Abstract—Facial Expression Recognition is a commercially important application, but one common limitation is that applications often require making predictions on out-of-sample distributions, where target images may have very different properties from the images that the model was trained on. How well, or badly, do these models do on unseen target domains? In this paper, we provide a systematic evaluation of domain adaptation in facial expression recognition. Using state-of-the-art transfer learning techniques and six commonly-used facial expression datasets (three collected in the lab and three “in-the-wild”), we conduct extensive round-robin experiments to examine the classification accuracies for a state-of-the-art CNN model. We also perform multi-source experiments where we examine a model’s ability to transfer from multiple source datasets, including (i) within-setting (e.g., lab to lab), (ii) cross-setting (e.g., in-the-wild to lab), (iii) mixed-setting (e.g., lab and wild to lab) transfer learning experiments. We find sobering results that the accuracy of transfer learning is not high, and varies idiosyncratically with the target dataset, and to a lesser extent the source dataset. Generally, the best settings for transfer include fine-tuning the weights of a pre-trained model, and we find that training with more datasets, regardless of setting, improves transfer performance. We end with a discussion of the need for more— and regular—systematic investigations into the generalizability of FER models, especially for deployed applications.

Index Terms—Affective Computing; Facial Emotion Recognition; Transfer Learning;

I. INTRODUCTION

Humans are emotionally expressive, and much of the most visible expressions occur on our faces [1]–[3]. In the past decade, recent technological advancements in machine learning has enabled the wide-spread use of facial expression recognition (FER) technology to automatically recognize expressions on people’s faces. Indeed, many of the most mature emotion recognition technologies offered on the market right now are services that claim to accurately classify emotions from static images of people’s faces, such as Affectiva’s AffDex, Amazon Rekognition, and Microsoft Azure Face API.

One limitation of FER is external generalizability. FER systems, especially many state-of-the-art deep learning-based systems [4], are trained using supervised learning, which involves minimizing the classification error on a dataset of faces and emotion labels. Even though it is standard practice to “hold out” a portion of the dataset to evaluate the test accuracy of a classifier, such cross-validation provides a measure of the expected classification accuracy only for new samples drawn from the same distribution as the training data. But in many applications, the data that one wishes to make predictions on may be very different—along both emotion-relevant and emotion-irrelevant dimensions—than the training distribution. Multiple factors such as differences in illumination, pose, and background objects—as well as changes in facial features like ethnicity—can contribute to what are called “dataset shifts” or “feature-space shifts”, which can impact the model’s performance on different datasets [5].

This is broadly known as transfer learning [6]–[8], where the knowledge in a trained model is “transferred”. Specifically in this paper, we focus on the particular case of transfer learning called domain adaptation [8]–[10], where the FER models are performing the same task (emotion recognition), but in a new domain—a target dataset that is different from the source dataset that it was trained on. Although previous papers have studied domain adaptation in FER, they often have examined transfer performance to only a small number (one [11]–[13], two [14]–[16], and three [10] datasets) of target datasets. The performance on the target dataset is used as evidence for the effectiveness of the proposed transfer learning technique. However, datasets differ along many dimensions, and so a proposed technique may work better for some target datasets but not others. Furthermore, the chosen target dataset may have its own idiosyncrasies, and thus, results based on one dataset may not be sufficient to allow generalizable conclusions about transfer learning.

In this work, we take a first step to address this gap by conducting a systematic empirical evaluation using six widely-used FER datasets, three of which are collected in the laboratory, and three are collected in-the-wild, i.e., with more naturalistic variation. We choose a state-of-the-art deep learning-based model (ResNet50 with pre-trained weights from VGGFace2) and state-of-the-art transfer-learning techniques, as our focus is on evaluate existing techniques on a larger, more diverse selection of datasets. First, we investigate the transfer performance of our model when trained on one source dataset and evaluated on a different target dataset—
with six datasets, this produces thirty permutations. We find that overall, single-source generalization performance is quite poor.

Next, we perform an evaluation of multiple-source training, where the model is trained on multiple source datasets, and which is common in fields like NLP [17]. This is to investigate if combining multiple FER datasets improves model generalization. We perform experiments in a variety of setups: 1) within-setting (e.g., train on in-lab data, test on in-lab data), 2) cross-setting (e.g., train on in-lab data, test on in-lab data), and 3) mixed-setting (e.g., trained on both setting, test on in-lab data, which we formulate in a Leave-one-out setup). As prior literature on multiple-source training in FER is quite limited, we also varied the amount of fine-tuning for the model, using four different fine-tuning approaches. We end with a discussion of our results, and outline implications of our work for the state of FER applications.

II. RELATED WORK

A. Transfer learning in FER

Facial expression recognition has been explored using many approaches [18], [19]. In recent years, the most successful approaches use various types of deep learning models [4], with the most popular architecture being Convolutional Neural Networks (CNNs) [20]–[22]. Deep learning models, like other supervised machine learning models, are trained to minimize some (classification) loss on a held-out portion of their training data. These models thus work well in scenarios in which the new data for which predicted labels are desired are drawn from the same feature distribution as the data the models were trained on. Specifically for FER, we focus our review on the problem of domain adaptation, where a model is trained on some source data from some domain (e.g., Dataset A), and is expected to produce labels on some new target data that could be from a different domain (e.g., Dataset B).

One technique in transfer learning is to start with a model that has been trained on a large amount of data, but not specifically on the task at hand. For example, [12] took a ResNet model that was pre-trained on object recognition on ImageNet, and then trained it to perform a new task—emotion classification of seven emotions—using AffectNet. In this case, the source domain is ImageNet (and the source task is object recognition), and the target domain is AffectNet (with target task being emotion classification). However, they only found a 3% improvement in the performance of the model pre-trained on ImageNet, compared to a ResNet model trained from scratch on AffectNet. Another example is [15], who used AlexNet pretrained on ImageNet, and then trained to predict on two target FER datasets, CFEE and RaFD. In addition to ImageNet, other papers have used models pre-trained on other datasets like VGG-Face, a large face recognition dataset, and fine-tuned them for FER [23]. Others have also used combinations: [22] used an ensemble network comprising of pre-trained weights from three models, ResNet-50, VGG-Face and DenseNet-121 to perform facial expression recognition.

Some papers attempt multiple rounds of fine-tuning. [13], which detailed a submission to the 2015 EmotiW emotion recognition challenge, started with an AlexNet model pre-trained on ImageNet data (object recognition). They then implemented progressive transfer learning, by which they sequentially trained their model on two different FER datasets, namely FER2013 and the training set of the EmotiW dataset. Their progressively-trained model came in third-best overall in the challenge, and outperformed the challenge baseline model by about 16% in accuracy.

Another commonly used technique during fine-tuning is "freezing" certain layers in the model, which means that the weights of these layers are not changed during the training process. This technique has not been as well-studied in FER, but is widely-used in other domains like medicine [24], [25], detecting malicious software [26] and speech recognition [27]. The intuition is that "early" layers of the pre-trained model (e.g., even if it was pre-trained on an unrelated task like object recognition) would extract important and more "basic" features, while later layers would transform these representations into more task-relevant features. For example, previous papers have found that freezing the weights of the early layers of the network give rise to better results on some target task, especially for smaller target datasets [28]. One interesting example relevant to FER is [29], who aimed to train personalized emotion recognition models for each subject in the AMIGOS dataset. They trained an AlexNet model on AffectNet, before fine-tuning subject-specific models on the AMIGOS dataset; during fine-tuning, they "froze" the convolutional layers of AlexNet, and modified only the weights of the non-convolutional layers.

Other approaches to domain adaptation in FER include using Generative Adversarial Networks (GANs) [14], training networks specifically to learn correlations between source and target datasets [11], [16] or using feature matching between source and target datasets [20], [31].

Finally, we note several assumptions made by previous papers that study transfer learning in FER. First, they assume that the target dataset is available during training, and hence they perform training on some subset of data from the target dataset. This is a potentially problematic assumption as target training data may not always be available, especially when an application is deployed in a new context. Evaluating a model on a target domain on which it has not previously seen any samples is also a more conservative test of generalizability. In this paper we choose to study the transfer of models not trained on any data from the target datasets.

A second feature common to many previous papers is to only use a small number of target datasets. Many papers use one [11]–[13], [32] or two [14]–[16] target datasets, with only the minority using more than two [10] target datasets. Individual datasets could have idiosyncrasies as a result of the data collection method which could affect how similar they are to other datasets, and consequently affect the transfer performance to these datasets. To avoid this issue, in this paper, we designed a "round-robin" approach where we utilize six
datasets as target datasets across our various experiments.

Aside from several papers that have tried progressive training (e.g., [13]), we are not aware of many FER papers that pool datasets, or otherwise study transfer learning from a multi-source setting. Some recent work in NLP [17] has suggested that by pooling and training on (ten) multiple reading comprehension datasets, one can train more robust and generalizable reading comprehension models. For facial expressions, a similar multi-source approach could allow us to pool information from different FER datasets—potentially training a model to be more robust to feature shifts and allowing more generalizable models. Thus, we designed a series of multi-source training setups to evaluate the efficacy of this approach.

![Sample images from 5 of the 6 datasets](image)

**Fig. 1.** Sample images from 5 of the 6 datasets. GEMEP does not allow publication of images. We note that CK+ does not have neutral images.

### III. Empirical Evaluation

We designed a set of experiments to investigate the generalization of models trained on one subset of datasets, to another subset of datasets (Fig. 2). In the simplest, single-source case, we train a model on a single dataset, and evaluate it on data from a new distribution. Another set of variations we shall explore is training a model using multiple datasets, which has been shown to improve generalization in other fields like NLP [17]. We primarily ground our experiments to test the generalization amongst two commonly-observed settings: in-lab datasets and “in-the-wild” datasets.

#### A. FER Datasets

There are many FER datasets that vary along a wide range of properties, and there are also many papers that survey these datasets [4], [18], [19]. One major distinction made by many researchers (e.g., [33]) between different FER datasets is whether they contained posed or spontaneous facial expressions: the former contains expressions which reflect the actors’ stereotypes of facial expressions and tend to be more exaggerated, while the latter are naturally-occurring and are presumably more representative of daily life. Datasets that contain posed expressions also usually tend to be collected in controlled, lab environments, with less heterogeneity in pose, lighting, background, and camera quality, while datasets with spontaneous expressions tend to be collated from the Internet via search engines or websites like YouTube. Here, our focus is on the generalizability of machine learning models for domain adaptation, which presumably is affected by visual properties like pose and lighting, and so we base our discussion on whether a dataset is collected in the lab or contains naturalistic examples (“in the wild”). We chose to use three in-lab datasets: CK+, IASLab, and GEMEP (which all contain posed expressions) and three in-the-wild datasets: AffectNet, Aff-Wild2, and FER2013 (see Fig. 1).

For all datasets, we use the same subset of 7 classes: \{anger, disgust, fear, happiness, sadness, surprise\} and \{neutral\}.

- **CK+ [34]** is an in-lab dataset with posed expressions. We select a subset of 309 images that are labelled with the six emotions above (CK+ does not have neutral).
- The IASLab face dataset contains 50 unique subjects in the lab, tasked to act out a total of 9 affective states. We use a subset of 1,449 images.
- The GEneva Multimodal Emotion Portrayals (GEMEP) [35] dataset has 10 actors portraying 18 affective states, in a lab controlled setting (GEMEP does not have happy or neutral).
- AffectNet [36] contains facial images collected by querying 3 major search engines using 1,250 emotion related keywords in six different languages. We use a subset of 460,039 images that are labelled with the six emotions and neutral.
- The Aff-Wild2 [37] dataset contains a “Expressions Classification” subset, which consists of 84 YouTube reaction videos labelled with the six emotions and neutral.
- **FER-2013 [38]** contains 35,887 images in the 7 categories, obtained using Google Search from a set of 184 emotion-related keywords like “blissful” and “enragéd”.

For the video datasets, Aff-Wild2 and GEMEP, we extracted single frames from the videos at fixed intervals of 1 frame per second, which allows us to form sizable datasets of static images, with substantial differences between images. For the labelling procedure, each frame of the videos in Aff-Wild2 is labelled with one of the 7 expressions, thus we used the corresponding labels given. On the other hand, each GEMEP video is labelled with one of the 7 expression, thus frames were assigned with the same label. We extracted 7,512 and 2,306 frames from AFF-Wild2 and GEMEP respectively. The distribution of labels are given in Table 1 and we note that some datasets are heavily imbalanced. All the datasets used provide images with frontal views of facial expression, which establishes a baseline for facial detection and feature learning, reducing the need for rigorous data preprocessing. For all datasets, we used a common pipeline to process the image data, using the OpenCV HaarCascade Classifier to identify and crop out facial features, before transforming the images to a 3-channel grayscale. This is done to remove any non-facial features and additional noise to improve the performance of FER.


**Datasets (Total Size)**

| Dataset       | Anger (Total%) | Disgust (Total%) | Fear (Total%) | Happy (Total%) | Sadness (Total%) | Surprise (Total%) | Neutral (Total%) |
|---------------|----------------|------------------|--------------|---------------|-----------------|-------------------|------------------|
| CK+ (309)     | 45 (14.6%)     | 59 (19.1%)       | 25 (8.1%)    | 69 (22.3%)    | 28 (9.1%)       | 83 (26.9%)        | 0 (0%)           |
| IASLab (1,449)| 199 (13.7%)    | 203 (14.0%)      | 212 (14.6%)  | 213 (14.7%)   | 214 (14.8%)     | 209 (14.4%)       | 199 (13.7%)      |
| GEMEP (2,306) | 637 (27.6%)    | 327 (14.2%)      | 451 (19.6%)  | 0 (0%)        | 600 (26.0%)     | 291 (12.6%)       | 0 (0%)           |
| AffectNet (460,039) | 27,964 (6.1%) | 890 (0.2%)      | 3,799 (0.8%) | 246,078 (53.5%) | 20,838 (4.5%) | 17,445 (3.8%) | 143,025 (31.1%) |
| Aff-Wild2 (7,512) | 759 (10.1%)  | 208 (2.8%)       | 398 (5.3%)   | 1,651 (22.0%) | 1,688 (22.5%)  | 995 (13.2%)      | 1,813 (24.1%)    |
| FER-2013 (35,887) | 4,953 (13.8%) | 547 (1.5%)      | 5,121 (14.3%) | 8,989 (25.0%) | 4,002 (11.2%) | 6,198 (17.3%) |

**TABLE I**

Breakdown of emotion labels across datasets. Values indicate number of data points (and percentage of total dataset).

B. **FER Model and Training Procedure**

![Single-Source Setup](image)

![Multi-Source Setup](image)

**Fig. 2.** Top: Single-source experiment, where we train the model with only one dataset and test on other datasets (Table I). Middle: Multi-source training setups (Table III). Bottom: Simplified schematic of ResNet50 architecture (39), along with the layers that are modified during fine-tuning.

We choose to use ResNet50 [39], which is a popular CNN-based model that won the ImageNet challenge in 2015 and is often used as the backbone model for transfer learning. In most computer vision transfer learning paradigms, the ResNet50 model is usually pre-trained on ImageNet, which is a benchmark for object classification (e.g., [12], [13], [15]). As FER consists of recognizing expressions on the same objects (i.e., faces), we reasoned that it might be better to pre-train the model on a facial recognition dataset instead. We chose to pre-train ResNet50 on VGGFace2, which is a facial recognition dataset that contains 3.31 million images of 9131 subjects downloaded from Google Images [40], and in preliminary tests we found better performance with pre-training on VGGFace2 than ImageNet.

For our multi-source experiments, we tested four types of transfer learning: total-freezing, partial fine-tuning, full fine-tuning, as well as a no-pre-training condition. For our single-source experiments, we used full fine-tuning (Fig 2).

For total-freezing, we start with ResNet50 with weights pre-trained on VGGFace2. The model is connected to a (randomly-initialized) output layer (Fig. 2) to predict the emotion labels. We train only the output layer weights on the source datasets to minimize cross-entropy loss, i.e., the weights of the pre-trained model are totally frozen. This training is done for a maximum of 15 epochs, and we use the model that gives the best performance on a validation set (from the source datasets).

For partial fine-tuning, similarly initialize the model with VGGFace2 weights. In addition to training the classification layer, we allow errors to backpropagate and modify the weights of half of the model layers, until the 3rd block (Conv3) in the ResNet50 architecture (i.e., the weights in Conv1 and Conv2 are kept frozen).

For full fine-tuning, we start with VGGFace2 weights, and during training on the source datasets, we allow all the model weights to be modified.

In addition to the above conditions using pre-trained weights, we also ran a no-pre-training condition in the multi-source experiments where we randomly initialized ResNet50 weights and trained the model from scratch. For this setup, we trained the model for up to 20 epochs, and we use the model that performs the best on the validation set.

To run our models, we used a Google Cloud instance equipped with a NVIDIA Tesla T4 GPU. We set the learning rate to .001 and batch size to 64. We trained our models using Stochastic Gradient Descent, with a momentum value of 0.9.

To evaluate our models, we report the top-1 accuracy, which is simply the proportion of correctly predicted samples. As most of the datasets are imbalanced, we also report the weighted F1-score which is given by:

\[
\text{weighted F1} = 2 \sum \frac{N_l \cdot \text{precision}_l \times \text{recall}_l}{N_l \cdot \text{precision}_l + \text{recall}_l} \tag{1}
\]

where \(N_l\) is the number of samples in class \(l\) and \(N_{\text{total}}\) is the total number of samples being evaluated.
C. Single-Source Transfer

In this experiment, we test whether models trained on one dataset generalize well to other datasets that have a different data distribution. We use full fine-tuning: we fine-tuned a pre-trained VGGFace2 on a source dataset and evaluate its performance on the rest of the datasets. The performance of all the combinations are given in Table II with both the top-1 accuracy for a seven-class classification and the weighted F1 score. The diagonal entries give the reference performance when the model is trained on the training partition and evaluated on the test partition of the same dataset with a 60%/20%/20% train/validation/test partitioning. The off-diagonal entries give the performance of a model trained and cross-validated (60%/20%) on a given source dataset (rows) evaluated on a target dataset (columns), where we use the entirety of the target dataset as a held-out evaluation set. Thus, for a given row in Table II the model is trained in the same manner, with the difference being what data is used as evaluation: the test-set of the source (diagonal entries) or the entire target dataset (off-diagonals).

We note that for five of the six datasets studied, the non-transfer performance is the highest (that is, the maximum performance in each column is the diagonal entry), which is not surprising as the evaluation data comes from the same distribution as the training data. The only dataset that strangely does not follow this pattern is CK+, whereby a model trained and tested on CK+ (37.1%) underperforms one trained on AffectNet and tested on CK+ (49.8%). We believe that this could be attributed to the relatively smaller size of CK+ which has 309 samples, of which 60 make up the test set; this is much smaller than the other datasets.

For single-source transfer learning performance, we see that performance is generally quite poor, especially compared to the non-transferred model performance. The model trained on AffectNet does the best overall, achieving the best transfer performance on four of the five other datasets (CK+, IASLab, Aff-Wild2, and FER-2013). But even still, it achieved top-1 accuracies of 45.1% on IASLab (compared to 88.2% non-transferred), and 57.6% on FER-2013 (compared to 70.0% non-transferred). We see also that there are dataset-idiomatic effects. For example, all the models did poorly on Aff-Wild2 as the target, with the maximum performance not higher than 24% (compared to 84.3% non-transferred). Another example is GEMEP—when GEMEP was the target, model performance was very variable and ranged from 0% (source: Aff-Wild2) to 98% (source: GEMEP). This highlights the dangers of studying transfer learning using only a single target dataset [11]–[13].

D. Multi-Source Transfer

Next, we sought to test if we can improve transfer performance by training models on multiple source datasets, which may increase the model’s ability to learn different data distributions. We used the following training setups:

- **Within-setting**: We train and test the model using the datasets that fall in the same setting. For example, we train on two in-lab datasets and evaluate performance on the third in-lab dataset. Because of our setup of three in-lab and three in-the-wild datasets, for each target dataset, the model will be trained on two source datasets.

- **Cross-setting**: To test the efficacy of domain adaptation across in-lab/in-the-wild settings, we trained the model on all three datasets of one setting, and evaluate it on a target dataset of a different setting. Thus, for each target dataset, the model would have been trained on three source datasets (all from the other setting).

- **Leave-one-out**: Finally, to test the efficacy of training on the maximum amount of data, and to investigate transfer when trained on both in-lab and in-the-wild settings, we perform a leave-one-out approach where we train on five source datasets and evaluate the model on the last dataset.

Because it is not clear how different types of fine-tuning would work on a multi-source FER setup, we decided to try with different common procedures. As mentioned in Sec. III-B we perform four fine-tuning approaches: initializing with VGGFace2 weights and either total-freezing, partial fine-tuning, full fine-tuning, as well as a no-pre-training condition where we train the ResNet50 model from scratch. Note that these four approaches are arranged in decreasing order of how much “prior knowledge” (from VGGFace2) is encoded into the final transferred model.

1We note too that GEMEP as a source dataset results in very low classification performance on the remaining target datasets, suggesting that transfer is difficult both to and from GEMEP.
First, if we compare the best single-source performances (bolded values in Table III) with the best multi-source performance (bolded values in Table III), we see that there are overall improvements of, on average 5.9% (range: 0.15% to 13.49%) in top-1 accuracy for 5 of the 6 datasets, except for GEMEP where there is a reduction of 4% from 27.62% to 23.68%. This result shows that training on multiple datasets increase transfer performance, especially on the smaller datasets (7.8% on CK+ and 13.49% on Aff-Wild2).

Next, we examine the training setups. We note up-front that GEMEP seems to be an outlier. Overall, except for GEMEP, the leave-one-out training setup seems to result in the best performance on all the datasets. The leave-one-out training setup also uses the most training data, in that each model is fine-tuned on five source datasets and evaluated on the sixth. Even though this setup has the most variance in the training data, the model does seem to learn task-relevant features and learn to generalize across task-irrelevant properties like visual differences in the datasets. If we consider the other two training setups (again, except for GEMEP), we see that generally the cross-setting setup works second-best when evaluated on In-Lab datasets, and the within-setting setup works second-best when evaluated on In-the-wild datasets. This suggests that in our experiments, the in-lab vs. in-the-wild distinction does not seem to be as important as the raw amount of training data. Thus, it seems like the general takeaway is: the more data on which to fine-tune, the better the transfer-learning performance.

The exception to this trend is GEMEP, for which the within-setting training setup yields the best performance (23.7%), which is, as mentioned earlier, less than the best single-source training performance (27.6%), which in turn is significantly less than training and testing on GEMEP itself (98.3%).

2Cross-setting training when evaluated on in-lab datasets mean that the model is fine-tuned on the three in-the-wild datasets. Similarly, Within-setting training evaluated on in-the-wild datasets mean that the model is fine-tuned on the other two in-the-wild datasets.

do not know why this is so, but speculate that it could be because GEMEP has no neutral or happy examples, which make up the bulk of examples in some of the other datasets. We also cannot rule out that this could be due to idiosyncratic differences in dataset collection.

Finally we examine the effect of different fine-tuning setups. Overall, we notice a trend that the performance seems to increase when going from total-freezing to partial fine-tuning to full fine-tuning. That is, the more we allow the model’s weights to vary during fine-tuning on the source dataset, the better they perform on an unseen target dataset. Yet, allowing the weights to vary from a pre-trained initialization on VGGFace2 still outperforms a no-pre-trained model which was initialized to random weights; this suggests that pre-training still improves performance, although we should allow the pre-trained weights to be fine-tuned to the specific task. There are, of course, individual exceptions to these general trends, such as Aff-Wild2, for which no-pre-trained is the best-performing setup.

IV. Discussion

The proliferation of commercial facial expression recognition technology has raised concerns both about the scientific basis of inferring emotions from expressions [41], as well as ethical concerns surrounding the potential misuses of such technology [42, 43]. The lack of systematic investigations—like this study—on the out-of-domain generalizability of deep learning models—a prerequisite for scientific validity—should be another cause for worry. Researchers all know that transfer learning performance on out-of-distribution examples will suffer, and by definition all deployed technology operates in this regime, where new data comes from a different distribution as the data they were trained on. But, until this study, we do not know just how much performance really suffers. Importantly, such technologies are already being deployed in a variety of applications, and if we cannot guarantee that our AI models can accurately detect facial expressions even on a mix of well-
controlled datasets, how much confidence can we have that they will work when deployed? Furthermore, the incentives in academic publishing are even set up against these types of investigations, because such well-controlled systematic research using existing models and existing datasets are “not novel”. As a field, we need to “slow down” to systematically assess the generalizability of our technologies—and we need to pressure commercial offerings to similarly demonstrate such generalizability—as it is a necessary pre-requisite for ethical deployment [43].

In this study, we conduct a systematic investigation of domain adaptation using six datasets and a comprehensive suite of experiments. The results are sobering. Generally, we conclude that single-source, single-target generalization performance for FER is poor. Our best-performing single-source models were those trained on AffectNet, which happened to be the largest dataset in our study. But even the best single-source models underperformed their non-transferred counterparts by margins between 12.5% (FER-2013: 70.0% to 57.5%) to 74.7% (GEMEP: 98.3% to 23.6%). When we examined multi-source, single-target performance, we again saw the same trend where the best performances came from models trained with the most data (our leave-one-out approach with five source datasets and one target dataset).

Our results also highlight the dangers of relying on a single target dataset for making general claims about transfer learning. Some of the datasets in our study looked quite different from other datasets, with very low transfer learning performance and/or very variable performance depending on the source dataset or the training method used. We specifically chose to do a “round-robin” format and to look for more general trends, rather than focusing on specific datasets. We believe that, moving forward, FER research should bear in mind the need to evaluate their models on a range of different datasets, in order to validate their generalizability.

Our study, though systematic, has several limitations. First, although we examined six datasets—the largest such examination in domain adaptation for FER, to the best of our knowledge—there were many other datasets that we were not able to include in this study (e.g., as they had different emotion categories). Second, FER datasets tend to have imbalanced emotion label distributions. Of the six datasets we examined, IASLab was the most well balanced (with each of seven classes making up between 13.7-14.8% of the dataset), while others like AffectNet were severely imbalanced (53.5% Happy and 31.1% Neutral, with the remaining 15.4% divided among 5 classes). These dataset imbalances are known to impact the robustness of the models trained on them, and this is a major concern especially if these (or similarly imbalanced) datasets provide the training data for commercially-deployed models. Third, as our focus was on datasets, we restricted our experiments to a single state-of-the-art model architecture, ResNet50 pretrained on VGGFace2, and four fine-tuning approaches, but without extensive optimization. Future work could examine other state-of-the-art FER models to assess their generalizability. It could be the case that adding more datasets, having more balanced datasets, or trying different model architectures / training protocols could change several of the quantitative results we found in this study. However, we believe that addressing these limitations will lead to differences in degree, but not differences of kind, and hence our generic conclusions should generalize.

V. CONCLUSION

In this work we presented an empirical study of domain adaptation using six facial expression recognition datasets using single-source and multi-source training setups. Our analysis shows that single-source, single-target generalization is poor, and we can improve the performance by training models on multiple source datasets. But even in our best-performing settings (train on five datasets and test on one dataset, and if we ignore the outliers of GEMEP and CK+), we find that mean transferred seven-class classification performance (of ∼55%) is still significantly lower than the mean non-transferred performance (∼84%, with a mean performance differential of 28%), and so we urge a great deal of caution when making or evaluating assertions about the generalization of current FER models.

We also want to stress that moving forward, there should be more—and regular—investigations like ours. Out-of-domain generalizability is a pre-requisite for developing scientifically-valid models. There needs to be more academic research into transfer learning and developing benchmark tests—and more acceptance among the academic community for such rigorous testing. And there needs to be incentives for industry developers to show that their commercial offerings are indeed generalizable.

ACKNOWLEDGMENT

We acknowledge the use of different FER datasets, including the IASLab[3]. We also acknowledge Jing Lin and Deng Jingyuan for their help on an earlier version of this project.

REFERENCES

[1] H. A. Elfenbein and N. Ambady, “On the universality and cultural specificity of emotion recognition: a meta-analysis.” Psychological Bulletin, vol. 128, no. 2, p. 203, 2002.
[2] J. A. Russell, J.-A. Bacherowski, and J.-M. Fernández-Dols, “Facial and vocal expressions of emotion.” Annual Review of Psychology, vol. 54, no. 1, pp. 329–349, 2003.
[3] P. Ekman, “Facial expression and emotion.” American Psychologist, vol. 48, no. 4, p. 384, 1993.
[4] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” IEEE Transactions on Affective Computing, 2020.
[5] J. Zhang, W. Li, P. Ogunbona, and D. Xu, “Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective,” ACM Computing Surveys (CSUR), vol. 52, no. 1, pp. 1–38, 2019.
[6] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[3] Development of the Interdisciplinary Affective Science Laboratory (IASLab) Face Set was supported by the National Institutes of Health Director’s Pioneer Award (DP1OD003312) to Lisa Feldman Barrett. More information is available online at www.affective-science.org.
[7] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, “A comprehensive survey on transfer learning,” Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2020.

[8] G. Csafka, “Domain adaptation for visual applications: A comprehensive survey,” arXiv preprint arXiv:1702.05374, 2017.

[9] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” Neurocomputing, vol. 312, pp. 135–153, 2018.

[10] R. Zhu, T. Zhang, Q. Zhao, and Z. Wu, “A transfer learning approach to cross-database facial expression recognition,” in 2015 International Conference on Biometrics (ICB), 2015, pp. 293–298.

[11] N. Kalishech, P. Thiam, P. Bellmann, and F. Schwenker, “Deep domain adaptation for facial expression analysis,” in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 317–323.

[12] T. Q. Ngo and S. Yoon, “Facial expression recognition on static images,” in International Conference on Future Data and Security Engineering. Springer, 2019, pp. 640–647.

[13] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, “Deep learning for emotion recognition on small datasets using transfer learning,” in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 443–449.

[14] B. Bozorgtabar, D. Mahapatra, and J.-P. Thiran, “Expandra: Adversarial domain adaptation for facial expression analysis,” Pattern Recognition, vol. 100, p. 107111, 2020.

[15] V. Mavari, S. Raman, and K. P. Miyapuram, “Facial expression recognition using visual saliency and deep learning,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2783–2788.

[16] K. Yan, W. Zheng, T. Zhang, Y. Zong, C. Tang, C. Lu, and Z. Cui, “Cross-domain facial expression recognition based on transductive deep transfer learning,” IEEE Access, vol. 7, pp. 108906–108915, 2019.

[17] A. Talmor and J. Berant, “MultiQA: An empirical investigation of generalization and transfer in reading comprehension,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019, pp. 4911–4921.

[18] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 6, pp. 1113–1133, 2015.

[19] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, “Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 8, pp. 1548–1568, 2016.

[20] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, “Adaptive deep metric learning for identity-aware facial expression recognition,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 522–531.

[21] H. Kaya, F. Gürpınar, and A. A. Salah, “Video-based emotion recognition in the wild using deep transfer learning and score fusion,” Image and Vision Computing, vol. 65, pp. 156–166, 2017.

[22] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using cnn-rnn and c3d hybrid networks,” in Proc. of the 18th ACM Int. Conf. on Multimodal Interaction., 2016, pp. 445–450.

[23] H. Ding, S. K. Zhou, and R. Chellappa, “Facenet2explain: Regularizing a deep face recognition net for expression recognition,” in 2017 12th IEEE Int. Conf. on Autom. Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 118–126.

[24] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttman, F.-E. de Leeuw, C. M. Temppany, B. Van Ginneken et al., “Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2017, pp. 516–524.

[25] R. K. Samala, H.-P. Chan, L. Hadjijiski, M. A. Helvie, C. D. Richter, and K. H. Cha, “Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets,” IEEE Transactions on Medical Imaging, vol. 38, no. 3, pp. 686–696, 2018.

[26] E. Rezende, G. Ruppert, T. Carvalho, F. Ramos, and P. De Geus, “Malicious software classification using transfer learning of resnet-50 deep neural network,” in 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017, pp. 1011–1014.

[27] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, “Transfer learning for speech recognition on a budget,” in Proceedings of the 2nd Workshop on Representation Learning for NLP, 2017, pp. 168–177.

[28] D. Soekhoo, P. Van Der Putten, and A. Plaat, “On the impact of data set size in transfer learning using deep neural networks,” in International Symposium on Intelligent Data Analysis. Springer, 2016, pp. 50–60.

[29] M. Rescigno, M. Spezialetti, and S. Rossi, “Personalized models for facial emotion recognition through transfer learning,” Multimedia Tools and Applications, vol. 79, no. 47, pp. 35 811–35 828, 2020.

[30] R. Zhu, G. Sang, and Q. Zhao, “Discriminative feature adaptation for cross-domain facial expression recognition,” in 2016 Int. Conf. on Biometrics (ICB). IEEE, 2016, pp. 1–7.

[31] H. Yan, “Transfer subspace learning for cross-dataset facial expression recognition,” Neurocomputing, vol. 208, pp. 165–173, 2016.

[32] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2017.

[33] K. R. Scherer and T. Bänziger, “On the use of actor portrayals in research on emotional expression,” in Blueprint for Affective Computing: A Sourcebook, pp. 166–176, 2010.

[34] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2010, pp. 94–101.

[35] T. Bänziger, M. Mortillaro, and K. R. Scherer, “Introducing the Geneva multimodal expression corpus for experimental research on emotion perception,” Emotion, vol. 12, no. 5, p. 1161, 2012.

[36] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2017.

[37] D. Kollia and S. Zafeiriou, “Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface,” arXiv preprint arXiv:1910.04855, 2019.

[38] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cui Kirstey, Y. Tang, D. Thaler, D.-H. Lee et al., “Challenges in representation learning: A report on three machine learning contests,” in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[40] B. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), 2018, pp. 67–74.

[41] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, “Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements,” Psychological Science in the Public Interest, vol. 20, no. 1, pp. 1–68, 2019.

[42] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. N. Sánchez, D. Raji, J. L. Rankin, R. Richardson, J. Schultz, and M. Whittaker, “AI Now 2019 Report,” New York, NY: AI Now Institute, 2019.

[43] D. C. Ong, “An ethical framework for guiding the development of affectively-aware artificial intelligence,” in 9th International Conference on Affective Computing and Intelligent Interaction (ACII), 2021.