CADResNet: Lightweight Multi-Task Learning for Multi-Label and Valence-Arousal-Dominance Prediction from Bodily Emotional Expressions

Tushar Shinde, Rohan Saha

Indian Institute of Technology Madras, Zanzibar, Tanzania
shinde@iitmz.ac.in

Abstract

Predicting continuous Valence–Arousal–Dominance (VAD) from bodily expressions in naturalistic videos remains a key challenge in affective computing, requiring models that capture subtle kinematic and appearance cues while remaining efficient. We present the first comprehensive multi-task study on the ABEE dataset from the BEEU Challenge 2025, comprising 3,538 short clips with multi-label discrete emotions (29 classes) and continuous VAD ratings (1–9 scale). We systematically evaluate visual features (ResNet-18, Swin-Tiny, ConvNeXt-Tiny) with temporal mean pooling, skeleton keypoints from YOLOv8-nano-pose, and their late fusion. Swin-Tiny features achieve the lowest VAD MAE (1.359) and competitive classification performance (micro F1 0.320). Skeleton fusion enables lightweight modeling (1.043M parameters) without sacrificing VAD accuracy. We introduce **CADResNet**, a compact dilated residual network with channel-wise attention and dimensionally-consistent multi-task regularization aligning categorical predictions with VAD distributions. Our results reveal that multi-label classification is moderately successful, whereas continuous VAD prediction remains challenging, highlighting the limitations of static, mean-pooled representations. We release a reproducible pipeline establishing strong baselines and efficiency-performance trade-offs for body-only affective computing.

Valence-Arousal-Dominance, Bodily-Expressed Emotion, Affective Computing, Video Emotion Recognition

Introduction

Recognizing human emotions from bodily movements in unconstrained video remains a fundamental challenge in affective computing, with applications spanning mental health monitoring, human-robot interaction, and socially intelligent systems (Picard 2000; Vinciarelli et al. 2015). While facial expressions have long dominated the field (Ekman 1992; Zen et al. 2016), bodily cues offer a complementary and often more robust channel, retaining interpretability under occlusion, low resolution, or extreme viewpoints (De Meijer 1989; Kleinsmith and Bianchi-Berthouze 2012). Psychological evidence underscores that posture and kinematics convey distinct affective states, frequently decorrelated from facial signals (Dael, Mortillaro, and Scherer 2012).

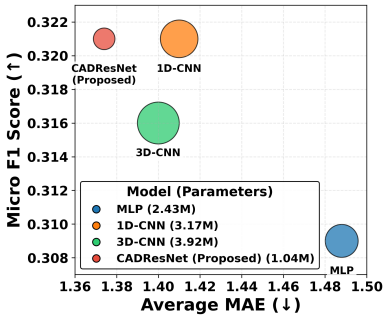

Figure 1: Comparison of model performance on the ABEE dataset. The plot shows micro F1 score versus average MAE for four models (MLP, 1D-CNN, 3D-CNN, and CADResNet). Circle size is proportional to the number of model parameters, highlighting the trade-off between accuracy and model complexity. The proposed CADResNet achieves competitive F1 while maintaining low MAE and minimal parameter count.

The Valence-Arousal-Dominance (VAD) framework (Russell 1980; Mehrabian 1996) represents affect in a continuous three-dimensional space, naturally accommodating blends, intensity variations, and individual differences, advantages over discrete categories for modeling naturalistic behavior. Regressing VAD from body motion alone, however, is substantially harder due to the subtle mapping between musculoskeletal dynamics and affective intensity.

Prior video-based emotion recognition has progressed from handcrafted descriptors (Zhao and Pietikainen 2007) to deep spatiotemporal models (Tran et al. 2015; Carreira and Zisserman 2017), yet remains heavily facial-centric (Li and Deng 2020). Body-focused efforts have leveraged pose features (Kleinsmith and Bianchi-Berthouze 2012) and skeleton graph networks (Yan, Xiong, and Lin 2018), with datasets like BoLD (Luo et al. 2020) advancing multi-label recognition but lacking continuous labels. The Annotated Bodily Expressed Emotion (ABEE) dataset from the BEEU Challenge 2025 addresses this by providing 3,538 short clips with 29 multi-label categories and VAD scores, enabling joint categorical-dimensional modeling.

We conduct the first comprehensive ablation on ABEE,

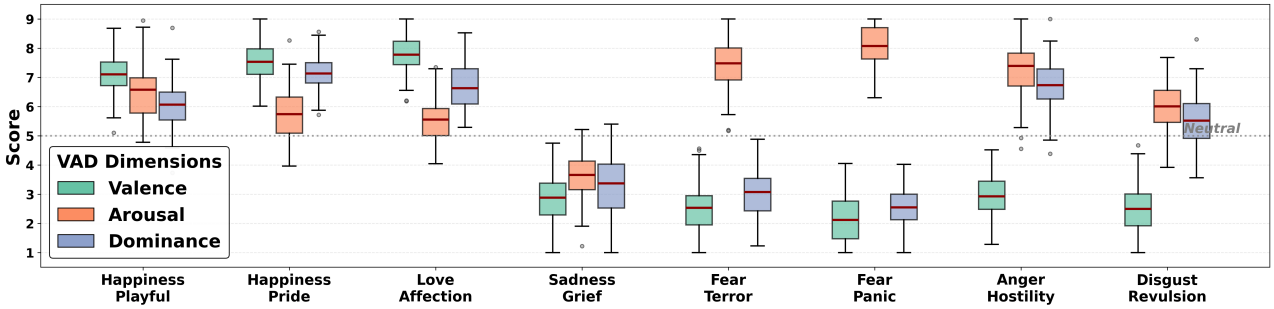

Figure 2: Grouped VAD box plots for each emotion category in the ABEE dataset. For each category, Valence (green), Arousal (orange), and Dominance (blue) distributions are shown. Positive emotions cluster at high Valence, negative emotions at low Valence, and high-arousal states (e.g., Anger, Fear) exhibit elevated Arousal. The dashed line at 5 indicates the neutral midpoint.

evaluating three pretrained visual backbones (ResNet-18 (He et al. 2016), Swin-Tiny (Liu et al. 2021), ConvNeXt-Tiny (Liu et al. 2022)) via TIMM (Wightman 2019), YOLOv8-nano-pose skeletons (Jocher, Chaurasia, and Qiu 2023), and fusions, across four multi-task architectures of controlled capacity. ConvNeXt-Tiny dominates single-modality performance (micro F1 0.324, MAE 1.370), while skeleton fusion aids VAD regression. Results reveal asymmetry: classification reaches F1 $\approx$ 0.32, but VAD MAE $\approx$ 1.37, indicating static representations capture prototypes yet miss fine-grained dynamics, motivating future temporal modeling.

Our contributions are:

- The first systematic study on ABEE, quantifying modality and architecture impacts.
- Identification of Swin-Tiny as the best modality to achieve lowest VAD MAE and competitive classification performance.
- A reproducible pipeline establishing strong baselines for body-only affective computing.

Background

The ABEE dataset comprises 3,538 short videos of spontaneous bodily expressions, divided into 2,462 training and 1,076 test clips. Each clip carries multi-label discrete emotion annotations (29 labels spanning 8 main and 20 sub-categories) and continuous Valence–Arousal–Dominance (VAD) ratings on a 1–9 scale. The training data exhibits moderate correlations between dimensions (Valence–Arousal: 0.45, Valence–Dominance: 0.35, Arousal–Dominance: 0.15) and an average of 3.96 labels per clip, reflecting naturalistic, blended emotional expressions.

Figure 1 summarizes model performance in terms of micro F1 score and mean absolute error (MAE) on the ABEE dataset. Each circle represents a model, with size indicating parameter count. We observe that CADResNet, our proposed model, achieves high F1 scores while maintaining a compact model size, demonstrating an effective balance between accuracy and efficiency.

Analysis of the dataset reveals strong alignment between discrete categories and VAD dimensions (Figure 2). Positive

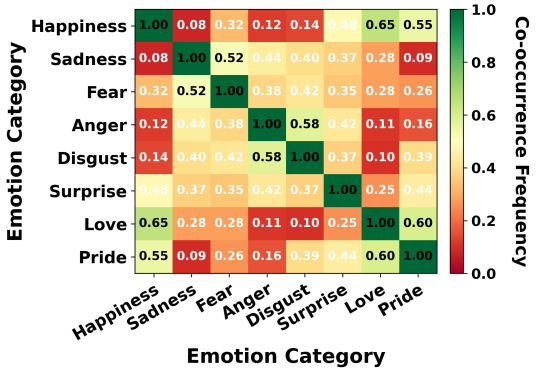

Figure 3: Emotion co-occurrence matrix across ABEE video clips. Each entry represents the frequency with which a pair of emotions co-occur in a single clip, highlighting naturalistic emotional blends. Strong diagonal values indicate frequent single-category expressions, while off-diagonal structure captures multi-label correlations.

emotions (e.g., Happiness, Love) cluster at high Valence, whereas negative emotions (e.g., Sadness, Disgust) concentrate at low Valence. High-arousal states such as Anger and Fear demonstrate elevated Arousal scores, consistent with affective theory. Dominance exhibits more moderate variability, yet provides additional discriminative power for high-intensity emotions.

Figure 3 shows the co-occurrence matrix of discrete emotions. Multi-label annotations are common: the average of 3.96 labels per clip indicates that expressions are often a combination of multiple affective states. The above structure motivates multi-task modeling, where predicting categorical probabilities is jointly aligned with continuous VAD regression resulting in naturalistic emotional blends and improving generalization in both discrete and continuous emotion prediction tasks.

Method

We design a multi-task learning framework to jointly predict multi-label emotions and continuous Valence–Arousal–Dominance (VAD) from body-only video.

Our approach systematically evaluates visual, skeletal, and fused modalities under controlled architectural variations.

Feature Extraction

Visual Features. We extract frame-level embeddings using three pretrained backbones from TIMM (Wightman 2019): ResNet-18 (He et al. 2016), Swin-Tiny (Liu et al. 2021), and ConvNeXt-Tiny (Liu et al. 2022). Frames are uniformly sampled every 5th frame, and features are obtained from the final pooling layer (with classification head removed). Temporal mean pooling produces video-level descriptors $\mathbf{f}_{\text{vis}} \in \mathbb{R}^{d_v}$, where $d_v = 512$ for ResNet-18 and $d_v = 768$ for Swin-Tiny and ConvNeXt-Tiny architectures.

Skeleton Features. We detect 17 body keypoints per frame using YOLOv8-nano-pose (Jocher, Chaurasia, and Qiu 2023). The person with the highest mean confidence is selected, and normalized (x, y, c) coordinates are flattened into a 51-dimensional vector. Mean pooling across frames is applied to the skeleton features, resulting in a fixed-size representation: $\mathbf{f}_{\text{skel}} \in \mathbb{R}^{51}$, where $\mathbf{f}_{\text{skel}}$ is the 51-dimensional vector derived from the keypoints of the human body. To combine the visual and skeleton modalities, we concatenate these representations, yielding the fused feature vector: $\mathbf{f} = [\mathbf{f}_{\text{vis}}; \mathbf{f}_{\text{skel}}]$, where $\mathbf{f}_{\text{vis}}$ is the visual feature vector, and $\mathbf{f}_{\text{skel}}$ is the skeleton feature vector.

Baseline Multi-Task Models

We establish three baselines sharing a learned embedding $h = g(\mathbf{f}; \theta_g)$: **MLP:** Three fully-connected layers ($1024 \to 1024 \to 512$) with ReLU and dropout 0.3, **1D-CNN:** Sequential 1D convolutions over the feature vector, followed by global average pooling, and **Fast 3D-CNN:** Lightweight 3D convolutions on pseudo-spatiotemporal volumes by repeating feature vectors along artificial axes.

Emotion logits $\mathbf{l} \in \mathbb{R}^C$ (where $C = 29$ is the number of emotion classes) and VAD predictions $\hat{\mathbf{v}} \in \mathbb{R}^3$ (for the continuous Valence, Arousal, and Dominance dimensions) are produced from the shared embedding $\mathbf{h}$. The models are trained using the following loss function:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{focal}}(\mathbf{l}, \mathbf{y}_{\text{emo}}) + \lambda \|\hat{\mathbf{v}} - \mathbf{v}_{\text{gt}}\|_2^2, \tag{1}$$

where $\mathcal{L}_{\text{focal}}(\mathbf{l}, \mathbf{y}_{\text{emo}})$ is the focal loss used for the multi-label emotion classification, $\mathbf{y}_{\text{emo}}$ is the ground-truth emotion label vector, and λ is a scalar hyperparameter that balances the classification and regression objectives. The second term represents the mean squared error (MSE) between the predicted VAD values $\hat{\mathbf{v}}$ and the ground truth VAD values $\mathbf{v}_{\text{gt}}$.

CADResNet Multi-Task Learning Method

Our analysis of the ABEE dataset reveals a moderate alignment between discrete emotion categories and continuous VAD (Valence-Arousal-Dominance) statistics. Specifically, each emotion class $c \in \mathcal{C}$ has an associated empirical VAD centroid $\boldsymbol{\mu}_c$, computed as the mean of all VAD vectors $\mathbf{v}_i$ for clips labeled with class c:

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \mathbf{v}_i, \tag{2}$$

where $\mathcal{I}_c$ is the set of instances (video clips) labeled with emotion class c, and $\mathbf{v}_i \in \mathbb{R}^3$ is the VAD vector for clip i. By exploiting this relationship between categorical emotion labels and continuous VAD vectors, our approach aims to enhance the prediction of both discrete emotions and continuous VAD scores. Given the predicted emotion probabilities $\mathbf{p} = \sigma(\mathbf{l})$ (where σ is the sigmoid function applied to the emotion logits $\mathbf{l}$), we define the *emotion-derived VAD* as the convex combination of class centroids:

$$\hat{\mathbf{v}}_{\text{emo}} = \mathbf{p}^\top \mathbf{M}, \tag{3}$$

where $\mathbf{M} \in \mathbb{R}^{C \times 3}$ stores the centroids for all C emotion classes. To enforce consistency between the direct VAD regression output $\hat{\mathbf{v}}$ and the emotion-derived VAD estimate $\hat{\mathbf{v}}_{\text{emo}}$, we introduce a consistency loss function:

$$\mathcal{L}_{\text{cons}} = \|\hat{\mathbf{v}}_{\text{emo}} - \mathbf{v}_{\text{gt}}\|_2^2, \tag{4}$$

where $\mathbf{v}_{\text{gt}}$ is the ground-truth VAD vector.

The full training objective is the weighted sum of the focal loss for emotion classification $\mathcal{L}_{\text{focal}}$, the mean squared error (MSE) loss for VAD regression, and the consistency loss:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{vad}} \|\hat{\mathbf{v}} - \mathbf{v}_{\text{gt}}\|_2^2 + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}, \tag{5}$$

where λ_{vad} and λ_{cons} are hyperparameters controlling the relative contributions of the VAD regression and consistency losses. We set $\lambda_{\text{vad}} = 2.0$ and $\lambda_{\text{cons}} = 0.3$, with a linear ramp of λ_{cons} over the first 100 epochs to stabilize training.

The above mentioned multi-task learning approach leverages the complementary information between discrete classification and continuous regression tasks, allowing for improved generalization across both emotion prediction tasks. The consistency loss ensures that the predicted VAD values are consistent with the categorical emotion distributions, further enhancing the robustness of the model for underrepresented emotions.

Experimental Setup

Datasets. We evaluate our multi-task framework on the Annotated Bodily Expressed Emotion (ABEE) dataset (ABE 2025), used in the BEEU Challenge 2025. ABEE contains 3,538 short videos depicting naturalistic bodily expressions, with multi-label emotion annotations spanning 29 classes (8 main + 20 sub-categories) and continuous Valence–Arousal–Dominance (VAD) ratings on a 1–9 scale. Following standard practice, we split the dataset into 2,462 training and 1,076 test videos, reserving 20% of the training set (493 clips) for validation stratified by dominant emotion.

Models. We evaluate four multi-task architectures sharing a learned embedding $h = g(\mathbf{f}; \theta_g)$: **MLP**: Three fully-connected layers ($1024 \to 1024 \to 512$) with ReLU and dropout ($p = 0.3$), **1D-CNN**: 1D convolutions with channels progressively increased ($256 \to 512 \to 512 \to 256$), followed by fully-connected layers ($256 \to 1024 \to 512$), **Fast 3D-CNN**: Lightweight 3D convolutions (channels 64 $\to 128 \to 128 \to 64$) on a pseudo-volume of size $(5, 8, d)$, followed by fully-connected layers ($2048 \to 1024 \to 512$), and proposed **CADResNet**: Dilated residual blocks (dilation

Table 1: Overall comparison for our framework with visual (Swin-Tiny) and skeleton features on the ABEE validation split.

Model	F1↑	IoU↑	Avg MAE↓	MAE_V↓	MAE_A↓	MAE_D↓	Params (M)↓	Size (MB)↓	Time (min)↓
MLP	0.309	0.177	1.488	1.619	1.503	1.342	2.430	9.27	**0.043**
1D-CNN	0.321	**0.194**	1.410	**1.547**	1.392	1.292	3.167	12.08	2.166
3D-CNN	0.316	0.168	1.400	1.628	1.317	1.256	3.923	14.96	26.223
Proposed	**0.321**	0.184	**1.374**	1.569	**1.313**	**1.240**	**1.043**	**3.98**	0.124

rates 1, 2, 4, 8) with channel attention, followed by a 512-dimensional bottleneck. To ensure fair comparisons, MLP, 1D-CNN, and Fast 3D-CNN have 2.4M trainable parameters; CADResNet remains lighter at 1.0M. All models produce emotion logits $\mathbf{l} \in \mathbb{R}^{29}$ and VAD predictions $\hat{\mathbf{v}} \in \mathbb{R}^3$ via independent linear heads.

Training Details. Training uses AdamW (Loshchilov and Hutter 2017) with a learning rate of 10^{-3} and weight decay 10^{-4}, up to 100 epochs with early stopping on validation micro F1. Batch size is 64, and features are standardized using statistics computed on the training set only.

Evaluation Protocol. For multi-label classification, micro-averaged F1 is reported as the primary metric, with sample-averaged IoU as complementary. Thresholds for each model are tuned on validation to maximize F1. For VAD regression, we report average MAE and per-dimension MAE. We also track model complexity (parameters, size) and training time to assess computational efficiency.

All experiments are implemented in PyTorch 2.1 on NVIDIA GPUs with CUDA 12.1.

Results and Discussion

In this section, we present and analyze the experimental results obtained from our multi-task learning framework for emotion classification and continuous Valence–Arousal–Dominance (VAD) prediction.

Valence–Arousal–Dominance (VAD) Prediction. Table 2 reports ablations across visual backbones and skeleton features. Among single-modality visual inputs, Fast 3D-CNN applied to Swin-Tiny and ConvNeXt-Tiny achieves the lowest average MAEs (1.359 and 1.370, respectively), while skeleton-only features remain competitive (1.365), highlighting the strong affective signal present in kinematics alone.

In the fused setting with Swin-Tiny, our CADResNet achieves the lowest average MAE of 1.374, improving over the strongest baseline (Fast 3D-CNN) by 1.9%. Gains are consistent across dimensions: Valence (1.569), Arousal (1.313), and Dominance (1.240). Importantly, these results are achieved with only 1.043M parameters and a 3.98 MB footprint, demonstrating an excellent efficiency-performance trade-off.

Multi-Label Emotion Classification. Classification performance is more modality-sensitive (Table 2). ConvNeXt-Tiny yields the highest micro F1 (0.324) and IoU (0.192), while skeleton features achieve 0.322 F1 with Fast 3D-CNN, surprisingly competitive with visual inputs. This confirms that body posture and motion provide highly diagnostic cues for emotion recognition.

In fused modalities, CADResNet with consistency regularization achieves micro F1 of 0.321 and IoU of 0.184, competitive with heavier convolutional baselines while using far fewer parameters. These results highlight that lightweight architectures can maintain strong performance without excessive computational cost.

Computational Complexity. Table 1 summarizes model efficiency. The proposed CADResNet achieves state-of-the-art trade-offs, with best VAD and classification performance with minimal parameters (1.043M) and memory footprint (3.98 MB). In contrast, the 3D-CNN baseline is heavier (3.923M, 14.96 MB) and incurs substantially higher training cost (26.22 min). These observations demonstrate that lightweight temporal modeling can outperform traditional 3D convolutions, making our approach suitable for resource-constrained and edge deployment scenarios.

Limitations. The continuous VAD regression and multi-label classification remains challenging (average MAE $\approx$ 1.37), highlighting the limitations of mean-pooled representations in capturing fine-grained temporal dynamics. Temporal averaging discards phase-specific kinematic patterns critical for subtle arousal and dominance variations. The absence of fully end-to-end temporal models (e.g., recurrent or Transformer-based architectures) limits performance on sequence-dependent affective signals. Future work should explore dynamic feature extraction, long-range temporal modeling, and explicit motion priors to close this gap.

Conclusion

We presented the multi-task learning study on the ABEE dataset, establishing reproducible baselines for multi-label emotion classification and continuous VAD regression from body-only video. Our analysis reveals that Swin-Tiny visual features provide a strong single-modality signal, achieving competitive classification and VAD performance, while skeleton-based representations are similarly informative, highlighting the value of posture and motion cues. Modality fusion enhances regression accuracy, demonstrating complementary strengths. We present a dimensionally-consistent regularization that aligns predicted emotion probabilities with empirical VAD centroids. When applied to a lightweight CADResNet on fused Swin-Tiny and skeleton features, this approach achieves competitive micro F1 (0.321) and the lowest average VAD MAE (1.374) with only 1.043M parameters, substantially outperforming heavier convolutional baselines in both efficiency and predictive performance. Future work would explore dynamic feature extraction, long-range temporal modeling, and explicit motion priors to close this intensity estimation gap.

Table 2: Ablation study on the ABEE validation split across visual backbones and skeleton features. All models are trained for 100 epochs under identical conditions. Micro F1 and MAE for Valence/ Arousal/ Dominance/ Average are reported. Bold indicates the best value per modality group.

Features	Model	F1 ↑	IoU ↑	MAE V/A/D/Avg ↓
ResNet18	MLP	0.301	0.169	1.644 / 1.557 / 1.452 / 1.551
	1D-CNN	0.319	0.176	1.647 / 1.335 / 1.264 / 1.415
	3D-CNN	**0.322**	**0.183**	**1.586 / 1.314 / 1.248 / 1.383**
Swin-Tiny	MLP	0.312	0.183	1.627 / 1.500 / 1.385 / 1.504
	1D-CNN	**0.320**	0.185	1.583 / 1.355 / 1.293 / 1.410
	3D-CNN	0.314	**0.187**	**1.530 / 1.309 / 1.239 / 1.359**
ConvNeXt-Tiny	MLP	**0.324**	**0.192**	1.569 / 1.557 / 1.395 / 1.507
	1D-CNN	0.320	0.187	1.564 / 1.310 / 1.255 / 1.377
	3D-CNN	0.322	0.191	**1.536 / 1.329 / 1.246 / 1.370**
Skeleton	MLP	0.315	0.165	1.583 / 1.370 / 1.270 / 1.408
	1D-CNN	0.317	0.179	1.553 / 1.336 / 1.251 / 1.380
	3D-CNN	**0.322**	**0.187**	**1.537 / 1.312 / 1.246 / 1.365**

References

2025. Annotated Bodily Expressed Emotion (ABEE) Dataset. BEEU Challenge 2025. Dataset and annotations available from the BEEU Challenge workshop. Release date: October 10, 2025.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Dael, N.; Mortillaro, M.; and Scherer, K. R. 2012. Emotion expression in body action and posture. *Emotion*, 12(5): 1085.

De Meijer, M. 1989. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal behavior*, 13(4): 247–268.

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. YOLO by Ultralytics.

Kleinsmith, A.; and Bianchi-Berthouze, N. 2012. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1): 15–33.

Li, S.; and Deng, W. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3): 1195–1215.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, Y.; Ye, J.; Adams Jr, R. B.; Li, J.; Newman, M. G.; and Wang, J. Z. 2020. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *International journal of computer vision*, 128(1): 1–25.

Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current psychology*, 14(4): 261–292.

Picard, R. W. 2000. *Affective computing*. MIT press.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Vinciarelli, A.; Esposito, A.; André, E.; Bonin, F.; Chetouani, M.; Cohn, J. F.; Cristani, M.; Fuhrmann, F.; Gilmartin, E.; Hammal, Z.; et al. 2015. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7(4): 397–413.

Wightman, R. 2019. Pytorch image models (timm).

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zen, G.; Porzi, L.; Sangineto, E.; Ricci, E.; and Sebe, N. 2016. Learning personalized models for facial expression analysis and gesture recognition. *IEEE transactions on multimedia*, 18(4): 775–788.

Zhao, G.; and Pietikainen, M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6): 915–928.