. To reliably transfer the task knowledge from the source model, CluP proposes a novel cluster-level pseudo-labelling strategy by refining the pseudo-labels using cluster statistics. We experimentally proved the effectiveness of CluP in improving the adaptation performance under various adaptation setups, scoring the new state-of-the-art in terms of FER under the SFUDA setting. As future work, we aim to extend our method towards online SFUDA, where the adaptation happens as the target data streams in.

Acknowledgement

This work was supported by the EU JPI/CH SHIELD project, by the PRIN project PREVUE (Prot. 2017N2RK7K), the EU H2020 MARVEL (957337) project, the EU ISFP PROTECTOR (101034216) project, the EU H2020 SPRING project (871245), and by Fondazione VRT. It was carried out under the "Vision and Learning joint Laboratory" between FBK and UNITN.

References

[1] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv preprint arXiv:2107.03107*, 2021.

[2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

[3] Manuel G Calvo and Lauri Nummenmaa. Perceptual and affective mechanisms in facial expression recognition: An integrative review. *Cognition and Emotion*, 30(6), 2016.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33, 2020.

[6] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, Lingbo Liu, and Liang Lin. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *TPAMI*, 2021.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[8] Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *JMLR*, 23, 2022.

[9] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *WACV*, 2021.

[10] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *NeurIPS*, 2013.

[11] Behzad Hasani, Pooran Singh Negi, and Mohammad Mahoor. Breg-next: Facial affect computing using adaptive residual networks with bounded gradient. *IEEE Transactions on Affective Computing*, 2020.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[13] Yunzhong Hou and Liang Zheng. Visualizing adapted knowledge in domain transfer. In *CVPR*, 2021.

[14] Yanli Ji, Yuhan Hu, Yang Yang, Fumin Shen, and Heng Tao Shen. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. *Neurocomputing*, 333, 2019.

[15] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *CVPR*, 2020.

[16] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *CVPR*, 2020.

[17] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *WACV*, 2021.

[18] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, 2019.

[19] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, 2020.

[20] Shan Li and Weihong Deng. Deep emotion transfer network for cross-database facial expression recognition. In *ICPR*, 2018.

[21] Shan Li and Weihong Deng. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, 2020.

[22] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

[23] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

[24] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *TPAMI*, 2021.

[25] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In *CVPR*, 2022.

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

[27] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 31, 2018.

[28] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9), 2021.

[29] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 2017.

[30] Roberto Pecoraro, Valerio Basile, Viviana Bono, and Sara Gallo. Local multi-head channel self-attention for facial expression recognition. *arXiv preprint arXiv:2111.07224*, 2021.

[31] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *International Symposium on Intelligent Systems and Informatics*, 2021.

[32] Jiawei Shi, Songhao Zhu, and Zhiwei Liang. Learning to amend facial expression representation via de-albino and affinity. *arXiv preprint arXiv:2103.10189*, 2021.

[33] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *TIP*, 29, 2020.

[34] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021.

[35] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, 2019.

[36] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *NeurIPS*, 2021.

[37] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *CVPR*, 2017.

[38] Marcus Vinicius Zavarez, Rodrigo F Berriel, and Thiago Oliveira-Santos. Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *SIBGRAPI*, 2017.

[39] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. *arXiv preprint arXiv:2101.02839*, 2021.

[40] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, 2018.

[41] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *NeurIPS*, 34, 2021.

[42] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5), 2018.

[43] Ronghang Zhu, Gaoli Sang, and Qijun Zhao. Discriminative feature adaptation for cross-domain facial expression recognition. In *International Conference on Biometrics (ICB)*, 2016.