TempT: Temporal consistency for Test-time adaptation

Onur Cezmi Mutlu, Mohammadmahdi Honarmand, Saimourya Surabhi, Dennis P. Wall
Stanford University
{cezmi, mhonar, mourya, dpwall}@stanford.edu

Abstract

*We introduce **Temp**oral consistency for **T**est-time adaptation (TempT), a novel method for test-time adaptation on videos through the use of temporal coherence of predictions across sequential frames as a self-supervision signal. TempT is an approach with broad potential applications in computer vision tasks, including facial expression recognition (FER) in videos. We evaluate TempT's performance on the AffWild2 dataset. Our approach focuses solely on the unimodal visual aspect of the data and utilizes a popular 2D CNN backbone, in contrast to larger sequential or attention-based models used in other approaches. Our preliminary experimental results demonstrate that TempT has competitive performance compared to the previous years' reported performances, and its efficacy provides a compelling proof-of-concept for its use in various real-world applications.*

1. Introduction

Affective computing aims to develop technologies with the capabilities like recognizing, interpreting, and simulating human affects. Expressions being one of the primary means of conveying emotion, facial expression recognition (FER) often constitutes an important part of human affective behavior analysis. There is an increasing number of use cases from driver safety applications to diagnosis and therapy of developmental problems of children [11]. With the continuous improvement in the computer vision field through extensive adoption of deep learning approaches, the real-world use of such algorithms is becoming easier and universal. However, the robustness and reliability of aforementioned algorithms tend to suffer from the domain shift phenomena which is still a prominent problem for computer vision models with limited generalization capability.

The domain shift problem becomes even more pronounced in the ”real world” scenarios due to uncontrollable environmental conditions. In the computer vision setting some examples to these conditions could be lighting, camera quality, motion, and resolution. Invariance and robustness against these variations is the main focus of domain adaptation and domain generalization research, with many successful algorithms already developed. In our work, we explore a specific subdomain of this field called Test-Time Adaptation (TTA), also referred to as Unsupervised Source-Free Domain Adaptation. In this setting, we assume no access to the target domain during training-time and no access to target domain labels in test-time. We treat each video as a new domain and our method adapts the trained model to a given video during test-time to improve its performance.

We investigate the performance of our approach on the Facial Expression Recognition (FER) task, where the goal is to classify each frame in a video for Ekman emotions. The task of video assessment at the frame level is a natural environment for machine learning models with spatiotemporal inductive biases since the ability to model inter-frame relations could potentially be useful. Examples of such models are 3D convolutional neural networks (CNN) [10], attention-based models [1], or hybrid approaches combining 2D CNNs with recurrent neural networks (RNN) [37]. The first two of these approaches usually suffer from greater computational requirements than 2D CNNs, whereas the last method has unstable training time behavior under inputs with longer duration. There are numerous solutions to these problems including more efficient architectures as well as well-studied training paradigms, but in our work, we focus on exploring an adaptive approach where a simple 2D CNN model, which lacks useful biases for the setting, uses temporal predictive consistency as a self-supervision signal to adapt at test-time. For benchmarking purposes, we use Affwild2 [14–22, 40] which is an invaluable FER dataset that contains over 500 videos and covers a wide variety of aforementioned variations. These qualities make it a suitable candidate for testing our algorithm.

2. Related Work

Facial Expression Recognition (FER) is a challenging task, especially in real-world scenarios. The difficulty arises from the fact that there is a significant amount of variation within each expression category, making it difficult to distinguish between different expressions. Additionally, there

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Other	**Total**
Affwild2	44676	32962	7851	9730	2622	3296	5540	31412	**138089**
Affectnet	55670	118605	19650	11647	5670	3626	19325	0	**234193**
RAF-DB	3096	5771	2390	1571	347	865	846	0	**14886**
Total	**103442**	**157338**	**29891**	**22948**	**8639**	**7787**	**25711**	**31412**	**387168**

Table 1. Label distribution of pretraining datasets

can be similarities between different expression categories, which further complicates the task of FER.

This challenge is even more pronounced in real-world settings, where the lighting conditions, poses, and identities of the individuals can vary significantly. In such scenarios, even individuals with the same identity, pose, and lighting conditions can exhibit different expressions, while individuals with different identities, ages, gender, and pose can express the same emotion.

Thus, FER is a task that requires robust algorithms that can effectively handle these intraclass variances and interclass similarities. In the past few years many Convolution Neural networks (CNN) based [4, 23, 27] and transformer-based architectures [39] have been proposed and significantly improve the performance of FER.

As far as we are aware, there has been no prior research on test-time adaptation (TTA) for facial expression recognition (FER). Our work is an attempt to explore the use of TTA on FER tasks. It represents a novel approach to FER that has the potential to improve accuracy and opens up new avenues for research into TTA.

Test-time Adaptation Early attempts for unsupervised domain adaptation were mainly based on updating running statistics of the batch normalization layers [26, 33] with the new information from test data. [34] was one of the early works to propose using an auxiliary self-supervised task to be used in the test-time with the purpose of adapting the backbone parameters. [35] proposed using entropy minimization as the main adaptation goal and limiting the set of parameters to be updated to the weights of batch normalization layers (as opposed to updating statistics as before) which are shown to be highly expressive in [5]. Originating from the close ties of domain adaptation with few-shot learning [42] introduces a meta-learning-based solution where the loss to be used for adaptation is meta-learned. Finally, [41] and [28] report impressive adaptation results by combining image augmentation and entropy minimization to overcome the shortcomings of the latter in scenarios with large domain shifts.

All of these works operate on static data that does not necessarily bear temporal correlations. Among them, only [35] explores continual adaptation to online data streams. [36] is a novel work that proposes a continual adaptation algorithm based on augmentation consistency. Yet, their algorithm makes an assumption of i.i.d. samples during test time, which may not always be correct. [6] addresses this issue and coins a new normalization layer that handles selective adaptation under non-iid data streams. To our knowledge, none of the works in the field exploits the temporal correlations in a given stream, and in our work, we aim to explore a possible direction for that.

3. Our Approach

3.1. Datasets and Preprocessing

Focusing on training a computer vision model that operates on images (rather than videos), we have numerous data sources that are popular in the FER literature. We combine Affwild2 with Affectnet [30] and Real-world Affective Faces Database (RAF-DB) [24, 25] to create a larger and more diverse training dataset. In our task, target classes are 7 basic emotions (also known as Ekman emotions [3]) plus an "other" class for expressions that do not fit into any category.

Affwild2 is significantly larger in comparison to the others and has a label imbalance as can be seen in Fig. 2 and Fig. 3. In order to overcome this, we perform a random sampling on it by limiting the number of frames to 300 per video per expression class basis. Detailed label distribution of the resultant dataset is given in Tab. 1.

We use provided cropped and aligned images in Affwild2, and others are only available in cropped versions, so we do not require any additional spatial preprocessing for any of the datasets. We then resize images to 112px × 112px with antialiasing. For training purposes, we use common image augmentation methods such as color jitter, brightness and contrast shift, histogram equalization, channel dropout, blur, and random horizontal flip.

3.2. Modeling

Our approach is based on individual predictions on video frames, which allows us to use popular image-processing architectures in the literature. Due to their proven performance and stability of training, we use models from Resnet [7] family, with variations such as aggregated residual transformations [38] and squeeze-and-excitation blocks [9]. Generated embeddings are processed by two fully-connected layers where the second, i.e. output, layer is

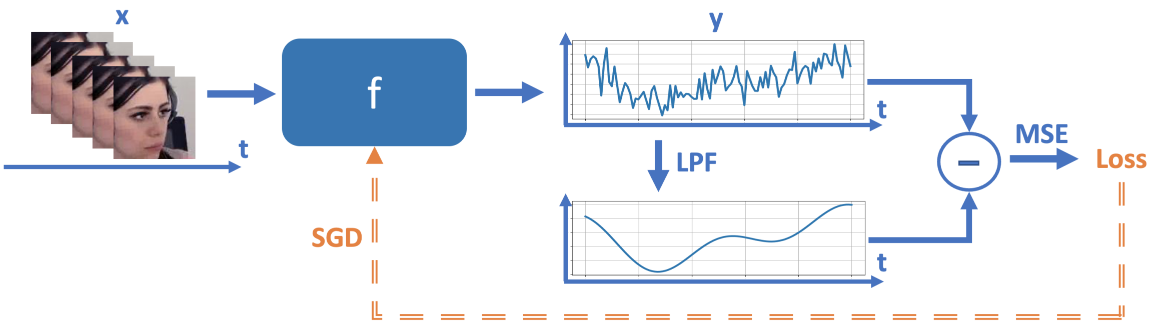

Figure 1. TempT algorithm

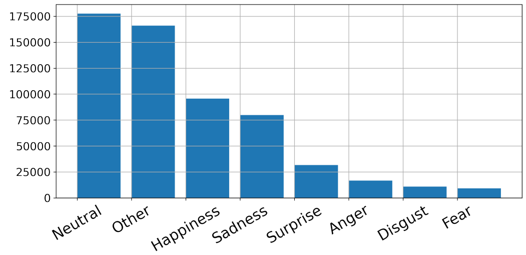

Figure 2. Label distribution of Affwild2 train set

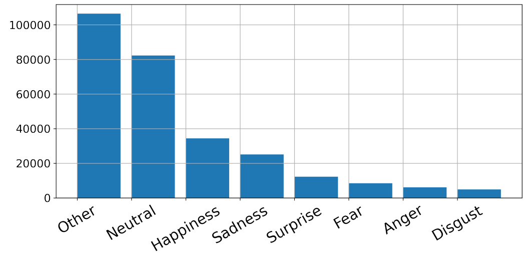

Figure 3. Label distribution of Affwild2 validation set

subject to weight and input normalization [32] to prevent overconfidence, improve smoothness, and generalization.

Significant class imbalance is a problem in this setting that needs to be addressed for successful supervised training. Label weighting, class up-sampling, and class down-sampling are classic methods to alleviate this issue, yet there are numerous scenarios where they fail to do so. We, therefore, adopt another approach namely Label-Distribution-Aware Margin Loss (LDAM) that was introduced in [2]. LDAM is similar to sample weighting in the sense that it modifies the loss depending on the class frequency but instead of using a multiplicative scaling, it intercepts with the class margins. The exact formulation is given in Eq. (1) where z is the unnormalized prediction vector,y is the ground truth class label, n_j is the number of samples in class j and C is a temperature-like hyperparameter that tunes the effect of margins. LDAM enforces larger margins on minority classes which in return increases the model robustness and prevents overfitting. For more details, we refer the reader to the original paper.

$$\begin{aligned} \mathcal{L}(z,y) &= -\log \frac{e^{z_y-\Delta_y}}{e^{z_y-\Delta_y}+\sum_{j\neq y} e^{z_j}} \\ \text{where } \Delta_j &= \frac{C}{n_j^{1/4}} \text{ for } j \in \{1,\dots,k\} \end{aligned} \tag{1}$$

Supervised training of the model is then performed with back-propagation algorithm using defined LDAM loss to account for the skewed label distribution. Adam [13] optimizer with weight decay [29] is used for optimization where learning rates were subject to a step-decay schedule. Modeling and training were performed using PyTorch [31] framework on NVIDIA V100 GPUs.

3.3. TempT: Temporal consistency for Test-time adaptation

Being trained on static images as opposed to videos, 2D CNN models do not carry the implicit bias for the smoothness and/or consistency in their predictions across frames. We found empirically that such models contain stronger high-frequency components at the output, and when they are subject to a low-pass filter, the results look more desirable. We propose using this fact to generate a supervision signal to tune the network and improve classification performance. In particular, we temporally smooth the model predictions using a low-pass filter and set it as the desired signal. Purpose of setting this filtered signal as the target is to enforce the model to make temporally consistent predictions. We then calculate the mean-squared error between the original and target signals and use back-propagation to update a subset of model parameters.

More formally, let $x^{(t)} \in \mathbb{R}^{112\times112\times3}$ be the t^{th} frame of video and $f(.) : \mathbb{R}^{112\times112\times3} \to \mathbb{R}^8$ be the trained neural network of interest. We hypothesize that predictive coherence between consecutive samples can be used as an im-

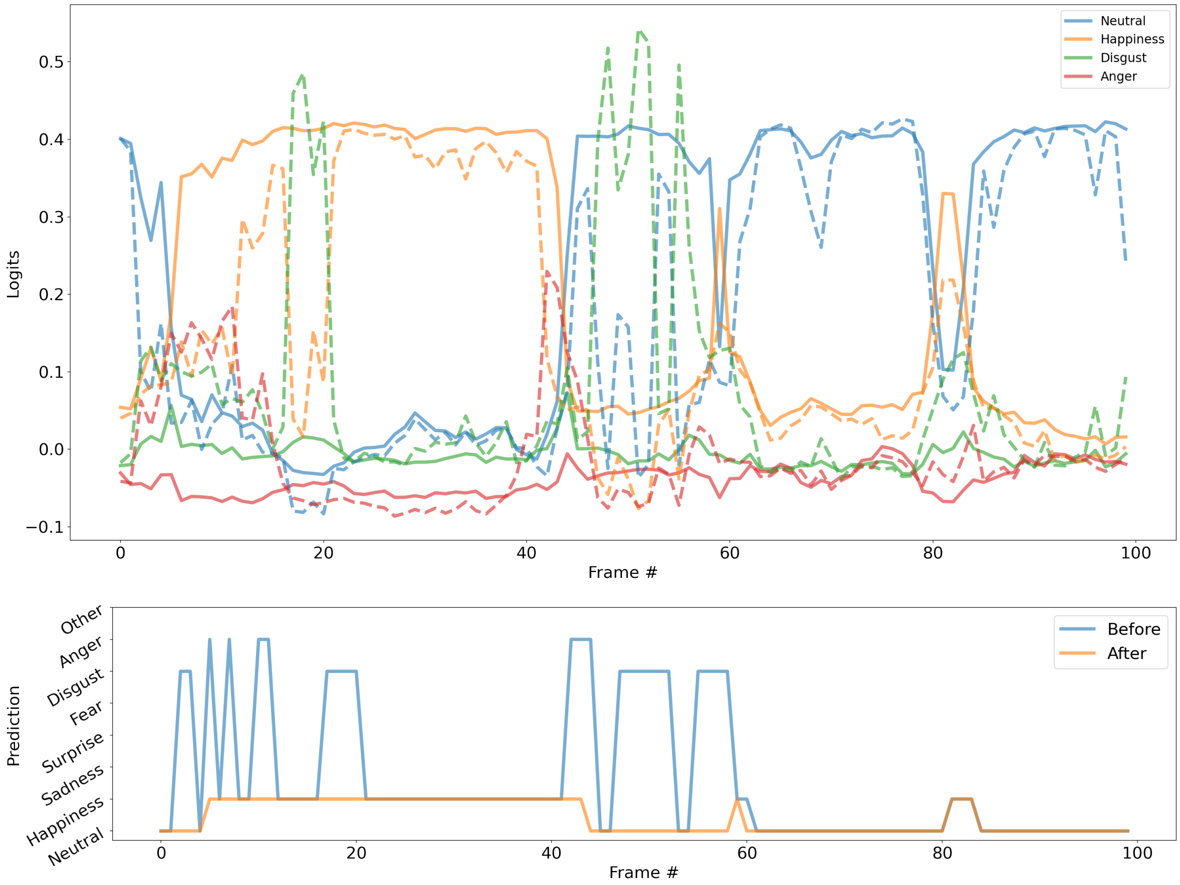

Figure 4. (top) Model outputs before (dashed) and after (solid) adaptation. (bottom) Model predictions before and after adaptation

plicit Jacobian regularizer. In [8], it has been shown that regularization on the Frobenius norm of input-output Jacobian of a neural network can help the network attain flatter minima with higher robustness against input variations. Now, consider the case when the frame rate of a video is high enough. We can then approximate the Jacobian as in Eq. (2).

$$J_{i,j}(x^{(t)}) = \frac{\partial f_i(x^{(t)})}{\partial x_j^{(t)}} \approx \frac{f_i(x^{(t)}) - f_i(x^{(t-1)})}{x_j^{(t)} - x_j^{(t-1)}} \tag{2}$$

Then minimizing the Frobenius norm of the Jacobian becomes equivalent to minimizing inter-frame prediction differences as in Eq. (3)

$$\begin{aligned} \min \|J(x^{(t)})\|_F &\equiv \min \sum_{i,j} J_{i,j}^2(x^{(t)}) \\ &\equiv \min \|f_i(x^{(t)}) - f_i(x^{(t-1)})\| \end{aligned} \tag{3}$$

We empirically found that the initial distribution of prediction differences is heavy-tailed, with the tail being caused by momentary jumps in predictions due to problems at input cropping and/or sharp changes in activations due to model imperfections. When we used the target in Eq. (3) these outliers made the training process unstable for a significant portion of the experiments. We, therefore, chose to use another equivalent formulation to minimize the target. We first pass all frames from the pipeline to obtain an initial set of unnormalized scores $y^{(t)} \in \mathbb{R}^8$. We then use the error signal in Eq. (4) as a self-supervision loss function to fine-tune the model. $LPF(.)$ can be any low pass filter; in our experiments, we use a median filter, due to its robustness to outliers.

$$\mathcal{L}(y) = \sum_t \|y^{(t)} - LPF(y)^{(t)}\| \tag{4}$$

Using the entire video for adaptation may not be computationally feasible when the video duration is long. To alleviate this, we count the number of changes in model predictions using a sliding window and select the regions with the most changes to be the training regions that will compose the training batch. The updated version of the loss signal can be examined in Eq. (5) where $\mathcal{R}$ is the set of selected regions, and r indicated the range of frames to be considered.

$$\mathcal{L}(y) = \sum_{r \in \mathcal{R}} \sum_{t \in r} \|y^{(t)} - LPF(y)^{(t)}\| \tag{5}$$

Being differentiable, this loss allows the use of backpropagation to update model parameters. The choice of parameters has an important effect on the performance of the adapted model since the selection defines the expressivity of the model and therefore the power of adaptive interventions. Following the analysis in [12], we select this subset to be the weight and bias terms in batch normalization layers while freezing the running statistics. This has been shown to yield enough expressivity while preventing overfitting. We then use AdamW optimizer with learning rate set to 0.0001, for the adaptation process and take 10 gradient steps, a number that has proven empirically optimal in our hyperparameter searches.

4. Experiments

We test TempT on the AffWild2 dataset and compare the results against the baseline model as well as another test time domain adaptation method, namely TENT [35]. We performed an extensive hyperparameter search on the adaptation parameters of TempT, such as the number of steps, learning rate, optimizer, etc., and report the performance of the best configuration in Tab. 2. Static models' performances are deterministic whereas for adaptation cases we report an average F1 score over 20 experiments to account for stochasticity arising from a random sampling of adaptation frames. We clearly see the positive effect of TempT on classification performance. One important observation of these results is the ability of adaptation to help a less complex model reach the performance of a much larger one. In this experiment, SE-ResNext-101 has 8 times the number of parameters of Resnet-18. Another observation of the results is the performance disruption that TENT introduces. It consistently hurt the performance of the baseline model and we argue that this is due to the highly correlated inputs that we have during test time, which is predicted in [6].

To further observe the changes that TempT induces, we also investigate time series generated by the model before and after adaptation. In Fig. 4 we provide such an example taken from 100 frame portion of a validation set video. On the top figure we provide unnormalized model outputs before and after adaptation, whereas bottom figure shows 'argmax' predictions. To create a cleaner top plot we omitted the classes that do not become the dominant prediction during this interval. From this visualization we can see, in a qualitative manner, that adaptation reduces the flickering behavior at the output and provides more coherent predictions over time, while increasing the F1 score for this particular video from 0.39 to 0.47. To have a quantitative understanding of this effect, we computed average number of decision changes before and after the adaptation on entire AffWild validation dataset. TempT reduces normalized number of changes (i.e. number of changes per frame) for a given video from 0.15 to 0.043.

	Supervised	TENT [35]	TempT
Resnet-18	0.307	0.277	0.323
SE-ResNext-101	0.325	0.269	0.345

Table 2. Average F1 Score performances on validation set

5. Conclusion and Future Work

In our work, we explored a novel model-agnostic algorithm that can have real-life applications for similar tasks and showed that this adaptive method could enhance model performance without any additional means of supervision. On the other hand, performance variance due to stochasticity in the frame sampling process is a problem that needs to be addressed to obtain a more deterministic understanding of the limits and behavior of the algorithm. With such increased stability and the performance boost it brings, TempT can potentially enable more reliable use of models on edge devices while protecting user privacy.

6. Acknowledgements

We would like to thank all members of the Wall Lab for providing valuable feedback. The work was supported in part by funds to DPW from the National Institutes of Health (1R01LM013364-01, 1R01LM013083), the National Science Foundation (Award 2014232), Lucile Packard Foundation (Auxiliaries Endowment) the ISDB Transform Fund, and program grants from Stanford's Human-Centered Artificial Intelligence Program, and from the Wu Tsai Neurosciences Institute's Neuroscience:Translate Program.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 3

[3] Paul Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999. 2

[4] Yingruo Fan, Victor O.K. Li, and Jacqueline C.K. Lam. Facial expression recognition with deeply-supervised attention network. *IEEE Transactions on Affective Computing*, 13(2):1057–1071, 2022. 2

[5] Jonathan Frankle, David J Schwab, and Ari S Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152*, 2020. 2

[6] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022. 2, 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[8] Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019. 4

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[10] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 1

[11] Haik Kalantarian, Khaled Jedoui, Peter Washington, Qandeel Tariq, Kaiti Dunlap, Jessey Schwartz, and Dennis P Wall. Labeling images with facial emotion and the potential for pediatric healthcare. *Artificial intelligence in medicine*, 98:77–86, 2019. 1

[12] Fahdi Kanavati and Masayuki Tsuneki. Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning. In *Medical Imaging with Deep Learning*, pages 338–353. PMLR, 2021. 5

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[14] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022. 1

[15] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800, 2021. 1

[16] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 1

[17] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[18] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 1

[19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 1

[20] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[21] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[22] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[23] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 2

[24] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 2

[25] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. 2

[26] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 2

[27] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2019. 2

[28] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 2

[29] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2017. 3

[30] 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):18–31, 2017. 2

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 3

[32] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016. 3

[33] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving

robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020. 2

[34] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 2

[35] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 2, 5

[36] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 2

[37] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 461–470, 2015. 1

[38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[39] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 2

[40] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1

[41] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642, 2022. 2

[42] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021. 2