2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)

Latent-Based Adversarial Neural Networks for Facial Affect Estimations

Decky Aspandi[1,2], Adria Mallol-Ragolta[2], Björn Schuller[2,3], and Xavier Binefa[1]

[1] Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain

[2] Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany

[3] GLAM – Group on Language, Audio & Music, Imperial College London, UK

Abstract— **There is a growing interest in affective computing research nowadays given its crucial role in bridging humans with computers. This progress has recently been accelerated due to the emergence of bigger dataset. One recent advance in this field is the use of adversarial learning to improve model learning through augmented samples. However, the use of latent features, which is feasible through adversarial learning, is not largely explored, yet. This technique may also improve the performance of affective models, as analogously demonstrated in related fields, such as computer vision. To expand this analysis, in this work, we explore the use of latent features through our proposed adversarial-based networks for valence and arousal recognition in the wild. Specifically, our models operate by aggregating several modalities to our discriminator, which is further conditioned to the extracted latent features by the generator. Our experiments on the recently released SEWA dataset suggest the progressive improvements of our results. Finally, we show our competitive results on the Affective Behavior Analysis in-the-Wild (ABAW) challenge dataset.**

I. INTRODUCTION

Affective computing has recently attracted the attention of the research community, due to its applications in multiple and diverse areas, including education [7] or healthcare [25], among others. Furthermore, the growing availability of affect-related datasets, such as SEWA [23] and the recently introduced Aff-Wild2 [21], enable the rapid development of deep learning-based techniques, which currently hold the state of the art [22], [23], [18].

In computer vision tasks, such as natural image generation[31] and image classification[29], adversarial learning techniques from the family of generative models have been extensively investigated [31], [29], [5]. This learning technique enables rapid progress, not only to create additional data, but also to improve the performance of predictive models. Nevertheless, in the context of affective computing-related applications, this technique is still young and confined to its usage for data augmentation purposes [13].

To expand the investigation of generative models in the field of affective computing, we investigate the use of latent features that are extracted in adversarial manners to improve the predictive capabilities of our model estimations. Specifically, we extract the visual latent features of the generator, which are then used to condition the discriminator on its estimations. Furthermore, we also aggregate the audio modality during training. We later show in our experiments on the SEWA [23] and Aff-Wild2 [21] datasets the benefits of our proposed approach with our competitive results. Specifically, the contributions of this work are:

1) We are the first to introduce the utilisation of latent features arranged in an adversarial way to improve affect-related model estimates.
2) We show the progressive improvements on our proposed works on the SEWA and Aff-Wild2 datasets and achieve competitive results on both datasets.

II. RELATED WORKS

Early approaches on automatic affect estimations involved the use of classical machine learning techniques with some degree of success. Several techniques explored include linear and partial least square regression [30], and support vector machines [28]. Furthermore, given the number of available modalities (e. g. video, audio, and bio-signals), several fusion techniques were also introduced to improve affect-related estimates. Different examples of these methods include early, late, model, and output-associative fusion [37]. Diverse affective information has been progressed, starting from Action Units detection, emotion detection, to more recently continuous valence and arousal estimation[22], [23].

Current progress relates to the emergence of big data that creates the opportunity to introduce large scale datasets in many fields, including affective computing. Examples of these datasets are SEMAINE [26], AVEC [32], AFEW [22], RECOLA [33], SEWA [23], and the recently introduced Aff-Wild2 dataset [21], [20], [36], [16]. These datasets enable the development of powerful deep learning models that improve the accuracy of current state of the art [22], [19], [18]. The investigations of deep learning-based techniques onto affective computing include the introduction of Convolutional Neural Networks (CNN) [4], incorporation of Recurrent Neural Network (RNN) [23], and recently the fusion with Tensor-based methods [27].

Adversarial learning [31] as a generative approach has been intensively studied in other machine learning research, especially in computer vision [5]. Given its potential, this method has also been explored in the field of affective computing, usually to augment the training data available for training [13]. However, there is another aspect of generative models that is largely unexplored in this field, which is the use of latent features to improve the models estimations, as shown in previous works from other fields, such as computer vision [35], machine learning [11], and bio-signal analysis [6]. This inspired us to investigate the use of adversarial learning to improve our proposed models' performance through extracted latent features.

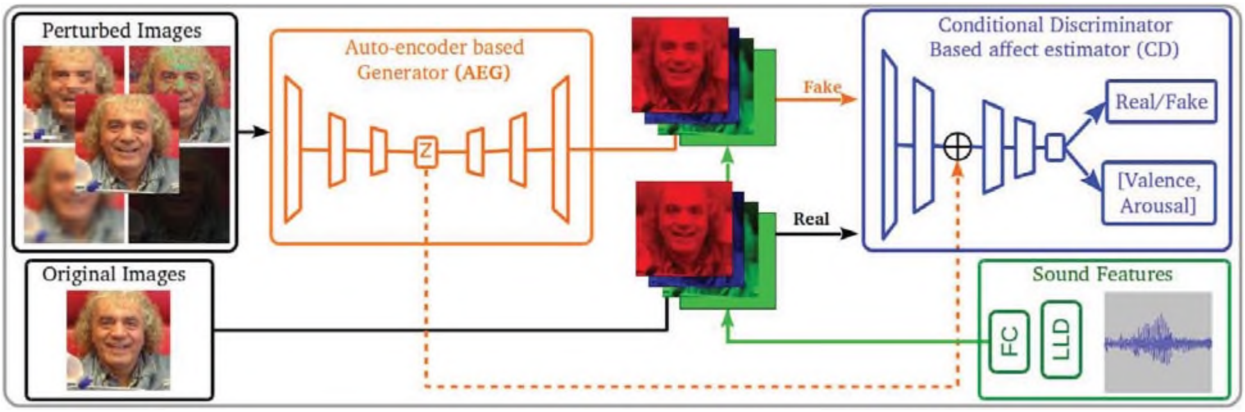

Fig. 1. Complete architecture of our proposed models which incorporate two main networks: first is an Auto-Encoder-based Generator (AEG) which denoises the image and creates robust latent features. Second is a Conditional Discriminator-based affect estimator (CD) that aggregates both sounds and image input which is conditioned by latent features from the CD to estimate both real/fake and valence/arousal values.

III. Latent-Based Adversarial Networks

We build our model based on the Star-GAN network[5], with architectural modifications to allow the extraction of latent features and use of the audio features. Figure 1 shows the overview of our proposed network. Our model operates by aggregating two main modalities: facial and audio features. There are two main sub-networks involved in our overall networks as already outlined above: the Auto-Encoder-based Generator(AEG), and the Conditional Discriminator-based affect estimator (CD) [24]. The main role of the AEG is to produce cleaned images from noisy images to fool the discriminator, while simultaneously extracting robust latent features. On the other hand, the CD tries to recognise the fake images created by the AEG, and, at the same time, estimates the actual valence and arousal values. We train the AEG and CD in an adversarial way as below:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_x\left[\log CD_{adv}(x)\right] \ + \\ &\mathbb{E}_x[\log\left(1 - CD_{adv}(AEG(\hat{x}))\right)],\end{aligned} \tag{1}$$

where x corresponds to the noisy image and $\hat{x}$ is the cleaned input image approximated by the AEG. We use similar noise introduction methods as in [3], which consist of four different types of artifacts: Gaussian blurring, Gaussian noise, image downsampling, and colour scaling.

A. Auto-encoder-based generator

Given the noisy input image, x, the AEG will approximate the cleaned version of the input image, $\hat{x}$. This is done by utilising coupled mirrored convolutions and deconvolutions with intermediate 2D bottleneck latent kernels; i. e. without skip connections. This scheme enforces the AEG to create latent robust features in order to effectively clean the input image. To improve the denoising and reconstruction process, we use the cycle loss [5], [15] defined below :

$$\mathcal{L}_{rec} = \mathbb{E}_x[||x - AEG(AEG(\hat{x}))||]. \tag{2}$$

B. Conditional discriminator-based affect estimator

The CD employs both facial and audio features to identify the real/fake status of the current input and the corresponding valence and arousal values of $\hat{\theta}$. The facial features correspond to the cleaned image (denoised or reconstructed from the generator and the corresponding latent features of z. From the audio modality, we use the low-level descriptors (LLDs) of the eGeMAPS feature set[8] (cf. Section III-D). Both latent and audio features are combined through late fusion [12], [34], [37] alongside the main RGB input images. Specifically, the audio features are merged by feeding them into a 1D fully connected layer to enlarge its dimension and converting it to a single 2D kernel, which is then concatenated with the denoised image. The latent features are combined in middle pipelines of the CD by concatenating them with intermediate kernels.

To detect both real and fake status and estimate valence and arousal values, we add another classifier[29] on top of the main classifier, which consists of a 2x2 pixels layer[14]. In the adversarial training, the CD will be optimised using real (r) and fake (f) images to minimise the affect loss $(\mathcal{L}_{afc})$ that judges the accuracy of the estimated valence and arousal values (cf. Equation 5). The corresponding loss of training the CD for both real $(\mathcal{L}^r_{va})$ and false examples $(\mathcal{L}^f_{va})$ can be seen below :

$$\mathcal{L}^r_{va} = \mathbb{E}_{x,\theta}[-\mathcal{L}_{afc}(\theta'|x)], \tag{3}$$

$$\mathcal{L}^f_{va} = \mathbb{E}_{x,\theta}[-\mathcal{L}_{afc}(D(\theta|G(x)))], \tag{4}$$

where $\hat{\theta}$ is the ground truth valence/arousal value, and the affect loss, $\mathcal{L}_{afc}$, corresponds to the amalgamations of multiple affect metrics: Mean Square Error(MSE) (Eq. 6), Correlation(COR) (Eq. 7), and Concordance Correlation Coefficients (CCC) (Eq. 8), [22], [21] :

$$\mathcal{L}_{afc} = \sum_{i=1}^{N} \frac{n_i}{N}(\mathcal{L}_{MSE} + \mathcal{L}_{COR} + \mathcal{L}_{CCC}) \tag{5}$$

$$\mathcal{L}_{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i, \theta_i)}, \tag{6}$$

$$\mathcal{L}_{COR} = \frac{\mathbb{E}[(\hat{\theta} - \mu_\theta) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}\sigma_\theta} \tag{7}$$

$$\mathcal{L}_{CCC} = 2x\frac{\mathbb{E}[(\hat{\theta} - \mu_\theta) - (\theta - \mu_\theta)]}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2}, \tag{8}$$

where n_i is the total number of instances of discrete valence/arousal class i, and N is the normalisation factor[1] for the total valence/arousal class. This normalisation factor is crucial given considerably unbalanced class instance on the Aff-Wild2 dataset[17].

C. Overall objective

Finally, the overall objective functions to train both AEG and CD are expressed as follows:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{afc}\mathcal{L}_{afc}^r, \tag{9}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{afc}\mathcal{L}_{afc}^f + \lambda_{rec}\mathcal{L}_{rec}, \tag{10}$$

in which λ_{afc} and λ_{rec} are the regulariser parameters for affect estimations and reconstruction loss.

D. Audio feature extraction

One of the first challenges when combining audio and video signals is the difference in term of sampling rates between both modalities. To overcome this issue, we first generate audio frames from the original audio signal by selecting the portions of the audio signal corresponding to one frame of video. We then enlarge the audio frame with the samples corresponding to the previous and future video frame to ensure information overlap between consecutive audio frames. We finally extract the LLDs of the EGEMAPS [8] feature set using OPENSMILE [9], and concatenate the first two sets of LLDs for further analysis. These LLDs are extracted from windows of 0.060 seconds with a step size of 0.010 seconds. Selecting the first two sets of LLDs only, we ensure the same dimensionality of the audio features in spite of videos recorded at different sampling rates.

E. Model training

To train our model, we use the respective training subset of each dataset. On the Sentiment Analysis in the Wild dataset (SEWA)[23], we followed original person-independence protocols, and apply the feature extraction techniques described on previous sections. Moreover, we also use the external tracker of [2] to refine the given bounding box. We also includes the experiments on Aff-Wild2 dataset as part of the Affective Behavior Analysis in-the wild (ABAW) 2020 Competition to provide more actual analysis of our models performance. In this dataset, we only utilise the training subset to obtain our validation results. We then use the full available data (training and validation) to train our final models to produce our test results. Specifically, we used the crop-aligned samples provided by the organisers as facial features, in addition to the audio signals from the available videos.

For both datasets, we trained our model progressively to allow us to analyse the impact of each proposed step. We first train both of generator and discriminator together, and proceed by adding the extracted latent z features alongside the audio features. Our models were trained using an NVIDIA Titan X GPU and it took approximately two days to converge. The source code of our models is available at our github page[1]

IV. EXPERIMENTS

A. Datasets and Experiment Settings

In this section, we describe our results on SEWA[23] and recently introduced Aff-Wild2 [17] datasets to confirm the advantages of each of our proposed approaches.

- The SEWA dataset[23] is a recently published affect dataset which consists of video and audio recording involving 398 subjects from multiple cultures. It is split into 538 sequences with various meta-data (e.g. subject id, culture etc) are available alongside the actual affect ground truth of valence/arousal and liking/disliking.
- The Aff-Wild2 challenge dataset is being published as part of the first ABAW 2020 competitions[17] which consists of three main challenges : valence-arousal, basic expression and eight action units. Aff-wild2 is considered to be the current, largest affect in the wild dataset with more than 558 videos and 458 total number of subjects. Specific on the valence and arousal challenge, there are 545 annotated videos with 2.786.201 frames which is split into three subsets : 346 videos of training, 68 videos of validation and 131 videos of test.

In each experiment, we provide the results from the variant of our models to highlight the important of each approach. First is the method Disc which corresponds to our results utilizing only plain Discriminator (CD) trained using standard ℓ^2 loss. Second is method AEG-CD that constitute to our model which uses adversarial training for both of AEG and CD. Lastly, the AEG-CD-ZS shows the results of our previous model trained with the inclusion of both latent features z from AEG and the mapped audio features.

We use MSE, COR and CCC metrics to evaluate the quality of each affect estimations[23], [32], [27], [17]. That on the Aff-Wild2 dataset, we compared our results on the validation stage against the baseline provided by the organizers [17]. While for the SEWA dataset, we report our results from original five cross validation settings [23] and compared them with the respective baseline[23] and recent state of the art of [27]

B. Experiment Results

Table I provides our results on the SEWA dataset, where we can see that our models able to produce quite competitive results, with a quite high accuracy on the arousal dimension. Specifically, we observe a relatively high accuracy obtained by our discriminator (Disc) that is enough to outperform the current baseline, albeit still lower than the results of [27].

[1] https://github.com/deckyal/ALN

TABLE I

EXPERIMENT RESULTS ON SEWA DATASET

Methods	MSE		COR		CCC	
	Val	Aro	Val	Aro	Val	Aro
Baseline [23]	-	-	0.322	0.4	0.195	0.427
Tensor [27]	0.334	0.380	**0.503**	0.439	**0.469**	0.392
Disc	0.336	0.399	0.395	0.457	0.349	0.379
AEG-CD	0.329	0.394	0.429	0.467	0.380	0.429
AEG-CD-SZ	**0.323**	**0.350**	0.442	**0.478**	0.405	**0.430**

TABLE II

EXPERIMENT RESULTS ON AFF-WILD2, ABAW CHALLENGE DATASET. THE VALUES IN PARENTHESES DENOTE THE TESTING RESULTS

Methods	MSE		COR		CCC	
	Val	Aro	Val	Aro	Val	Aro
Baseline [17]	-	-	-	-	**0.14** (0.11)	0.24 (**0.27**)
Disc	0.44	0.30	0.07	0.19	0.07	0.20
AEG-CD	0.42	0.28	0.10	0.22	0.08	0.22
AEG-CD-SZ	0.42	0.28	0.11	0.29	0.10 (**0.17**)	**0.26** (0.16)

Using the adversarial training further improve the accuracy which conforms the previous findings of the benefit in using the adversarial learning upon standard l2 loss[5], [29], [10]. Another potential explanation of this improvement can be attributed to the generated images that may reduce the available noises on the input images (cf. Section IV-C). Finally, incorporating both of the latent and audio features improves the overall accuracy of our results, surpassing the current state of the art on this dataset on the arousal dimension, highlighting the benefit of incorporating such features.

We also found similar findings in our results on the Aff-Wild2 dataset as shown in Table II. In this dataset, we observe identical improvements toward our results on the validation stage, with the lowest accuracy produced by our Disc model and progressively increased to the best accuracy of AEG-CD-SZ, attaining superior accuracy on arousal domain compared to the baseline. In the test set however, we found that AEG-CD-SZ produces a quite balanced accuracy for both valence and arousal, with higher accuracy against the baseline on the valence domain. This may be a result from the incorporation of the validation split in our training, that further altered the distribution of valence and arousal instances.

C. Latent Feature and Visual Analysis

In this section, we further visualize the learned latent kernel features to explain the observed progressive improvement of our models. Figure 2 shows the examples of original input images and their denoised (cleaned) versions, followed by randomly selected six latent kernels. In regards to the denoising quality, we found that our model manages to clean the underlying noise on the input image considerably well. Furthermore, we notice the consistency of the learned latent features, i.e their spatial structures are not drastically altered, regardless the input conditions. These robust representations help the Discriminator in its inference as complementary features [6] resulting in overall higher accuracy, as shown on the previous section.

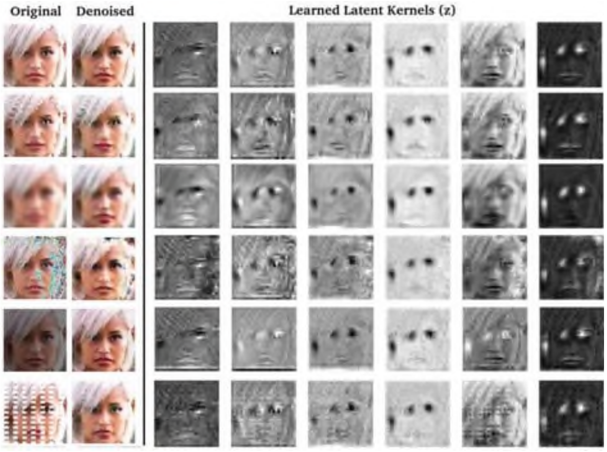

Fig. 2. Example of denoised images and the corresponding latent kernels (selected randomly). As we can see, the denoised images are quite cleaned, and the latent kernels are also consistent across different input conditions.

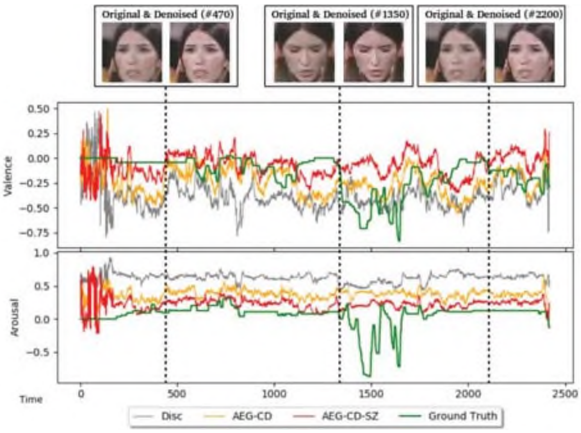

Fig. 3. Visualization of the results from of our model variants. Notice that the results from AEG-CD-ZS, both in Valence and Arousal domain are the most closely resemble to the ground truth compared to the others. Furthermore, the denoised image appear to be clearer than the original input

Furthermore in Figure 3, we can also see the continous results of our model variants. By observing this, we could concur that AEG-CD-SZ produces the most accurate predictions compared to the rest indicated by their resemblances to the ground truth. Also notice that, the denoised input images are also clearer and sharper compared to the original input, which demonstrates the real-life applicability of denoising functions of our model. Finally, these cleaned inputs, may also further aids the discriminator training in conjunction with learned latent features, hence improved its overall results.

V. CONCLUSION

In this paper, we presented the first investigation using latent features extracted through adversarial learning in Affective Computing domain. Specifically, we performed progressive training on our generator to extract robust features given noisy inputs paired with a discriminator through adversarial learning. Then, we employed a conditional discriminator to aggregate several modality's inputs to achieve our affect estimations. We tested the performance of our models on two datasets: SEWA and Aff-wild2. In our experiments, we observed progressive improvements made by each of our approaches ultimately leading to competitive results on both datasets. In the future, we seek to incorporate temporal modelling to further increase the accuracy of the proposed models.

VI. ACKNOWLEDGMENTS

This work is partly supported by the Spanish Ministry of Economy and Competitiveness under project grant TIN2017-90124-P, the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), and the donation bahi2018-19 to the CMTech at UPF. Further funding has been received from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

REFERENCES

[1] D. Aspandi, O. Martinez, and X. Binefa. Heatmap-guided balanced deep convolution networks for family classification in the wild. In *2019 14th IEEE FG 2019*, pages 1–5, May 2019.

[2] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa. Fully end-to-end composite recurrent convolution network for deformable facial tracking in the wild. In *2019 14th IEEE FG*, pages 1–8, May 2019.

[3] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa. Robust facial alignment with internal denoising auto-encoder. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 143–150, May 2019.

[4] P. Cardinal, N. Dehak, A. L. Koerich, J. Alam, and P. Boucher. Ets system for av+ec 2015 challenge. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, page 17–23, New York, NY, USA, 2015. ACM.

[5] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE CVPR*, pages 8789–8797, 2018.

[6] J. Comas, D. Aspandi, and X. Binefa. End-to-end facial and physiological model for affective computing and applications. In *15th IEEE FG*, In Press 2020.

[7] S. Duo and L. Song. An e-learning system based on affective computing. *Physics Procedia*, 24, 01 2010.

[8] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, April 2016.

[9] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor. In *Proceedings of the* 18^{th} *ACM International Conference on Multimedia*, pages 1459–1462, Firenze, Italy, 2010. ACM.

[10] Q. Gan, Q. Guo, Z. Zhang, and K. Cho. First step toward model-free, anonymous object tracking with recurrent neural networks. *arXiv preprint arXiv:1511.06425*, 2015.

[11] U. Garciarena, R. Santana, and A. Mendiburu. Expanding variational autoencoders for learning and exploiting latent representations in search distributions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '18, page 849–856, New York, NY, USA, 2018. Association for Computing Machinery.

[12] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE SYS MAN CYBERN*, volume 4, pages 3437–3443. IEEE, 2005.

[13] J. Han, Z. Zhang, and B. Schuller. Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. *IEEE Computational Intelligence Magazine*, 14(2):68–81, 2019.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE CVPR*, pages 1125–1134, 2017.

[15] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *34th ICML 70*, pages 1857–1865. JMLR. org, 2017.

[16] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou. Recognition of affect in the wild using deep neural networks. In *IEEE CVPRW, 2017*, pages 1972–1979. IEEE, 2017.

[17] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. *arXiv preprint arXiv:2001.11409*, 2020.

[18] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. 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.

[19] D. Kollias and S. Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[20] D. Kollias and S. Zafeiriou. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv preprint arXiv:1811.07771*, 2018.

[21] D. Kollias and S. Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[22] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[23] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *arXiv preprint arXiv:1901.02839*, 2019.

[24] A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5534–5544. Curran Associates, Inc., 2017.

[25] C. Liu, K. Conn, N. Sarkar, and W. Stone. Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE T Robot*, 24:883 – 896, 09 2008.

[26] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *2010 IEEE Int Con Multi*, pages 1079–1084. IEEE, 2010.

[27] A. Mitenkova, J. Kossaifi, Y. Panagakis, and M. Pantic. Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE FG 2019*, pages 1–7. IEEE, 2019.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE T Affect Comput*, 2(2):92–105, 2011.

[29] A. Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[30] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel. Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 75–82, New York, NY, USA, 2016. Association for Computing Machinery.

[31] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[32] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019.

[33] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE FG*, pages 1–8, April 2013.

[34] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[35] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *ECCV*, September 2018.

[36] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *IEEE CVPRW, 2017*, pages 1980–1987. IEEE, 2017.

[37] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T Pattern Anal*, 31(1):39–58, 2009.