Jack of many Faces: A Step Towards Facial Expression and Physiological State Analysis with a Single Network

Abdullah Tariq[1]
abdulla6@our.ecu.edu.au
Martin Masek[1]
m.masek@ecu.edu.au
R. Muhammad Atif Azad[2]
atif.azad@bcu.ac.uk
Zulqarnain Gilani[1,3]
s.gilani@ecu.edu.au

[1] Centre of AI & ML (CAIML), School of Science, Edith Cowan University WA, Australia
[2] Department of Computer Science, Birmingham City University UK
[3] Dept of Computer Sciences, The University of Western Australia Australia

Abstract

Facial feature analysis, particularly dynamic facial expression recognition, is essential in computer vision for understanding human emotions, behaviors, and physiological states. However, existing approaches often exhibit limited performance, stemming from inadequate modelling of facial dynamics, noise sensitivity, ambiguous expression semantics, and are generally specific to single-task scenarios. To address these issues, we propose a compact 3D spatio-temporal network capable of handling both expression recognition and physiological state analysis. Our network includes two custom modules: (1) Contrastive Adversarial Efficient Local Channel Attention (ConAdv-ELCA), which extracts and disentangles fine-grained local facial features, and (2) Efficient Global Channel Attention (EGCA), to capture local-global interactions. Unlike prior work, which predominantly evaluates models on similar datasets within single-task domains, our work has demonstrated the ability to generalize across different tasks that are based on facial analysis. Experimental results demonstrate that our model consistently achieves state-of-the-art or near-state-of-the-art performance on blood alcohol concentration estimation, dynamic facial expression recognition, and driver fatigue detection.

1 Introduction

Human face encodes a wealth of information, providing critical cues related to identity, emotion, and cognitive states. Facial feature analysis has wide-ranging applications, including driver safety monitoring [54, 62, 65], mental health assessment [3, 24, 44], and human-computer interaction [18, 25, 35]. Although facial feature analysis has been extensively explored by different researchers in recent years, most existing approaches are designed for single tasks and/or rely on task specific architectures. Researchers often develop separate models for physiological state analysis [2, 21, 50, 63] and emotion detection [20], each tailored to a specific dataset, which limits their scalability and reusability across domains.

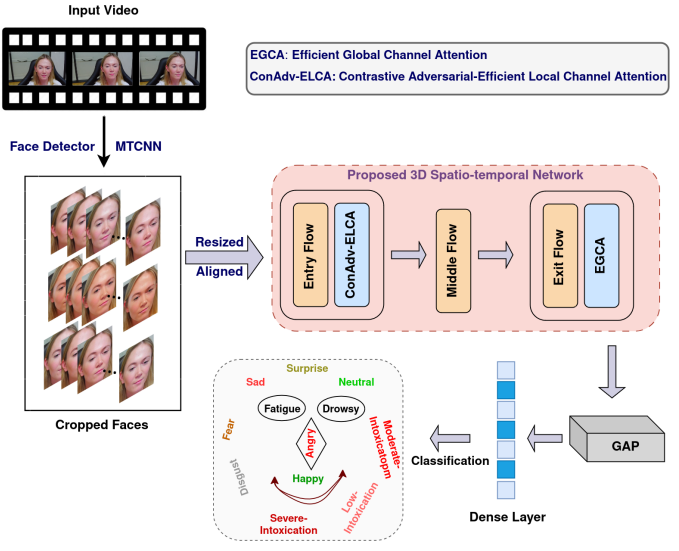

Figure 1: Pipeline of our unified 3D spatio-temporal network for facial analysis, integrating custom attention modules for expression, drowsiness, and intoxication recognition.

Developing a single facial representation model that can generalize across diverse facial analysis tasks remains an open and underexplored challenge.

Recent advancements in physiological state analysis and emotion recognition [6, 12, 43, 47] predominantly rely on multimodal fusion techniques, combining modalities such as audio, textual data, or wearable sensors to achieve improved performance. However, these methods introduce significant architectural complexity and domain specific dependencies, limiting their flexibility across diverse tasks. For instance, audio-based emotion recognition often suffers performance degradation in noisy or uncontrolled environments, and sensor-driven intoxication detection methods are frequently constrained by domain-specific sensor configurations, hindering generalizability. This highlights the need for a unified, adaptable model that can generalize across diverse physiological and emotional state recognition tasks without major architectural changes or auxiliary inputs.

In this work, we propose a unified spatio temporal framework as shown in Figure 1 designed for multiple facial understanding tasks with minimal task-specific modifications. Our key insight is that both local and global spatio temporal facial patterns are shared across a wide spectrum of affective and cognitive states. By learning from these shared cues, our model develops a task agnostic facial representation that transfers effectively across disparate tasks, including emotion recognition, driver fatigue detection, and alcohol induced impairment. To achieve this, we design a novel dual-attention mechanism, (1) Contrastive Adversarial Efficient Local Channel Attention (ConAdv-ELCA) to extract discriminative local patterns and (2) an Efficient Global Channel Attention (EGCA) module to model long range dependencies efficiently. Unlike conventional task specific methods, our proposed spatiotemporal framework exclusively leverages visual face data to generate a shared facial representation applicable across emotional, cognitive, and physiological tasks. We rigorously validate our framework through extensive experiments on five benchmark datasets covering emotion recognition [10, 19, 58], driver drowsiness detection [45], and alcohol-induced im-

pairment estimation [21]. Furthermore, for cross task and cross dataset generalization, we finetune only on the DFEW [19] dataset and use t-SNE visualization to demonstrate that our model effectively clusters samples from unseen tasks, such as drowsiness and intoxication detection, into meaningful semantic groups. Our contributions are as follows:

- We design a single facial feature analysis framework that generalizes across multiple tasks (we demonstrate three, see Sec. 5) in physiological state analysis and dynamic facial expression recognition. To the best of our knowledge, no one has demonstrated such SOTA-competitive versatility across these areas with a single network.

- We introduce. novel ConAdv-ELCA module (Sec. 3.2), designed to capture crucial local features, enhance feature disentanglement, and improve generalizability. This approach also presents an innovative technique for feature alignment and fusion in complex, dynamic scenarios.

- We propose an EGCA module (Sec. 3.3) that extends non-local networks with dynamic channel wise interactions, enabling efficient modeling of inter channel context to enhance global feature representation in facial video analysis.

2 Related Work

2.1 Dynamic Facial Expression Recognition

Recent studies [5, 27, 29, 30, 36, 59, 68] have significantly advanced dynamic facial expression recognition (DFER). Specifically, methods such as [5, 27] exploit hierarchical grouping and external transformer guidance to enhance robustness against noisy frames and semantic ambiguity among emotions. Nevertheless, these models remain specialized, often overlooking integrated local-global feature relationships, limiting their generalization across broader facial analysis tasks. Moreover, some previous works [5, 12, 31] utilize multimodal inputs to increase performance. For example, Chen et al.[5] use human intervention to generate descriptions and relevant prompts related to different emotion classes to perform DFER. We have compared our results with these multimodal methods; however, we believe that a direct comparison may not be entirely fair, as we use only the visual modality. Nonetheless, our framework outperforms most of them, as shown in Supplementary Material Table9.

2.2 Physiological State Analysis via Facial Features

Recent research [11, 21, 23, 26, 34, 41] has transitioned from recognizing overt facial expressions to detecting subtle physiological states like fatigue, drowsiness, and intoxication from facial cues. Although these approaches target critical real-life scenarios such as drunk driving and workplace safety, most studies remain highly task-specific, with limited efforts on cross-task generalization. Notably, methods addressing intoxication detection typically rely on traditional manual feature selection and demonstrate modest performance [21]. Our proposed approach significantly surpasses these existing methods, achieving state-of-the-art performance in both intoxication and drowsiness detection tasks.

3 Methodology

Our proposed network, illustrated in Figure 2, is a compact 3D spatiotemporal network designed for dynamic emotion recognition and physiological state analysis using only visual

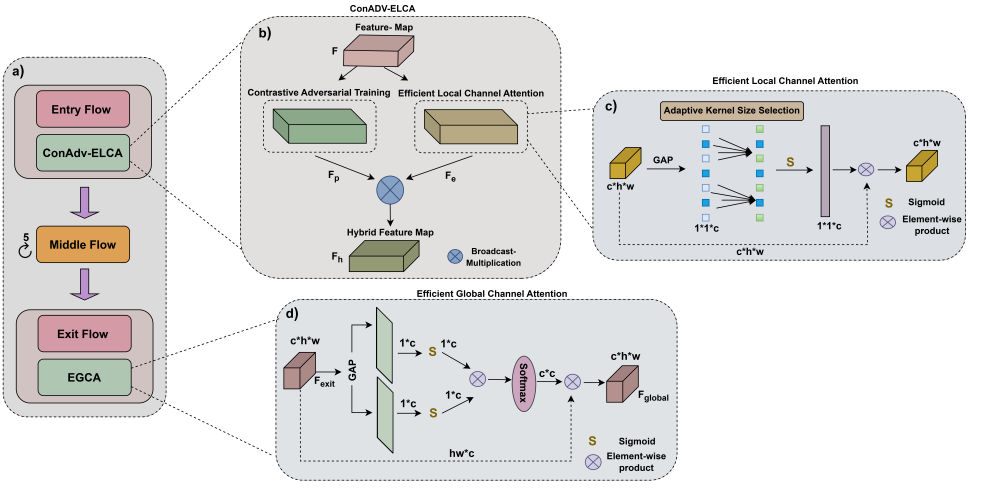

Figure 2: Our 3D spatio-temporal model (a) integrates custom designed ConAdv-ELCA module (b) after the entry-flow for early disentanglement of facial features. ConAdv-ELCA includes the Efficient Local Channel Attention (ELCA) module (c), while customly designed Efficient Global Channel Attention (EGCA) module (d) is applied after the exit-flow to capture global context.

data. As a backbone of our framework, we choose Xception model [6], known for its compact structure and computational efficiency in static image-based tasks. To enable saptiotemporal learning, we extend it to a 3D by replacing all 2D convolutions with 3D convolutions. Converting a 2D model into a 3D architecture typically results in an increase in the model parameters. To maintain computational efficiency (see Table 1) while preserving the model's structure, we reduced the number of blocks in the middle-flow[1] of the Xception network [6] from 8 to 5 (see supplementary material, Table 7 for the rationale behind this choice).

Table 1: Comparison of the efficiency matrix of the proposed network with SOTA.

Model	Params (M)	Inference Time (ms)
Xception (Original)	22.91	10–20
MAE-DFER [41]	85.00	–
CLIPER [31]	88.00	–
Fine-CLIPER [5]	20.00	–
Ours	23.87	12–20

Each middle-flow block in Xception consists of three convolutional layers. This choice ensures that the network remains computationally efficient even after being converted into 3D and integrating our two custom-designed modules, ConAdv-ELCA and EGCA. Table 1 presents a comparison of the parameters and inference time between our developed 3D spatio-temporal network, the baseline 2D Xception network, and several SOTA networks in DFER. To motivate the need for our proposed two different attention mechanisms, we first revisit the prevalent attention mechanisms and argue that they predominantly focus on local features [46] while ignoring global information, which limits the generalization of the model.

[1]The terms entryflow, middleflow, and exitflow were originally introduced in the Xception network paper [6] as shown in Figure 2. We use these terms as they are, to stay in line with the existing literature.

3.1 Revisiting Channel Attention Mechanisms

Channel attention mechanisms enhance CNN feature representations by weighting channel importance. Given a feature tensor $T \in \mathbb{R}^{c\times h\times w}$, channel attention weights $Ch_{att} \in \mathbb{R}^{c}$ are computed as:

$$Ch_{att} = \sigma(f_c(\text{compress}(T))), \tag{1}$$

where σ is a sigmoid function, f_c denotes fully connected or convolutional mapping, and 'compress' typically involves Global Average Pooling (GAP). The tensor channels are then scaled accordingly:

$$\tilde{T}:,i,:,: = Chatt_i \cdot T_{:,i,:,:}, \quad i \in 0,\ldots,c-1. \tag{2}$$

However, GAP tends to emphasize low-frequency (local) features [42], limiting the ability to capture global contextual relationships [46] critical for comprehensive facial understanding. Consequently, local attention alone becomes inadequate for higher-level semantic tasks that require long-range dependencies.

To address this limitation, we propose a dual-attention strategy: ConAdv-ELCA focuses on refining early layer local feature representations, while EGCA captures global channel dependencies in deeper layers. This synergetic approach effectively balances detailed and contextual facial information, significantly improving cross task generalization (validated by ablations in Table 5).

3.2 ConAdv-ELCA

We introduce ConAdv-ELCA (see Figure 2 block (b, c)), a novel dual-path attention mechanism with contrastive adversarial supervision, designed to disentangle semantically relevant facial features from early feature maps. Feature maps extracted after the entry-flow of our network are simultaneously processed to enhance relevant important channel similar to [56], which improves local feature discrimination F_e, and projected gradient descent (PGD), which introduces controlled perturbations to generate adversarial feature maps F_p. Given that adversarial training is most effective in early layers [61], where low-level features such as edges and textures are more susceptible to perturbations, we apply PGD at this stage to improve robustness. However, combining adversarially perturbed features F_p with their clean counterparts F_e increases the size of the feature map and introduces redundancy. To address this, Barlow-Twins loss is utilized to reduce redundancy between two feature maps by aligning the adversarially perturbed features with their clean counterpart.

A key challenge in adversarial training is balancing robustness and accuracy. While larger epsilon values (e.g., $\varepsilon = \frac{8}{255}$) improve robustness, they often degrade model performance [22, 39]. However, in our approach, we mitigate this trade-off by using Barlow Twins to align clean and perturbed features not only to reduce redundancy but also to enhance generalization in the early layers by utilizing the benefits of adversarial perturbations. See (supplementary material, Table 8) which demonstrates that, while direct adversarial training can harm model performance, combining it with self-supervised alignment loss [64] allows our model to become generalized across various downstream tasks without hurting accuracy.

To generate adversarial perturbations, we employ PGD as:

$$\delta = \varepsilon \cdot \text{sign}\left(\nabla_{F_{\text{entry}}} L(f_\theta, F_{\text{entry}})\right) \tag{3}$$

where F_{entry} is the feature map after the entry-flow, f_θ is the model trained in the entry-flow and L is its classification loss. This perturbation δ is applied to create a perturbed feature map F_{p}, which is subsequently aligned with F_{e}, using Barlow Twins.

The Barlow Twins loss works by comparing these two feature maps, encouraging alignment between the clean and adversarial feature maps by computing their cross-correlation matrix and minimizing discrepancies between the diagonal entries while decorrelating the off-diagonal entries to promote discriminative representations. The cross-correlation matrix C is computed between the clean feature map F_{e} and perturbed feature map F_{p} as follows:

$$C_{ij} = \frac{1}{N} \sum_{n=1}^{N} F_{\mathrm{e}}^{(n)}(i) \cdot F_{\mathrm{p}}^{(n)}(j) \tag{4}$$

where N is the batch size. The objective here is to ensure that the two feature maps are highly correlated along corresponding dimensions (diagonal elements) while remaining diverse across other dimensions (non-diagonal elements), thus enhancing discriminability.

Following adversarial alignment, we fuse F_{e} and F_{p} to obtain a hybrid feature map F_{h} with broadcast multiplication. This representation preserves both robust local features and generalized adversarially invariant features. This hybrid feature F_{h} is then passed to modified middle-flow of our network.

3.3 EGCA

Figure 2 (block d) provides an overview of our EGCA module. The idea of EGCA is to enhance non-local neural network [57] by incorporating 1D convolution similar to [56]. Let $F_{\mathrm{exit}} \in \mathbb{R}^{c \times h \times w}$ represent the input feature map from the exit-flow, where c is the number of channels, and h and w are the height and width of the tensor, respectively. The feature map from the exit-flow undergoes global average pooling, followed by a 1D convolution to obtain the query and key tensors, each sized $1 \times c$. A softmax operation is then applied to produce a global channel attention map of size $c \times c$. Meanwhile, the value tensor is obtained by reshaping F_{exit} into $hw \times c$ without undergoing GAP, as depicted in Figure 3, where an arrow directly connects F_{exit} to the multiplication step after the softmax operation. Finally, attention map $c \times c$ is then applied to $hw \times c$, and the resulting product is reshaped back to $c \times h \times w$, yielding the globally attended feature map, denoted as F_{global}.

4 Experiments

To evaluate our proposed approach, we apply it to a set of tasks using five distinct facial video datasets. The tasks include facial expression recognition on three datasets; the Dynamic Facial Expression in-the-Wild (DFEW) [19] dataset (with 16,372 video clips across seven expression classes), the Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos (FERV39K) [58] (with 38,935 clips spanning 22 scenes and seven expression classes), and the Acted Facial Expressions in the Wild (AFEW) [10] dataset (with 1,832 clips). For physiological state analysis, we use the National Tsinghua University Driver Drowsiness Detection (NTHU-DDD) dataset [45], which includes 360 video sequences classified into drowsy and nondrowsy states, and the MiX Telematics alcohol intoxication dataset [21], comprising 180 ten-minute clips segmented into three levels of intoxication based on blood alcohol concentration (BAC). Whereas other approaches tend to present results from a single task domain, we include this more comprehensive evaluation to assess the suitability of the proposed approach for different facial analysis tasks.

We also noted that the FERV39K and DFEW datasets contain captions embedded in some frames of certain videos. To prevent potential data leakage, we ensured that only cropped face regions were passed to the model. Interestingly, no previous study has highlighted this issue. To ensure a fair comparison, we followed previous studies and utilized the

Table 2: Comparison with previous work on estimating BAC through facial features.

Ref.	Data	Class type	AUROC	Accuracy	F1
[23]	IR Videos/Eye tracking system	Binary 0.05	0.88	0.76	0.75
		Binary 0.08	-	-	-
		Multi-class	-	-	-
[29]	IR Videos/Eye tracking system	Binary 0.05	-	-	-
		Binary 0.08	0.80	0.82	-
		Multi-class	-	-	-
[21]	RGB Videos/Conventional camera	Binary 0.05	-	-	-
		Binary 0.08	-	-	-
		Multi-class	-	0.75	0.75
Ours	RGB Videos/Conventional camera	Binary 0.05	**0.91**	**0.87**	**0.88**
		Binary 0.08	**0.93**	**0.90**	**0.89**
		Multi-class	**0.90**	**0.87**	**0.86**

Binary 0.05: *Sober [BAC < 0.05] or Drunk [BAC ≥ 0.05] g/dL*
Binary 0.08: *Sober [BAC < 0.08] or Drunk [BAC ≥ 0.08] g/dL*
Multi-class: *Sober [BAC = 0.00], Moderate [0.05 ≤ BAC ≤ 0.07], or Severe [BAC ≥ 0.08] g/dL*

predefined training and testing splits provided within the datasets for FERV39K, DFEW, AFEW, and NTHU-DDD. For MiX Telematics, results are reported using 5-fold cross-validation following [21]. For details on dataset statistics, please refer to the supplementary material section 4.

4.1 Implementation Details

Preprocessing. All input frames were resized to 299×299 pixels. For the DFEW dataset, which provides face cropped clips, only resizing was applied. For NTHU-DDD, MiX-Telematics, FERV39K, and AFEW, face detection and alignment were performed using MTCNN [66], followed by cropping and resizing. We adopted evaluation metrics consistent with prior work for fair comparison. Accuracy and AUC were used for the MiX-Telematics dataset; Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) for DFEW, FERV39K, and AFEW; and Accuracy for NTHU-DDD. UAR reflects the average per-class recall, while WAR accounts for class imbalance via instance-weighted recall.

Training and Inference. The model was trained using SGD with a learning rate of 0.001 (adapted from Xception's JFT-based setup [6]), momentum of 0.9, and weight decay of 1×10^{-5}. We applied `StepLR` scheduling with a step size of 35 and gamma of 0.1. Categorical Cross-Entropy was used for multi-class datasets; Binary Cross-Entropy for NTHU-DDD. All training was conducted in PyTorch on an NVIDIA RTX A6000 GPU.

Table 3: Accuracy Comparison of various methods on NTHU-DDD dataset for driver drowsiness detection. Highest results are in bold while the 2nd highest are underlined.

Methods	Accuracy (%)
CNN + ConvCGRNN [53]	84.81
Hierarchical Temporal DBN-HMM [50]	84.82
CNN + TSC-LSTM-Refinement [15]	84.85
ResNeXt3D-101 + LSTM [52]	88.60
2s-STGCN [1]	92.70
IsoSSL-MoCo [10]	<u>93.71</u>
VBFLFLA [63]	93.10
Ours	**95.04**

5 Results and Discussion

5.1 Comparison on Physiological State Analysis Tasks

We evaluated our proposed 3D spatiotemporal framework on driver drowsiness detection using the NTHU-DDD dataset [60] and alcohol-induced impairment classification using the MiX-Telematics dataset [21]. As summarized in Table 3, our method significantly outperforms current state-of-the-art approaches for driver fatigue detection, achieving an accuracy of **95.04%**, surpassing the closest competitor IsoSSL-MoCo [40] by approximately 1.3%. Similarly, on the challenging intoxication estimation task (as shown in Table 2), our approach demonstrates superior performance compared to existing models, highlighting its effectiveness in capturing subtle physiological facial cues. It is important to highlight that, while numerous studies analysed drunk driving through driving behavior [33, 48], this work instead analysed video of drivers; such studies are relatively scarce, making these results particularly noteworthy.

Table 4: Performance comparison of methods on the FERV39K dataset in terms of UAR and WAR. Highest results are in bold and 2nd highest are underlined.

Methodology	DFEW		FERV39k		AFEW	
	UAR (%)	WAR (%)	UAR (%)	WAR (%)	UAR (%)	WAR (%)
Uni-modal						
I3D-RGB [9]	-	-	-	-	41.86	45.41
Resnet18+GRU [8, 27]	51.68	64.02	-	-	43.75	46.72
EmotiW-2019 Baseline [9]	-	-	-	-	-	38.81
Resnet18+LSTM [24, 27]	51.32	63.85	-	-	43.96	48.82
3D ResNet18 [44]	-	-	-	-	42.14	45.67
AEN [27]	56.66	<u>69.37</u>	<u>38.18</u>	47.88	50.88	54.60
NR-DFERNet [29]	<u>55.77</u>	68.01	33.99	45.97	48.37	53.54
GCA+IAL [30]	55.71	69.24	-	-	-	-
IAL [30]	-	-	35.82	<u>48.54</u>	-	-
STT [46]	54.58	66.65	37.76	48.11	<u>49.11</u>	<u>54.23</u>
C3D [31]	-	-	-	-	43.75	46.72
R(2+1)D [32]	-	-	-	-	42.89	46.19
M3DFEL [36]	-	-	35.94	47.67	-	-
DPCNet [35]	55.13	66.32	-	-	47.86	51.67
Freq-HD [20]	-	-	33.07	45.26	-	-
Former-DFER [38]	53.69	65.70	37.20	46.85	47.42	50.92
Ours	**60.89**	**71.98**	**44.08**	**53.78**	**54.02**	**56.77**

Table 5: Ablation study results across five datasets. UAR: Unweighted Average Recall, WAR: Weighted Average Recall, AUC: Area Under the Curve.

ELCA, EGCA, ConAdv	DFEW		FERV39K		AFEW		NTHU-DDD	MiX-Telematics	
	UAR	WAR	UAR	WAR	UAR	WAR	Accuracy	Accuracy	AUC
(none)	50.12	58.77	35.01	38.97	48.60	50.95	85.98	79.29	81.34
ELCA (after exit)	55.21	63.98	36.73	41.11	49.59	53.61	86.95	80.45	82.26
ELCA (after entry)	56.69	65.31	38.33	44.89	50.94	54.88	87.71	83.25	84.51
ELCA (after entry), EGCA (at/after exit)	**61.17**	68.37	42.39	50.09	53.21	55.62	93.59	85.67	88.21
ELCA (entry), EGCA (exit), ConAdv	60.89	**71.98**	**44.08**	**52.35**	**54.02**	**56.77**	**95.04**	**87.71**	**90.87**

5.2 Evaluation on Dynamic Facial Expression Recognition

To validate the generalizability of our model, we conducted experiments on three standard DFER datasets: DFEW, FERV39k, and AFEW. Table 4 presents a comprehensive comparison of our method against several competitive baselines and recent SOTA techniques. Our method consistently achieves the highest performance across all three datasets. Specifically, on DFEW, our model attains **60.89%** UAR and **71.98%** WAR, significantly exceeding previous approaches such as AEN [27] and NR-DFERNet [29]. On FERV39k and AFEW, similar

trends were observed, with our model achieving clear margins over prior methods, underlining the efficacy of our dual-attention mechanisms in capturing both local and global facial dynamics.

5.3 Cross-task/Dataset Generalization Analysis

To assess whether the proposed model learns transferable facial representations that generalize across tasks with disjoint label spaces (e.g., 7-class emotion recognition in DFEW vs. 3-class intoxication detection in MiX Telematics), we evaluated representation level generalization of our framework. Specifically, we finetuned the model on DFEW and extracted the features from unseen tasks (NTHU-DDD and MiX Telematics), and visualize the learned embeddings using t-SNE. As shown in Figure 3, the embeddings form semantically meaningful clusters aligned with target labels, indicating robust cross task encoding. To complement this, we conduct cross dataset evaluations: training on one dataset (e.g., DFEW) and testing on others (FERV39K, AFEW) and vice versa without modifying the architecture. Despite different datasets, the model retains good performance (as shown in Table 6), confirming the efficacy of proposed framework.

5.4 Ablation Study

To assess the impact of our custom designed ConAdv-ELCA and EGCA modules, we conducted ablation studies on MiX-Telematics, DFEW, FERV39K, AFEW, and NTHU-DDD, testing configurations with and without local/global channel attention and contrastive adversarial training (Tables 5). The results revealed that: (a) applying local channel attention after entry-flow is more beneficial than after exit-flow, (b) adding global channel attention after exit-flow further enhanced the performance, and (c) integrating adversarial training with contrastive learning consistently improved the results, except for DFEW, where WAR increases but UAR decreases.

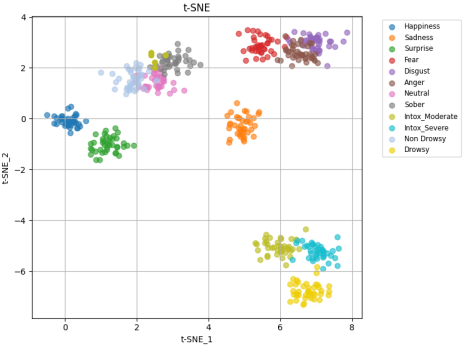

Figure 3: t-SNE visualization of features from a model trained on DFEW only. Emotion classes form clear clusters, indicating strong within task learning, while unseen NTHU-DDD and MiX-Telematics samples also cluster distinctly, demonstrating transferable facial representations across tasks.

Table 6: Generalizability of the proposed model across datasets. UAR: Unweighted Average Recall, WAR: Weighted Average Recall.

Finetuning Dataset	**FERV39k**	**DFEW**	**AFEW**	**NTHU-DDD**	**Mix Telematics**
	UAR/WAR	UAR/WAR	UAR/WAR	Accuracy	Accuracy
FERV39k	44.08 / 53.78	55.09 / 62.75	52.87 / 54.13	-	-
DFEW	41.63 / 48.91	60.89 / 71.98	53.79 / 55.02	-	-
Mix Telematics	-	-	-	90.48	90.01

As discussed in Section 3.2, we further examined the impact of each module in ConAdv-ELCA configurations. The results (in supplementary material Table 8) confirmed that direct

adversarial training leads to performance degradation, likely due to overfitting on adversarial samples. In contrast, our contrastive adversarial training with ECA effectively enhanced generalization across all datasets.

Table 5 highlights that incorporating ELCA, EGCA, and ConAdv improves accuracy on MiX-Telematics by 8.42%. Beyond MiX-Telematics, Tables 5 demonstrate similar gains across other datasets: WAR increases by 13.21% on DFEW, while UAR and WAR improve on FERV39K (9.07% and 5.82%), AFEW (5.42% and 5.67%), and NTHU-DDD (9.06%). These consistent improvements validate the effectiveness of our approach. Interestingly, while adversarial training improves WAR on DFEW, it leads to a decrease in UAR. This suggests that adversarial augmentation may disproportionately impact class-wise balance, particularly affecting underrepresented emotion classes. Further investigation is needed to understand and mitigate this effect.

5.5 Conclusion

We proposed a compact, unified 3D spatiotemporal neural network tailored for facial video analysis across multiple domains: specifically, dynamic facial expression recognition (DFER), driver drowsiness detection, and intoxication level estimation. The model integrates two novel modules: ConAdv-ELCA, which disentangles local facial features using contrastive adversarial learning, and EGCA, which enhances global contextual understanding via efficient channel-wise attention. Through extensive experiments on five benchmark datasets, our method consistently outperformed state-of-the-art uni-modal models all while relying solely on visual facial data. Our results demonstrate that the proposed architecture learns compact and transferable facial representations that generalize across tasks without architectural changes or modality fusion. This positions our model as a promising unified backbone for real world affective and physiological state understanding.

5.6 Acknowledgments

SZG was partially funded by a WANMA grant funded by the Western Australian Future Health Research and Innovation Fund

References

[1] Jing Bai, Wentao Yu, Zhu Xiao, Vincent Havyarimana, Amelia C Regan, Hongbo Jiang, and Licheng Jiao. Two-stream spatial–temporal graph convolutional networks for driver drowsiness detection. *IEEE Transactions on Cybernetics*, 52(12):13821–13833, 2021.

[2] Salah Eddine Bekhouche, Yassine Ruichek, and Fadi Dornaika. Driver drowsiness detection in video sequences using hybrid selection of deep features. *Knowledge-Based Systems*, 252:109436, 2022.

[3] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 18(8):5619–5627, 2022.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[5] Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. *arXiv preprint arXiv:2407.02157*, 2024.

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017.

[7] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4673–4682, 2024.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[9] Abhinav Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pages 546–550, 2019.

[10] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516, 2013.

[11] Guanglong Du, Linlin Zhang, Kang Su, Xueqian Wang, Shaohua Teng, and Peter X Liu. A multimodal fusion fatigue driving detection method based on heart rate and perclos. *IEEE transactions on intelligent transportation systems*, 23(11):21810–21820, 2022.

[12] Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expression recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024.

[13] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

[14] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[15] Jing-Ming Guo and Herleeyandi Markoni. Driver drowsiness detection using hybrid convolutional neural network and long short-term memory. *Multimedia tools and applications*, 78:29059–29087, 2019.

[16] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Andrew Hundt, Gabrielle Ohlson, Pieter Wolfert, Lux Miranda, Sophia Zhu, and Katie Winkle. Love, joy, and autism robots: A metareview and provocatype. *arXiv preprint arXiv:2403.05098*, 2024.

[19] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. DFEW: A large-scale database for recognizing dynamic facial expressions in the wild. In *ACM MM*, pages 2881–2889, 2020.

[20] Mohan Karnati, Ayan Seal, Debotosh Bhattacharjee, Anis Yazidi, and Ondrej Krejcar. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Transactions on Instrumentation and Measurement*, 72:1–31, 2023.

[21] Ensiyeh Keshtkaran, Brodie von Berg, Grant Regan, David Suter, and Syed Zulqarnain Gilani. Estimating blood alcohol level through facial features for driver impairment assessment. In *IEEE Winter Conf. on App. of Comput. Vis.*, pages 4539–4548, 2024.

[22] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in neural information processing systems*, 33:2983–2994, 2020.

[23] Kevin Koch, Martin Maritsch, Eva Van Weenen, Stefan Feuerriegel, Matthias Pfäffli, Elgar Fleisch, Wolfgang Weinmann, and Felix Wortmann. Leveraging driver vehicle and environment interaction: Machine learning using driver monitoring cameras to detect drunk driving. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–32, 2023.

[24] Puneet Kumar, Alexander Vedernikov, and Xiaobai Li. Measuring non-typical emotions for mental health: A survey of computational approaches. *arXiv preprint arXiv:2403.08824*, 2024.

[25] Ali Ladak, Jamie Harris, and Jacy Reese Anthis. Which artificial intelligences do people care about most? a conjoint experiment on moral consideration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2024.

[26] Alexandre Lambert, Aakash Soni, Assia Soukane, Amar Ramdane Cherif, and Arnaud Rabat. Artificial intelligence modelling human mental fatigue: a comprehensive survey. *Neurocomputing*, page 126999, 2023.

[27] Bokyeung Lee, Hyunuk Shin, Bonhwa Ku, and Hanseok Ko. Frame level emotion guided dynamic facial expression recognition with emotion grouping. In *CVPR*, pages 5681–5691, 2023.

[28] John D Lee, Dary Fiorentino, Michelle L Reyes, Timothy L Brown, Omar Ahmad, James Fell, Nic Ward, and Robert Dufour. Assessing the feasibility of vehicle-based sensors to detect alcohol impairment. *Washington, DC: National Highway Traffic Safety Administration*, 2010.

[29] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022.

[30] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *AAAI*, volume 37, pages 67–75, 2023.

[31] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[32] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024.

[33] ZhenLong Li, HaoXin Wang, YaoWei Zhang, and XiaoHua Zhao. Random forest–based feature selection and detection method for drunk driving recognition. *International Journal of Distributed Sensor Networks*, 16(2):1550147720905234, 2020.

[34] Fan Liu, Delong Chen, Jun Zhou, and Feng Xu. A review of driver fatigue detection and its advances on the use of rgb-d camera and deep learning. *Engineering Applications of Artificial Intelligence*, 116:105399, 2022.

[35] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE CAA J. Autom. Sinica*, 4(4):668–676, 2017.

[36] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022.

[37] Fuyan Ma, Bin Sun, and Shutao Li. Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[38] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18177–18186, 2022.

[39] Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

[40] Luntian Mou, Chao Zhou, Pengtao Xie, Pengfei Zhao, Ramesh Jain, Wen Gao, and Baocai Yin. Isotropic self-supervised learning for driver drowsiness detection with attention-based multimodal fusion. *IEEE TMM*, 25:529–542, 2021.

[41] Ercheng Pei, Meshia Cedric Oveneke, Yong Zhao, Dongmei Jiang, and Hichem Sahli. Monocular 3d facial expression features for continuous affect recognition. *IEEE TMM*, 23:3540–3550, 2020.

[42] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021.

[43] Alec Radford and et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[44] Surjodeep Sarkar, Manas Gaur, L Chen, Muskan Garg, Biplav Srivastava, and Bhaktee Dongaonkar. Towards explainable and safe conversational agents for mental health: A survey. *arXiv preprint arXiv:2304.13191*, 2023.

[45] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 27, 2014.

[46] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global-local, spatial-channel attention for image retrieval. In *IEEE Winter Conf. on App. of Comput. Vis.*, pages 2754–2763, 2022.

[47] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023.

[48] Yifan Sun, Jinglei Zhang, Xiaoyuan Wang, Zhangu Wang, and Jie Yu. Recognition method of drinking-driving behaviors based on PCA and RBF neural network. *Promet-Traffic & Transportation*, 30(4):407–417, 2018.

[49] Zeng Tao, Yan Wang, Junxiong Lin, Haoran Wang, Xinji Mai, Jiawen Yu, Xuan Tong, Ziheng Zhou, Shaoqi Yan, Qing Zhao, et al. A^3lign-DFER: Pioneering comprehensive dynamic affective alignment for dynamic facial expression recognition with CLIP. *arXiv preprint arXiv:2403.04294*, 2024.

[50] Masoumeh Tashakori, Ali Nahvi, and Serajeddin Ebrahimian Hadi Kiashari. Driver drowsiness detection using facial thermal imaging in a driving simulator. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of engineering in medicine*, 236(1):43–55, 2022.

[51] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[52] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.

[53] Toan H Vu, An Dang, and Jia-Ching Wang. A deep neural network for real-time driver drowsiness detection. *IEICE TRANSACTIONS on Information and Systems*, 102(12): 2637–2641, 2019.

[54] Chen Wang, Liang Shao, Jun Liu, and Jiawei Xiang. Driver abnormal behavior detection system using two-stage object detection. In *13th International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE 2023)*. IET, 2023.

[55] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17958–17968, 2023.

[56] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, pages 11534–11542, 2020.

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[58] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022.

[59] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *ACMMM*, pages 101–110, 2022.

[60] Ching-Hua Weng, Ying-Hsiu Lai, and Shang-Hong Lai. Driver drowsiness detection via a hierarchical temporal deep belief network. In *Computer Vision–ACCV 2016 Workshops*, pages 117–133. Springer, 2017.

[61] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 501–509, 2019.

[62] Shaoqi Yan, Yan Wang, Xinji Mai, Qing Zhao, Wei Song, Jun Huang, Zeng Tao, Haoran Wang, Shuyong Gao, and Wenqiang Zhang. Empower smart cities with sampling-wise dynamic facial expression recognition via frame-sequence contrastive learning. *Computer Communications*, 216:130–139, 2024.

[63] Lie Yang, Haohan Yang, Henglai Wei, Zhongxu Hu, and Chen Lv. Video-based driver drowsiness detection with optimised utilization of key facial features. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[64] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[65] Buyuan Zhang, Haoyang Zhang, Tao Zhen, Bowen Ji, Liang Xie, Ye Yan, and Erwei Yin. A two-stage real-time gesture recognition framework for uav control. *IEEE Sensors Journal*, 2024.

[66] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[67] Lei Zhao, Zengcai Wang, Guoxin Zhang, and Huanbing Gao. Driver drowsiness recognition via transferred deep 3d convolutional network and state probability vector. *Multimedia Tools and Applications*, 79(35):26683–26701, 2020.

[68] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *ACMMM*, pages 1553–1561, 2021.

[69] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023.

Supplementary Material

1 Rationale behind reducing 3 blocks from middle-flow

Table 7 shows our rationale behind reducing 3 blocks from the middle-flow of the Xception network. We made this choice in order to maintain a balance between the network's computational cost and accuracy. Our empirical studies indicate that using 5 middle-flow blocks provides the best balance, as further increasing the blocks beyond 5 leads to only marginal performance improvements, while the number of parameters increases by 3.12 million for increasing 1 block, leading to a higher computational cost without significant gains.

Table 7: The results indicate that using 5 middle flow blocks in the Xception network provides the best balance between performance and the number of parameters.

Blocks	Params (M)	NTHU-DDD (Acc)	Mix Telematics (Acc)	DFEW (UAR/WAR)	FERV39k (UAR/WAR)
3	17.63	82.31%	78.63%	47.21/55.02%	33.85/36.98%
4	20.75	84.36%	80.19%	49.21/57.35%	34.41/37.92%
5	23.87	85.98%	81.34%	50.12/58.77%	35.01/38.97%
6	26.99	86.74%	81.89%	50.61/58.93%	35.08/39.31%
8	33.23	86.79%	81.94%	50.70/58.98%	35.58/39.60%

2 Ablation Study on ConADV-ELCA

To validate the effectiveness of our contrastive alignment strategy for local facial features, we conduct experiments under various configurations of the ConAdv-ELCA module. As shown in Table 8, although adversarial perturbation and alignment losses such as Barlow Twins [64] operate on different principles, their combination as a regularization mechanism improves performance compared to using adversarial training or efficient channel attention alone.

Table 8: Comparison of proposed ConAdv-ELCA configurations on various datasets. AT denotes Adversarial Training, ECA represents Efficient Channel Attention, and CL refers to Contrastive Learning. The table highlights that integrating these modules directly into network does not significantly improve performance. It is worth noting that applying Adversarial Training (AT) alone leads to a decline in the model's performance.

AT	ECA	CL	DFEW		FERV39K		NTHU-DDD	Mix-Telematics
			UAR (%)	WAR (%)	UAR	WAR	Accuracy (%)	Accuracy (%)
-	-	-	50.12	58.77	35.01	38.97	85.98	81.34
✓	-	-	48.61	55.29	30.74	35.90	78.87	80.04
-	✓	-	56.69	65.31	38.33	44.89	87.71	83.25
✓	✓	✓	57.40	67.55	39.21	47.41	89.93	84.29

3 Comparison with multimodal approaches

Multi-modal approaches include [5, 7, 12, 28, 31, 47, 69], and are probably increasing in popularity, due in large part to the obvious fact that if one uses more information, one can potentially perform better. Curiously, just as with uni-modal (and particularly visual face only) systems, we cannot find a system that has been demonstrated to perform well across

more than one application. As mentioned in the introduction, the utility (not to mention the likely availability) of different modes of information may vary considerably across different application areas: for instance, textual cues aid expression recognition [8, 12, 31], while physiological signals enhance fatigue [11] and intoxication detection. For those reasons, a multi-modal approach may have to adjust the proportion of contribution from the various modalities, in response to changes in application area. In the extreme case, an extra modality that is very useful for one application area may be useless or even hinder in another area. We also, in the introduction, alluded to some of the drawbacks of multi-modal approaches (such as compute costs). More complete discussion can of course be found in the literature: e.g., [13, 32]. See [38] for a particular illustration of performance degradation in the absence of a modality. Thus, direct comparison of these mutlimodal approaches are not fair with unimodal approaches. Nevertheless, Table 9 shows that our framework surpasses most of them.

However, our purpose is not to argue against multi-modal approaches: as we have stated, we only argue that a well designed uni-modal system can achieve SOTA or near SOTA performance without relying on multi-modality. Further, it is useful to know how well one can do this and to know the size of the margin gained by multi-modal approaches, in cases where they lead.

We make passing note of the fact that the use of VLMs are, in some sense, blurring or changing the traditional definition of multi-modality. A multi-modal system would be commonly assumed to be evidenced by different explicit input streams, of differing modality, on the inputs to the system. A system incorporating a VLM, inside the system, can conceivably still only draw one explicit input (mode) and yet within the system the VLM is implicitly providing input information of a cross-modal nature (as well as other real world information captured by the model and not explicitly provided by the explicit single mode input stream).

Table 9: Performance comparison of methods on the FERV39K dataset in terms of UAR and WAR. Highest results are in bold and 2nd highest are underlined.

Methodology	DFEW		FERV39k		AFEW	
	UAR (%)	WAR (%)	UAR (%)	WAR (%)	UAR (%)	WAR (%)
FineCliper [8]	<u>64.89</u>	<u>75.05</u>	**45.22**	<u>53.98</u>	-	-
EmoCLIP [10]	58.04	62.12	39.35	41.60	44.62	46.19
MAE-DFER [42]	63.41	74.43	43.12	52.07	-	-
CLIPER [54]	57.56	70.84	41.23	51.34	52.00	56.43
MMA-DFER [8]	**67.01**	**77.51**	44.11	**58.52**	-	-
LOGO-Former [57]	54.21	66.98	38.22	48.13	-	-
Align-DFER [45]	-	-	41.87	51.77	-	-
DFER-CLIP [59]	59.61	71.25	41.27	51.65	-	-
Ours (Unimodal)	60.89	71.98	<u>44.08</u>	52.35	**54.02**	**56.77**

4 Datasets

In this section, we will provide a description of the datasets employed in our study.

4.1 FERV39K

FERV39K [58] is the largest publicly available in-the-wild dataset for Dynamic Facial Expression Recognition (DFER), comprising 38,935 video clips captured across 22 diverse scenes within four distinct scenarios. Similar to DFEW, FERV39K also categorised into

seven classes. It is the first large-scale DFER dataset to introduce scenario-scene categorization, support cross-domain adaptability, and offer an extensive collection nearing 39K video clips. To ensure a fair comparison, we utilized the predefined training and testing splits provided within the dataset.

4.2 AFEW

It comprises 1,832 video clips sourced from TV programs and movies. The dataset is provided as separate training, validation, and testing sets. Following the methodology of previous approaches [27], we train our model on the training set and assess its performance on the validation set.

4.3 DFEW

This dataset comprises of 16,372 video clips, each annotated for expression recognition. The source of the clips are various movies and include variations in illumination, occlusions and pose. Expressions are categorised into seven classes (“happy", “sad", “neutral", “angry", “surprise", “disgust" and “fear"). The dataset provides emotion distribution labels as soft annotations. The dataset comes divided into five equal non-overlapping sets, which we have used as folds in five fold cross validation, following [27].

4.4 MiX Telematics

It comprises of 180 clips of 10 minutes (60 for each class); it captures simulated driving behavior under controlled blood alcohol concentration (BAC) levels in 60 participants aged 19 to 76. Driving sessions are segmented by BAC levels of 0.00 g/100ml (sober), 0.05–0.07 g/100ml (low intoxication), and above 0.08 g/100ml (high intoxication). These levels represent real-world thresholds of alcohol impairment based on WHO recommendations and are used to evaluate model performance in identifying driver impairment. Results are reported on 5-fold cross validation as in [21].

4.5 NTHU

This dataset comprises of video sequences from the simulated driving of 36 subjects. The recordings were made in both day and night conditions with active infrared (IR) illumination. The dataset consists of 360 video clips from 36 subjects with 18 subjects are in training, 4 subjects in validation and 14 subjects in testing set. The dataset comprises two categories, drowsy and non-drowsy. We utilized the training, validation, and testing data exactly as provided in the publicly available dataset.