UD-Net: Uncertainty-Guided Label Distribution Learning for Facial Depression Severity Estimation

Su Yu[3, 2], Zhenyu Yang[3, 2], Yu Zheng[1, 2], and Jing Wang[1, 2 *]

[1]*School of Computer Science and Engineering, Southeast University*, Nanjing, China
[2]*Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education*, China
[3]*Chien-Shiung Wu College, Southeast University*, Nanjing, China

***Abstract*—Facial depression severity estimation is an important problem in affective computing, but learning reliable predictors is difficult when annotations are inherently subjective and available training data are limited, as in AVEC 2014. Standard regression pipelines optimize point-wise losses and implicitly assume that the provided scores are noise-free, which overlooks the heteroscedastic label ambiguity in the target labels. Although Label Distribution Learning (LDL) alleviates this issue by learning distributions instead of single values, most LDL-based methods adopt a constant distribution width for all samples, failing to reflect sample-dependent difficulty. In this work, we present UD-Net, a lightweight ambiguity-aware LDL framework that incorporates a novel Error-Driven Variance Rectification strategy. During training, the rectified variance serves as a self-calibrated, curriculum-style supervision signal, encouraging wider distributions for hard cases and sharper ones for easy cases. On the AVEC 2014 test set, UD-Net achieves an MAE of 6.88 and RMSE of 8.86. Despite using only a ResNet-18 backbone, it delivers competitive accuracy compared with heavier 3D-CNN approaches while significantly reducing computational cost.**

***Index Terms*—Depression estimation, label distribution learning, label ambiguity, efficient deep learning, curriculum learning**

I. Introduction

Depression is a prevalent and debilitating mental health disorder. Early diagnosis and continuous monitoring are crucial for effective treatment. While traditional clinical diagnosis relies heavily on self-report questionnaires, these methods are often subjective, resource-intensive, and susceptible to biases [1]. Consequently, automated depression estimation based on non-verbal behavioral cues, particularly facial expressions, has emerged as a vital research topic in affective computing [2], [3]. By leveraging computer vision, researchers aim to develop objective, non-invasive, and scalable screening tools [3].

Despite rapid progress in deep learning, estimating depression severity from facial images remains a formidable task, fundamentally distinct from standard regression problems. This difficulty stems from two primary issues: **label ambiguity** and **data scarcity**. Unlike objective tasks where the ground truth is absolute, depression severity scores are annotated by human experts based on subjective impressions. This process inevitably introduces annotation noise. Moreover, typical datasets like AVEC 2014 provide only a single global depression score for an entire video clip. Broadcasting this single label to every frame ignores the dynamic fluctuations of human expressions. This inherent noise—often characterized as *heteroscedastic label ambiguity*—makes the ground truth labels intrinsically ambiguous. Furthermore, benchmark datasets [2], [4] are relatively small in scale, posing a high risk of overfitting.

Existing approaches predominantly formulate depression estimation as a standard point-to-point regression problem [3], assuming target labels are precise and noise-free. Minimizing the Mean Squared Error (MSE) on such noisy data often forces the model to be overconfident. To mitigate this, **Label Distribution Learning (LDL)** [5] was introduced. By transforming single label values into probability distributions, LDL allows the model to predict a range of adjacent scores, providing richer, "soft" supervision signals robust to noise.

However, a critical limitation persists in most existing LDL-based methods: they typically assume a **fixed variance** for the target distributions across all samples. This static strategy implicitly assumes *homoscedasticity*—that every sample is equally difficult to learn. This overlooks the **heteroscedastic nature** of real-world data, where prediction difficulty varies due to occlusion, illumination changes, or ambiguous expressions (Fig. 1(a)).

To address this, we propose **UD-Net**, a lightweight **Ambiguity-Aware Label Distribution Learning** framework featuring an **Error-Driven Variance Rectification** mechanism (Fig. 1(b)). During training, we utilize prediction error as a dynamic proxy for sample difficulty. Larger errors trigger wider target distributions to express higher uncertainty, relaxing penalties for noisy cases. This naturally forms a difficulty-aware curriculum learning signal [6], enabling more effective learning from noisy data.

The main contributions are summarized as follows:

- We propose an ambiguity-aware LDL framework that dynamically models label noise, addressing the limitations of fixed-variance LDL methods in heteroscedastic affective data.

This research was supported by the China Postdoctoral Science Foundation (2022M720028, 2025T180432), the Fundamental Research Funds for the Central Universities (2242025K30024), and the National Undergraduate Training Programs for Innovation.

*Corresponding author: Jing Wang (wangjing91@seu.edu.cn).

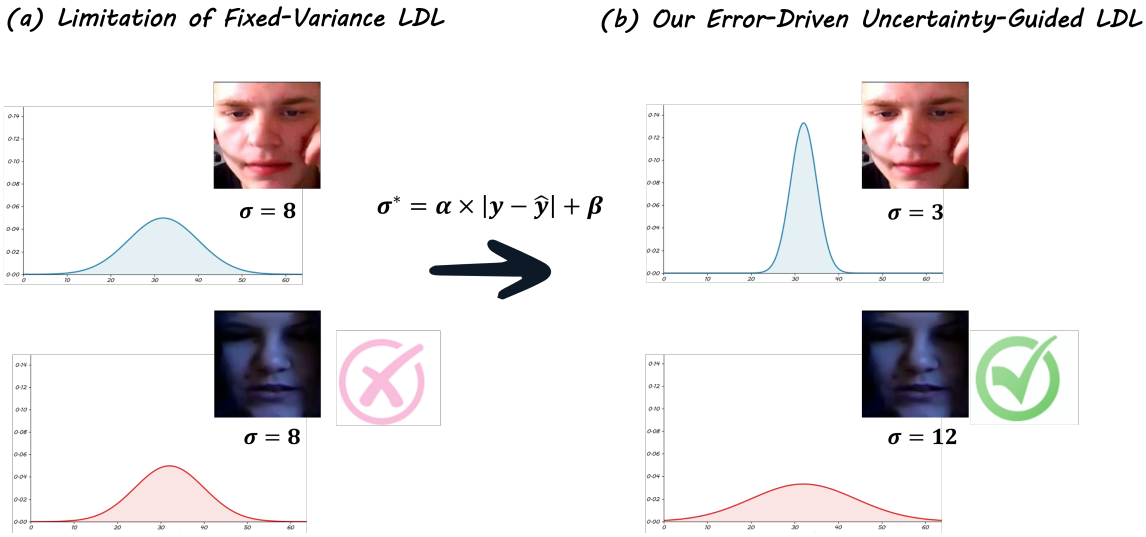

Fig. 1. **Conceptual comparison of our Ambiguity-Aware LDL versus conventional Fixed-Variance LDL. (a) Left:** Conventional LDL methods apply a "one-size-fits-all" strategy, assigning a fixed-width target distribution (σ) regardless of sample difficulty. **(b) Right:** Our proposed UD-Net introduces an **Error-Driven Variance Rectification** mechanism. The target variance (σ^*) is dynamically calibrated based on the model's prediction error ($|y-\hat{y}|$), assigning wider distributions to hard samples and sharper ones to easy samples, forming an adaptive, curriculum-style learning signal.

- We introduce a novel Error-Driven Variance Rectification strategy, leveraging training errors to adaptively adjust target distribution width, providing self-calibrated curriculum guidance without extra annotations.
- Extensively evaluated on AVEC 2014, UD-Net achieves competitive performance against computationally expensive 3D-CNN models using a lightweight ResNet-18, offering a superior efficiency-accuracy trade-off.

II. Related Work

A. Visual Depression Analysis

The field of automated depression analysis has evolved significantly with benchmark datasets such as the AVEC series [2], [4], [7], [8]. Early visual approaches primarily relied on hand-crafted features [2] or foundational CNNs. More recently, the focus has shifted towards heavier spatial-temporal architectures [9], [10], including End-to-End 3D-CNNs [11] and Transformers, aiming to capture long-range facial behavior dynamics. For a comprehensive recent review, readers can refer to [12].

Critique: While these sophisticated architectures have pushed performance boundaries, they often come at the cost of massive parameter counts. Moreover, existing visual methods focus heavily on *feature space engineering* while treating the regression target as a rigid point value. In contrast, our work focuses on *target representation optimization*. We aim to achieve competitive performance using a lightweight backbone (ResNet-18 [13]) by refining the supervision signal itself.

B. Label Distribution Learning (LDL)

Label Distribution Learning (LDL) [5] is designed to handle label ambiguity by assigning a degree of description to each label. It has proven highly effective in tasks where ground truth is inherently subjective, such as facial age estimation [14] and attractiveness computation [15]. In depression estimation, most prior LDL studies employ a static strategy, assigning a fixed standard deviation to all samples. Our method advances this paradigm by making the distribution width adaptive and error-driven, providing a robust solution to heteroscedastic noise.

C. Uncertainty Estimation

Prior works have explored uncertainty estimation through sophisticated probabilistic frameworks, such as Bayesian Neural Networks (BNN) or sampling-based methods like Monte Carlo Dropout [16]. In this work, we do not aim to explicitly disentangle uncertainty types (e.g., epistemic vs. aleatoric) using computationally expensive probabilistic ensembles. Instead, we focus on modeling sample-dependent label ambiguity in a deterministic, highly efficient manner suitable for real-time applications.

III. Methodology

A. Data Preprocessing

Given raw video frames from AVEC 2014 [17], we utilize MTCNN [18] to detect and align faces. Cropped images are resized to 224×224 pixels. Standard data augmentations (random horizontal flip, rotation $\pm 10°$, color jittering) are applied during training to enhance robustness. Input tensors are normalized using the ImageNet mean and standard deviation.

B. Framework Overview

The overall architecture of UD-Net is illustrated in Fig. 2. We adopt a lightweight ResNet-18 [13] as the visual backbone. To prevent overfitting on the small-scale AVEC 2014 dataset, we employ a **partial freezing strategy**. The initial layers (*conv1* to *layer2*) are frozen to retain generic texture

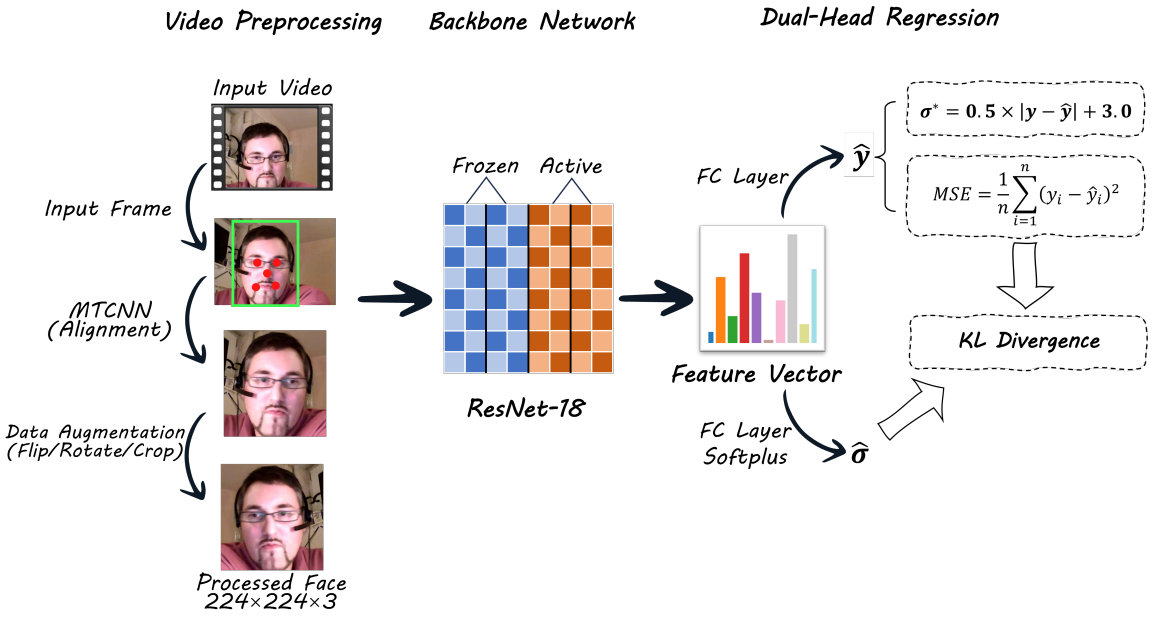

Fig. 2. The overview of UD-Net. Extracting deep features via a partial-freezing ResNet-18, the network utilizes dual heads for score prediction ($\hat{y}$) and variance estimation ($\hat{\sigma}$). Training is supervised by an Error-Driven Variance Rectification mechanism.

features learned from ImageNet, while high-level semantic layers (*layer3* and *layer4*) and fully connected layers are fine-tuned.

The backbone maps an input image to a feature vector, feeding into two parallel fully connected heads: 1) **Prediction Head**: Outputs the estimated depression score $\hat{y}$. 2) **Variance Head**: Outputs the estimated distribution width $\hat{\sigma}$.

C. Label Distribution Construction

Following the LDL paradigm, we discretize the continuous depression score range into a label space of K intervals, denoted as $\mathcal{K} = \{0, 1, \ldots, K-1\}$. Here, $K = 64$ to cover the BDI-II score range. Given ground truth y and a target variance σ^*, the target probability distribution $D_y(k)$ over $k \in \mathcal{K}$ is:

$$D_y(k) = \frac{1}{Z} \exp\left(-\frac{(k-y)^2}{2(\sigma^*)^2}\right) \tag{1}$$

where $Z = \sum_{j \in \mathcal{K}} \exp\left(-\frac{(j-y)^2}{2(\sigma^*)^2}\right)$ is the normalization constant.

D. Error-Driven Variance Rectification

A key innovation of UD-Net is the adaptive determination of variance.

a) Network Output Constraint: To ensure physical validity, the predicted variance $\hat{\sigma}$ is constrained using a Softplus activation followed by clamping:

$$\hat{\sigma} = \text{Clamp}\left(\text{Softplus}(W_\sigma x + b_\sigma), \sigma_{min}, \sigma_{max}\right) \tag{2}$$

Empirically, we set $\sigma_{min} = 3.0$ and $\sigma_{max} = 15.0$.

b) The Self-Calibrated Target Signal: During training, we construct a dynamic "Target Sigma" σ^* to guide the network:

$$\sigma^* = \sigma_{base} + \gamma \cdot |y - \text{sg}(\hat{y})| \tag{3}$$

where $\sigma_{base} = 3.0$ represents intrinsic ambiguity, $\gamma = 0.5$ is a sensitivity coefficient, and $\text{sg}(\cdot)$ denotes the stop-gradient operator.

In subjectively annotated datasets such as AVEC 2014, prediction errors are often associated with ambiguous visual cues or inherent annotation noise. Therefore, we utilize the prediction error as a **heuristic proxy** to adaptively model sample-dependent label ambiguity. This formulation dynamically rectifies σ^*, offering an efficient deterministic alternative to complex probabilistic modeling.

Gradient Isolation: By explicitly detaching $\hat{y}$ from the computational graph ($\text{sg}(\cdot)$) when calculating σ^*, we prevent a *degenerate solution* where the model deliberately degrades prediction accuracy simply to inflate σ^* and minimize the loss trivially.

E. Objective Function

The total objective function $\mathcal{L}$ combines Distribution Matching and Variance Regression:

$$\mathcal{L} = \mathcal{L}_{KL} + \lambda \mathcal{L}_{var} \tag{4}$$

where $\mathcal{L}_{KL} = \sum_{k \in \mathcal{K}} D_y(k; \sigma^*) \log \frac{D_y(k;\sigma^*)}{P(k;\hat{y},\hat{\sigma})+\epsilon}$, P is the network's softmax distribution, and $\mathcal{L}_{var} = ||\hat{\sigma} - \sigma^*||^2$. We set $\lambda = 0.2$.

F. Theoretical Interpretation of Variance Rectification

Assuming a simplified case where the network approximates the Gaussian mean directly, the gradient magnitude of the KL loss with respect to $\hat{y}$ is inversely proportional to the variance:

$$\left|\frac{\partial \mathcal{L}_{KL}}{\partial \hat{y}}\right| \propto \frac{1}{(\sigma^*)^2}|y - \hat{y}| \tag{5}$$

For easy samples (small error), σ^* remains small, resulting in a large gradient that forces precise fitting. For hard/noisy samples (large error), σ^* increases, dampening the gradient and preventing overfitting to outliers (a "soft" noise rejection mechanism).

Furthermore, while this gradient-modulating effect conceptually resembles adaptive sample re-weighting (e.g., Focal Loss), our fundamental distinction lies in operating within the **Target Representation Space**. Instead of rigidly scaling the loss scalar, our ambiguity-aware LDL explicitly broadens the target distribution. This not only imposes a soft penalty on difficult samples but also preserves the continuous semantic correlation among adjacent scores, thereby constructing a smoother and more robust optimization manifold.

G. Inference Phase

During testing, the ground truth y is unavailable. The network directly yields the distribution $P(k)$. The final score $\hat{y}_{final}$ is calculated as the expectation: $\hat{y}_{final} = \sum_{k=0}^{K-1} k \cdot P(k; \hat{y}, \hat{\sigma})$.

IV. Experiments

A. Experimental Setup

Dataset and Protocol: We evaluate our method on the rigorous AVEC 2014 benchmark [2], widely adopted in recent state-of-the-art visual depression studies. We utilize the *Training* set for training, the *Development* set for validation, and report performance on the *Test* set. We apply a uniform sampling strategy, extracting approximately 100 frames per video.

Implementation Details: The model is optimized using the Adam optimizer [19] with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. We train the network for 50 epochs with a batch size of 32, utilizing an early stopping strategy based on the Development set performance to prevent overfitting.

B. Comparison with State-of-the-Art

Table I presents the comparison. To ensure statistical significance, the reported metrics for our UD-Net are averaged over three independent runs with different random seeds.

Comparison with Visual-Only Methods: UD-Net **significantly outperforms** existing spatial-temporal architectures, including the Two-Stream Network [20] and TDCA-Net [21]. This validates that refining the *target representation* via our ambiguity-aware LDL contributes more to prediction accuracy than merely stacking complex temporal modules. Furthermore, our method remains highly competitive with the heavy End-to-end 3D-CNN [11], operating with significantly lower computational cost.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON AVEC 2014 TEST SET

Method	Modal	Backbone	MAE ↓	RMSE ↓
Multimodal Methods (For Reference)				
DepressionMIGNN [22]	A+V	GNN+MIL	5.25	6.75
MDDR [23]	A+V+M	Dual-Attn	5.75	7.12
Single-Modal Methods				
LQGDNet [9]	V	Hybrid	6.08	7.84
3D-CNN+attention [24]	V	3D-CNN	6.78	8.42
Two-Stream Net [20]	V	CNN+RNN	6.86	8.78
TDCA-Net [21]	A	Time-CNN	7.08	8.90
DCNN Baseline*	V	ResNet-50	7.47	9.82
UD-Net (Ours)**	**V**	**ResNet-18**	**6.88±0.06**	**8.86±0.02**

* Self-implemented baseline. ** Results are reported as mean ± std over 3 runs.

C. Ablation Study on Modules

Table II validates our module design. Fixed-LDL (B) improves over baseline regression (A). If the network predicts variance but targets a fixed $\sigma^* = 3.0$ (C), performance degrades, proving that simply adding a variance head without proper supervision introduces noise. Finally, UD-Net (E) provides the best performance, confirming the error-driven curriculum signal is essential.

TABLE II
ABLATION STUDY OF DIFFERENT MODULES

Var	Method Description	Rectification	MAE	RMSE
A	ResNet-18 Regression	-	7.33	9.56
B	Fixed LDL ($\sigma = 3.0$)	No	7.09	9.31
C	Adaptive Head (No Guide)	No	7.15	9.40
D	UD-Net (Full Fine-tuning)	Yes	7.27	9.15
E	**UD-Net (Ours)**	**Yes**	**6.88**	**8.86**

D. Impact of Partial Freezing Strategy

An experiment with **full fine-tuning** (Table II, Row D) resulted in performance degradation (MAE increases to 7.27). This confirms that on the small-scale AVEC dataset, updating the entire feature extractor leads to overfitting. Our partial freezing strategy acts as a necessary regularizer.

E. Sensitivity of γ

When $\gamma = 0$ in Eq. 3, the method degenerates to Fixed LDL (MAE 7.09). When $\gamma = 0.5$, MAE is minimized at 6.88. When $\gamma = 1.0$, the target distribution becomes too flat for erroneous samples, causing under-fitting (MAE 6.95). Thus, $\gamma = 0.5$ is optimal.

F. Evaluation of CCC Indicator

For completeness regarding modern affective computing metrics, our UD-Net achieves a **CCC of 0.5709** [25]. This indicates a moderate-to-strong positive correlation between the predicted and ground truth trajectories.

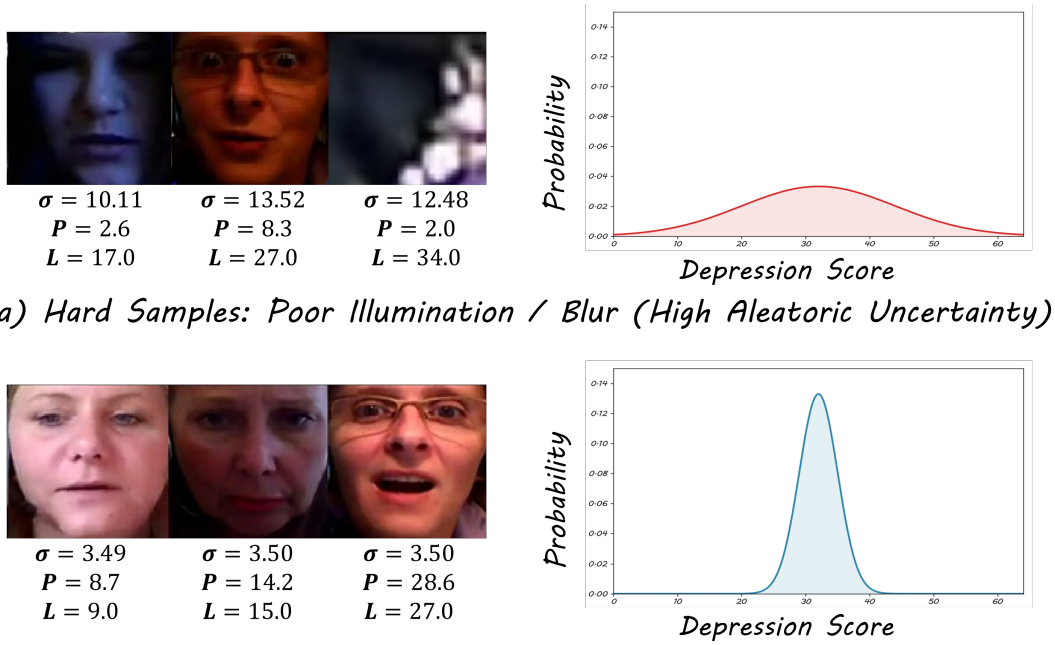

Fig. 3. Visualization of label ambiguity modeling. **(a)** For hard samples with poor illumination or blur, UD-Net predicts wide distributions (high $\hat{\sigma}$). **(b)** For easy samples with clear facial features, the model outputs sharp distributions (low $\hat{\sigma}$). The text overlays indicate the Ground Truth Label (L) and the Model Prediction (P), with the red values highlighting the predicted variance $\hat{\sigma}$.

G. Error-Variance Correlation Analysis

To quantitatively validate our assertion that the predicted variance effectively captures sample ambiguity, we analyzed the correlation between the absolute prediction error $|y - \hat{y}|$ and the estimated distribution width $\hat{\sigma}$ on the unseen Test set.

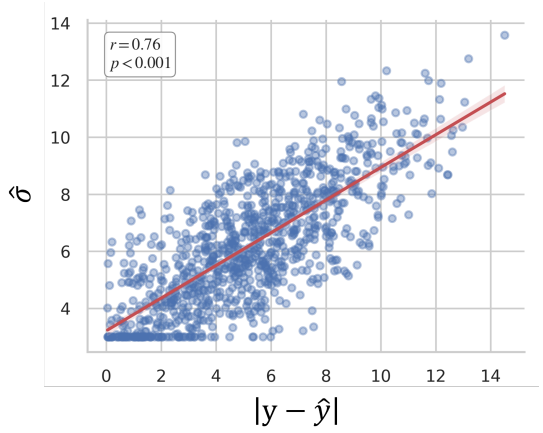

Fig. 4. Correlation between absolute prediction error and the predicted variance $\hat{\sigma}$ on the Test set.

As depicted in Fig. 4, a strong positive correlation is observed. This empirical evidence demonstrates that our dual-head architecture successfully learns to calibrate its output distribution, assigning appropriately wider variances (lower confidence) to samples that inherently deviate from the mean trajectory due to annotation noise or visual occlusion.

H. Qualitative Analysis

We visualize the predicted label distributions in Fig. 3. As shown in Fig. 3(a), "Hard" samples suffer from severe illumination issues or blur. In these cases, UD-Net adaptively outputs a wider distribution ($\hat{\sigma} \approx 9.0$–12.0). This indicates the model is "aware" of the insufficient visual information. Conversely, Fig. 3(b) displays "Easy" samples with clearly visible facial features, where the model predicts a sharp distribution ($\hat{\sigma} \approx 3.5$), reflecting high confidence.

V. Discussion

A. Efficiency vs. Performance Trade-off

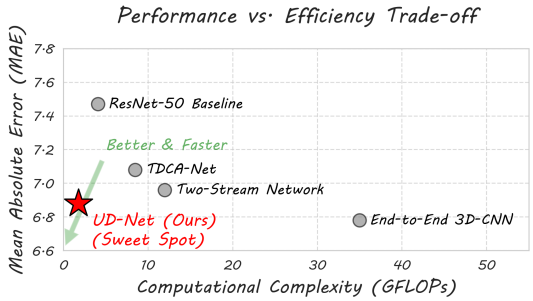

Fig. 5. Efficiency Analysis (GFLOPs vs. MAE). UD-Net achieves a superior trade-off compared to heavy 3D-CNNs.

As illustrated in Fig. 5, heavy 3D-CNNs require massive computational resources (> 30 GFLOPs). UD-Net challenges this trend by achieving comparable accuracy (MAE 6.88)

using a lightweight ResNet-18 backbone (≈ 1.8 GFLOPs). By transforming noisy single-scalar labels into rich distributions, UD-Net effectively "cleans" the data dynamically. Optimizing the *target space* yields a sustainable solution for resource-constrained edge devices.

B. Practical Usage for Trustworthy AI

Unlike "black-box" regression, UD-Net provides an interpretable confidence measure via $\hat{\sigma}$. This serves as a cornerstone for **Trustworthy AI** in healthcare: high predicted ambiguity ($\hat{\sigma} > \tau$) can trigger a "refer-to-expert" protocol, facilitating a safe Human-in-the-Loop workflow. This deterministic approach bridges the gap between deep learning and the rigorous safety requirements of clinical screening.

C. Limitations and Future Work

Despite its efficacy, UD-Net relies on static frames, potentially missing long-term behavioral dynamics. Additionally, extreme occlusion might inflate σ^* due to input degradation rather than true label ambiguity. Future work will focus on: 1) **Temporal Extension**: Integrating lightweight temporal modules (e.g., TCNs) to smooth frame-level predictions; 2) **Multimodal Fusion**: Using UD-Net alongside audio models; and 3) **Stabilized Rectification**: Exploring moving-average error signals to mitigate training instability.

VI. Conclusion

We presented UD-Net, an efficient visual depression estimator. The core innovation lies in the Error-Driven Variance Rectification, which transforms training errors into a dynamic curriculum signal, enabling the model to adaptively calibrate its target distributions. Extensive experiments demonstrate that UD-Net yields performance comparable to complex 3D architectures while requiring a fraction of the computation, highlighting that optimizing target representation is a highly resource-efficient pathway.

References

[1] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current Opinion in Psychology*, vol. 4, pp. 75–79, 2015.

[2] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*. Association for Computing Machinery, 2014, pp. 3–10.

[3] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang, X. Pan, and W. Dang, "Deep learning for depression recognition with audiovisual cues: A review," *Information Fusion*, vol. 80, pp. 56–86, 2022.

[4] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13)*. ACM Press, 2013, pp. 3–10.

[5] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. Association for Computing Machinery, 2009, pp. 60–67.

[7] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, emotion, and physiologic state recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[8] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "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*. ACM, 2019, pp. 3–12.

[9] Y. Shang, Y. Pan, X. Jiang, Z. Shao, G. Guo, T. Liu, and H. Ding, "Lqgdnet: A local quaternion and global deep network for facial depression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2557–2563, 2023.

[10] Y. Pan, Y. Shang, T. Liu, Z. Shao, G. Guo, H. Ding, and Q. Hu, "Spatial–temporal attention network for depression recognition from facial videos," *Expert Systems with Applications*, vol. 237, p. 121410, 2024.

[11] W. C. de Melo, E. Granger, and A. Hadid, "Depression detection based on deep distribution learning," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4544–4548.

[12] X. Cao *et al.*, "Deep learning-based depression recognition through facial expression: A systematic review," *Neurocomputing*, vol. 627, no. 11, p. 129605, 2025.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[15] S. Liu, B. Li, Y.-Y. Fan, Z. Guo, and A. Samal, "Facial attractiveness computation by label distribution learning with deep CNN and geometric features," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 1344–1349.

[16] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5574–5584.

[17] J. Gratch, R. Artstein, G. Lucas, D. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2014, pp. 3123–3128.

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[20] M. A. Uddin, J. B. Joolee, and Y. K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 864–870, 2022.

[21] C. Cai, M. Niu, B. Liu, J. Tao, and X. Liu, "Tdca-net: Time-domain channel attention network for depression detection," in *INTERSPEECH 2021*, 2021, pp. 2511–2515.

[22] S. Zhao, Y. Zhang, Y. Su, K. Su, J. Liu, T. Wang, and S. Yu, "Depressionmignn: A multiple-instance learning-based depression detection model with graph neural networks," *Sensors*, vol. 25, no. 14, p. 4520, 2025.

[23] W. Zhang, E. Zhu, J. Chen, and Y. Li, "Mddr: Multi-modal dual-attention aggregation for depression recognition," in *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, 2024, pp. 321–329.

[24] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 10 140–10 156, 2022.

[25] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.