Enhancing facial landmark detection with ControlNet-based data augmentation

Kritaphat Songsri-in[1], Munlika Rattaphun[1], Sopee Kaewchada[2], Sunisa Kidjaideaw[2], Sangjun Ruang-On[2], Wichit Sookkhathon[2], Patompong Chabplan[2]

[1]Department of Computer Science, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Tha Ngio, Thailand
[2]Department of Information Technology and Digital Innovation, Faculty of Science and Technology,
Nakhon Si Thammarat Rajabhat University, Tha Ngio, Thailand

Article Info

Article history:

Received Feb 3, 2025
Revised Jun 12, 2025
Accepted Jul 3, 2025

Keywords:

ControlNet
Deep learning
Face image generation
Face landmark detection
Machine learning

ABSTRACT

Facial landmark detection plays a pivotal role in various computer vision applications, including face recognition, expression analysis, and augmented reality. However, existing approaches often struggle with accuracy due to the variations in lighting, poses, and occlusion. To address these challenges, this study explores the integration of ControlNet with Stable Diffusion to enhance facial landmark detection via data augmentation. ControlNet, an advanced extension of diffusion models, improves image generation by conditioning outputs on structured inputs such as landmark coordinates, enabling precise control over image attributes. By leveraging annotated landmark data from the 300W dataset, ControlNet synthesizes diverse facial images that supplement traditional training datasets. Experimental results demonstrate that ControlNet-based augmentation reduces the interocular normalized mean error (INME) in landmark detection from a baseline of 4.67 to a range of 4.63 to 4.74, with optimal parameter tuning yielding further accuracy gains. These findings highlight the potential of generative models in complementing discriminative approaches and improving robustness and precision in facial landmark detection. The proposed method offers a scalable solution for enhancing model generalization, particularly in applications requiring high-fidelity facial analysis. Future research can extend this framework to broader computer vision tasks that demand detailed feature localization and structured data augmentation.

Corresponding Author:

Patompong Chabplan
Department of Information Technology and Digital Innovation, Faculty of Science and Technology,
Nakhon Si Thammarat Rajabhat University
Tha Ngio, Thailand
Email: patompong_cha@nstru.ac.th

1. INTRODUCTION

Facial landmark detection is a critical area in computer vision, supporting numerous applications, including facial recognition [1]–[4], expression analysis [5]–[7], 3D facial modeling [8]–[10] and augmented reality [11], [12]. These applications rely on accurately identifying specific facial points, or landmarks, that represent essential facial features. Over the past decades, a variety of algorithms have been proposed to localize facial keypoints accurately under diverse conditions. Early approaches were often built on statistical shape models or graphical representations of facial structure. Active shape models (ASM) [13] and active appearance models (AAM) [14] are seminal model-based frameworks that iteratively fit a parametric shape and appearance to face images by enforcing learned shape constraints. These methods and other deformable models provided a foundation for face alignment, but their performance degrades on unconstrained images with large pose or expression variation. To better handle such variability, part-based graphical models were introduced. For example, the mixture-of-trees model by [15] represented facial landmarks as tree-structured parts with global and local mixtures, enabling joint face detection, pose estimation, and landmark localization in wild images. While these graphical techniques increased robustness to pose, their accuracy was limited by the rigidity of the underlying shape assumptions.

Subsequently, direct regression methods gained popularity for their efficiency and accuracy, bypassing explicit shape modeling. Cascaded shape regression frameworks [16] emerged as a dominant approach, where an initial coarse landmark estimate is iteratively refined by a sequence of learned regressors. By learning shape update transformations, these methods can rapidly converge to the target landmarks. First demonstrated an explicit shape regression that directly maps image features to landmark displacements without any parametric model [17]. Numerous enhancements followed: formulated the supervised descent method (SDM) to minimize a nonlinear least-squares alignment objective [18], applied random forests with conditional regressors to predict facial keypoints in real time while accounting for head pose [19]. Later, ensemble-based regressors were introduced which further improved reliability. Employed an ensemble of regression trees, enabling one-millisecond face alignment with competitive accuracy [20]. To reduce overfitting and improve generalization, combined gradient-boosted trees with Gaussian processes in a cascade (cGPRT) [16], which acted as a form of regularized ensemble that achieved state-of-the-art results on challenging benchmarks. These regression and ensemble methods significantly improved alignment speed and accuracy, yet their data-driven nature meant that generalization to extreme poses or expressions was still constrained by the availability and diversity of training data.

With the rise of deep learning, convolutional neural network (CNN) approaches have dramatically advanced the state-of-the-art in many vision tasks [21]–[23], including facial landmark detection [24], [25]. Deep neural networks can learn robust feature representations and implicit shape constraints from large datasets. First demonstrated a CNN cascade for facial point detection, outperforming earlier cascaded regressors by a large margin [26]. Subsequent works leveraged increasingly sophisticated deep models and training strategies. Multi-task learning frameworks were introduced to improve robustness: for example, Zhang *et al.* [27] trained a CNN to predict landmarks together with head pose and facial attributes, gaining resilience to occlusions and pose changes through shared feature learning. Other researchers integrated 3D face modeling into the learning process to handle profile views. Combined a cascaded CNN with a 3D Morphable Model to align faces across large poses [28], and proposed a 3D-assisted solution that fits a dense 3D face to 2D landmarks, thereby improving alignment of self-occluded [29]. Fully convolutional architectures and heatmap regression techniques have also yielded excellent accuracy. A very deep residual network for landmark localization by study [25] nearly saturated the performance on several 2D and 3D face alignment datasets, achieving remarkably low normalized mean errors. In addition, improved loss functions and data handling have enhanced CNN-based alignment. Notably, Feng *et al.* [30] introduced the Wing loss to better penalize small errors while tolerating outliers, leading to more robust convergence. Incorporated boundary-aware features to explicitly model face contour information, which boosted landmark accuracy on challenging cases like profiled faces and exaggerated expressions [31]. Thanks to these advances, modern neural methods can achieve high accuracy under controlled conditions. However, their performance can still degrade in unconstrained environments due to the inherent diversity of real-world faces.

A key remaining challenge is the reliance of deep models on abundant and varied labeled data. In practice, collecting and manually annotating a sufficiently diverse facial landmark dataset is costly and labor-intensive. Many existing datasets have biased distributions, such as limited extreme poses, occlusions or ethnic diversity, causing models trained on them to generalize poorly to new domains. Data augmentation is therefore crucial to improve model robustness [32]. Conventional augmentation techniques such as random cropping, flipping, rotation and noise injection can expand a dataset but only produce limited perturbations of existing images and may not introduce truly novel face appearances or geometries. This has motivated the use of generative models to synthetically enlarge training data. More recently, diffusion models [33], [34] have emerged as a powerful class of generative models, achieving state-of-the-art image quality and diversity in synthesis tasks. By leveraging a pretrained diffusion prior, one can guide image synthesis using additional inputs such as text, sketches, or keypoint maps [35]. This suggests a tantalizing opportunity: by conditioning a generative model on facial landmark configurations, we can produce synthetic face images that come with free landmark labels, thereby creating virtually unlimited training data with precise ground truth.

In this work we present a novel data augmentation framework that integrates ControlNet with Stable Diffusion to synthesize photorealistic face images conditioned on input landmark layouts. Our contributions are threefold. First, we develop the first diffusion model that uses conditional augmentation for facial landmarks. Second, we provide empirical evidence that our method reduces normalized mean error compared to baseline models. Third, we show how structural generative augmentation can apply to other vision tasks such as human pose estimation and hand keypoint detection where labeled data are scarce. By providing a scalable way to create large volumes of accurately labeled data, our method enables the training of more robust and generalizable models in facial analysis and related fields.

2. METHOD

To optimize facial landmark detection, this method integrates ControlNet with Stable Diffusion for synthetic data augmentation. By conditioning the image generation process on predefined facial landmark configurations, this approach generates varied training images to enhance the robustness and accuracy of facial landmark detection. The following subsections describe the dataset, model architecture, loss functions, training strategy, and implementation details.

2.1. Datasets

This study utilizes two primary datasets for training and evaluating the facial landmark detection model: the 300 W dataset [36], a widely established benchmark for facial landmark detection, and a ControlNet-based augmented dataset. The ControlNet-based augmented dataset generates synthetic images conditioned on facial landmarks from the 300 W dataset. These datasets together provide both real and synthetic data, allowing for a systematic examination of model performance across various data configurations.

2.1.1. The 300 W dataset

The 300 W dataset is a crucial benchmark in the facial landmark detection domain, offering a diverse collection of facial images curated to challenge and evaluate detection algorithms effectively. It includes various subsets designed to simulate real-world scenarios, capturing a broad spectrum of facial conditions, such as different lighting environments, facial expressions, and levels of occlusion. This dataset serves as the primary source of annotated real-world data for training and evaluating facial landmark detection models. It includes 3,148 training images and 600 testing images, providing a substantial volume of data for robust model training and analysis. Figure 1 displays sample images from the 300 W dataset, illustrating the diversity of facial features and landmarks that make this dataset invaluable for rigorous testing and validation. Figure 1(a) illustrates examples from the 300 W dataset, highlighting the diversity of facial variations and the detailed annotation of facial landmarks.

2.1.2. ControlNet-based augmented dataset

To supplement the 300 W dataset, a synthetic dataset was created using ControlNet, an advanced image generation model capable of producing realistic facial images conditioned on specific landmark configurations. ControlNet was applied to the 300 W landmark annotations to generate synthetic images that closely adhere to the structural features of the original dataset, enhancing diversity in training data by introducing new variations in lighting, pose, and facial expressions. This augmented dataset was generated at varying ratios λ relative to the original dataset, from 0% to 100% in steps of 10%, allowing for experimental evaluation of different real-to-synthetic data combinations. By integrating ControlNet-based synthetic images, the augmented dataset provides a scalable solution to boost model generalization and robustness across a range of facial landmark detection scenarios. Figure 1(b) showcases examples from the ControlNet-based augmented dataset, illustrating how this synthetic data closely resembles real-world conditions and enhances training diversity.

2.2. Model architecture

For efficient computations, our model is designed specifically to handle the single objective of facial landmark detection with precision. The network begins processing with a 64×64×3 color image as input. This input is sequentially passed through five 3×3 convolutional layers, each using a rectified linear unit (ReLU) activation function to introduce non-linearity, addressing challenges like the vanishing gradient. After each convolutional layer, a max-pooling operation reduces the spatial dimensions by half, which enhances the model’s translational invariance and condenses information. Each of the five convolutional layers is structured with kernels defined by Width×Height×Input×Output, where the kernel size specifies each layer’s input and output channels, ensuring efficient feature extraction. Following these foundational layers, the network includes fully connected layers to process the extracted features. These fully connected layers transform the spatial information into a final output vector of 2L values, where each pair of values represents the x and y coordinates of each of L facial landmarks. In this setup, L is configured for 68 landmark points to capture detailed facial features accurately. This structure allows the model to excel in precise landmark localization, effectively capturing the essential details required for facial analysis. The architecture of the model is depicted in Figure 2.

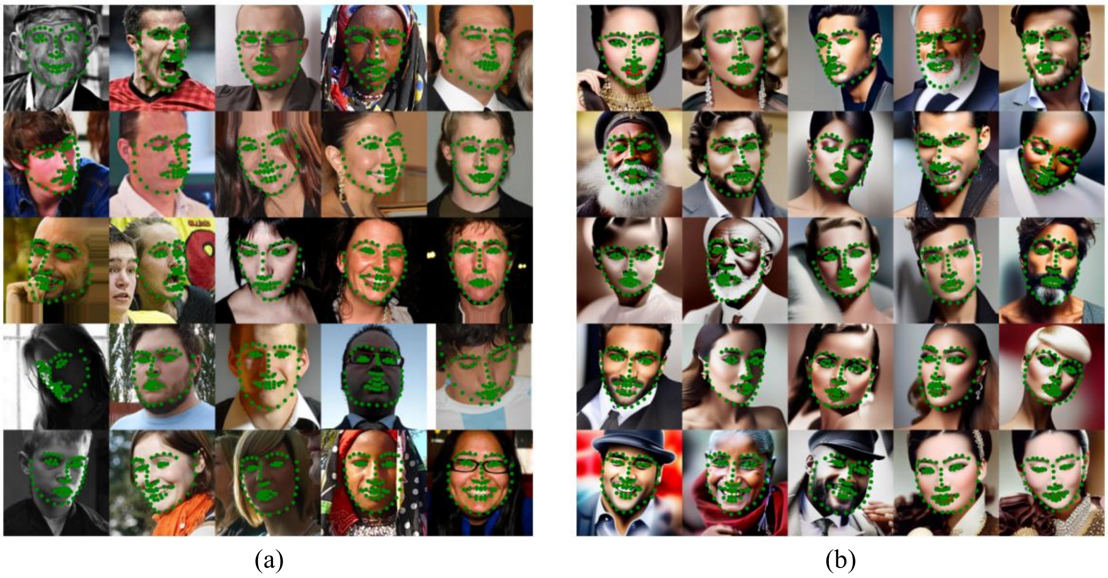


Figure 1. Examples from the datasets used for training and evaluation: (a) sample images from the 300 W dataset displaying diverse facial expressions, lighting conditions, and occlusions with annotated landmarks. (b) synthetic images from the ControlNet-based augmented dataset, generated using 300 W landmark configurations to introduce additional variations in pose, lighting, and expression

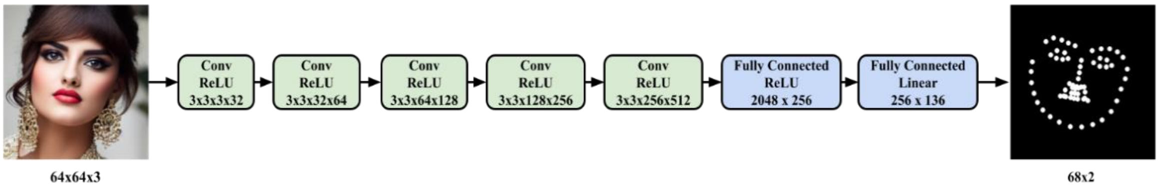


Figure 2. Overall architecture: a sequence of five 3×3 conv+ReLU+max-pool blocks, followed by fully connected layers that output 2×68 landmark coordinates

2.3. Loss function

The model’s training objective focuses on minimizing localization error for facial landmark detection. The mean absolute error (MAE) is used to quantify the discrepancy between the predicted and actual landmark positions, ensuring accuracy in facial landmark localization. The loss function is defined in (1):

$$L_{Landmark} = \frac{1}{NL}\sum_i^N \sum_j^L \left| l_{ij} - \hat{l}_{ij} \right| \qquad (1)$$

where N is the number of images, L represents the total landmarks in each image, l_{ij} is the ground truth location of the j-th landmark in image i, and $\hat{l}_{ij}$ is the predicted location generated by the model. This MAE-based loss function ensures accurate localization by linearly penalizing errors across the predicted coordinates.

2.4. Model training strategy

To assess the effects of synthetic data on facial landmark detection, the model was trained with datasets containing different ratios λ of ControlNet-generated images to original images, ranging from 0.0 to 1.0 in steps of 0.1. Each ratio was treated as a separate experiment, with the proportion of synthetic to real images held constant throughout the training process. By systematically varying these ratios, this approach enables a comparative analysis of how different levels of synthetic data influence model performance, providing insights into the optimal dataset composition for enhancing accuracy and robustness in facial landmark detection.

As illustrated in Figure 3, each experimental setup represents a unique dataset composition by balancing real and synthetic data according to the designated ratio. This structure allows the model to learn from both natural and augmented facial variations, examining how synthetic data contributes to generalization across diverse facial conditions. By comparing performance across these configurations, the experiments aim to identify the most effective ratio of synthetic augmentation for enhancing the model's ability to accurately detect facial landmarks.

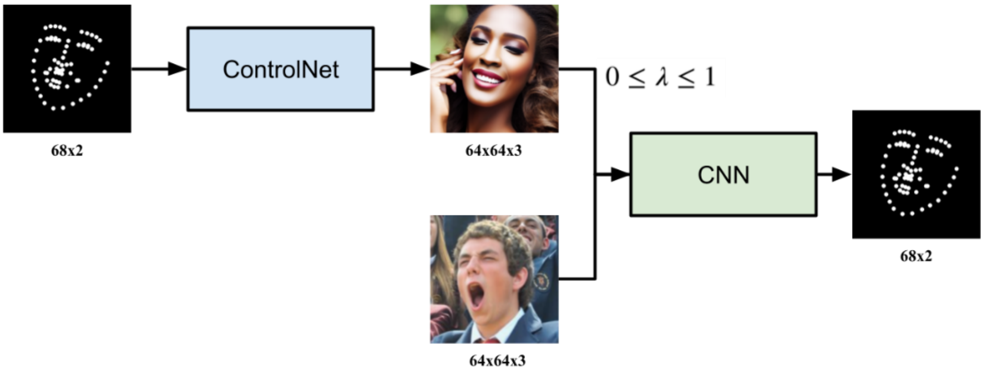


Figure 3. Dataset augmentation strategy: for each experiment, a fraction λ of ControlNet-generated synthetic images is additionally added on top of the real 300 W images

2.5. Implementation details

This facial landmark detection model was implemented using Python and TensorFlow, leveraging its flexibility for deep learning tasks. Input images were normalized to a range of 0 and 1 by dividing pixel values by 255. The model was trained using the Adam optimizer, with a piecewise constant learning rate schedule. The initial learning rate of 1×10^{-3} was reduced to 1×10^{-4} after the first third of the training epochs and further to 1×10^{-5} after the second third, ensuring gradual refinement of model parameters. Training was conducted for 1000 epochs with a batch size of 64. Regularization was applied using L2 weight decay 5×10^{-4} to mitigate overfitting. Augmentation techniques, including random rotations, flipping, cropping, and Gaussian blurring, were employed to enhance data diversity and robustness. The implementation strategy, combining efficient architecture, adaptive learning rates, and augmentation, facilitated accurate and robust prediction of facial landmarks under varied conditions.

3. RESULTS AND DISCUSSION

This section presents an experimental evaluation of the proposed method, focusing on the impact of ControlNet-based synthetic data augmentation on facial landmark detection performance. The interocular normalized mean error (INME) is employed as the primary evaluation metric, providing a scale-independent assessment of landmark localization accuracy. Comparative analyses are conducted across various augmentation ratios and parameter settings to determine the optimal configurations for achieving robust and precise facial landmark detection.

3.1. Metrics

The INME provides a refined metric specifically suited for evaluating facial landmark detection. This measure calculates the average difference between the predicted and actual landmark positions, with normalization based on the interocular distance, defined as the distance between the two outermost points of the eyes. This normalization ensures a scale-independent assessment. The formula for INME is presented in (2):

$$INME = \frac{1}{N}\sum_{i}^{N}\frac{\sqrt{\sum_{j}^{L}(l_{ij}-\hat{l}_{ij})^2}}{D_i} \tag{2}$$

where N represents the number of images, L is the total number of landmarks in each image, l_{ij} and $\hat{l}_{ij}$ are the ground truth and predicted landmark positions, respectively, and D_i is the distance between the outer corners of the eyes in each image.

3.2. Methods comparison

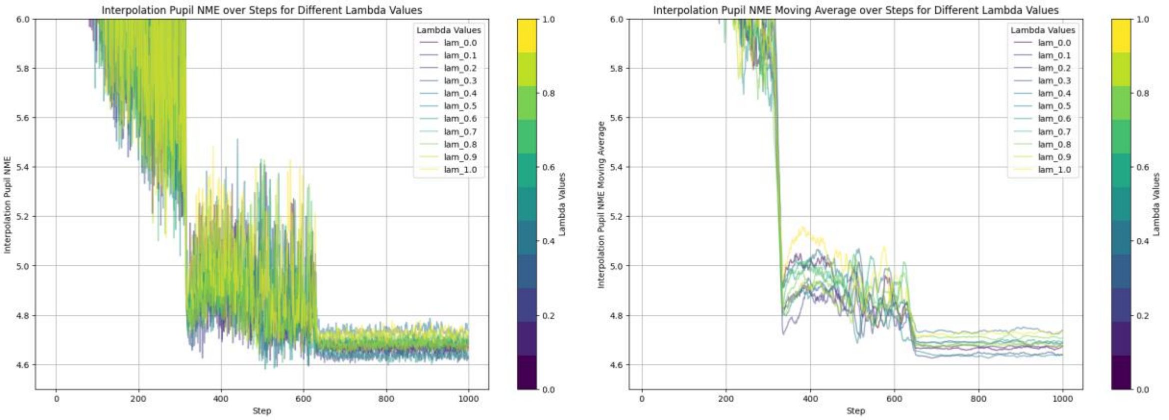
The results of the experiments, presented in Table 1, show the performance of the facial landmark detection model with varying ratios of ControlNet-based augmented data, ranging from 0 to 1. The INME is used as a key performance indicator, where lower INME values indicate higher accuracy in landmark prediction. From Table 1, it can be observed that the baseline model, without any synthetic augmentation, achieves an INME of 4.67. As the augmentation ratio increases from 0.1 to 1, the INME fluctuates slightly between 4.63 and 4.74, indicating that different levels of augmented data have varied effects on model accuracy.

Further insight into the effect of ControlNet-augmented data on model learning is illustrated in Figure 4(a) and 4(b), which display raw and moving average of INME values over training iterations, clarifying long-term performance trends. During the initial third of the training iterations, INME decreases sharply from approximately 6.0, demonstrating that the model rapidly adapts to the training data. After this initial drop, INME stabilizes between 4.8 and 5.4 during the second third of the iterations, with a general downward trend, indicating continued model improvement.

The impact of varying the Lambda parameter on INME is also notable. Lower Lambda values (below 0.5) are associated with lower INME, suggesting that selecting an optimal Lambda value can significantly enhance model performance. After the final third of the training iterations, INME converges to a steady range of 4.6 to 4.8 across all Lambda values, demonstrating that the model has achieved stable landmark prediction accuracy. The moving average in Figure 4(b) effectively smooths out raw INME fluctuations, making the trend of performance improvement more apparent. The experimental results highlight the effectiveness of using ControlNet-augmented data and the importance of tuning Lambda to achieve optimal performance in facial landmark detection. The analysis underscores that the integration of carefully chosen synthetic data ratios, along with an optimal Lambda, can enhance model robustness and precision in landmark localization.

Table 1. Interocular normalized mean error (INME) of the facial landmark detection model for varying ControlNet-based augmentation ratios (λ). Lower INME indicates higher landmark prediction accuracy

Ratios	INME↓
0 (Baseline)	4.67
0.1	4.68
0.2	**4.63**
0.3	4.68
0.4	4.74
0.5	**4.63**
0.6	4.69
0.7	4.71
0.8	4.70
0.9	4.68
1	4.73



(a) (b)

Figure 4. Impact of Lambda (λ) on INME during training: (a) raw INME values per iteration for different Lambda settings, and (b) corresponding moving-average curves, highlighting how tuning Lambda influences convergence and landmark localization accuracy

In summary, the experimental results highlight the effectiveness of using ControlNet-augmented data and the importance of tuning Lambda to achieve optimal performance in facial landmark detection. The analysis underscores that the integration of carefully chosen synthetic data ratios, along with an optimal Lambda, can enhance model robustness and precision in landmark localization. Additionally, balancing the amount of synthetic and real data ensures diverse training samples without introducing excessive noise, further stabilizing model convergence.

4. CONCLUSION

This study highlights the effectiveness of ControlNet-based data augmentation in enhancing the accuracy and robustness of facial landmark detection. By integrating ControlNet-generated synthetic images with real data from the 300 W dataset, the proposed approach addresses critical challenges in landmark detection, including variations in lighting, pose, and facial expressions. The experimental results demonstrate that augmenting training datasets with synthetic data significantly reduces the INME, thereby improving landmark localization accuracy.

Furthermore, the findings emphasize the importance of optimizing the ratio of synthetic to real data and fine-tuning model parameters, such as Lambda, to achieve maximum performance gains. Careful selection of synthetic-to-real data proportions ensures that the model learns from diverse conditions without being overwhelmed by artificial samples. In addition, adjusting Lambda allows for controlling the trade-off between reconstruction accuracy and regularization, which ultimately helps stabilize training and prevents overfitting.

This methodology holds considerable promise for broader applications in computer vision tasks that require precise feature localization. In particular, fields such as facial expression recognition benefit from reliable landmark positioning, and improved 3D facial modeling depends on accurate feature correspondence. Future research should focus on refining synthetic data generation techniques, exploring more advanced generative models, and extending this approach to other areas of facial analysis to further validate its generalizability.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the use of service and facilities of the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University.

FUNDING INFORMATION

This study receives funding from the National Science and Technology Development Agency, Thailand, under the Prototype Research Grant Scheme JRA-CO-2565-17792-TH.

REFERENCES

[1] M. Bilal, S. Razzaq, N. Bhowmike, A. Farooq, M. Zahid, and S. Shoaib, “Facial recognition using hidden Markov model and convolutional neural network,” *AI*, vol. 5, no. 3, pp. 1633–1647, Sep. 2024, doi: 10.3390/ai5030079.

[2] M. Rane *et al.*, “Face recognition using convolutional neural network (CNN),” in *Smart Trends in Computing and Communications. SMART 2023*, T. Senjyu, C. So–In, and A. Joshi, Eds., Lecture Notes in Networks and Systems, vol. 645. Singapore: Springer, 2023, pp. 203–214, doi: 10.1007/978-981-99-0769-4_20.

[3] T. Kuarkamphun and C. Ratanavilisagul, “Face recognition using skin color segment and modified binary particle swarm optimization,” *ICSEC 2022 - International Computer Science and Engineering Conference 2022*, pp. 66–71, 2022, doi: 10.1109/ICSEC56337.2022.10049354.

[4] P. A. Javier Orlando, J. M. Robinson, and J. E. M. Baquero, “Comparison of convolutional neural network models for user’s facial recognition,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 192–198, 2024, doi: 10.11591/ijece.v14i1.pp192-198.

[5] A. Halim *et al.*, “Facial expressions analysis to evaluate the level of students’ understanding,” in *2023 Intelligent Methods, Systems, and Applications (IMSA)*, Jul. 2023, pp. 424–429, doi: 10.1109/IMSA58542.2023.10217489.

[6] D. Zhao, J. Wang, H. Li, and D. Wang, “Landmark-based adaptive graph convolutional network for facial expression recognition,” *IEEE Access*, vol. 12, pp. 136088–136102, 2024, doi: 10.1109/ACCESS.2024.3463176.

[7] Kavita and R. S. Chhillar, “Performance analysis of deep unified model for facial expression recognition using convolution neural network,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 4, pp. 4046–4054, 2024, doi: 10.11591/ijece.v14i4.pp4046-4054.

[8] R. Bharadwaj, R. Borse, P. Ahire, P. Aldhar, and A. Hoshing, “Dynamic 3D facial expression reconstruction from images,” in *Proceedings - 2024 5th International Conference on Image Processing and Capsule Networks, ICIPCN 2024*, 2024, pp. 66–72, doi: 10.1109/ICIPCN63822.2024.00020.

[9] M. Wang and H. Wang, “Facial photo-guided head anatomy modeling based on deep learning and 2D/3D shape prior model registration,” *Smart Innovation, Systems and Technologies*, vol. 374, pp. 247–257, 2024, doi: 10.1007/978-981-99-7011-7_20.

[10] J. Ling, Z. Wang, M. Lu, Q. Wang, C. Qian, and F. Xu, “Semantically disentangled variational autoencoder for modeling 3D facial details,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 8, pp. 3630–3641, 2023, doi: 10.1109/TVCG.2022.3166666.

[11] A. Acquisti, R. Gross, and F. Stutzman, “Face recognition and privacy in the age of augmented reality,” *Journal of Privacy and Confidentiality*, vol. 6, no. 2, 2014, doi: 10.29012/jpc.v6i2.638.

[12] 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, doi: 10.1109/TIP.2019.2899267.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models—their training and application,” *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995, doi: 10.1006/cviu.1995.1004.

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, Jun. 2001, doi: 10.1109/34.927467.

[15] Xiangxin Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2879–2886, doi: 10.1109/CVPR.2012.6248014.

[16] D. Lee, H. Park, and C. D. Yoo, “Face alignment using cascade Gaussian process regression trees,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 4204–4212, doi: 10.1109/CVPR.2015.7299048.

[17] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, “Face alignment by explicit shape regression,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2887–2894, doi: 10.1109/CVPR.2012.6248015.

[18] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 532–539, doi: 10.1109/CVPR.2013.75.

[19] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, “Human pose estimation using body parts dependent joint regressors,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3041–3048, doi: 10.1109/CVPR.2013.391.

[20] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1867–1874, doi: 10.1109/CVPR.2014.241.

[21] B. Jin, X. Xu, and Y. Zhang, “Thermal coal futures trading volume predictions through the neural network,” *Journal of Modelling in Management*, vol. 20, no. 2, pp. 585–619, Feb. 2025, doi: 10.1108/JM2-09-2023-0207.

[22] S. Ittisoponpisan, C. Kaipan, S. Ruang-on, R. Thaiphan, and K. Songsri-in, “Pushing the accuracy of Thai food image classification with transfer learning,” *Engineering Journal*, vol. 26, no. 10, pp. 57–71, Oct. 2022, doi: 10.4186/ej.2022.26.10.57.

[23] B. Jin, X. Xu, and Y. Zhang, “Peanut oil price change forecasts through the neural network,” *foresight*, vol. 27, no. 3, pp. 595–612, Apr. 2025, doi: 10.1108/FS-01-2023-0016.

[24] K. Songsri-in, M. Rattaphun, S. Kaewchada, and S. Ruang-on, “DualFaceNet: augmentation consistency for optimal facial landmark detection and face mask classification,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 3, pp. 3228–3239, Sep. 2024, doi: 10.11591/ijai.v13.i3.pp3228-3239.

[25] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks),” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 1021–1030, doi: 10.1109/ICCV.2017.116.

[26] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3476–3483, doi: 10.1109/CVPR.2013.446.

[27] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Lecture Notes in Computer Science. Cham: Springer, 2014, pp. 94–108, doi: 10.1007/978-3-319-10599-4_7.

[28] A. Jourabloo and X. Liu, “Large-pose face alignment via CNN-based dense 3D model fitting,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA, 2016, pp. 4188-4196*,* doi: 10.1109/CVPR.2016.454.

[29] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, “Face alignment across large poses: A 3D solution,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 146–155, doi: 10.1109/CVPR.2016.23.

[30] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 2235–2245, doi: 10.1109/CVPR.2018.00238.

[31] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, “Look at boundary: A boundary-aware face alignment algorithm,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 2129–2138, doi: 10.1109/CVPR.2018.00227.

[32] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[33] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 2020-December, 2020.

[34] D. Ryu and J. C. Ye, “Pyramidal denoising diffusion probabilistic models,” *arXiv:2208.01864*, 2022.

[35] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 3813–3824, doi: 10.1109/ICCV51070.2023.00355.

[36] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: the first facial landmark localization challenge,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 397–403, doi: 10.1109/ICCVW.2013.59.

BIOGRAPHIES OF AUTHORS

Kritaphat Songsri-in finished MEng and Ph.D. in computing from Imperial College London in 2011 and 2020, respectively. Currently, he is an assistant professor in the Department of Computer Science at Nakhon Si Thammarat Rajabhat University, Thailand. His research interests include machine learning, deep learning, and computer vision. He has published in and is a reviewer for multiple international conferences and journals such as IEEE Transactions on Image Processing and IEEE Transactions on Information Forensics and Security. He was a recipient of the Royal Thai Government Scholarship covering his undergraduate and postgraduate degrees in 2010. He received the Best Student Paper Awards at the IEEE 13th International Conference for Automatic Face and Gesture Recognition (FG2018) and the 6th National Science and Technology Conference (NSCIC2021). In 2021, his Ph.D. thesis received an award from the National Research Council of Thailand (NRCT). He can be contacted at email: kritaphat_son@nstru.ac.th.

Munlika Rattaphun received the B.S. degree in computer science from Thaksin University, Songkhla, Thailand, in 2009, the M.S. degree in computer science from Prince of Songkla University, Songkhla, Thailand, in 2011, and the Ph.D. degree in computer science and information engineering from National Chiayi University, Chaiyi, Taiwan, in 2022. She is currently a lecturer at the Department of Computer Science, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Nakhon Si Thammarat, Thailand. Her current research interests include machine learning, nearest-neighbor search, and recommender systems. She can be contacted at email: munlika_rat@nstru.ac.th.

Sopee Kaewchada received the B.Sc. degree in computer science from Rajabhat Phetchaburi Institute, Thailand, in 1997 the M.S. degree in management of information technology from Walailak University, Thailand, in 2003, and the Ph.D. degree in creative innovation in science and technology, Nakhon Si Thammarat Rajabhat University, Thailand, in 2023. Currently, she is an assistant professor at the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. She can be contacted at email: sopee_kae@nstru.ac.th.

Sunisa Kidjaideaw received the B.Sc. degree in computer science from Nakhon Si Thammarat Rajabhat University, Thailand, in 2006 the M.S. degree in management of information technology from Walailak University, Thailand, in 2010, and the Ph.D. degree in information technology, Sripatum University, Thailand, in 2020. Currently, she is an assistant professor at the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. She can be contacted at email: sunisa_kid@nstru.ac.th.

Sangjun Ruang-On received the B.B.A. degree in business computer from Sripatum University, Thailand, in 1994, the M.Sc. degree in information technology from Sripatum University, Thailand, in 2003. Currently, she is an assistant professor at the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. She can be contacted at email: sangjun_rua@nstru.ac.th.

Wichit Sookkhathon received the B.Sc. degree in computer science from Rajabhat Songkhla Institute, Thailand, in 1996, the M.Sc. degree in information technology from Universiti Utara Malaysia, in 2004, and Ph.D. degree in quality information technology from Phetchaburi Rajabhat University, in 2013. Currently, he is an assistant professor at the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. He can be contacted at email: wichit_soo@nstru.ac.th.

Patompong Chabplan received the B.Sc. degree in computer science from Nakhon Si Thammarat Rajabhat University, Thailand, in 2005 the M.S. degree in management of information technology from Walailak University, Thailand, in 2010, and the Ph.D. degree in information technology, King Mongkut's University of Technology North Bangkok, Thailand, in 2020. Currently, he is a lecturer at the Department of Information Technology and Digital Innovation of the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. He can be contacted at email: patompong_cha@nstru.ac.th.