2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-5142-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIPR62202.2024.00103

Robust Light-Weight Facial Affective Behavior Recognition with CLIP

Li Lin[1], Sarah Papabathini[1], Xin Wang[2], Shu Hu[1]*
[1]Purdue University, {lin1785, Slpapaba, hu968}@purdue.edu
[2]University at Albany, State University of New York xwang56@albany.edu

***Abstract*—Human affective behavior analysis aims to delve into human expressions and behaviors to deepen our understanding of human emotions. Basic expression categories (EXPR) and Action Units (AUs) are two essential components in this analysis, which categorize emotions and break down facial movements into elemental units, respectively. Despite advancements, existing approaches in expression classification and AU detection often necessitate complex models and substantial computational resources, limiting their applicability in everyday settings. In this work, we introduce the first lightweight framework adept at efficiently tackling both expression classification and AU detection. This framework employs a frozen CLIP image encoder alongside a trainable multilayer perceptron (MLP), enhanced with Conditional Value at Risk (CVaR) for robustness and a loss landscape flattening strategy for improved generalization. Experimental results on the Aff-wild2 dataset demonstrate superior performance in comparison to the baseline while maintaining minimal computational demands, offering a practical solution for affective behavior analysis. The code is available at https://github.com/Purdue-M2/Affective_Behavior_Analysis_M2_PURDUE.**

***Index Terms*—Robust, Expression Classification, Action Unit Detection**

I. Introduction

Basic expression categories (EXPR) and Action Units (AUs) are key approaches in affective behavior analysis, a field that is essential for enhancing human-computer interaction by delving into the nuances of human emotions. This exploration is becoming increasingly important for creating more intuitive human-computer interactions. Basic expression categories divide expressions into a limited number of groups according to the emotion categories, *e.g.*, happiness, sadness, etc. AU depicts the local regional movement of faces which can be used as the smallest unit to describe the expression [1].

The sixth Competition on Affective Behavior Analysis in-the-wild (ABAW6) [2] is organized with the objective of addressing the challenges associated with human affective behavior analysis. It makes great efforts to construct large-scale multi-modal video datasets Aff-wild and Aff-wild2 [3]–[14]. Introducing these datasets has advanced the field of facial expression analysis in the wild and accelerated the practical implementation of related industries.

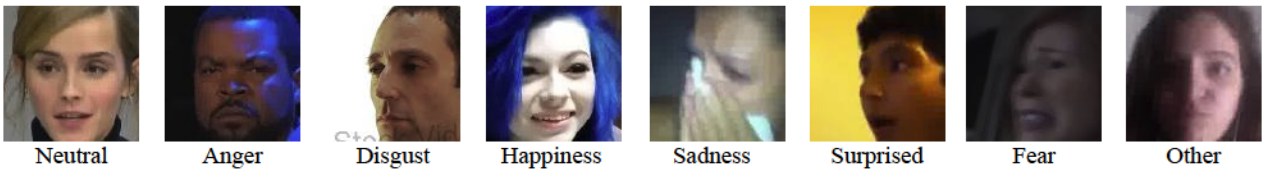
Deep learning has emerged as a pivotal technology in affective behavior analysis, demonstrating remarkable success in enhancing the detection and classification performance. By leveraging complex neural network architectures, deep learning models can automatically learn from vast amounts of face video data, including audio and images, to identify the expression patterns and subtle facial muscle movements. Specifically, in *Expression Classification*, Some works [15], [16] have applied extensive fine-tuning to large-scale deep learning models like the Masked Autoencoder (MAE) [17] and Swin transformer [18] for expression recognition. Although these methods have delivered promising outcomes, they typically require significant computational resources. Other works [19], [20] used pre-trained models, but their framework is complex for downstream tasks (*e.g.*, [19] trained a Transformer [21]). Some of the above approaches can also be applied to *Action Unit Detection*, and there are methods [22], [23] specifically designed for AU. Those works use extra annotations related to the face landmarks, which may not be practical for real-world scenarios lacking such detailed annotations.

In this work, we proposed the first lightweight, straightforward framework for expression classification and action unit detection. This framework comprises three modules: frozen Contrastive Language-Image Pre-training (CLIP) [24] ViT as feature extractor, trainable 3-layer multilayer perception (MLP), and optimization. We utilize the CLIP ViT L/14 [25] as an image encoder to capture high-level representations of face images, primarily because it has been trained on an expansive dataset comprising 400M image-text pairs. This extensive training enables the encoder to develop a nuanced understanding of facial features and expressions. The extracted features are processed through the 3-layer MLP, specifically trained for the given task. To bolster the model's precision, particularly in difficult cases, we integrate Conditional Value at Risk (CVaR) into the respective loss functions: cross-entropy for expression classification and binary cross-entropy for action unit detection. The optimization component is designed to smooth the loss landscape, improving the model's generalizability and performance. Our contributions are summarized as follows:

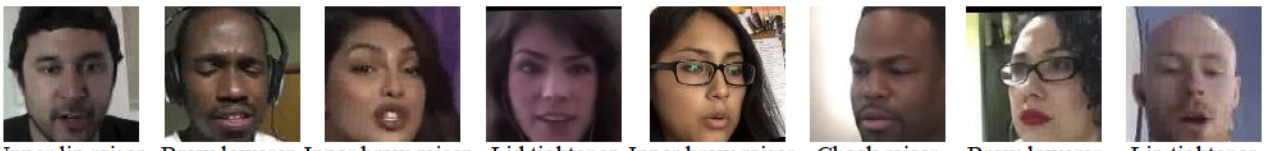
1) We propose the first lightweight efficient framework suitable for expression classification and action unit detection.
2) We incorporate CVaR into the loss functions, improving the accuracy and reliability of predictions, especially in challenging scenarios for both tasks.
3) Our method outperforms the baseline in both tasks, as demonstrated in experiments on the Aff-wild2 dataset.

*Corresponding author

2770-4319/24/$31.00 ©2024 IEEE
DOI 10.1109/MIPR62202.2024.00103



(a) Expression Classification



(b) Action Unit Detection

Fig. 1: *Displayed samples from Aff-wild2 dataset. (a) Expression Classification is a multi-class task, which has 8 categories in total. (b) Action Unit Detection is a multi-label task, each image contains annotations in terms of 12 AUs.*

II. Method

Our proposed method can be applied to both *Expression Classification Challenge* and *Action Unit Detection Challenge* [2]–[14]. The adaptability of our model lies in the configuration of the neurons in the final output layer. Specifically, for the Expression Classification Challenge, the last layer is designed with 8 neurons (8 expressions) with a softmax function. Conversely, for the Action Unit Detection Challenge, this layer is expanded to contain 12 neurons (12 action units) with a sigmoid function. Below, we describe the commonalities that are shared in these two tasks.

Feature Space Modeling. We propose a simple procedure based on features extracted from the image encoder of CLIP ViT L/14 [25]. CLIP ViT is trained on an extraordinarily large dataset of 400M image-text pairs, so the high-level feature extracted from it is sufficient exposure to the visual world. Additionally, since ViT L/14 has a smaller starting patch size of 14×14 (compared to other ViT variants), we believe it can also aid in modeling the low-level image details (*i.e.*, texture and edge). Given a dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ with size n, where X_i is the i-th sample and Y_i is the i-th sample label . Feed CLIP visual encoder with dataset $\mathcal{D}$, and use its final layer to map the training data to their feature representations (of 768 dimensions). We get the resulting feature bank $\mathcal{C} = \{(F_i, Y_i)\}_{i=1}^n$ and further use the feature bank, which is our training set, to design an MLP classifier.

MLP Classifier. After constructing the feature bank, we use those feature embeddings to train our classifier, a straightforward 3-layer Multilayer Perceptron (MLP). To foster a stable learning process and enhance the model's ability to generalize, we incorporate batch normalization after each linear transformation. This is followed by a ReLU activation, allowing the model to effectively capture intricate data patterns. To further combat the risk of overfitting, a dropout layer is included following the activation function.

Objective Function. To obtain a robust model, we apply a distributionally robust optimization (DRO) technique called *Conditional Value-at-Risk* (CVaR) [26]–[33]. By integrating CVaR into the cross-entropy (EC) loss (EC for expression classification, BEC for action unit detection), the model is encouraged to pay more attention to the riskiest predictions. For example, in expression classification, this methodology emphasizes prioritizing expressions such as fear, which not only are less prevalent in the dataset but also closely resemble other expressions, like surprise, making them challenging to distinguish accurately. Addressing these cases is vital because errors here could drastically affect the correct interpretation of emotional states. Similarly, within Action Unit Detection, this enhanced focus on uncertain predictions is pivotal. It ensures the model pays particular attention to subtle or complex facial muscle movements where inaccuracies could lead to significant misinterpretations.

To this end, in what follows, we assume that $\mathcal{C} = \{(F_i, Y_i)\}_{i=1}^n$ consists of i.i.d. samples from a joint distribution $\mathbb{P}$, F_i is the i-th data point's feature, and Y_i is the i-th point's label. Given some variant of minibatch gradient descent, in these two tasks, we are minimizing the empirical risk of the loss $\mathcal{L}_{avg}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, F_i, Y_i)$ for $\theta \in \Theta$, instead of minimizing the true unknown risk $\mathcal{R}_{avg}(\theta) = \mathbb{E}_{(F,Y)\sim\mathbb{P}}[\ell(\theta; F, Y)]$, where ℓ is the individual loss function (*e.g.*, cross-entropy in expression classification, binary cross-entropy in action unit detection) of the model, which has parameters θ for MLP.

However, the average loss is not robust to the imbalanced data, a common characteristic of the Aff-Wild2 dataset (*e.g.*, Neutral expression has 468,069 images, whereas Fear only contains 19,830 images). In addition, the training data distribution is usually inconsistent with the testing data distribution, which is called the domain shift problem that widely exists in real-world scenarios. For example, training data and testing data come from two different datasets. Therefore, we explore

a DRO technique CVaR for handling imbalanced data, which can be formulated as:

CVaR –

As a reminder, is the individual loss function where in expression classification is cross-entropy loss, and in action unit detection is binary cross-entropy loss.
is the hinge function, the conditional value at risk at level . As , we are concerned about minimizing the risk of 'hard' samples. In contrast, as , it becomes minimizing the $\mathcal{R}$. Inspired by [27], we can minimize a loss function that aims to minimize an upper bound on the worst-case risk by employing the CVaR. In practice, we minimize an empirical version of CVaR . This gives us the following optimization problem:

$\mathcal{L}$ — (1)

Suppose for a moment that we have obtained the optimal value of in (1), then the only training points that contribute to the loss are the 'hard' ones with a loss value greater than , whereas the 'easy' training points with low loss smaller than are ignored. To this end, a robust model for expression classification and action unit detection is obtained.

Optimization. Last, to further improve the model's generalization capability, we optimize the model by utilizing the sharpness-aware minimization (SAM) method [34] to flatten the loss landscape. As a reminder, the model's parameters are denoted as , flattening is attained by determining an optimal for perturbing to maximize the loss, formulated as:

$$\mathcal{L} \quad (\quad) \qquad \mathcal{L} \quad \texttt{sign} \quad \mathcal{L} \tag{2}$$

where is a hyperparameter that controls the perturbation magnitude. The approximation is obtained using first-order Taylor expansion assuming that is small. The final equation is obtained by solving a dual norm problem, where sign represents a sign function and $\mathcal{L}$ being the gradient of $\mathcal{L}$ with respect to . As a result, the model parameters are updated by solving the following problem:

$$\mathcal{L} \quad (\quad) \tag{3}$$

Perturbation along the gradient norm direction increases the loss value significantly and then makes the model more generalizable while classifying expression and detecting action units.

End-to-end Training. In practice, we first initialize the model parameters and then randomly select a mini-batch set from , performing the following steps for each iteration on :

Fix and use binary search to find the global optimum of since (1) is convex w.r.t. .
Fix , compute based on Eq. (2).
Update based on the gradient approximation for (3): $\mathcal{L}$, where is a learning rate.

III. Experiments

A. Experimental Settings

Datasets. Aff-Wild2 [2]–[14] is used in both Expression Classification Challenge and the Action Unit Detection Challenge. In the Expression Classification Challenge, the dataset consists of 548 videos from Aff-Wild2, annotated for the six basic expressions, neutral state, and an 'other' category representing non-basic expressions. Seven videos feature two subjects, both of whom are annotated. In total, annotations are provided for 2,624,160 frames from 437 subjects. The training, validation, and testing sets consist of 248, 70, and 230 videos, respectively. The dataset for the Action Unit Detection Challenge comprises 542 videos annotated for 12 AUs corresponding to the inner and outer brow raiser, the brow lowerer, the cheek raiser, the lid tightened, the upper lip raiser, the lip corner puller and depressor, the lip tightener and pressor, lips part and jaw drop. The training, validation, and testing sets contain 295, 105, and 142 videos, respectively.

Evaluation Metrics. The performance measure is the average F1 Score. In the Expression Classification Challenge, the average F1 Score across ('macro' F1 score) is 8 categories; in Action Unit Detection, it is across 12 AUs.

Baseline Methods. For our research, we compare our methods with the standard baseline provided by the competition's organizer [2], referred to as 'Official.' In the Expression Classification Challenge, the 'Official' baseline uses a VGG16 model that's been pre-trained on the VGGFACE dataset. This model's convolutional layers are set and unchangeable, and it's fine-tuned for our specific task using only its three fully connected layers. It's designed to identify eight facial expressions using a softmax function in the output layer. Data augmentation is applied using a technique called Mixaugment. In the Action Unit Detection Challenge, the baseline is similarly constructed with a fixed VGG16 architecture. Still, it differs in its output layer, which uses a sigmoid function to pinpoint twelve distinct facial action units.

Implementation Details. All experiments are based on the PyTorch and trained with an NVIDIA RTX A6000 GPU. For training, we fix the batch size 32, and epochs 32, and use Adam optimizer with an initial learning rate at . Additionally, we employ a Cosine Annealing Learning Rate Scheduler to modulate the learning rate adaptively across the training duration, aiming to bolster the model's path to convergence. Experimentally, we find the best is 0.3, and we set the to .

B. Results

Comparison with Baseline. Table I presents a comparative evaluation of our method against the baseline technique regarding the 'macro' F1 score for both the Expression Classification Challenge and the Action Unit Detection Challenge. Our method outperforms the official baseline, achieving a 'macro' F1 score improvement of 11% in the Expression

Task	Method	F1 Score
Expression Classification	Official [2]	0.25
	Ours	**0.36**
Action Unit Detection	Official [2]	0.39
	Ours	**0.43**

TABLE I: *Comparison with the baseline method in terms of 'macro' F1 score on the validation set in Expression Classification Challenge and Action Unit Detection Challenge, respectively. The best results are shown in* ***Bold****.*

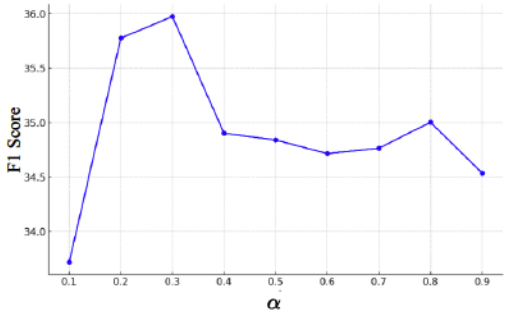


Fig. 2: *'Macro' F1 score to different α values in expression classification.*

Classification Challenge, rising from 0.25 to 0.36. In the Action Unit Detection Challenge, our approach exhibits a superior performance as well, enhancing the 'macro' F1 score by 4%, from 0.39 to 0.43. These results underscore the effectiveness of our proposed method, particularly in its ability to more accurately classify expressions and detect action units, highlighting our method's superior performance in both challenges.

Sensitivity Analysis. Fig.2 illustrates the 'macro' F1 score across different α values in Eq. (1) of expression classification. The α in CVaR controls how much the model prioritizes the tail end of the loss distribution, which corresponds to the most challenging cases during training. The optimal performance observed at $\alpha = 0.3$ implies that targeting the most difficult 30% of the data enhances the model's ability to handle critical instances, yielding the highest F1 Score. Beyond this point, as α increases, there is a noticeable decline in performance, signaling that too much focus on the tail may lead to over-prioritization of complex examples, potentially reducing overall model effectiveness.

IV. Conclusion

Existing methods for human affective behavior analysis utilizing complex deep learning models are often associated with high computational costs, limiting their feasibility and scalability in real-world applications. To overcome this limitation, we introduce the first efficient lightweight that employs a frozen CLIP image encoder and a trainable MLP, augmented with CVaR and a loss landscape flattening strategy. The CVaR integration bolsters our model's robustness, while the loss flattening strategy enhances generalization. Experimental results showcase that our framework not only outperforms the baseline but does so with a streamlined and resource-efficient architecture, establishing a new benchmark for efficient yet effective affective behavior analysis.

Acknowledgments. This work is supported by the U.S. National Science Foundation (NSF) under grant IIS-2348419 and the National Artificial Intelligence Research Resource (NAIRR) Pilot and TACC Lonestar6. Sarah Papabathini is supported by the Office of Naval Research (ONR) under award N00014-21-1-2691. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of NSF, NAIRR Pilot, and ONR.

References

[1] E. B. Prince, K. B. Martin, D. S. Messinger, and M. Allen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 2015.

[2] D. Kollias, P. Tzirakis, A. Cowen, S. Zafeiriou, C. Shao, and G. Hu, "The 6th affective behavior analysis in-the-wild (abaw) competition," *arXiv preprint arXiv:2402.19344*, 2024.

[3] D. Kollias, P. Tzirakis, A. Baird, A. Cowen, and S. Zafeiriou, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5888–5897, 2023.

[4] D. Kollias, "Multi-label compound expression recognition: C-expr database & network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5589–5598, 2023.

[5] D. Kollias, "Abaw: Learning from synthetic data & multi-task learning challenges," in *European Conference on Computer Vision*, pp. 157–172, Springer, 2023.

[6] D. Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2328–2336, 2022.

[7] D. Kollias and S. Zafeiriou, "Analysing affective behavior in the second abaw2 competition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3652–3660, 2021.

[8] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the first abaw 2020 competition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pp. 794–800.

[9] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution matching for heterogeneous multi-task learning: a large-scale face study," *arXiv preprint arXiv:2105.03790*, 2021.

[10] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.

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

[12] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," *arXiv preprint arXiv:1910.11111*, 2019.

[13] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 1–23, 2019.

[14] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild'challenge," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pp. 1980–1987, IEEE, 2017.

[15] W. Zhang, B. Ma, F. Qiu, and Y. Ding, "Multi-modal facial affective analysis based on masked autoencoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5792–5801, 2023.

[16] J.-H. Kim, N. Kim, and C. S. Won, "Facial expression recognition with swin transformer," *arXiv preprint arXiv:2203.13472*, 2022.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.
[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.
[19] K. N. Phan, H.-H. Nguyen, V.-T. Huynh, and S.-H. Kim, "Expression classification using concatenation of deep neural network for the 3rd abaw3 competition," *arXiv preprint arXiv:2203.12899*, vol. 5, 2022.
[20] W. Zhou, J. Lu, Z. Xiong, and W. Wang, "Leveraging tcn and transformer for effective visual-audio fusion in continuous emotion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5755–5762, 2023.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[22] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Jaa-net: joint facial action unit detection and face alignment via adaptive attention," *International Journal of Computer Vision*, vol. 129, pp. 321–340, 2021.
[23] Y. Tang, W. Zeng, D. Zhao, and H. Zhang, "Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12899–12908, October 2021.
[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.
[25] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Open clip." https://github.com/mlfoundations/open_clip, 2021.
[26] S. Hu, Z. Yang, X. Wang, Y. Ying, and S. Lyu, "Outlier robust adversarial training," in *Asian Conference on Machine Learning*, pp. 454–469, PMLR, 2024.
[27] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu, "Improving fairness in deepfake detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4655–4665, 2024.
[28] L. Lin, X. He, Y. Ju, X. Wang, F. Ding, and S. Hu, "Preserving fairness generalization in deepfake detection," *CVPR*, 2024.
[29] S. Hu, X. Wang, and S. Lyu, "Rank-based decomposable losses in machine learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[30] S. Hu and G. H. Chen, "Distributionally robust survival analysis: A novel fairness loss without demographics," in *Machine Learning for Health*, pp. 62–87, PMLR, 2022.
[31] S. Hu, L. Ke, X. Wang, and S. Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7649–7657, 2021.
[32] S. Hu, Y. Ying, X. Wang, and S. Lyu, "Sum of ranked range loss for supervised learning," *Journal of Machine Learning Research*, vol. 23, no. 112, pp. 1–44, 2022.
[33] S. Hu, Y. Ying, S. Lyu, *et al.*, "Learning by minimizing the sum of ranked range," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21013–21023, 2020.
[34] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2020.