Efficient Pretrained Representations for Video-based Fine-Grained Violence Detection, Action Unit Detection, and Valence-Arousal Estimation

Andrey V. Savchenko[1, 2] **Kseniia Tsypliakova**[1, 3]

[1]Sber AI Lab, Moscow, Russia and [2]ITMO University, Saint Petersburg, Russiaand [3]HSE University, Moscow, Russia

Abstract

Understanding human affect in unconstrained videos remains challenging due to subtle dynamics and domain shifts. This work presents our approach to three ABAW-10 tasks: fine-grained violence detection (VD), action unit (AU) detection, and valence-arousal (VA) estimation. We show that strong ImageNet-pretrained 2D encoders with lightweight temporal modeling outperform complex video architectures (macro F1 0.783 on VD validation). For facial analysis, efficient EmotiEffLib-based pipelines with calibration techniques achieve AU F1 0.547 and VA CCC 0.562 on Aff-Wild2. Our team ranked **1st in VD** and **2nd in AU** challenges.

Introduction & Motivation

- Affective behavior analysis in-the-wild requires robustness to recording conditions, temporal cues, and class imbalance.
- The ABAW competition provides large-scale datasets (Aff-Wild2, DVD) for benchmarking AU, VA, and violence detection.
- **Key idea**: Simple, well-calibrated models on strong pretrained features achieve competitive performance without heavy architectures.
- We switched from TensorFlow to PyTorch for reproducibility.

Valence-Arousal (VA) Estimation

Setup:

- **Dataset:** Aff-Wild2 VA challenge data (356 train / 76 val videos with 1,653,710 and 376,323 frames) provided by ABAW-10 organizers.
- **Labels:** it is required to predict valence V and arousal A, both in a range $[-1; 1]$ (multi-output regression task).

Approach:

- Pretrained MT-DDAMFN features (EmotiEffLib) + MLP regression head.
- Loss combines MSE and Concordance Correlation Coefficient (CCC) penalty.
- Post-processing: sliding-window smoothing (window size 5).

Loss (MSE + CCC):

$$\mathcal{L}_{VA} = (z_V - y_V)^2 + (z_A - y_A)^2 - \big(CCC(z_V, y_V) + CCC(z_A, y_A)\big)$$

Valence-Arousal (Validation Results)

Method	CCC_V	CCC_A	P_{VA}
Baseline ResNet-50 (Kollias et al., CVPRW 2025)	0.24	0.20	0.22
MAVEN (Ahire et al., CVPRW 2025)	0.307	0.305	0.3061
DDAMFN+LSTM (Cabacas-Maso et al., CVPRW 2025)	-	-	0.479
Best EfficientNet (Savchenko, CVPRW 2024)	0.490	0.596	0.543
Time-aware Gated Fusion (Lee et al., ICCVW 2025)	0.427	0.676	0.552
GRJCA (Rajasekhar et al., CVPRW 2025)	0.459	0.652	0.555
MultiScaleTCN (Yu et al., CVPRW 2025)	0.467	0.663	0.565
CLIP+TCN (Zhou et al., arxiv 2025)	0.562	0.612	0.587
wav2vec 2.0	-0.011	0.244	0.116
LibreFace, MLP	0.307	0.418	0.362
MT-DDAMFN, pretrained	0.413	0.229	0.321
MT-DDAMFN, MLP	0.469	0.538	0.503
MT-DDAMFN, MLP, smoothing	0.510	0.615	0.562

The final model outperforms the baseline by 34%.

Action Unit (AU) Detection

Setup:

- **Dataset:** Aff-Wild2 AU challenge data (295 train / 105 val videos with 1,356,694 and 445,836 frames) provided by ABAW-10 organizers.
- **Labels:** each frame is associated with 12 AUs (multi-label frame-level classification task). For each frame, a binary label is available for each AU.

Approach:

- Facial embeddings: EmotiEffNet-B0 (from EmotiEffLib) pretrained on AffectNet.
- Classifier: MLP (1 hidden layer, 12 sigmoid outputs) with weighted BCE loss.
- Calibration: per-AU threshold tuning (search [0.1,...,0.9]) + temporal smoothing.
- Ensemble with LightAutoML-based classifier for logits from pre-trained EmotiEffNet-B0

Loss (weighted BCE):

$$\mathcal{L}_{AU} = -\sum_{c=1}^{12}\big(w_c^{pos} y_c \log \sigma(z_c) + (1 - y_c)\log(1 - \sigma(z_c))\big)$$

Action Units (Validation Results)

Method	F1-score P_{AU}
Baseline VGGFACE (Kollias et al., CVPRW 2025)	39.0
CLIP (Lin et al, MIPR 2024)	43.0
DDAMFN+LSTM, threshold 0.5 (Cabacas-Maso et al, CVPRW 2025)	41.1
DDAMFN+LSTM, best thresholds (Cabacas-Maso et al, CVPRW 2025)	45.1
AUD-TGN (Yu et al, CVPRW 2024)	53.7
Best EfficientNet (Savchenko et al, CVPRW 2025)	54.5
ConvNeXt + Whisper+STN +TCN + Sliding Window (Yu et al, CVPRW 2025)	56.1
MAE+Transformer Fusion (Zhang et al, CVPRW 2024)	56.9
MAE ViT + TCN (Zhou et al, CVPRW 2024)	57.6
CLIP+TCN (Zhou et al, CVPRW 2025)	58.0
wav2vec 2.0	31.3
EmotiEffNet, logits, threshold 0.5	47.5
EmotiEffNet, logits, best thresholds	48.3
EmotiEffNet, logits, smoothing, threshold 0.5	48.6
EmotiEffNet, logits, smoothing, best thresholds	49.3
EmotiEffNet, embeddings, threshold 0.5	52.7
EmotiEffNet, embeddings, best thresholds	52.9
EmotiEffNet, embeddings, smoothing, threshold 0.5	54.0
EmotiEffNet, embeddings, smoothing, best thresholds	54.2
EmotiEffNet, logits + embeddings, threshold 0.5	53.7
EmotiEffNet, logits + embeddings, best thresholds	53.9
EmotiEffNet, logits + embeddings, smoothing, threshold 0.5	54.6
EmotiEffNet, logits + embeddings, smoothing, best thresholds	54.7

Each component contributes: best thresholds (+0.8), embeddings vs logits (+4.6), smoothing (+1.8). The final model outperforms the challenge baseline by 15.7%.

Fine-Grained Violence Detection (VD)

Setup:

- **Dataset:** DVD dataset (103 train / 17 val videos with 772K and 77K frames) provided by ABAW-10 organizers.
- **Labels:** for each frame, it is required to predict binary label - violent or non-violent (binary classification task).

Approach:

- 2D backbone: ConvNeXt-T (ImageNet-1K pretrained) extracts 768-d per-frame features.
- Temporal head: 5-layer dilated TCN (or BiLSTM) for sequence modeling.
- Training: weighted cross-entropy (positive weight 1.15), OneCycleLR, TrivialAugment.

Our VD Pipelines

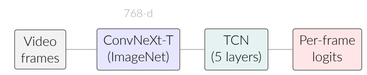

Figure 1. Our best single-stream VD pipeline: ConvNeXt-T + TCN.

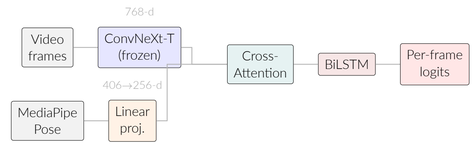

Figure 2. Multimodal RGB+Skeleton fusion with cross-attention and BiLSTM.

VD: Experimental Results (Validation)

Method	$F1_V$	$F1_{NV}$	Macro $F1$
Baseline			
ResNet-50 + BiLSTM (Kollias et al., ICCV 2025)	0.56	0.71	0.640
3D / video backbones			
SlowFast-R50	0.303	0.661	0.482
R3D-18 + Conv1d	0.186	0.797	0.491
S3D + Conv1d	0.268	0.781	0.524
R(2+1)D-18 + Conv1d	0.335	0.717	0.526
VideoMAE-Small	0.544	0.659	0.602
S3D + BiLSTM	0.462	0.752	0.607
R(2+1)D + BiLSTM	0.517	0.778	0.647
2D backbone (single-stream RGB)			
ResNet-18 + Conv1d	0.556	0.769	0.663
Video Swin-T	0.530	0.812	0.671
EfficientNet-B0 + BiLSTM	0.681	0.757	0.719
ConvNeXt-T + BiLSTM	0.745	0.820	0.782
ConvNeXt-T + TCN	0.738	0.828	**0.783**
Two-stream: RGB + Optical Flow			
TwoStr. concat + BiLSTM	0.658	0.801	0.730
TwoStr. gated + BiLSTM	0.669	0.800	0.735
Multi-modal: RGB + Skeleton			
ConvNeXt-T + Skel. (early)	0.692	0.810	0.751
Skel. attn + TCN	0.723	0.797	0.760
Skel. attn + BiLSTM	0.715	0.828	0.772

Comparison with State-of-the-Art (Test Sets)

Task	Baseline's Score	Our Score	Top Score (ABAW-10)
VD (macro F1)	0.504	0.5866	1st place
AU (F1)	36.5%	48.8%	51.0% (1st)
VA (P_{VA})	0.20	0.52	0.62 (1st)

Conclusion: We ranked first in the VD challenge and second in AU detection, demonstrating that well-calibrated models with strong pretrained features offer an excellent trade-off between accuracy and simplicity.

Key Takeaways & Design Principles

- **Strong pretrained representations** (ConvNeXt, EmotiEffNet) are more critical than complex spatiotemporal backbones.
- **Lightweight temporal modeling** (TCN, BiLSTM) suffices when frame-level features are rich.
- **Calibration matters**: per-class threshold tuning and temporal smoothing consistently improve AU.
- **Multimodality** (RGB + skeleton) gives moderate gains in VD but adds complexity.
- Simple pipelines are reproducible, efficient, and competitive for ABAW-style benchmarks.

Acknowledgments: Ministry of Economic Development of the Russian Federation (IGK 000000C313925P4C0002), agreement No. 139-15-2025-010.