JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 19, No. 2, June 2025, pp. 72-80

SESAME: Selective Enhancement of Soft-argmax Estimation for Challenging Facial Landmark Detection via Adaptive Masking-Based Representation Learning

Yeeun Choi
Division of Business, Chungnam National University, Daejeon, Korea
ye20039@o.cnu.ac.kr

Hyeonmin Jeong and Chaemin Yoo
Department of Artificial Intelligence, Chungnam National University, Daejeon, Korea
hyunmin9808@o.cnu.ac.kr, 202102552@o.cnu.ac.kr

Gwanghee Lee and Kyoungson Jhang*
Department of Computer Engineering, Chungnam National University, Daejeon, Korea
manggu251@o.cnu.ac.kr, sun@cnu.ac.kr

Abstract

Facial landmark detection is a task that involves estimating landmark points such as eyes, nose, and mouth in facial images, and is utilized in various applications including facial recognition, emotion analysis, expression recognition, and person identification. Among implementation methods, coordinate regression faces a significant challenge with notably lower prediction accuracy in specific areas such as the jaw and below the ears. We propose a selective enhancement of soft-argmax estimation (SESAME) as a technique to address these limitations. SESAME consists of three stages. In the preliminary step, we train a randomly initialized facial landmark detection model using the WFLW facial landmark dataset to identify areas of poor performance. During the pre-training step, we focus on enhancing representation learning by applying intensive masking to poorly performing landmark regions in random facial input images from WFLW, followed by reconstruction. In the fine-tuning step, we further train the enhanced model using the WFLW dataset. Experimental results show significant improvement in the prediction accuracy of traditionally challenging areas, particularly in jawline regions where the normalized mean error (NME) substantially decreased from 6.547% to 6.184%. The overall average NME across all regions improved from 4.397% to 4.351%, demonstrating enhanced overall performance.

Category: Computer Vision and Graphics

Keywords: Deep learning; Computer vision; Facial landmark detection

Open Access **http://dx.doi.org/10.5626/JCSE.2025.19.2.72** **http://jcse.kiise.org**

Received 6 February 2025; **Accepted** 16 October 2025

*Corresponding Author

pISSN: 1976-4677 eISSN: 2093-8020

I. INTRODUCTION

Facial landmark detection is a task that predicts landmark coordinates for specific facial regions from input facial images [1]. Traditionally, heatmap regression has been the predominant approach, which outputs landmark positions as probability maps and estimates final coordinates through post-processing steps [2, 3]. Coordinate regression, on the other hand, incorporates the post-processing steps of heatmap regression into the neural network itself, directly predicting landmark coordinates [4-6].

While coordinate regression offers advantages such as elimination of post-processing steps and easier reduction of computational complexity, it suffers from significant performance variations across different landmarks, resulting in lower overall performance.

To address this limitation, we propose focusing on intensive learning of underperforming regions. Our approach, SESAME (selective enhancement of soft-argmax estimation), leverages masked autoencoder (MAE) [7] to selectively enhance the representation of underperforming regions, thereby improving the overall performance of coordinate regression methods.

The remainder of this paper is organized as follows. Section II reviews related works, including coordinate regression and MAE-based representation learning. Section III details the proposed methodology. Section IV presents experimental results, including quantitative evaluations and performance comparisons of SESAME.

II. RELATED WORKS

A. Heatmap Regression and Coordinate Regression

Heatmap regression has been the most extensively studied approach in facial landmark detection. Notable works include adaptive wing loss (AWing) [8] and sparse local patch transformer (SLPT) [9], which primarily focus on methods for heatmap generation. AWing is a study on loss functions aimed at improving coordinate prediction accuracy. Building upon the existing wing loss, it is designed to be more sensitive to samples with small errors, thereby encouraging the model to learn more precise predictions. SLPT is a transformer-based network model that learns part-based structure of objects. Unlike traditional convolutional neural network (CNN)-based models, SLPT is designed to simultaneously learn both global and local object information.

On the other hand, coordinate regression has been relatively less explored in facial landmark detection. A notable example from the field of human pose estimation is the soft-argmax layer (SAM) [10]. SAM proposed a differentiable coordinate transformation layer using soft-argmax. Although coordinate regression models directly predict landmark coordinates, it has been observed that SAM implicitly generates intermediate indirect heatmaps within the network, revealing a latent heatmap-based representation.

B. Representation Learning with Masked Autoencoder

Research has been conducted on how machine learning models learn meaningful representations from data [11]. These studies focus on enabling models to learn the inherent structure of data and effectively represent various factors such as lighting, orientation, and texture through disentanglement [12]. A fundamental method in this domain is the autoencoder, which compresses an input image into a lower-dimensional latent space and reconstructs it, embedding the learned representations within the compressed dimensions [13].

MAE enhances representations by masking certain regions of the input image and learning to reconstruct them using contextual information. We utilized this approach by focusing the masking and reconstruction process on landmarks with lower performance, enabling intensive learning of representations for these specific regions. This helps mitigate the limitation of traditional coordinate regression methods where certain landmarks show degraded prediction performance, and is designed to enable the model to understand the overall facial structure in a more consistent manner.

Among existing research, SCE-MAE [14] proposed a method for learning representations from largescale unlabeled data using self-supervised learning-based MAE techniques. While such methods enhance generalization through additional data exposure, our approach diverges by applying selective masking to underperforming landmarks. To achieve this, we utilize a labeled dataset, WFLW.

C. Hourglass Networks

We adopted hourglass networks as our neural network architecture since it is widely used for landmark coordinate prediction and enables appropriate comparison with existing research. Hourglass networks excel at learning complex features at various scales through repeated upsampling and downsampling operations [2, 15]. As shown in Fig. 1, multiple hourglass modules are stacked together, allowing the model to progressively learn more complex and detailed features. Each stack extracts feature maps from the input image, and the final landmark coordinates are predicted based on the feature maps generated by the last stack.

Hourglass networks traditionally operate by generating heatmaps for predicting facial landmark coordinates. These heatmaps represent the probability distribution of a landmark's position in the form of a matrix, with one heatmap corresponding to each landmark [2]. For instance,

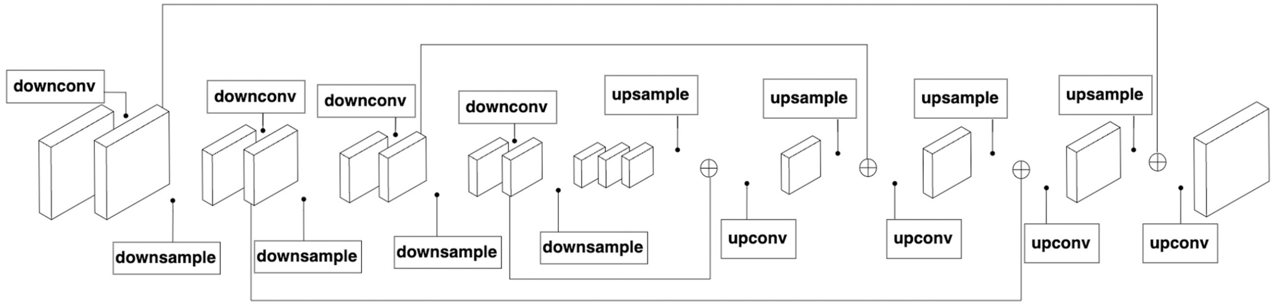

Fig. 1. Hourglass Module. This module represents the basic structure of the hourglass network and consists of a symmetrical encoder-decoder architecture. The left path repeatedly applies downsampling and convolution operations, learning representations while progressively reducing spatial resolution. The right path restores resolution through upsampling, and features extracted during the encoding process are transmitted to corresponding upsampling stages through skip connections.

if there are 98 facial landmarks, the model outputs a total of 98 heatmaps. These heatmaps capture crucial facial features by synthesizing information extracted from various resolutions, forming the foundation for coordinate prediction.

Although hourglass networks are typically used for heatmap regression, in this study, we employed them for coordinate regression using SAM [10].

D. Soft-argmax for Coordinate Estimation

Soft-argmax, the most commonly used post-processing method for converting heatmaps to coordinates, estimates landmark coordinates based on differentiable probability distributions [10]. Soft-argmax converts all pixel values in the heatmap into probability distributions and then calculates the final landmark coordinates through weighted averaging. SAM [10] proposed using soft-argmax as a learnable neural network layer, and we employ a coordinate regression model that directly predicts coordinates using SAM.

Specifically, the formula for calculating the predicted coordinate values is shown in Eq. (1).

$$\text{Soft-argmax}\,(X) = \sum_i \frac{e^{x_i}}{\sum_j e^{x_j}} i \tag{1}$$

III. METHODOLOGY: SESAME

The proposed methodology, SESAME, leverages a Gaussian noise masking technique to improve the performance of facial landmark detection. This approach is structured into three main stages—the preliminary step, the pre-training step, and the fine-tuning step.

In the preliminary step, we obtain region-specific normalized mean error (NME) values from a preliminarily trained model. In the pre-training step, these region-specific NME values are normalized into probability distributions to set the masking probability for each region. Regions with higher NME values (i.e., lower performance) are masked more frequently using a Gaussian noise mask. The model is then trained to reconstruct the masked regions, encouraging it to focus on enhancing the representation of the poorly performing landmarks.

In the fine-tuning step, the pre-trained MAE is used as a backbone to train the model to predict landmark coordinates. This enables SESAME- based on the initial NME values of landmarks.

A. Preliminary Step: Identifying Low-Performance Landmarks

Fig. 2 shows the neural network architecture used in this step, which integrates the SAM into an hourglass

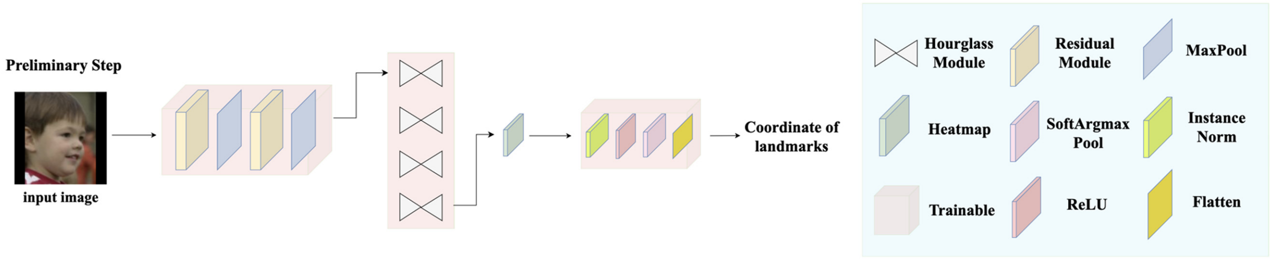

Fig. 2. SAM-based model (preliminary step). This step performs initial training of a neural network that directly predicts facial landmark coordinates. After learning spatial features through multiple hourglass modules, it applies soft-argmax operations on the generated heatmaps to predict the final landmark coordinates.

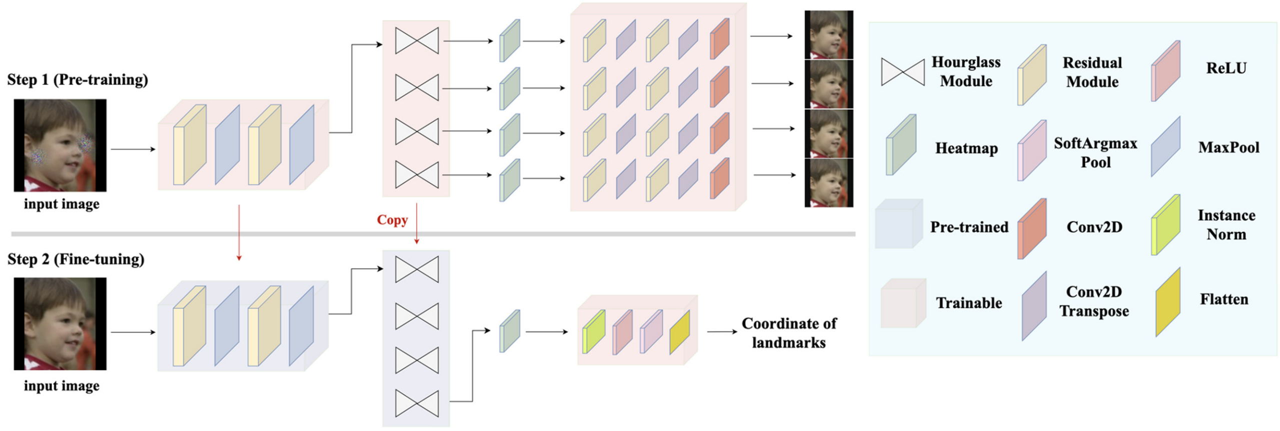

Fig. 3. Selective enhancement of soft-argmax estimation (SESAME; two steps after the preliminary step). This method consists of pre-training and fine-tuning stages, enhancing performance by separating the facial landmark detection training process into two phases. In the pre-training stage, masked image reconstruction is utilized to train the hourglass network's reconstruction capability for specific regions. In the fine-tuning stage, the entire network is fine-tuned based on the pretrained model to ultimately predict landmark coordinates. The enhanced representations learned during the pre-training stage contribute to more accurate predictions, particularly in regions that originally showed lower performance.

network to directly predict facial landmark coordinates. In this preliminary step, we initially train a randomly initialized neural network to identify regions with low performance. The region-specific NME values obtained from this training process are normalized into probability distributions, which are then used as masking probabilities in the pre-training step as shown in Fig. 3. We will refer to the model trained in this step as the SAM-based model.

B. Pre-training Step: Masked Image Reconstruction

The intermediate layers of hourglass networks are trained by configuring their outputs into three-channel outputs that reconstruct masked images. The pre-training step is shown in Step 1 of Fig. 3. Regions with high NME values obtained from the preliminary step are probabilistically masked more frequently, leading to more intensive learning of representations for these regions during the reconstruction process. SESAME shares the general principle of MAE [7] in that both methods mask specific regions of an image and train the model to reconstruct them, but their designs are fundamentally different. MAE divides an image into fixed-size patches (e.g., 16×16) and randomly masks a large portion of them. In contrast, SESAME applies Gaussian noise masks (Fig. 4(a)) to regions around facial landmarks. The masking probability for each landmark region is adaptively assigned according to its prediction error (NME). Furthermore, SESAME employs a CNN-based hourglass backbone optimized for facial landmark regression, explicitly guiding the network to strengthen feature learning for underperforming landmarks. These aspects distinguish SESAME from the random patch masking paradigm of conventional MAE and demonstrate that its pre-training process is specifically tailored for the facial landmark detection task.

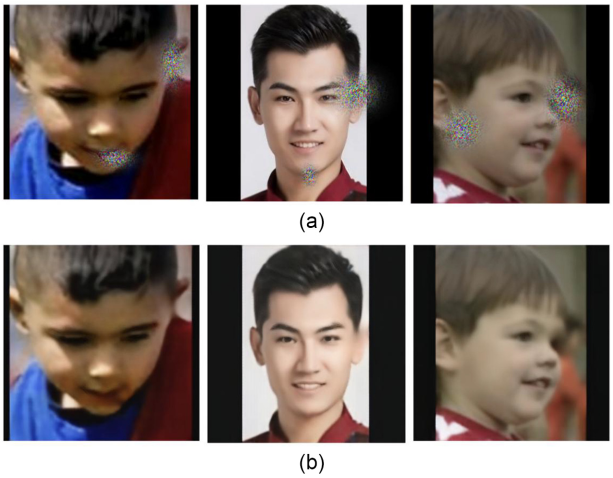

Fig. 4. Reconstructions of some photos of masked WFLW dataset: (a) Gaussian noise has been selectively applied to obscure specific facial landmarks, (b) the corresponding reconstructed images where the model successfully restores the occluded regions.

1) Gaussian Noise Mask

As shown in Fig. 4(a) and 4(b), we applied Gaussian noise to mask specific landmark regions. To ensure smooth transitions between masked and unmasked areas, we used Gaussian blending. We randomly varied the standard deviation (σ) of the Gaussian distribution, designing the masks to appear in different sizes and intensities. This causes the model to learn to reconstruct landmark regions with lost information using surrounding contextual information [16].

2) Reconstruction Loss Function

We used mean squared error (MSE) loss to reconstruct the masked images. To focus the model on reconstructing the masked regions, we incorporated a mask map (λ) into the loss calculation. We reduced the contribution of the unmasked regions by multiplying their errors by 0.01 while maintaining the full error contribution for the masked areas. The final loss is obtained by averaging these two values after adding them. This enables the model to focus on accurately reconstructing the masked regions [17]. The modified MSE loss is presented in Eq. (2).

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_i\left(y_i - \hat{y}_i\right)^2 + 0.01\left(1-\lambda_i\right)\left(y_i - \hat{y}_i\right)^2\right) \quad (2)$$

C. Fine-tuning Step: Landmark Coordinate Estimation

In the fine-tuning step, the model trained for masked image reconstruction in the pre-training step (as shown in Fig. 3) is further trained to predict landmark coordinates. We removed the intermediate layers responsible for image reconstruction and inserted a SAM to enable direct landmark coordinate estimation. By fine-tuning the model with the enhanced representations learned during the pre-training phase, the backbone network significantly improves the prediction accuracy, particularly in regions that previously exhibited poor performance.

IV. EXPERIMENTS

A. Experimental Setup

In the preliminary stage, the hyperparameters were set as follows: initial learning rate of 0.001, 300 epochs, Adam optimizer with β1 = 0.9, β2 = 0.999, and ε = 1e-07. Exponential decay was applied for learning rate reduction, and MAE loss was used as the loss function.

In the pre-training stage, a customized MSE loss function was used to train only the representations of noise-masked image regions. The initial learning rate was set to 0.001, with the same Adam optimizer parameters: β1 = 0.9, β2 = 0.999, and ε = 1e-07.

Finally, in the fine-tuning stage, the hyperparameters were configured with an initial learning rate of 0.001, 300 epochs, and Adam optimizer (β1 = 0.9, β2 = 0.999) with exponential decay applied. MAE loss was used as the loss function for this training phase as well.

B. Wider Facial Landmarks in-the-Wild (WFLW)

The WFLW dataset [18] is designed for facial landmark detection, consisting of facial images and corresponding landmark coordinates for each image. It contains a total of 10,000 images with 98 annotated facial landmarks per image. The dataset is divided into 7,500 images for training and 2,500 images for testing.

C. Data Augmentation Strategy

All images are cropped according to the facial bounding boxes provided in the dataset. For the training set, the bounding box is randomly expanded by 10%–50% to include more surrounding context, whereas for the validation set it is consistently expanded by 20%. The cropped regions are then resized and padded to a fixed resolution of 256×256. After this process, additional augmentations were applied only to the training set as follows:

- Color jitter: adjust saturation and contrast
- Brightness shift: pixel intensity ±0.3
- JPEG artifact: random quality between 20 and 100
- Channel transfer: grayscale conversion with $p = 0.2$
- Occlusion: rectangular mask filled with noise or solid color
- Horizontal flip: $p = 0.5$ with 98-point mirror mapping
- Rotation: random angle within ±20°
- Translation: random shift up to 5% of image size
- Scaling: preserve aspect ratio, resize and pad

D. Metric

As shown in Eq. (3), we used NME [19] as the evaluation metric for facial landmark detection to measure the difference between the actual landmark coordinates and those predicted by the model.

$$\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|y_i - \hat{y}_i\|_2}{d} \quad (3)$$

The NME is defined as shown in Eq. (3) and serves as an indicator of how closely the predicted landmark coordinates align with the ground truth coordinates. The difference between the actual coordinates and the predicted coordinates is normalized by dividing by the interocular distance d (the distance between the eyes) to minimize variations due to the size of the face, as defined in Eq. (4).

$$d = \sqrt{\left(x_{\text{right}} - x_{\text{left}}\right)^2 + \left(y_{\text{right}} - y_{\text{left}}\right)^2} \quad (4)$$

This measurement is calculated using the coordinates of the outer corners of both eyes, specifically defined as the Euclidean distance between the coordinates of the outer corner of the left eye and the outer corner of the right eye.

E. Experimental Results

Fig. 5 compares the NME values between the SAM-based model (preliminary step) and SESAME. The results indicate a significant improvement in prediction accuracy, particularly in the facial contour regions where the initial performance was lower. This substantial reduction in NME demonstrates that representation learning through landmark masking and reconstruction processes worked effectively.

Table 1 presents the NME comparison results for each keypoint category, providing a quantitative evaluation of the difference in NME between the SAM-based model and SESAME. In this study, we categorized facial landmarks into 10 major groups: jawline and facial outline, eyebrows, eyes, nasal bridge, nose base, outer and inner lips, and pupils. We analyzed the NME for each category to assess the proposed method's impact.

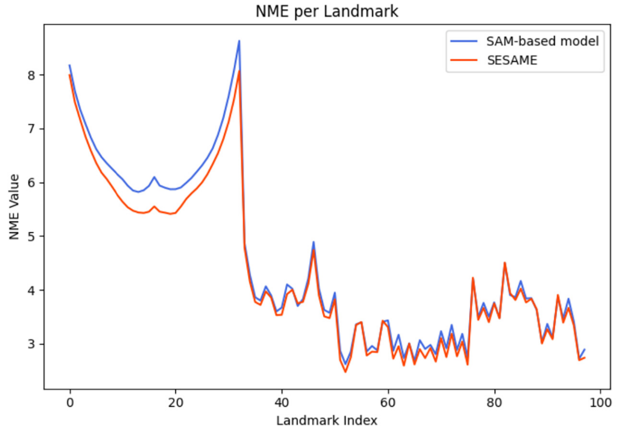

Fig. 5. Comparison of normalized mean error (NME) between the model in SAM-based model and SESAME. The blue curve represents the SAM-based model (preliminary step), and the red curve represents SESAME.

The jawline and facial contour region (marks 0–32) showed a significant reduction in NME with a change rate of -5.537%. This improvement is related to the characteristics of the jawline region. Because the jawline has a low texture contrast with the background, it showed high error rates in the baseline model. SESAME applied error-driven masking to this challenging region, forcing the model to reconstruct and learn it repeatedly, which reduced prediction errors. Additionally, the left and right eyes (landmarks 60–75) and nose bridge (landmarks 51–54) showed NME change rates of -4.317%, -4.335%, and -3.706%, respectively. These results demonstrate that the proposed methodology effectively mitigates the performance variability across different landmarks commonly encountered in traditional coordinate regression approaches. By focusing on enhancing the representation of underperforming landmarks, SESAME achieves more consistent and accurate landmark detection across various facial regions.

Furthermore, as shown in Table 1, landmarks 40–60, which performed well in the SAM-based model (preliminary step), maintained good performance in SESAME. Ultimately, our proposed method achieved a lower overall NME, decreasing from 4.397% to 4.351%.

Table 2 compares the performance of our method with existing approaches. While the baseline methods are primarily based on heatmap regression, both the SAM-based model and SESAME utilize a coordinate regression

Table 1. NME comparison by keypoint

Part of face	Number of labels	NME (%)		NME change	Change rate (%)
		Traditional	Proposed		
Facial outline	0–32	6.55	6.18	-0.363	-5.54
Left eyebrow	33–41	4.02	3.92	-0.1	-2.48
Right eyebrow	42–50	3.98	3.90	-0.083	-2.08
Left eye	60–67	2.98	2.85	-0.129	-4.32
Right eye	68–75	3.01	2.88	-0.13	-4.34
Nose bridge	51–54	2.92	2.81	-0.108	-3.71
Nose below	55–59	3.10	3.06	-0.038	-1.21
Outer lips	76–87	3.86	3.82	-0.041	-1.06
Inner lips	88–95	3.46	3.41	-0.052	-1.49
Pupils	96–97	0.47	0.45	-0.015	-3.16

This is the comparison of normalized mean error (NME) values between the traditional coordinate regression method and the coordinate regression method applying our proposed technique for each keypoint. The NME change amount represents the degree of NME reduction in the proposed method compared to the traditional method, while the change rate shows this reduction as a percentage relative to the traditional method.

Table 2. Normalized mean error (NME) on the WFLW (Fullset) dataset compared to hourglass networks

Method	NME (%)
LAB [18]	5.27
DeCaFa [20]	4.62
Awing [8]	4.36
SLPT [9]	4.14
STAR loss [21]	4.02
RHT-R [22]	4.01
KeyPosS [23]	4.00
SAM-based model (preliminary step)	4.40
SESAME	4.35

framework. As shown in Table 2, SESAME recorded a slightly higher NME (0.211 percentage points) compared to the more recent SLPT [9]. SLPT learns landmark-to-landmark correlations by masking feature maps extracted from a pre-trained CNN at specific landmark regions, rather than directly regressing coordinates.

Instead, it generates heatmaps, from which landmark coordinates must be subsequently extracted through an additional post-processing step. This design also relies on a backbone network to capture patch-level features around each landmark. In contrast, SESAME does not employ explicit correlation modeling or a Transformer backbone; instead, it adopts a CNN-based hourglass architecture within a coordinate regression framework. Similar to SLPT, SESAME emphasizes challenging landmarks, but it achieves this through the application of Gaussian masks centered on high-error landmark regions. While this design choice explains SESAME's slightly lower accuracy compared to SLPT, it simultaneously demonstrates that even with a simpler architecture, selectively improving underperforming landmarks can enhance overall generalization performance. Notably, SESAME outperformed other methods, achieving 0.919 percentage points lower NME than LAB [18], 0.269 points lower than DeCaFa [20], and a further 0.009 points lower than AWing [8].

Meanwhile, recent studies such as STAR loss [21], RHT-R [22], and KeyPosS [23] show incremental performance improvements within a narrow range of 0.01 percentage points. Considering that SESAME improved the performance from 4.397% in the SAM-based model to 4.351%—a 0.046 percentage point reduction—our method demonstrates an effective enhancement in facial landmark detection accuracy.

V. CONCLUSIONS

In this study, we proposed a novel approach to improving facial landmark detection performance by selectively enhancing landmark-specific representations using MAE. First, we extract initial NME for each landmark through training a coordinate regression neural network model that uses hourglass networks as a backbone and SAM for final coordinate prediction. Then, we attempted to improve performance by selectively enhancing the hourglass network's representations through selective Gaussian noise masking and reconstruction around landmarks with high NME on WFLW dataset images, followed by fine-tuning. Experimental results confirmed significant improvements in NME performance for landmark regions that initially showed poor NME performance compared to the initial neural network model.

Ultimately, the proposed method overcame the limitations of existing approaches and enabled more precise predictions by enhancing the representation of landmarks that were difficult to learn. We confirmed that facial landmark detection performance could be effectively improved by introducing a Gaussian noise masking technique using NME-based probability distributions and a pre-training method through landmark target reconstruction. This approach shows high applicability in various applications such as facial recognition and emotion analysis, and is expected to contribute to future research in pose estimation and facial analysis.

Furthermore, in future research, we plan to attempt additional performance improvements by utilizing self-supervised learning techniques to leverage all unlabeled facial images. Through this, we aim to explore ways to build more generalized facial landmark detection models. However, the fundamental performance limitations of coordinate regression methods compared to heatmap regression methods remain a challenge to be addressed.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

ACKNOWLEDGEMENTS

This work was supported by the research fund of Chungnam National University.

REFERENCES

1. R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," 2016 [Online]. Available:
2. J. Zhang, H. Hu, and G. Shen, "Joint stacked hourglass network and salient region attention refinement for robust face alignment," *ACM Transactions on Multimedia Computing,*

Communications, and Applications, vol. 16, np. 1, article no. 10, 2020. https://doi.org/10.1145/3374760

3. A. Bulat, E. Sanchez, and G. Tzimiropoulos, “Subpixel heatmap regression for facial landmark localization,” 2021 [Online]. Available: https://arxiv.org/abs/2111.02360.
4. A. P. Fard and M. H. Mahoor, “ACR loss: adaptive coordinate-based regression loss for face alignment,” in *Proceedings of 2022 26th International Conference on Pattern Recognition (ICPR)*, Montreal, Canada, 2022, pp. 1807-1814. https://doi.org/10.1109/ICPR56361.2022.9956683
5. F. Nian, T. Li, B. K. Bao, and C. Xu, “Relative coordinates constraint for face alignment,” *Neurocomputing*, vol. 395, pp. 119-127, 2020. https://doi.org/10.1016/j.neucom.2017.12.071
6. J. Wan, H. Xi, J. Zhou, Z. Lai, W. Pedrycz, X. Wang, and H. Sun, “Robust and precise facial landmark detection by self-calibrated pose attention network,” *IEEE Transactions on Cybernetics*, vol. 53, no. 6, pp. 3546-3560, 2023. https://doi.org/10.1109/TCYB.2021.3131569
7. K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 15979-15988. https://doi.org/10.1109/CVPR52688.2022.01553
8. X. Wang, L. Bo, and L. Fuxin, “Adaptive wing loss for robust face alignment via heatmap regression,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 6970-6980. https://doi.org/10.1109/ICCV.2019.00707
9. J. Xia, W. Qu, W. Huang, J. Zhang, X. Wang, and M. Xu, “Sparse local patch transformer for robust face alignment and landmarks inherent relation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 4042-4051. https://doi.org/10.1109/CVPR52688.2022.00402
10. D. C. Luvizon, H. Tabia, and D. Picard, “Human pose regression by combining indirect part detection and contextual information,” *Computers & Graphics*, vol. 85, pp. 15-22, 2019. https://doi.org/10.1016/j.cag.2019.09.002
11. Y. Bengio, A. Courville, and P. Vincent, “Representation learning: a review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013. https://doi.org/10.1109/TPAMI.2013.50
12. X. Wang, H. Chen, S. A. Tang, Z. Wu, and W. Zhu, “Disentangled representation learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9677-9696, 2024. https://doi.org/10.1109/TPAMI.2024.3420937
13. K. T. Baghaei, A. Payandeh, P. Fayyazsanavi, S. Rahimi, Z. Chen, and S. B. Ramezani, “Deep representation learning: fundamentals, perspectives, applications, and open challenges,” 2022 [Online]. Available: https://arxiv.org/abs/2211.14732.
14. K. Yin, V. Rao, R. Jiang, X. Liu, P. Aarabi, and D. B. Lindell, “SCE-MAE: selective correspondence enhancement with masked autoencoder for self-supervised landmark estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, pp. 1313-1322. https://doi.org/10.1109/CVPR52733.2024.00131
15. A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Computer Vision – ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 483-499. https://doi.org/10.1007/978-3-319-46484-8_29
16. M. Ding and G. Fan, “Articulated and generalized Gaussian kernel correlation for human pose estimation,” *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 776-789, 2016. https://doi.org/10.1109/TIP.2015.2507445
17. D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2536-2544. https://doi.org/10.1109/CVPR.2016.278
18. W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, “Look at boundary: a boundary-aware face alignment algorithm,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 2129-2138. https://doi.org/10.1109/CVPR.2018.00227
19. 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 *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 1021-1030. https://doi.org/10.1109/ICCV.2017.116
20. A. Dapogny, K. Bailly, and M. Cord, “DeCaFa: deep convolutional cascade for face alignment in the wild,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 6892-6900. https://doi.org/10.1109/ICCV.2019.00699
21. Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji, “Star loss: reducing semantic ambiguity in facial landmark detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 15475-15484. https://doi.org/10.1109/CVPR52729.2023.01485
22. J. Wan, J. Liu, J. Zhou, Z. Lai, L. Shen, H. Sun, P. Xiong, and W. Min, “Precise facial landmark detection by reference heatmap transformer,” *IEEE Transactions on Image Processing*, vol. 32, pp. 1966-1977, 2023. https://doi.org/10.1109/TIP.2023.3261749
23. X. Bao, Z. Q. Cheng, J. Y. He, W. Xiang, C. Li, J. Sun, et al., “KeyPosS: plug-and-play facial landmark detection through GPS-inspired true-range multilateration,” in *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, Canada, 2023, pp. 5746-5755. https://doi.org/10.1145/3581783.3612366

Yeeun Choi https://orcid.org/0009-0009-1473-1035

She is pursuing the Bachelor's degree in Business with a double major in Computer Science at Chungnam National University, Republic of Korea, since 2022. Her research interests include computer vision, representation learning, facial landmark detection, image segmentation, and active learning. She is also interested in multi-modal and vision-language models (VLMs) for visual understanding.

Hyeonmin Jeong

He is pursuing the Bachelor's degree in Artificial Intelligence at Chungnam National University since 2021. His research interests are in computer vision and deep learning. His research interests include computer vision, deep learning, 2D/3D object detection and model generalization.

Chaemin Yoo https://orcid.org/0009-0003-7224-9964

He is an undergraduate student in the department of Artificial Intelligence at Chungnam National University since 2021. His research interests are in computer vision, deep learning and natural language processing. He is currently conducting research on metric learning and large language model (LLM) fine-tuning.

Gwanghee Lee

He is pursuing the Ph.D. degree in Computer Science at Chungnam National University since 2021. His research interests are in computer vision and deep learning. Specifically, his research interests include computer vision deep learning, face alignment, pose estimation, object detection, keypoint detection, semantic segmentation, instance segmentation, and face recognition.

Kyoungson Jhang

He received B.S., M.S., and Ph.D. degrees in Department of Computer Engineering from Seoul National University in 1986, 1988, and 1995, respectively. Since September 2001, he has been working as a professor for the Department of Computer Science and Engineering at Chungnam National University, Daejeon, Korea. His research focuses on computer vision and deep learning.