Handling partial occlusions in facial expression recognition with variational autoencoder

Abdelaali Kemmou[1], Adil El Makrani[1], Ikram El Azami[1], Moulay Hafid Aabidi[2]
[1]Laboratory of Research in Informatics, Faculty of Science, Ibn Tofail University, Kenitra, Morocco
[2]Department of Computer and Mathematical Engineering, Higher School of Technology, Sultan Moulay Slimane University, Khenifra, Morocco

Article Info

Article history:

Received Dec 11, 2024
Revised Aug 19, 2025
Accepted Sep 1, 2025

Keywords:

Facial expressions
Facial occlusions
Optical flow
Recognition
Variational autoencoder

ABSTRACT

Facial expression recognition (FER) is essential in various domains such as healthcare, road safety, and marketing, where real-time emotional feedback is crucial. Despite advancements in controlled settings such as well-lit, frontal, and unobstructed conditions, FER still faces significant challenges in natural, unconstrained environments. One of the most difficult issues is the presence of occlusions, which obscure key facial features. To overcome this, multiple strategies have been proposed, generally falling into two categories: those focused on analyzing visible facial regions and those aimed at reconstructing hidden facial features. In this study, we present a variational autoencoder (VAE)-based solution designed to reconstruct facial features obscured by occlusions. Experimental results show our VAE model optimized with the structural similarity index measure (SSIM) cost function achieves superior performance, with recognition rates of 91.2% for eye occlusions and 89.7% for mouth occlusions. The SSIM-optimized VAE effectively reconstructs occlude facial features while preserving structural details, demonstrating significant improvements over conventional approaches. This VAE-based solution proves particularly robust for real-world scenarios involving common facial obstructions like masks or sunglasses, making it valuable for applications in healthcare monitoring, driver safety systems, and human-computer interaction.

Corresponding Author:

Abdelaali Kemmou
Laboratory of Research in Informatics, Faculty of Science, Ibn Tofail University
Kenitra, Morocco
Email: abdelaali.kemmou@uit.ac.ma

1. INTRODUCTION

Although, facial expression recognition (FER) has improved a lot under controlled environments to face challenges like low resolution images, lighting conditions variations or head movements [1], [2]. This study, however brings us to focus on a common issue: occlusions in faces. While encouraged for security and human-computer interaction applications, in these settings FER performance can greatly degrade when critical facial regions are occluded by items like scarves, sunglasses, masks (Figure 1), or a hand raised to the chin. Simple everyday occlusions can hide important parts of the face, making it difficult to obtain an accurate measurement and introduce noise. Solving this issue is important to build robust, and reliable FER systems in the wild, where occlusions of this type are common.

In contrast, this study addresses the problem by proposing a new method to reconstruct occluded facial regions based on an adapted variational auto-encoder (VAE) architecture. Our aim is to leverage the VAE's ability to generate latent representations of hidden facial information, creating a robust system that can restore missing features. Our goal is to improve robustness of FER systems so that they are accurate and reliable despite various occlusions, through comprehensive experimentation and analysis.

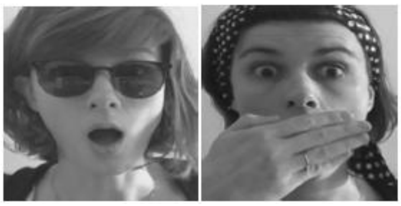

Figure 1. Examples of occlusions that are frequently observed in real-world

In addition the paper is centered around implementing an FER task on partial occluded face images using VAE-based facial feature reconstruction. The work therefore underscores the need to fill in missing features due specifc occlusions, with an emphasis on VAE-like methods and models. The tasks required to accomplish this research goal include:

- Justify the selection and generation of occlusions in the CK+ database images.
- Analyze methods for reconstructing obscured facial features.
- Develop a VAE-based model for reconstructing occluded facial features.
- Optimize the model’s hyperparameters to improve recognition rates.
- Validate the developed method.
- Analyze the experimental results.

This research offers three main contributions. First, it proposes a model based on the VAE approach to reconstruct obscured facial features. Second, the reconstructed features are used as inputs to train a convolutional neural network (CNN) classifier. Finally, an experimental study on the CK+ dataset provides empirical evaluation of the method's effectiveness.

Section 2 reviews the relevant literature and state-of-the-art methods addressing the occlusion challenge in FER tasks. Section 3 details the methodology used to develop the neural network models, including the VAE for feature reconstruction and the CNN for classification. Section 4 presents the results of our approach and compares them with previous studies. The conclusion summarizes the key findings of this research.

2. RELATED WORKS

Facial occlusions pose a significant challenge to automatic FER. In response, two primary research directions have emerged: methods that focus on analyzing unoccluded facial regions and those that attempt to reconstruct occluded areas.

Early approaches divided the face into predefined regions, prioritizing visible areas while disregarding occluded zones. However, these static segmentation strategies lacked adaptability to variable occlusion patterns. More recent solutions, particularly those based on deep learning, have introduced attention mechanisms capable of dynamically identifying and emphasizing the most informative unoccluded regions, thereby enhancing recognition performance [3]-[5]. However, these approaches still depend largely on the quality and location of visible attributes and can perform poorly in severe or misaligned occlusions.

In contrast, reconstruction-based techniques aim to retrieve facial information lost due to occlusions. While prior work highlighted the importance of modeling partial occlusions to reasoning about critical facial traits, recent works adopted generative models to synthesize occluded information which was also applicable [6]-[9]. For example, Lu *et al.* [10] proposed a Wasserstein generative adversarial networks (WGAN)-based framework to increase robustness by generating realistic facial expression. In terms of lighting or distorted facial expressions, WGANs will vary in difficulty and sometimes will produce unrealistic results or blurs. Similarly, generative adversarial networks (GAN)-based inpainting methods by Chen *et al.* [11] and Borges *et al.* [12] achieved strong results on benchmark datasets, but their performance relies on precise occlusion localization and can drop when facing real-world, non-uniform occlusions.

Denoising autoencoders also demonstrate an ability to recover facial features from incomplete inputs. A recent contribution by Kemmou *et al.* [13] used motion-guided autoencoders to reconstruct expression-relevant optical flow, though such motion-based methods may be less effective for static images or weak motion cues. These findings align with the work of [14] and [15], who showed that deep CNNs and hybrid architectures can benefit from occlusion-aware preprocessing, though this often comes at the cost of greater model complexity and longer training times. Variational autoencoders (VAEs) have also gained traction as a powerful tool for handling occlusions. Gui *et al.* [16] highlighted their expanding role in visual understanding tasks.

Our approach builds on these recent insights by combining optical flow cues with a VAE-based latent reconstruction model, specifically designed to handle static occlusions in FER tasks. This hybrid formulation leverages the temporal motion patterns between neutral and apex frames to infer occluded features, while addressing the limitations of both GAN and attention-based methods. Overall, the field is moving toward hybrid and generative solutions that jointly model structure and motion, enabling more robust FER under a variety of occlusion scenarios.

3. METHOD

In this section, we present a reconstruction-based approach aimed at restoring occluded faces to conditions that are more suitable for accurate analysis by compensating for the impact of occlusions. Our strategy involves leveraging the similarity in motion patterns, as shown in Figure 2, by restoring the missing motion information caused by occlusions.

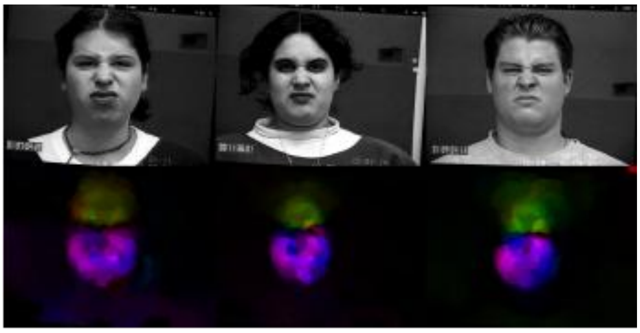

Figure 2. Disgust samples from CK+ (top) and corresponding DeepFlow optical flow maps between neutral and apex frames (bottom)

To reconstruct the missing data using the similarity property, we propose a novel approach that reconstructs the optical flow derived from sequences of occluded facial videos. The method, illustrated in Figure 3 is based on a VAE architecture to reconstruct the optical flows calculated from occluded data [17]. A classifier, trained on optical flows from unoccluded data, then uses the reconstructed optical flows as input to recognize facial expressions in the classification step.

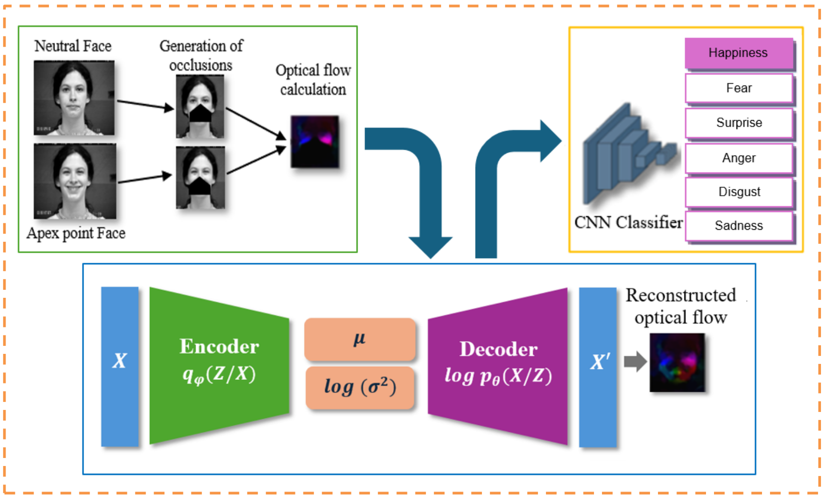

Figure 3. Overview of the proposed VAE-based reconstruction framework for FER under occlusion

3.1. Data preparation

Effective training of a VAE model relies on two crucial components: ground truth optical flow from unoccluded images and the matching optical flow calculated from the same images with occlusions. These elements are critical for the training process, allowing the model to learn and accurately reconstruct occluded facial features.

3.1.1. Generation of occlusions

To simulate occlusions in the most crucial regions for FER, we introduce various occlusion patterns affecting the eyes and mouth, as illustrated in Figure 4. The eyes and mouth are common areas for occlusion when evaluating methods designed to tackle the issue of occlusions. To simulate these occlusions, static black boxes are overlaid over the images in the video sequence.

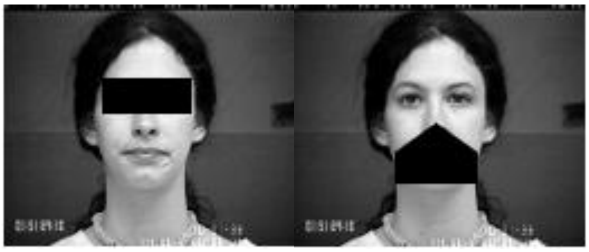

Figure 4. Chosen occlusions for evaluating the approach, applied to the CK+ database

3.1.2. Optical flow calculation

Given the limited availability of datasets in existing literature, evaluating the approach requires restricting the number of parameters learned, which leads to the use of shallow architectures. Additionally, utilizing smaller optical flow maps at the system input is crucial. To ensure the accuracy of flow computation, the optical flow is initially calculated on high-resolution images, then downsampled, as shown in Figure 5, to achieve a uniform size that meets the input requirements of the feature extraction model. This process helps optimize the retention of computational quality.

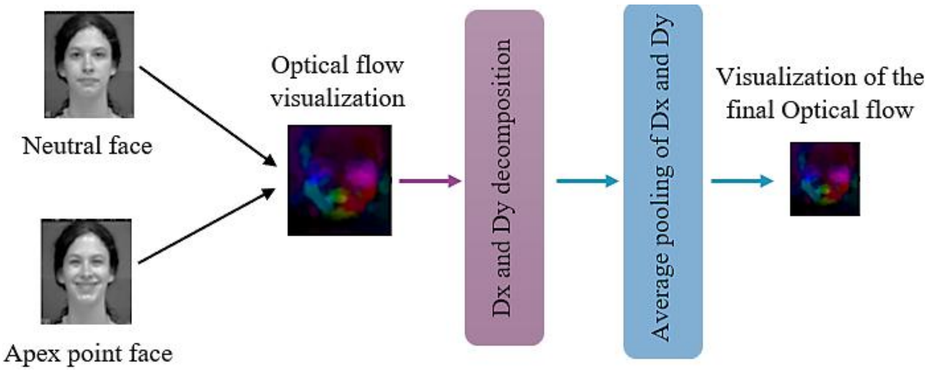
Neutral face
Apex point face
Optical flow visualization
Dx and Dy decomposition
Average pooling of Dx and Dy
Visualization of the final Optical flow

Figure 5. Suggested method for computing and reducing optical flow size to optimize its use in the approach

3.2. Reconstruction of optical flow

Our approach employs a VAE architecture to reconstruct optical flow calculated from occluded data. These optical flows, detailed in the previous section, serve as inputs to the model. Here, we first present the VAE architecture for reconstruction, then explain the model’s probabilistic aspects and latent space, and finally review the various cost functions used to train the VAE.

3.2.1. Variational auto-encoder architecture

The architecture used in our method as shown in Figure 6 consists of an encoder and a decoder designed for reconstructing optical flows. The encoder processes the input optical flow data using consecutive layers of convolutions with 4×4 kernels, progressively downsampling the input to a latent space. Each convolutional layer is succeeded by rectified linear unit (ReLU) activations and batch normalization to stabilize the training. At the end of the encoder, the latent space is represented by two fully connected layers, one for the mean and one for the log variance, allowing the VAE to encode the input into a probabilistic latent space.

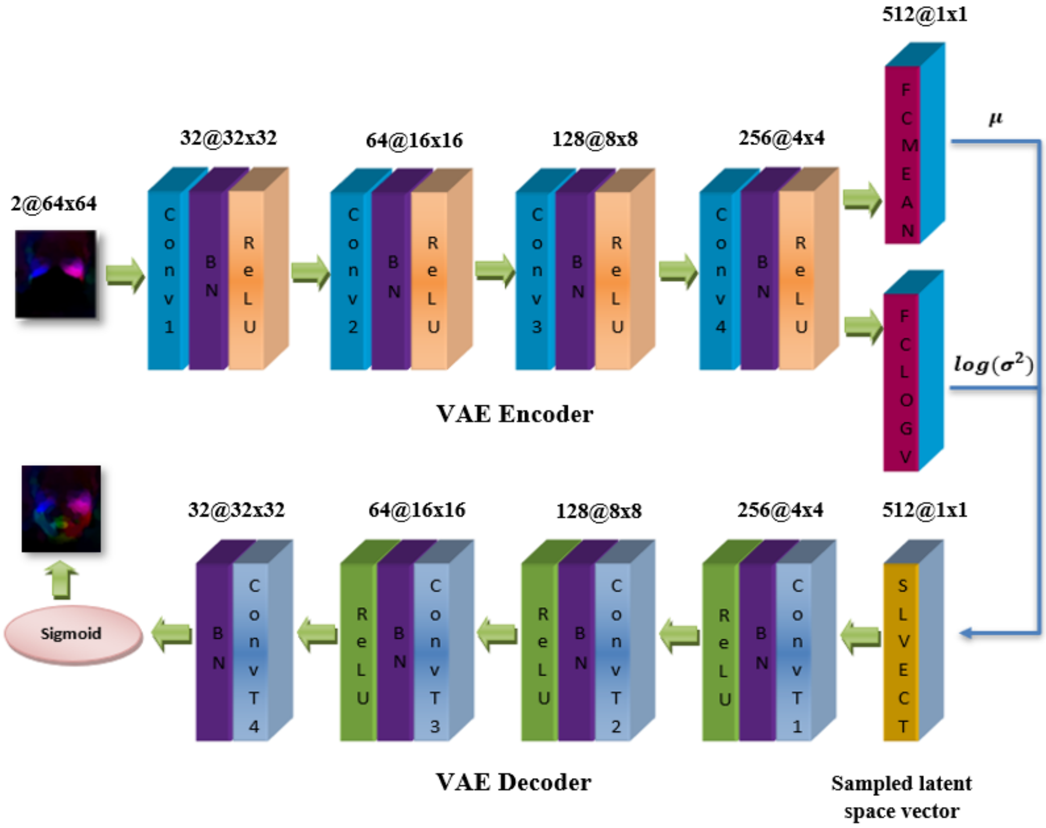

Figure 6. VAE architecture for reconstructing optical flow from occluded facial inputs

The decoder takes a latent vector sampled from this learned distribution and reconstructs the optical flow. The decoder consists of several layers of transposed convolutions, each followed by ReLU activations and batch normalization to upsample the latent vector back to the original optical flow size (2×64×64). The final layer employs a Sigmoid activation to produce the output, which corresponds to the reconstructed optical flow, while ensuring that the range of pixel values is normalized between 0 and 1.

Since optical flow consists of both positive and negative displacements in the x and y directions, the last transposed convolution layer is not followed by ReLU, allowing the network to reconstruct both positive and negative flow values.

3.2.2. Probabilistic model and Latent space

In a VAE model, unlike traditional autoencoders, the encoder does not directly map the input to a fixed latent vector. Instead, the input is mapped to a distribution, typically a Gaussian distribution, defined by two parameters: mean (μ) and variance (σ^2). These parameters define the probability distribution that the model learns for each input in the latent space. This probabilistic setup helps keep the space smooth and consistent, making it easier to move between points and generate realistic new samples.

With a VAE, the latent space isn't one single fixed point it's a continuous space. When reconstructing, the decoder samples latent vectors from this space using the reparameterisation trick to sample from a Gaussian distribution with mean and variance being learned. This forces the latent space to be well-organized, with nearby inputs linking directly to one another which produces soft interpolations and high quality reconstructions.

A VAE is made up of two main components: the encoder and the decoder. The encoder takes the occluded optical flow and compresses it into a smaller latent space, keeping only the most important features for reconstruction. The decoder then uses this latent representation to rebuild the optical flow image. Before decoding, the optical flow vector is reshaped to fit the expected dimensions, producing images of size 2×64×64 pixels. Thanks to its probabilistic design, the VAE can manage variations in the input and generate realistic optical flow reconstructions.

3.2.3. Reparameterization trick

To keep the VAE differentiable, we apply the reparameterization trick [17]. We sample z from the latent space by adding Gaussian noise to the mean μ and scaling it by the standard deviation σ:

$$z = \mu + \sigma.\epsilon \tag{1}$$

where, $\sigma = \exp\ (0.5 \log(\sigma^2))$ and $\epsilon \sim N(0, I)$ (random noise sampled from a standard normal distribution).

3.2.4. Cost functions

The loss function includes both the reconstruction loss (binary cross-entropy) and the KL divergence, which ensures that the latent space distribution remains close to a standard normal distribution.

a. Reconstruction loss

This measures how accurately the decoder can reconstruct the original image from the latent vector. This is typically computed by different loss functions as follows:

– Binary cross-entropy loss between the reconstructed and original images [18].

$$BCE\ Loss = \sum_{i=1}^{n}[x_i \log(x_i') + (1 - x_i)\log(1 - x_i')] \tag{2}$$

where $x_i$and x_i' are the pixels of the original and reconstructed images, respectively, and the sum is over all pixels.

– Mean squared error (MSE) loss

MSE is commonly used when the pixel values of the optical flow are continuous, and it measures the squared differences between the true and predicted values [19]. It works well when optical flow data is not normalized.

$$MSE\ Loss = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i') \tag{3}$$

– Structural similarity index measure (SSIM) loss

SSIM compares the perceptual similarity between the true and predicted images by evaluating structural details (e.g., textures and edges) [20], which is particularly useful for optical flow reconstruction to preserve the movement's structure. SSIM measures luminance, contrast, and structure.

$$SSIM\ Loss = \frac{(2\mu_x\mu_{x'}+C_1)(2\sigma_{xx'}+C_2)}{(\mu_x^2+\mu_{x'}^2+C_1)(\sigma_x^2+\sigma_{x'}^2+C_2)} \tag{4}$$

where, μ_x and $\mu_{x'}$ are the means of the true and reconstructed optical flow images; σ_x^2 and $\sigma_{x'}^2$ are the variances of the true and reconstructed optical flow; $\sigma_{xx'}$ is the covariance between the true and reconstructed optical flow; and C_1 and C_2 are small constants to avoid division by zero.

b. Kullback-Leibler divergence (KL divergence)

This regularizes the learned latent space by ensuring that the distribution learned $q_\varphi(Z|X)$ by the encoder remains close to a standard normal distribution $N(0, I)$.

$$D_{KL} = -\frac{1}{2}\sum_{i=1}^{d}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2) \tag{5}$$

where d is the dimension of the latent space, μ_i and σ_i^2 are the mean and variance for the i-th latent variable.

c. Total VAE loss

The total loss for training the VAE is the sum of the reconstruction loss and the KL divergence.

$$VAE\ Loss\ =\ Reconstruction\ Loss\ +\ KL\ Divergence \tag{6}$$

This loss is minimized to train the VAE, optimizing the encoder and decoder parameters to generate accurate reconstructions while maintaining a smooth latent space distribution.

4. RESULTS AND DISCUSSION

We test our method mainly on how well it recognizes facial expressions using the reconstructed data, not just on reconstruction accuracy. We start by explaining the experiment setup, including the database and classifier used to measure the effect of reconstruction. Next, we adjust key parameters and carry out a step-by-step analysis.

4.1. Experimental protocol

The CK+ database is widely recognized in the literature for evaluating recognition methods in scenarios involving partial facial occlusions [21]. It is particularly suited for studying occlusions, as it is a fully controlled database. Additionally, CK+ fits our method well since it is a dynamic dataset with 374 labeled video sequences, making it perfect for calculating optical flow in our experiments.

4.2. Experimental protocol for automatic facial expression recognition

We propose employing the architecture suggested by Allaert *et al.* [22] (Figure 7), as it has demonstrated its effectiveness in FER tasks, particularly when utilizing optical flow-based learning methods.

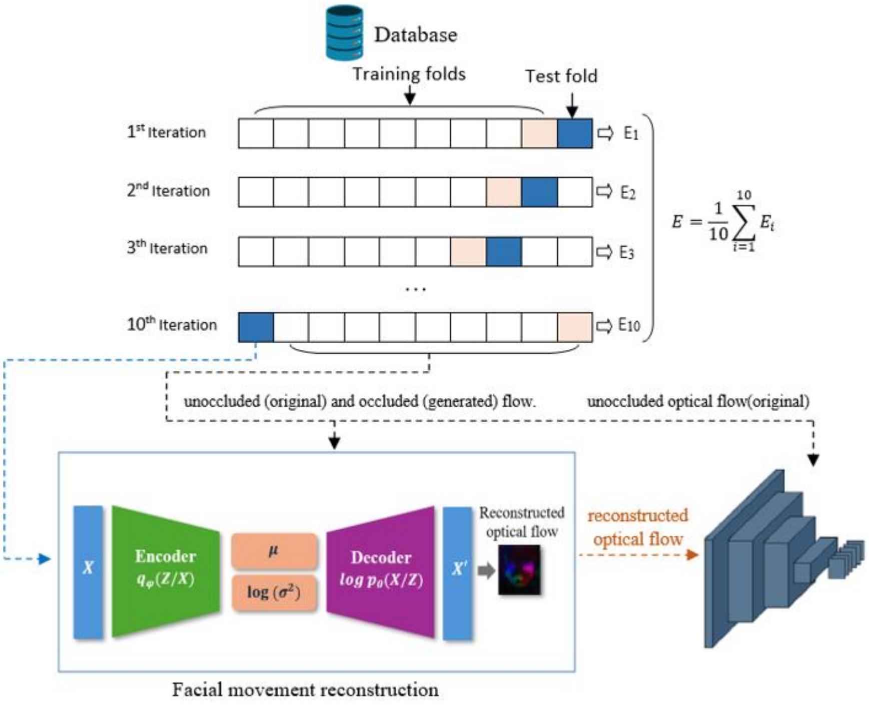
Database
Training folds
Test fold
1st Iteration
2nd Iteration
3rd Iteration
10th Iteration
E1
E2
E3
E10
$E = \frac{1}{10}\sum_{i=1}^{10} E_i$
unoccluded (original) and occluded (generated) flow.
unoccluded optical flow(original)
X
Encoder $q_\varphi(Z/X)$
μ
$\log(\sigma^2)$
Decoder $\log p_\theta(X/Z)$
X'
Reconstructed optical flow
reconstructed optical flow
Facial movement reconstruction

Figure 7. Performance assessment based on 10-fold cross-validation

To train CNN classifier, each fold is evaluated sequentially by training the CNN on 8 folds, validating on the 9th, and testing on the 10th, as illustrated in Figure 7. The FER rates provided represent the average recognition rates across the 10 successive test folds.

4.3. Parameterization of the recognition process

To optimize the classification process, we explored the size of the optical flow images to balance recognition performance and computational complexity. Various optical flow sizes 24×24, 48×48, 64×64, 96×96, and 128×128 were evaluated. As shown in Figure 8, we found that the 64×64 size produced the highest recognition scores, making it the optimal choice for further analysis.

In the previous section, we detailed the step-by-step process used to generate the scores. These results are summarized in Table 1, providing an initial benchmark for evaluating our method. We compared these results with scores obtained using the same CNN architecture, trained on images of faces without occlusions. The evaluation was performed in two scenarios: first, with no occlusions present and second, with varying degrees of partial occlusions. This comparison helps us assess the effectiveness of our method in handling occluded facial data versus unoccluded data.

4.3.1. Evaluation of the method based on the various cost functions

During evaluation, optical flows were computed between the neutral (start) and apex (end) frames for each CK+ sequence. As shown in Table 2, different cost functions were tested within the VAE reconstruction framework. The SSIM cost function delivered the best results, achieving a recognition rate of 91.2% for eye occlusions and 89.7% for mouth occlusions. These results highlight how SSIM excels at preserving structural details in reconstructed optical flows, making it more effective than other cost functions for handling occluded facial data.

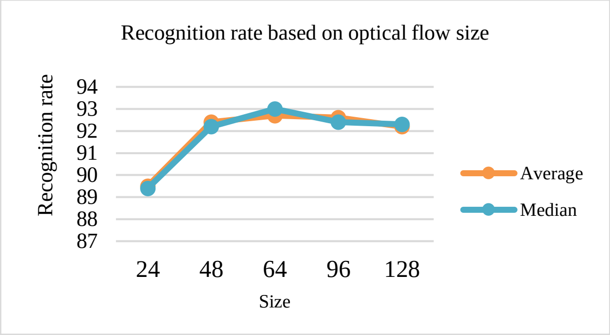

Figure 8. Recognition CNN performance for different optical flow input sizes, averaged over 100 seeds

Table 1. Baseline CNN results on non-occluded and partially occluded faces

Without occlusion (%)	Eyes region occlusion (%)	Mouth area occlusion (%)
92.8	73.8	71.1

Table 2. Results based on the cost function applied during backpropagation in the VAE reconstruction architecture

Cost function	Eyes region occlusion (%)	Mouth area occlusion (%)
BCE	84.6	73.4
MSE	87.3	83.8
SSIM	**91.2**	**89.7**

Table 3 shows the gains achieved with different cost functions, measured by the performance changes resulting from the proposed method. Comparing the improvements from different cost functions allows us to determine which one best enhances recognition rates. This analysis clarifies how various cost functions influence the method’s performance and supports choosing the best one for the task.

Table 3. Improvements gained from different cost functions, calculated as the difference in results using the proposed method

Cost function	Eyes region occlusion (%)	Mouth area occlusion (%)
BCE	10.8	2.3
MSE	13.5	12.7
SSIM	**17.4**	**18.6**

4.3.2. State-of-the-art comparison

Table 4 compares the results of our optimized approach with those of other leading methods evaluated on the CK+ dataset. In the table, the best results for each occlusion type are highlighted in bold, while the second best outcomes are underlined. Additionally, we show the losses caused by various occlusions, indicating the difference between the results without occlusions and those obtained using different methods. The comparison demonstrates the strong performance of our proposed method, particularly with significantly better results for mouth occlusions compared to other state-of-the-art techniques.

Our comparative study reveals that architectural differences fundamentally explain the performance variations among occlusion-handling methods for FER. While traditional approaches like [23] and [24] rely on static image analysis, their inability to model temporal dynamics leads to significant performance degradation (-17.4% for eye occlusion in Dapogny *et al.* [24]). Our VAE-based method overcomes these limitations through two key innovations: i) optical flow integration that captures motion patterns to infer occluded regions (-1.6% drop for eyes) and ii) probabilistic latent modeling with KL divergence that enables robust feature reconstruction (-3.1% drop for mouth, outperforming AE's -9.6% and DCGAN's -5.2%).

Table 4. Comparison of the proposed method’s results with leading approaches on the CK+ dataset for eye and mouth occlusion scenarios

	Without occlusion (%)	Eyes region occlusion (%)	Mouth area occlusion (%)
Huang *et al.* [23]	93.2	**93/-0.2**	73.5/-19.7
Dapogny *et al.* [24]	**93.4**	76/-17.4	67.1/-26.3
AE-based method [13]	92.8	89.7/-3.1	83.2/-9.6
DCGAN-based method [25]	92.8	90.5/-2.3	87.6/-5.2
Our VAE-based method	92.8	91.2/-1.6	**89.7/-3.1**

The evaluation on CK+ dataset under standardized protocols confirms our method's consistent superiority across occlusion types. By synergistically combining motion-aware feature extraction with probabilistic reconstruction, our framework achieves state-of-the-art performance while addressing the critical challenge of occlusion robustness in real-world applications. This dual advantage of temporal modeling and generative reconstruction positions our approach as particularly effective for practical FER scenarios where occlusions are frequent and variable.

However, we acknowledge that CK+ is a controlled dataset with posed expressions and limited variability in lighting and background conditions. As such, generalization to real-world ("in-the-wild") settings remains an open challenge. In future work, we plan to evaluate the proposed method on more diverse datasets such as AffectNet and RAF-DB to assess its robustness under natural occlusions, spontaneous expressions, and complex environments.

4.3.3. Computational efficiency and inference time

To assess the practical feasibility of our VAE-based reconstruction method, we measured inference time and model size on a machine equipped with an NVIDIA Quadro P600 GPU and 32 GB RAM. The average inference time per sample (including optical flow computation, VAE reconstruction, and CNN classification) is approximately 95 ms, allowing near real-time processing at around 10 FPS. The full model (VAE+CNN) occupies 34 MB, confirming its suitability for memory-constrained environments. These findings indicate that our approach maintains a good trade-off between reconstruction quality and computational efficiency, making it applicable to real-time or low-latency FER tasks with lightweight hardware.

5. CONCLUSION

In summary, we have introduced an innovative approach to addressing occlusions in FER systems. Our method focuses on reconstructing occluded facial regions within the optical flow domain by leveraging the natural similarity in motion patterns between individuals. Utilizing a VAE-based architecture specifically designed for handling noisy data, we aim to recover and reconstruct critical information related to facial expressions. The study primarily concentrates on static occlusions occurring between the initial and final frames of video sequences. However, we have not yet explored dynamic occlusions, such as passing hands, which present additional challenges. Future research will focus on analyzing the impact of dynamic occlusions on optical flows and understanding the complexities they introduce in correlating these movements with facial expressions. From a scientific perspective, this work advances the integration of generative reconstruction with motion-based features, offering a robust solution for expression recognition under occlusion. Practically, our method can enhance the reliability of FER systems in real-world applications such as driver monitoring, telemedicine, and security, where occlusions are common.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Abdelaali Kemmou	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Adil El Makrani		✓		✓	✓	✓		✓		✓		✓	✓	
Ikram El Azami		✓		✓	✓	✓		✓		✓		✓	✓	
Moulay Hafid Aabidi				✓	✓	✓				✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors confirm that they have no known financial, personal, or non-financial competing interests that could have influenced the work reported in this manuscript.

DATA AVAILABILITY

The data used in this study are publicly available. Specifically, we used the Extended Cohn-Kanade Dataset (CK+), a complete dataset for action unit and emotion-specified facial expressions. Access details can be found at [21].

REFERENCES

[1] M. D. Putro, J. Litouw, and V. C. Poekoel, “Low-resolution facial emotion recognition on low-cost devices,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 2, pp. 2199-2209, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp2201-2211.

[2] Z. Zhao, Q. Liu, and F. Zhou, “Robust lightweight facial expression recognition network with label distribution training,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 3510–3519, doi: 10.1609/aaai.v35i4.16465.

[3] Y. Li, J. Zeng, S. Shan, and X. Chen, “Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism,” *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019, doi: 10.1109/TIP.2018.2886767.

[4] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[5] X. Zhu, Z. He, L. Zhao, Z. Dai, and Q. Yang, “A cascade attention based facial expression recognition network by fusing multi-scale spatio-temporal features,” *Sensors*, vol. 22, no. 4, p. 1350, 2022, doi: 10.3390/s22041350.

[6] S. Ge, C. Li, S. Zhao, and D. Zeng, “Occluded Face Recognition in the Wild by Identity-Diversity Inpainting,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3387–3397, 2020, doi: 10.1109/TCSVT.2020.2967754.

[7] D. Brown and I. Mollah, “Enhanced Human Face Recall by Reconstruction of Real and Synthetic Occlusions,” in *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, 2024, pp. 1105–1112, doi: 10.1109/ICDICI62993.2024.10810768.

[8] N. Zhang, N. Liu, J. Han, K. Wan, and L. Shao, “Face De-Occlusion With Deep Cascade Guidance Learning,” *IEEE Trans Multimedia*, vol. 25, pp. 3217–3229, 2023, doi: 10.1109/TMM.2022.3157036.

[9] S. Du and L. Zhang, “FRNet: Improving Face De-occlusion via Feature Reconstruction,” in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV),* Singapore: Springer Nature Singapore, 2024, pp. 313–326, doi: 10.1007/978-981-99-8552-4_25.

[10] Y. Lu, S. Wang, W. Zhao, and Y. Zhao, “WGAN-Based Robust Occluded Facial Expression Recognition,” *IEEE Access*, vol. 7, pp. 93594–93610, 2019, doi: 10.1109/ACCESS.2019.2928125.

[11] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang, “Occlusion-aware face inpainting via generative adversarial networks,” in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, 2017, pp. 1202-1206, doi: 10.1109/ICIP.2017.8296472.

[12] T. M. Borges, T. E. de Campos, and R. de Queiroz, “Towards robustness under occlusion for face recognition,” *arXiv*, Sep. 2021, doi: 10.48550/arXiv.2109.09083.

[13] A. Kemmou, A. El Makrani, I. El Azami, and M. H. Aabidi, “Automatic facial expression recognition under partial occlusion based on motion reconstruction using a denoising autoencoder,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, pp. 276–289, Apr. 2024, doi: 10.11591/ijeecs.v34.i1.pp276-289.

[14] I. Lee, E. Lee, and S. B. Yoo, “Latent-OFER: Detect, Mask, and Reconstruct with Latent Vectors for Occluded Facial Expression Recognition,” *arXiv*, Jul. 2023, doi: 10.48550/arxiv.2307.11404.

[15] C. Ge, “Overcoming occlusions in complex environments to achieve robust perception of human emotions,” *Engineering Research Express*, vol. 6, no. 4, Dec. 2024, doi: 10.1088/2631-8695/ad9fd6.

[16] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, “A review on generative adversarial networks: Algorithms, theory, and applications,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3313–3332, 2021, doi: 10.1109/TKDE.2021.3130191.

[17] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” Dec. 2013, doi: 10.48550/arXiv.1312.6114.

[18] Z. Pan *et al.*, “Loss functions of generative adversarial networks (gans): Opportunities and challenges,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 500–522, Aug. 2020, doi: 10.1109/TETCI.2020.2991774.

[19] T. Over and S. Foks, “Mean Squared Error, Deconstructed,” *Journal of Advances in Modeling Earth Systems*, vol. 13, Oct. 2021, doi: 10.1029/2021MS002681.

[20] B. Ghojogh, F. Karray, and M. Crowley, "Theoretical insights into the use of structural similarity index in generative models and inferential autoencoders," in *International Conference on Image Analysis and Recognition*, 2020, pp. 112–117, doi: 10.1007/978-3-030-50516-5_10.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

[22] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Optical flow techniques for facial expression analysis: Performance evaluation and improvements," *arXiv*, 2019, doi: 10.48550/arXiv.1904.11592.

[23] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2181–2191, Dec. 2012, doi: 10.1016/j.patrec.2012.07.015.

[24] A. Dapogny, M. Cord, and P. Pérez, "The missing data encoder: Cross-channel image completion with hide-and-seek adversarial network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 10688–10695, doi: 10.1609/aaai.v34i07.6696.

[25] A. Kemmou, A. El Makrani, I. Elazami, F. Lehlou, and M. H. Aabidi, "Improved Facial Expression Recognition Through Occluded Optical Flow Reconstruction Using Deep Convolutional Generative Adversarial Network," in *2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, 2024, pp. 1–7, doi: 10.1109/WINCOM62286.2024.10657959.

BIOGRAPHIES OF AUTHORS

Abdelaali Kemmou received the Diploma of Master at the Faculty of Science, Ibn Tofail University, Kenitra. Morocco. In 2020 he joined the doctoral study center of Ibn Tofail University, Kenitra, Morocco. He is a member of the laboratory of research in computer science (L@RI). He is currently a Ph.D. Researcher on the study and development of deep learning algorithms for big data. He can be contacted at email: abdelaali.kemmou@uit.ac.ma.

Adil El Makrani is a Research Professor in Computer Science at the Faculty of Science, Ibn Tofail University, Kenitra. He received the Master in a computer science, computer graphics and imagery, and Ph.D. in Computer Science, Sidi Med Ben Abdellah University in 2009 and 2015, respectively. He affiliated to the Research in Informatics laboratory (L@RI). His research currently focuses on artificial intelligence technologies, big data analytics, and their applications. He can be contacted at email: adil.elmakrani@uit.ac.ma.

Ikram El Azami currently works at the Department of Informatics, Université Ibn Tofail. He does research in databases, machine learning, data mining and distributed computing. His most recent publication is "AraTrans the new transformer model to generate Arabic text". He can be contacted at email: ikram.elazami@uit.ac.ma.

Moulay Hafid Aabidi is Professor of Computing Science Research at the University Sultan Moulay Slimane, Higher School of Technology, Khenifra. His study focuses on the optimization of NP-Hard issues with the metaheuristic approaches in artificial intelligence and big data for decision-making in diverse fields. Includes big data analytics, artificial intelligence, and information system. He can be contacted at email: myhafidaabidi@yahoo.fr.