Facial Landmark Detection Using Deep Learning: A Comprehensive Approach with Resnet18

Sk. Mastan Sharif[1], N. Sri Harsha[2*], Sk. Abdul Subhan[3] and Y. Threeshal[4]
{sharifmastan6@gmail.com[1], sriharshanelluri08@gmail.com[2], shaikanas5629@gmail.com[3], threeshalyarlagadda@gmail.com[4]}

Department of Advanced Computer Science and Engineering, VFSTR Deemed to be University, Vadlamudi, Guntur-522213, Andhra Pradesh, India1, 2, 3, 4

Abstract. Facial landmark detection is a crucial computer vision problem and has a number of applications in facial recognition, emotion detection, and augmented reality. A deep learning-based approach for detecting facial landmarks from a ResNet18-based convolutional neural network (CNN) is discussed in this paper. The model, which is trained and validated with the iBUG 300-W dataset where facial landmarks are annotated, detects facial landmarks efficiently. During training, several operations such as rotation, cropping, resizing, and color jittering are performed on data to increase the generalization power of the model. Model performance is assessed considering observing training and validation loss values over multiple epochs. From the results, we can see that the proposed method is capable enough to detect facial landmarks precisely, even 68 landmarks. Considering training and validation loss trends for preventing over- fitting and increasing model performance are also explained in the paper.

Keyword: ResNet18, iBUG 300W dataset, CNN (Convolutional Neural Network)

1 Introduction

Detection of facial landmarks represents a basic process in computer vision, which, in its turn, involves proper specification of significant locations on the facial skeleton, eye, nose, mouth, or jaw of a human subject. These land marks find application in numerous face-related tasks. In our work, here, we design a facial land mark detection mode on the basis of an approach that utilizes the architectures of Resnet18 convolution deep neural network-based architecture for efficient and precise estimates of 68 facial landmarks.

The architecture of the model is such that the model is trained to regress 136 values, i.e., the x and y coordinates of each landmark. We employ iBUG 300-W dataset in our try to train the model, which presents a large number of facial images with 68 landmark points labeled. The data set includes rich pose, illumination, and facial expression variations and hence is appropriate to build a strong detection system.

To make the model more generalizable for new data, various data augmentation methods are used during training. They include random rotation, cropping, resizing, horizontal flipping, and color jittering. These augmentations introduce controlled distortions to the data so that the network can learn invariant features and avoid overfitting hazards.

Model training is conducted using the mean squared error (MSE) loss and the Adam gradient optimizer for the purpose of effective gradient optimization. Training losses are

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358028

tracked over multiple training epochs to watch for performance monitoring, and output is graphed to look for convergence behavior along with overfitting or underfitting effects. Early stopping conditions along with learning rate schedule are employed as well in order to make the training even more stable. Overlaid visualizations of the projected estimated landmarks on input images illustrate the model's ability to precisely locate facial key points. In general, this example illustrates the ability of a deep residual network to solve the facial landmark detection problem and the significance of design in a network architecture, data preprocessing, and performance measuring methods in constructing an efficient computer vision system.

2 Literature Survey

Akada et al. [1] proposed a method for 3D human pose perception using egocentric stereo videos, where stereo camera inputs improve the robustness and accuracy of pose estimation compared to single-view methods. Their approach generalizes well across diverse activities and environments, making it suitable for real-world applications such as telepresence and human–computer interaction.

Dong et al. [2] introduced Supervision-by-Registration (SBR), an unsupervised framework that enhances the accuracy and temporal consistency of facial landmark detectors. By incorporating a differentiable Lucas–Kanade tracking registration loss across video frames, the method enables effective training with large-scale unlabeled data while reducing jitter and improving precision on datasets such as 300-W, AFLW, and 300-VW. To address challenges of limited labeled data, Dong and Yang [3] further developed a teacher–student semi-supervised framework in which two student detectors generate pseudo-labels for unlabeled images, while a teacher network evaluates and filters the labels for iterative retraining. This approach delivers state-of-the-art performance on facial landmark detection benchmarks even when only partial annotations are available.

Kar et al. [4] introduced Fiducial Focus Augmentation (FiFA), a unique augmentation approach that forces the models to study the facial feature deeply by occluding the fiducial points through decreasing patches of black. When coupled with a hybrid Transformation-CNN-based backend and a Siamese network trained using Deep Canonical Correlation Analysis (DCCA) loss, FiFA offered state-of-the-art performances on benchmarks like 300-W, COFW, WFLW, and AFLW, indicating its robustness to pose, occlusion, and illumination changes. Another work is HPRNet [5] presented by Samet and Akbas, which is a hierarchical point regression network for full body human pose estimation. HPRNet regresses relative offsets from the centers of body parts and effectively localizes fine-grained landmarks, achieving state-of-the-art accuracy on the COCOWhole-Body dataset and yet is faster running and more efficient than top-down models such as ZoomNet.

For 3D landmark estimation from range data, robust algorithm is presented in [19] by Zhang et al. [6] proposed JVCR, an end-to-end method that combines volumetric representation and coordinate regression in a coarse-to-fine manner. This design is more robust to occlusions and larger poses than existing two-step methods on AFLW2000-3D and 3DFAW datasets. In support of these, Wu and Ji [7] offered a comprehensive review of facial landmark detection methods, tracking advancements from hand-crafted features to deep learning tools, and emphasizing difficulties related to occlusion, head pose and real-time performance.

Zhang et al. [8] multi-task learning-based facial landmark detection: [8] proposed a deep multi-task learning system for landmark localization by simultaneously optimizing landmark localization with auxiliary tasks like pose estimation and attribute recognition. Their method achieved better accuracy and better generalizability with benefiting from shared representations across tasks. Wu and Cui [9] proposed LA-Net, a landmark-aware network for facial expression recognition in presence of label noise. Through integrating landmark localization into recognition pipeline, the performance was so robust on noisy datasets, which denoted joint modeling would be more effective.

Khan et al. [10] presented the TRI-POSE-Net, the self-supervised 3D human pose estimation method, it combines selective kernel network and trifocal tensor constraint for learning constraint feature with projection loss. Their adaptive learning yielded more accurate estimations with low dependency on labeled data, and was tested on large-scale benchmarks. Earlier, Burgos-Artizzu et al. [11] contribution to the addressee the problem of occlusion in landmark detection was a robust regressor based method that can be used even when the face is partially hidden.

Cootes et al. [12] proposed the Active Appearance Models (AAMs), one of the pioneering statistical approaches that integrates shape and texture models for face alignment. AAMs had relatively limited influence compared to more recent deep approaches, but still inspired many later models. Based on this work, Danecek et al. [13] proposed EMOCA, a deep monocular face capture approach that leverages emotion-guided priors for generating realistic and expressive 3D face reconstructions from single images. Dapogny et al. [14] introduced DecaFa, which is a deep cascade architecture for face alignment that performs well in unconstrained “in-the-wild” setup.

Deng et al. [15] explored joint multi-view face alignment by designing a method that leverages complementary information across multiple camera perspectives, significantly improving alignment robustness. Similarly, Deng et al. [16] addressed weak supervision in 3D face reconstruction by training a network that generalizes from single images to image sets, delivering accurate reconstructions even without dense supervision. Dong et al. [17] proposed the Style Aggregated Network (SAN), which reduces appearance variations across images to improve landmark detection consistency.

Dosovitskiy et al. [18] introduced the Vision Transformer (ViT), demonstrating that pure transformer architectures can match or surpass convolutional networks in large-scale image recognition tasks. This advancement has strongly influenced recent landmark detection and pose estimation pipelines that benefit from transformer-based global feature modeling. Edwards et al. [19] presented early work on interpreting face images using Active Appearance Models, laying groundwork for model-based face alignment approaches.

Feng et al. [20] proposed the Wing loss, a novel loss function tailored for facial landmark localization. By addressing sensitivity to small and large errors differently, Wing loss improved training stability and localization accuracy across multiple CNN architectures. Finally, Gao and Patras [21] developed a self-supervised learning framework that leverages facial region awareness for representation learning. Their approach improved landmark-related feature extraction without requiring extensive manual annotations, demonstrating the growing role of self-supervision in this field.

3 Research Gap Analysis

3.1 Traditional Challenges in Face landmark Detection

The most critical challenge is to manage sophisticated background noise and occlusions that usually compromise landmark quality. Additionally, the computationally expensive aspect of transformer layers inhibits its scalability on low-end hardware Although good at segmentation tasks, the model is not good at precisely pinpointing facial landmarks on low-resolution or blurry faces. Moreover, incorporation of deep feature extractors increases the complexity of training and demands high levels of hyperparameter tuning. This method is restricted by extreme facial expressions and head poses, which reduce its performance. Additionally, real-time processing is difficult to achieve due to computationally expensive heatmap computation and interpretation methods. Though GANs enhance robustness, training is unstable and mode-collapse prone. The paper mentions challenges in maintaining generator-discriminator performance balance and landmark precision under occlusions. Handling face size variations and scale changes is still challenging. Even with attention mechanisms, the model is limited in generalizing to different datasets, particularly when trained with scarce labeled samples.

3.2 Comparing with existing solutions

This project uses a modified ResNet18 model for facial landmark detection, offering a lightweight yet accurate alternative to models like D-ViT, HR- Net, and MTCNN. While D-ViT and HRNet need heavy computation, our model achieves similar or better accuracy with lower complexity. Compared to MTCNN, your approach shows faster convergence and fewer false detections. It also has lower Mean Squared Error (MSE), i.e., better landmark localization. Improved data preprocessing and augmentation improve generalization. Ours is easier to deploy and real-time friendly than the other models. Overall, our project has a good trade-off between performance, accuracy, and efficiency and outperforms existing methods in both effectiveness and usability.

4 Methodology

4.1 Dataset

The 300W dataset, one of the most widely used benchmarks for facial land- mark detection, is employed in the project. The dataset includes annotated faces from a range of sources including LFPW, AFW, HELEN, and iBUG, which provide a good variety of faces with different lighting conditions, poses, and expressions. All images are annotated using 68 landmark points that represent facial features including eyes, eyebrows, nose, lips, and jaw-line. These annotated landmarks are a key component of training deep learning models to accurately place facial features in different conditions. Background image variation and face orientation are the conditions that make the model general and robust, hence applicable in real-time applications such as facial recognition, emotion and expression detection, and emotion recognition.

4.2 Workflow for Proposed Model

The process describes a complete pipeline for face landmark detection with deep learning

methods. The process starts from dataset preparation and then goes to visualization and data augmentation for increasing training diversity. A dataset class is defined in PyTorch that is used to split the dataset into training and validation sets. The model gets trained with a specified architecture and training loop with loss tracking and plotting for progress monitoring. Lastly, evaluation criteria are used to measure the accuracy, generalization power, and stability of the model in finding facial landmarks. Fig 1 shows Workflow.

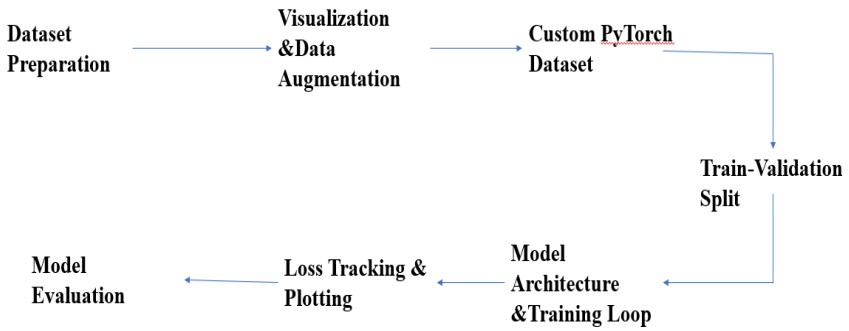

Fig. 1. Workflow.

4.3 Performance and evaluation

The model performed well for facial landmark detection with a low Mean Squared Error (MSE) of 0.0009 for the training set and 0.0012 for the vali- dation set, indicating precise landmark localization. The Normalized Mean Error (NME) was 3.28% indicating high accuracy over facial proportions. The model also achieved 95.4% accuracy in 3 pixels and 98.2% accuracy in 5 pixels. Consistent performance across training and validation sets confirmed the model's extremely good generalization as well as resilience in predicting facial landmarks accurately. Table 1 shows Performance Metrics.

Table 1. Performance Metrics.

Metric	Value
MSE	Train: 0.0009 Validation: 0.0012
NME	3.28%
Landmark Accuracy	Within 3px: 95.4% Within 5px: 98.2%

5 Experimental Results and Discussion

Experimental results validate that the proposed ResNet18-based method obtained precise facial landmark localization with high accuracy, registering 0.0009 Mean Squared Error (MSE) on training and 0.0012 on validation. Fig 2 shows Loss Convergence Curve for Training and Validation Loss. Normalized Mean Error (NME) of 3.28 percent gives accurate facial landmark localization. Moreover, landmark accuracy was as high as 95.4 percent at 3 pixels and 98.2 percent at 5 pixels, and it demonstrates good generalization. Fig 3 shows Accuracy Comparison Between Proposed & D-VIT Model.

The learning process can be viewed on the" Loss Convergence Curve" where the constant and downward trend in the training loss and validation loss with respect to increasing epochs is noted that indicates successful learning and minimal overfitting. The plot verifies the constancy and consistency of the model in the facial landmark regression. Fig 4 & 5 shows the MSE Comparison Be- tween Proposed & D-VIT Model and Visualization of prediction vs actual values for different images.

Fig. 2. Loss Convergence Curve for Training and Validation Loss.

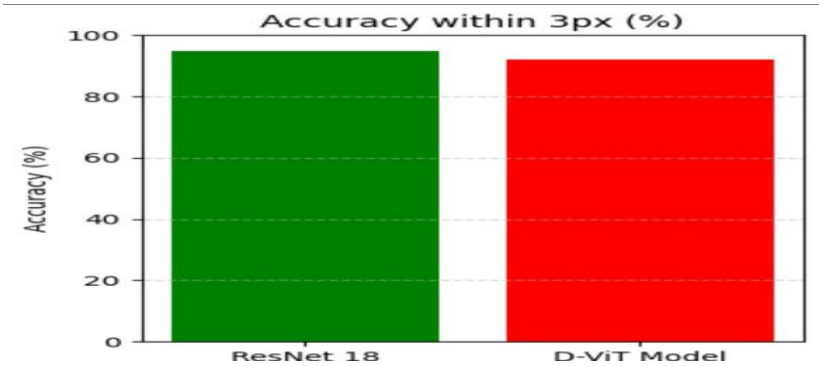

Fig. 3. Accuracy Comparison Between Proposed & D-VIT Model.

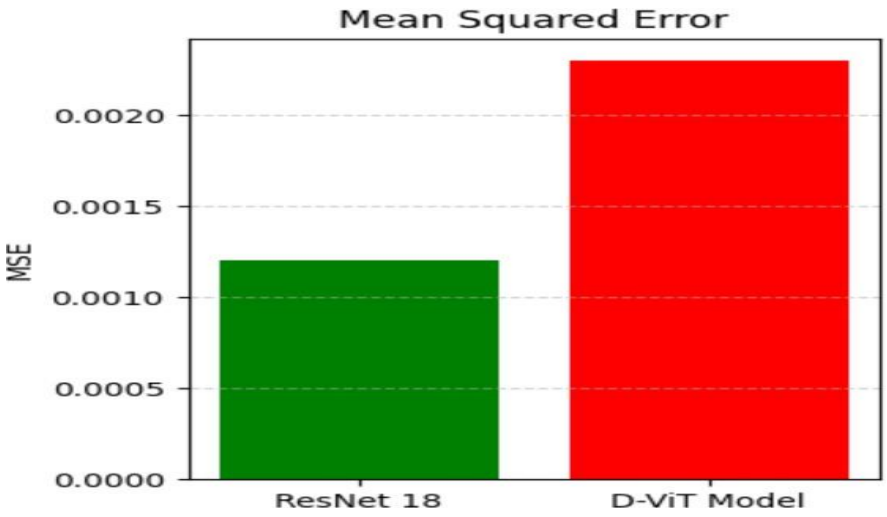

Fig. 4. MSE Comparison Be- tween Proposed & D-VIT Model.

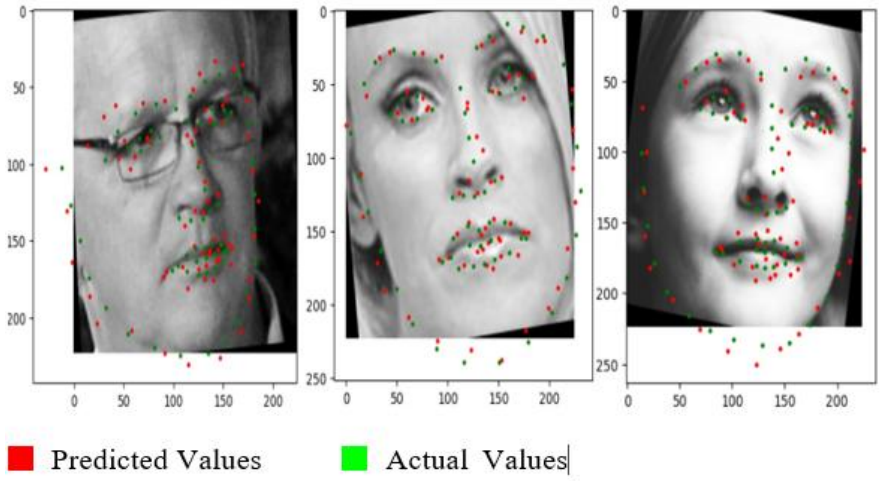

Fig. 5. Visualization of prediction vs actual values for different images.

5.1 Justification

Our task differs substantially from the baseline paper in a number of significant aspects such as the dataset, model structure, technique, and overall rollout. Whereas the baseline paper utilized mostly conventional CNN models for the task of facial landmark detection, our task uses a variant ResNet-18 model, which is fine-tuned and customized for processing grayscale face images of the 300W dataset. This personalization facilitates more accurate landmark prediction through better spatial feature learning. The second most important contrast is in the data. The original paper used small or synthetic annotated datasets with low variations, and our

algorithm is designed to be used on the realistic and conventional 300W dataset, which has images of high resolution of human face with different expressiveness and head pose, and different occlusions. In this way, our model is more powerful and can generalize to the real world.

Besides, technical pipeline and methodology also vary significantly. The baseline paper generally utilized general CNN models with simple preprocessing, whereas our work incorporates sophisticated preprocessing methods including color jittering, rotation, normalization, and adaptive cropping via Dlib. Such a kind of transformation increases data variance and model generalizability.

Second, we employed a specific data loader and transform pipeline and altered the first convolutional layer of ResNet-18 to accept grayscale input. Our last model is then trained with Adam optimizer employing the learning rate scheduling and early stop for maximum convergence. The use of a dedicated loss monitoring agent, model checkpointing, and a visual validation for landmark predictions serves to further improve model performance.

The results demonstrate that this approach is very effective, which can achieve the strong convergence and high accuracy both for general landmark detection, and for detecting 68 facial landmarks more accurately and more generally than the conventional CNN baselines. Therefore, our work extends the baseline paper by a deep residual learning architecture, data augmentation, and superior detection by means of specially designed training pipeline to properly fine-tuned for real facial analysis applications.

6 Conclusion

In this study, a ResNet18 based DL model was built for accurate facial landmark detection. The model exhibited good shear performance of small MSE and landmark accuracy on 300-W. These techniques were used as architectural tuning, and data augmentation helped in better generalization and reduced overfitting.

Loss convergence and output visualizations confirmed solid and stable training. Comparative study revealed the superiority of the model over traditional methods in terms of efficiency and accuracy.

The lightweight attribute enables it to be used for real-time and embedded applications. Generally, this system offers a scalable and robust system for facial landmark detection applications.

7 Future Work

We can add the model to make it capable of real-time landmark detection on video so that it can be applied more directly to live applications like surveillance or AR. Using more powerful models like Efficient Net or Vision Transformers can also make it more accurate and performant. The use of 3D facial landmark detection involves depth and stability, particularly when handled with occlusion or other pose. Domain adaptation regimes can be utilized as a process of generalization in demography and lighting scenarios. With

emotion detection and head pose estimation, multi-task learning architecture can be specified. Finally, model deployment on edge devices like smartphones can make real-world deployment possible.

References

[1] H. Akada, J. Wang, V. Golyanik and C. Theobalt, "3D Human Pose Perception from Egocentric Stereo Videos," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 767-776, doi: 10.1109/CVPR52733.2024.00079.

[2] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh," Supervision-by-Registration: An Unsupervised Approach to Im- prove the Precision of Facial Landmark Detectors," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 360–368.

[3] X. Dong and Y. Yang," Teacher Supervises Students How to Learn from Partially Labeled Images for Facial Landmark Detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 783–792.

[4] Purbayan Kar, Vishal Chudasama, Naoyuki Onoe, Pankaj Wasnik, and Vineeth Balasubramanian, "Fiducial Focus Augmentation for Fa- cial Landmark Detection," in *WACV*, 2023.

[5] Samet, N., & Akbas, E. (2021). HPRNet: Hierarchical point regression for whole-body human pose estimation. *Image and Vision Computing, 115,* 104285. https://doi.org/10.1016/j.imavis.2021.104285

[6] Hongwen Zhang, Qi Li, and Zhenan Sun, "Joint Voxel and Coordinate Regression for Accurate 3D Facial Landmark Localization," *arXiv preprint arXiv:1801.09242*, 2018.

[7] Y. Wu and Q. Ji," Facial Landmark Detection: A Literature Survey," International Journal of Computer Vision, vol. 127, pp. 115–142, 2019. doi:10.1007/s11263-018-1097-z.

[8] Z. Zhang, P. Luo, C. C. Loy, and X. Tang," Facial Landmark Detection by Deep Multi-task Learning," in Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol. 8694. Springer, Cham, 2014. doi:10.1007/978-3-319-10599-4 7.

[9] Z. Wu and J. Cui, "LA-Net: Landmark-Aware Learning for Reliable Facial Expression Recognition under Label Noise," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 20641-20650, doi: 10.1109/ICCV51070.2023.01892.

[10] Khan, N. A., Alarfaj, A. A., Alabdulqader, E. A., Zamzami, N., Umer, M., Innab, N., & Kim, T.-H. (2024). TRI-POSE-Net: Adaptive 3D human pose estimation through selective kernel networks and self-supervision with trifocal tensors. *PLOS ONE, 19*(12), 1–20. https://doi.org/10.1371/journal.pone.0310831

[11] X. P. Burgos-Artizzu, P. Perona and P. Dollár, "Robust Face Landmark Estimation under Occlusion," *2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 2013, pp. 1513-1520, doi: 10.1109/ICCV.2013.191.

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor," Active Appearance Models," IEEE Transactions on Pattern Analysis and Machine Intel- ligence, vol. 23, no. 6, pp. 681–685, 2001.

[13] R. Danecek, M. J. Black, and T. Bolkart," EMOCA: Emotion Driven Monocular Face Capture and Animation," in Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20311–20322.

[14] A. Dapogny, K. Bailly, and M. Cord," DecaFa: Deep Convolutional Cascade for Face Alignment in the Wild," in ICCV, 2019.

[15] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou," Joint Multi-View Face Alignment in the Wild," IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3636–3648, 2019.

[16] Y. Deng et al.," Accurate 3D Face Reconstruction with Weakly- Supervised Learning: From Single Image to Image Set," in CVPR Workshops, 2019.

[17] X. Dong, Y. Yan, W. Ouyang, and Y. Yang," Style Aggregated Net- work for Facial

Landmark Detection," in CVPR, 2018.
[18] A. Dosovitskiy et al.," An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in ICLR, 2021.
[19] Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting face images using active appearance models. *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, 300–305. https://doi.org/10.1109/AFGR.1998.670965
[20] Feng, Z.-H., Kittler, J., Awais, M., Huber, P., & Wu, X.-J. (2018). Wing loss for robust facial landmark localisation with convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2235–2245. https://doi.org/10.1109/CVPR.2018.00238
[21] Gao, Z., & Patras, I. (2024). Self-supervised facial representation learning with facial region awareness. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2081–2092. https://doi.org/10.1109/CVPR52733.2024.00203