Latent-OFER: Detect, Mask, and Reconstruct with Latent Vectors for Occluded Facial Expression Recognition

Isack Lee, Eungi Lee, Seok Bong Yoo*
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
{sackda24, 181061, sbyoo}@jnu.ac.kr

Abstract

Most research on facial expression recognition (FER) is conducted in highly controlled environments, but its performance is often unacceptable when applied to real-world situations. This is because when unexpected objects occlude the face, the FER network faces difficulties extracting facial features and accurately predicting facial expressions. Therefore, occluded FER (OFER) is a challenging problem. Previous studies on occlusion-aware FER have typically required fully annotated facial images for training. However, collecting facial images with various occlusions and expression annotations is time-consuming and expensive. Latent-OFER, the proposed method, can detect occlusions, restore occluded parts of the face as if they were unoccluded, and recognize them, improving FER accuracy. This approach involves three steps: First, the vision transformer (ViT)-based occlusion patch detector masks the occluded position by training only latent vectors from the unoccluded patches using the support vector data description algorithm. Second, the hybrid reconstruction network generates the masking position as a complete image using the ViT and convolutional neural network (CNN). Last, the expression-relevant latent vector extractor retrieves and uses expression-related information from all latent vectors by applying a CNN-based class activation map. This mechanism has a significant advantage in preventing performance degradation from occlusion by unseen objects. The experimental results on several databases demonstrate the superiority of the proposed method over state-of-the-art methods. This code is available at https://github.com/leeisack/Latent-OFER.

1. Introduction

Facial expression recognition (FER) has undergone remarkable advancements in recent years and is now widely used across various industries. However, the ability of FER models in the presence of occlusion remains a challenge.

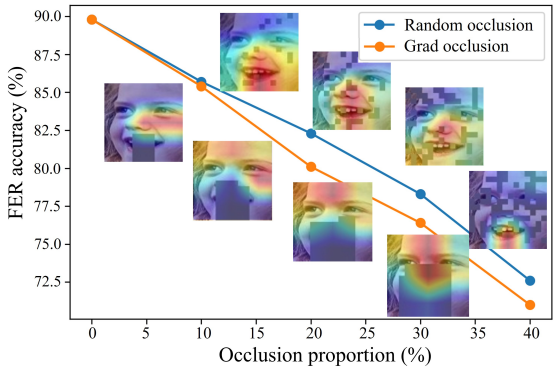

Figure 1. Facial expression recognition (FER) performance on RAF-DB according to the occluded proportions.

Figure 1 illustrates the accuracy of the previous state-of-the-art model for FER [58] for various occlusion proportions. In this study, we evaluated the robustness of the typical model in recognizing facial expressions using two types of occlusions: random sampling occlusion and grad occlusion. Random sampling occlusion divides the entire image into 196 patches and randomly masks them according to the proportion. Grad occlusion processes an image with an intentionally occluded area that affects FER using gradient-weighted class activation mapping (Grad-CAM) [46]. This study reveals a more substantial decrease in performance with the second type of occlusion, particularly when the occluded area is crucial for accurate FER. This finding has been objectively measured using Grad-CAM [46].

Previous studies on FER [1, 2, 3, 5, 7, 13, 14, 15, 21, 23, 29, 31, 38, 43, 44, 49, 50, 54, 57, 60, 62, 63, 68, 69, 70] have not given adequate attention to the influence of occlusions. However, addressing this challenge is crucial for enhancing the practical applications of FER in real-world scenarios. Although research on occluded FER (OFER) is relatively scarce, its importance is increasingly recognized [66].

Currently, several approaches address OFER. The occlusion-robust feature extraction approach [9, 53] aims to identify an occlusion-insensitive and discriminative representation, but it is challenging because the types and locations of occlusion are often unknown. The sub region anal-

*Corresponding author

ysis approach [26, 55] divides regions based on facial landmarks and uses attention mechanisms to focus on crucial areas. However, the inability to detect facial landmarks due to occlusions can lead to errors in the FER process. The unoccluded image network assist approach [35, 61] uses two distinct networks: one trained on unoccluded images and the other trained on occluded images. This approach leverages unoccluded images as privileged information to assist in expression recognition in the presence of occlusion. It is unsuitable in real-world situations because it cannot differentiate between occluded and unoccluded images.

Hence, the proposed approach is the occlusion recovery-based approach, which aims to transform occluded images into complete images through a deocclusion process. We propose the deocclusive autoencoder for reconstructing facial images. The deocclusive autoencoder can be functionally divided into an occlusion detector and reconstruction module. The occlusion detector uses the vision transformer (ViT) support vector data description (SVDD) for the reconstruction module. This approach enables the model to detect occlusion caused by unseen objects, an essential step in accurately generating deoccluded facial images. The reconstruction network consists of the ViT structure and convolutional neural network (CNN) structure for facial image reconstruction. We leverage the strengths of the ViT in generating realistic facial images despite varying poses and further refine them using the CNN. We call it the hybrid reconstruction network. The hybrid reconstruction network enhances FER performance by generating deoccluded images that express detailed and vivid facial expression attributes. This enhancement is achieved by incorporating a self-assembly layer and semantic consistency loss. In contrast, previous work on image reconstruction has primarily focused on achieving naturalness, which can result in dull facial expressions. Additionally, we use informative ViT-latent vectors obtained from the reconstruction process. We combine the CNN features and ViT-latent vectors for enhanced facial expression predictions. The main contributions in this work are summarized as follows:

• We propose an expression-relevant feature extractor that uses spatial attention to assign a higher weight to specific facial features, allowing us to identify critical positions for FER. We can retrieve expression-relevant latent vectors from the ViT-latent space to extract valuable information using these positions as critical values.

•We propose ViT-SVDD, a patch-based occlusion detection module optimized for ViT-based networks. As a self-supervised local classifier, the ViT-SVDD module is trained only on latent vectors of unoccluded facial images. This method accurately classifies occlusions caused by unseen objects for subsequent reconstruction.

• We propose a hybrid reconstruction network that combines the strengths of the ViT and CNN architectures with a self-assembly layer and semantic consistency loss to generate facial images naturalness and rich in expression. This approach enhances the quality of deoccluded images and improves the accuracy of FER in challenging conditions.

2. Related Work

2.1. Occluded Object Detection

Several early studies [8] have attempted to detect occlusions in images for OFER tasks. The conventional methods provide location information about the occlusions and are called occlusion-aware methods [26, 55]. However, these methods are not practical for real-world scenarios. Another approach [30] is to train models using synthesized images with various occlusions to detect the occluded positions. Although this approach can determine occlusion positions, it requires a diverse set of object images, which can be challenging to obtain. Moreover, the performance of these models [26, 30, 55] degrades when they encounter unseen objects as input. We drew inspiration from anomaly detection tasks [10, 45, 47, 64] with one-class classification algorithms [20, 32, 39, 40, 52] and proposed a model that learns only from unoccluded datasets and can generalize well to unseen datasets.

2.2. Occluded Facial Expression Recognition

We have no prior knowledge of where occlusions may appear on a facial image or how large or complex they might be. These occlusions can significantly reduce the accuracy of FER by either increasing the intra-class variability or inter-class similarity. Four categories address this issue: occlusion-robust feature extraction, subregion analysis, unoccluded image network-assisted, and occlusion recovery.

Occlusion-robust feature extraction approach. This approach aims to extract features less affected by occlusions while maintaining discriminative capability. Wang et al. [53] used self-supervised and contrastive learning to explore robust representations with synthesized occlusions.

Subregion analysis approach. This approach excludes the occluded parts from recognition. By focusing only on the unoccluded facial parts, the influence of occlusion on recognition performance can be reduced. Li et al. [26] proposed a gate unit with an attention mechanism that allows the model to focus on informative unoccluded facial areas.

Unoccluded image network-assisted approach. This approach employs unoccluded facial images as guidance to assist with OFER. Pan et al. [35] trained two deep neural networks from occluded and unoccluded facial images. Then, the unoccluded network guides the occluded network. Xia et al. [61] measured the complexity of unoccluded data using distribution density in a feature space. The classifier can be guided by unoccluded data and subsequently leverage more meaningful and discriminative samples.

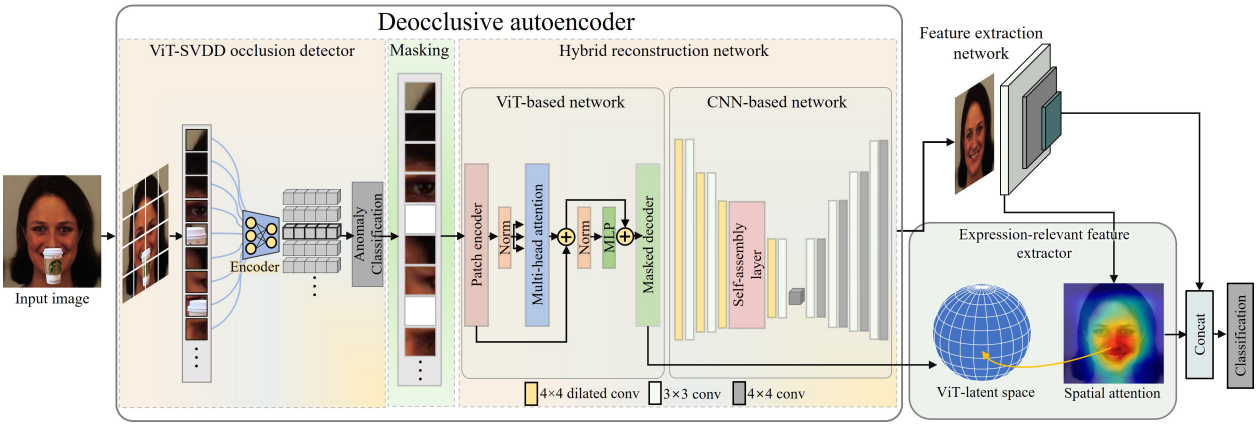

Figure 2. Framework of the Latent-OFER that creates a deoccluded image that erases the occlusion. In this process, the expression-relevant latent vectors are extracted by mapping the vision transformer (ViT)-latent vector and the convolutional neural network (CNN) class activation map. Latent-OFER predicts facial expressions that combine CNN-based features with specific VIT-based latent vectors.

Occlusion recovery-based approach. This approach uses face completion as a basis to deal with occlusion. To reconstruct occluded faces, Lu et al. [30] suggested using a WGAN consisting of an autoencoder-based generator and discriminators. By reconstructing the occlusion, these approaches provide necessary information while avoiding the interference of noisy information and obtaining informative appearance. Therefore, we adopt an image-inpainting technique as a recovery-based approach, to deal with occlusion. A notable point of differentiation between our method and existing approaches lies in our specific emphasis on the reconstruction of occluded regions while simultaneously retaining the facial expression information. Furthermore, we employ ViT-based latent vectors extracted throughout the reconstruction process to enhance the performance of FER.

2.3. Image inpainting

Several approaches based on CNN have been developed to generate semantically coherent content. Pathak et al. [37] employed context encoders to generate plausible features for the area requiring restoration. Yu et al. [65] introduced a contextual attention module to refine by referencing the surrounding features. Song et al. [51] used a patch-swap layer that replaces each patch within the masking areas of a feature map with the most comparable patch in the unoccluded areas. Liu et al. [28] developed a coherent semantic attention layer to ensure semantic relevance between the swapped features.

In recent years, transformer-based methods have also significantly contributed to computer vision. The ViT [11] achieves better results for image recognition than CNN-based methods. Furthermore, MAE [17] demonstrated that transformer-based models are well-suited for image-inpainting tasks. However, transformer-based algorithms have a limitation in that they may produce blurry results when restoring large block areas [17]. Another work [25] presents MAT, a transformer-based model for large-hole inpainting. Nevertheless, in cases where the eyes or mouth are fully occluded, the ViT cannot generate images in detail because it is significantly less biased toward local textures [34]. Therefore, the challenge remains.

3. Proposed method

As shown in Figure 2, we propose multi-stage approach to address OFER, involving detecting, masking, and reconstructing occlusions to recognize facial expressions. The proposed approach enhances recognition accuracy through cooperative learning ViT-latent vectors extracted from the image reconstruction process and the existing CNN features. We divided the facial image into patches, classified each patch as occluded or unoccluded, and reconstructed the occluded patches to be deoccluded. Subsequently, we leveraged the reconstructed image and expression-relevant latent vectors to predict facial expressions.

3.1. Occlusion detection module: ViT-SVDD

General object detection and segmentation models may not be suitable for real-world scenarios because they may be unable to detect unseen objects. To address this limitation, we use one-class classification, which is often employed in anomaly detection. This approach makes it possible to classify unused occlusion during training through self-supervised learning, which only uses unoccluded patches for learning, providing a more effective solution for real-world applications.

One-class classification methods for classifying normal or abnormal can operate at various levels of granularity, ranging from low-level anomaly detection [47] at the pixel level to high-level anomaly detection [40] at the image level. Detection and classification are performed based on the size of the unit, which can be customized to suit the user's needs. We used a ViT-based reconstruction method;

thus, we propose a middle-level anomaly detector specifically optimized for ViT. We divided the image to match the size of the ViT patch and created ViT-latent vectors. These patches are encoded with informative features to produce ViT-latent vectors. To generate the smallest feature space for unoccluded patches, we used the deep SVDD algorithm [39]. One-class deep SVDD employs quadratic loss to penalize the distance of every network representation. This objective is define as

$$min \frac{1}{n} \sum_{i=1}^{n} \left\| \Phi\left(x_i; \mathcal{W}\right) - c \right\|^2 + \frac{\lambda}{2} \sum_{l=1}^{L} \left\| w^l \right\|_F^2 . \quad (1)$$

where n denotes number of training data, L denote number of layer, $\mathcal{W}$ denotes set of weights $\mathcal{W} = \left\{w^1, ..., w^L\right\}$ and c represents a hypersphere characterized by the center. The first term induces the features of all normal images to converge at the center point c, whereas the last term is a weight decay regularizer on the network parameter W with hyperparameter $\lambda > 0$, where $\|\cdot\|_F$ denotes the Frobenius norm. Eq (1) simply employs a quadratic loss for penalizing the distance of every network representation $\Phi\left(x_i; \mathcal{W}\right)$ to $c \in \mathcal{F}$, Where $\mathcal{F}$ is output feature space. The network learn parameters $\mathcal{W}$ such that data points are closely mapped to c of the hypersphere. To determine whether a patch is occluded, we calculated the distance between the new input information and the center c of the feature space for each patch. If the distance exceed the pre-defined radius, the corresponding patch is classified as occluded and masked. The optimal value of the radius is automatically determined in the SVDD procedure. Through this process, occlusion patch detection is possible for unseen objects. The proposed ViT-SVDD approach allows for validating the performance of synthetic images with occlusion patch annotations. By detecting occlusion patches, we can improve the accuracy of the reconstruction method and make it more suitable for real-world applications.

3.2. Image reconstruction module: hybrid reconstruction network

The facial image reconstruction process employs an occlusion-masked image generated by the occlusion detector. The hybrid reconstruction network is designed to cooperate by fusing ViT-based and CNN-based networks. Through this mechanism, we leverage both the strengths of ViT and CNN. The ViT-based method employs 16×16 patches as input image; however, we used the outputs of the occlusion detector as input because the image has already been partitioned into patch units by the occlusion detector.

3.2.1 Network structure

The ViT-based approach encodes the input patches and positively embeds all tokens. The occluded patch reconstruction is achieved through correlation with other patches.

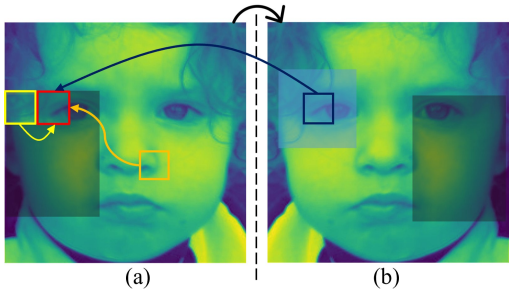

Figure 3. Illustration of the self-assembly operation. (a) Feature map in the self-assembly layer. (b) Flip of (a). The patch in masked region is generated by combining three patch information.

The ViT has low inductive bias and a high degree of freedom [34], enabling it to generate credible images despite diverse occlusion shapes, positions, and facial poses. As discussed in Section 2.3, the ViT-based approach may sometimes not supply detailed results. To address this limitation, we combined the ViT and CNN. The network consists of a U-Net architecture. Additionally, we added a self-assembly layer inside the encoder to generate a detailed representation. This multi structure approach effectively combines the strengths of the ViT and CNN-based networks to generate high-quality facial image reconstructions that represent facial expressions well.

3.2.2 Self-assembly layer

We implemented a self-assembly layer to improve image reconstruction for FER. We reconstructed masked regions like [17] but made enhancements specific to facial images. Based on the concept that the left and right characteristics of a person's face are symmetrical[42, 48, 67], we used the feature information present in the corresponding location of the horizontally flipped image when reconstructing a masked region. We expanded the range of candidate patches used in the generation process by incorporating information from three sources, the previously generated patch, the most similar patch in the unmasked region, and patches located in the corresponding position of the horizontally flipped image. In this process, the masked region contains the reconstruction results with the ViT network. We assigned weights to each based on the similarity value with the current patch. The weight calculation is based on a cross-correlation metric:

$$S(p, p_x) = \frac{\langle p, p_x \rangle}{\|p\| \cdot \|p_x\|}, \quad (2)$$

where p represents a patch in the masked region, and the patch p_x is the comparison target. The similarity value between p and p_x is denoted by S.

The self-assembly operation is defined in Eq (3), which generates a patch value p_i. Figure 3 depicts the operation process, where p_s denotes the patch value symmetrically positioned with respect to p and averaged by considering the peripheral patches. The S_{sym} is calculated as Eq (2) through p_i and p_{s_i}. In addition, p_k is the most similar patch

to p among the unmasked region, and S_{known} is calculated as $S(p_i, p_{k_i})$. Further, p_{i-1} represents the previously generated patch, and S_{i-1} is obtained by $S(p_i, p_{i-1})$. The similarity value S is normalized to be used as the weight.

In Figure 3, the red patch in (a) is p_i, which is a combined result of p_k (the orange patch), p_{i-1} (the yellow patch), and p_s (the blue patch in (b)). Because p_1 has no prior generated patch, S_0 is zero. In certain cases, such as side-face images, the patches in symmetrical positions may not be relevant for generating the patch. Thus, the value of S_{sym} in this situation is minimal and is rarely used to generate the patch.

$$\begin{aligned} p_i = &\frac{S_{sym}}{S_{sym} + S_{known} + S_{i-1}} \times p_{s_i} \\ &+ \frac{S_{known}}{S_{sym} + S_{known} + S_{i-1}} \times p_{k_i}. \\ &+ \frac{S_{i-1}}{S_{sym} + S_{known} + S_{i-1}} \times p_{i-1} \end{aligned} \tag{3}$$

3.2.3 Objective

The aim of image reconstruction is to fill in the masked portion to provide supplementary information for FER. In pursuit of this, we incorporated a semantic consistency loss that enables optimizing the task while maintaining the reconstruction loss L_{re}, consistency L_c [28], feature patch discriminator L_{d_f} [19], and patch discriminator L_d [36].

The semantic consistency loss L_{sc} emphasizes facial expression attributes. The L_{sc} has the effect of reducing intra-class variability and can be defined as:

$$L_{sc} = \sum_{c=1}^{7} p_c\left(z_{gt}\right) log\left(p_c\left(z_{rec}\right)\right), \tag{4}$$

where c represents seven basic expressions, $p_c\left(z_{gt}\right)$ denotes the predicted probability of c in the ground-truth image, and $p_c\left(z_{rec}\right)$ indicates the predicted probability of c in the reconstruction result. The predicted probability distributions are obtained via the pretrained FER network. During training, the overall loss function is defined as:

$$L = \lambda_{re} L_{re} + \lambda_c L_c + \lambda_{sc} L_{sc} + \lambda_d (L_d + L_{d_f}), \tag{5}$$

where $\lambda_{re}, \lambda_c, \lambda_{sc}, \lambda_d$ denote the trade-off parameters for the reconstruction, consistency, semantic consistency, discriminator losses, respectively. Furthermore, the FER network is trained with the same feature extraction architecture using probability distributions about the ground-truth label and prediction of FER.

3.3. Facial expression recognition network

The proposed FER network is designed as an attention-based model for predicting facial expressions. We employed spatial and channel-attention mechanisms [59].

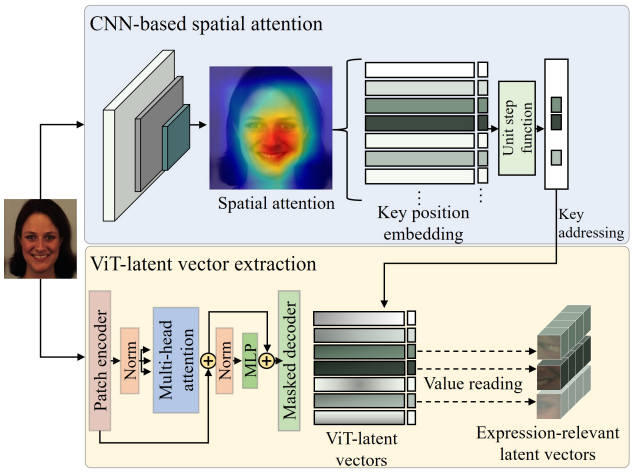

Figure 4. Expression-relevant vision transformer (ViT)-latent vector extraction method.

We obtained the refined feature map and CAM using the attention-based model. In addition, we obtained the expression-relevant latent vectors from the ViT using the CAM. As depicted in Figure 2, the Latent-OFER cooperatively uses a CNN-based feature and ViT-based latent vectors. Thus, the model performs better.

Expression-relevant ViT-latent vectors. The proposed method employs only expression-relevant latent vectors rather than the entire latent space to improve FER performance. During the reconstruction process, ViT-based latent vectors are extracted by embedding the input image. We used a CAM to identify spatially significant areas within the image for FER, and the class activation map is generated through a CNN.

The position of the area is stored as a key, and the attention weight for each space is recorded. The key of the region where the weight of the spatial attention exceeds the top 50% is used. This key is retrieved from the entire ViT-latent vectors, and the corresponding value is read. The activation map is used to identify expression-relevant latent vectors, as presented in Figure 4. This process enables the selection of positions that are relevant to FER while avoiding extraneous details, such as appearance information irrelevant to expressions, which can increase inter-class dissimilarity and lead to more accurate and effective learning results.

In cases where patch detection fails, the occluded patch's latent vectors are used for training and inference. However, spatial attention is not focused on the occluded area, and the latent vector of the occlusion patch is neither searched nor used for training and inference.The proposed extractor is not significantly affected in this situation.

4. Experiments

In this part, we describe three FER benchmark datasets and several occluded FER datasets used to evaluate this

	Accuracy (%)	Precision (%)	Recall (%)
One class SVM [39]	91.1	90.2	89.4
Patch-SVDD [64]	85.2	80.1	94.1
ViT-SVDD	98.3	94.1	98.7

Table 1. Comparison of occlusion patch detection performance.

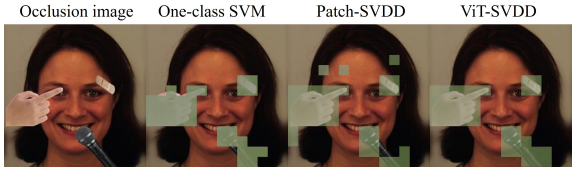

Figure 5. Occlusion patch classification results.

model and explain the results. We demonstrate that the proposed method performs better on both the FER task and occluded FER task. Last, an ablation study demonstrates how each model element contributes to the final performance.

4.1. Dataset

RAF-DB [24] is a large-scale facial expression database with around 30,000 images. We used data on seven basic expressions, 12,271 images as the training set, and 3,068 images as the testing set.

AffectNet [33] is the largest facial expression database with annotations. We used seven basic expressions facial images, about 287,568 training and 3,500 testing images.

KDEF [4] is a set of 4,900 images of facial expressions. Each expression is viewed from five viewpoints.

Syn-AffectNet and Syn-RAF-DB consist of FER benchmark datasets that synthesize a real object to occlude.

Occlusion-AffectNet and Occlusion-RAF-DB are real-world occlusion datasets selected from the AffectNet validation set and RAF-DB testing set by Wang et al. [53].

FED-RO dataset is also a real-world occlusion dataset collected by Li et al. [26]. It contains 400 images labeled with seven basic emotions. We trained the proposed model using RAF-DB and AffectNet when testing the FED-RO datasets, following the protocol suggested in [26].

4.2. Implementation details

We trained a self-supervised occlusion detection module with the KDEF dataset [4] synthesized by randomly copy-pasting objects such as hands and cups for occlusion. We used a ViT-based network pretrained with ImageNet [6] and a CNN-based network trained on several FER training datasets [24, 33] for detailed representation in the image reconstruction module. The FER network uses ResNet-18 [18] backbone architecture. It is pretrained on MS-Celeb-1M [16]. The trade-off parameters in hybrid reconstruction module are set as λ_{re}=1, λ_{c}=0.01, λ_{sc}=1, λ_{d}=0.002. The experimental source code is implemented with PyTorch, and the models are trained with a GTX-3090 GPU.

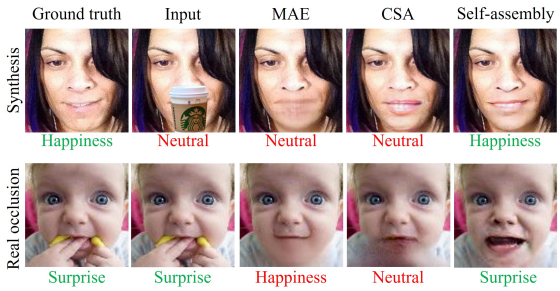

Figure 6. Qualitative comparison of synthesis and real occlusion. Facial expression recognition prediction results according to image reconstruction. Labels highlighted in green indicate matching the correct expression, whereas red indicates a misprediction.

4.3. Comparison of occlusion detection module

The proposed ViT-SVDD detects occlusion for each patch using a suitable detector for ViT-based image reconstruction. We compared the module with existing one-class classification methods trained solely on unoccluded images to evaluate its effectiveness. We used a test set consisting of occluded and unoccluded images.

Patch-SVDD, which trains in a hierarchical structure from small pixels to large pixels and segments anomaly positions into small pixel units, was also included in the comparison. The proposed module achieved the highest occlusion detection accuracy among all methods, as demonstrated in Table 1.

Based on these results, the proposed method is particularly well-suited for the patch-by-patch processing rather than being the optimal solution for all anomaly detection tasks. While several proposed anomaly detection methods are suitable for various units, such as whole images and pixels, this approach is specifically designed to fit the ViT-SVDD framework and detect anomalies in a patch-specific manner. As illustrated in Figure 5, the results demonstrate improved performance in patch-specific occlusion detection, indicating that the proposed method is best suited for special situations that require this processing type.

4.4. Comparison of reconstruction module

The hybrid reconstruction network is compared to MAE [17] and CSA [28]. All reconstruction results are the direct output of the model without post processing. Figure 6 compares the results of deocclusion for synthesis and real occlusion on RAF-DB and AffectNet. As displayed in Figure 6 and Table 2, we present qualitative and quantitative comparisons and the FER result. The self-assembly layer enlarge the facial expression information and increase the number of patch candidates involved in the generation process to reconstruct visually pleasing and natural, and semantic consistency loss induces rich in expression of facial images, providing an advantage for FER. While other com-

	MAE [17]	CSA [28]	Self-assembly
Accuracy (%)	72.6	75.6	77.3

Table 2. Facial expression recognition prediction results according to image reconstruction on Syn-RAF-DB.

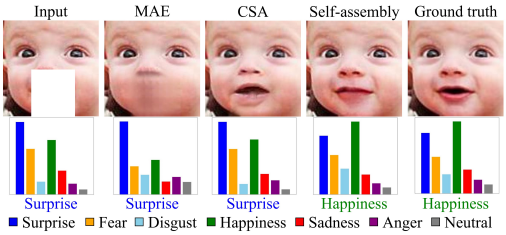

Figure 7. Comparison of reconstruction results and the results of each expression probability.

Method	Affect Net	Syn-Affect Net	RAF DB	Syn-RAF DB	KDEF	Syn-KDEF
DACL [15]	62.5	49.6	88.8	76.0	87.9	68.9
RUL [68]	62.6	52.2	88.9	77.1	83.1	51.2
DAN [58]	63.8	51.1	89.7	76.1	86.8	70.8
EAC [69]	63.2	54.9	88.0	78.6	87.5	67.6
Latent -OFER	63.9	56.1	89.6	80.1	88.3	86.7

Table 3. Accuracy (%) comparison to the state-of-the-art results on the various facial expression recognition datasets.

parative models generate reasonable content, the proposed approach offers a superior representation of FER.

As presented in Figure 7, the first row shows a partially masked facial image, ground truth image, the corresponding reconstruction results of MAE, CSA and self-assembly. The second row is depicted for better visibility of the influence on FER. The proposed method display the highest probability of a ground truth label.

4.5. Comparison of FER accuracy in occlusion

Comparison with Typical FER-model. As listed in Table 3, we compared the proposed model with several state-of-the-art methods for FER on the AffectNet (C7), RAF-DB, and KDEF datasets, all available as open source. We conducted experiments by separately testing these models with synthesized occlusion images and the original dataset for each dataset. In the field of FER, recent models [15, 58, 68, 69, 41, 22] demonstrated satisfactory performance on the original dataset. However, they fail to achieve acceptable accuracy on occluded datasets. In contrast, the proposed model was designed to be robust to occlusion and outperformed other models in terms of accuracy on occluded images. Moreover, the proposed model also achieved nearly the best performance on the original images compared to typical models.

Comparison with the OFER-model. We compared the performance of the proposed FER model targeting occlusion with other state-of-the-art models using FED-RO. As the code for other models is unavailable, we used the reported accuracy from the respective papers. Table 4 reveals

	gACNN [26]	Pan's [35]	Xia's [61]	RAN [55]	OADN [9]	Wang's [53]	Latent -OFER
FED-RO	66.5	69.3	70.5	68.0	68.1	70.0	71.8

Table 4. Accuracy (%) comparison to the state-of-the-art results on the FED-RO dataset.

	BoostGAN [12]	RAN [55]	OADN [9]	Latent -OFER
Occlusion-AffectNet	43.4	58.5	64.0	66.1

Table 5. Accuracy (%) comparison to the state-of-the-art results on the occlusion-AffectNet dataset.

	BoostGAN [12]	RAN [55]	Wang's [53]	Latent -OFER
Occlusion-RAF-DB	55.4	82.7	82.5	84.2

Table 6. Accuracy (%) comparison to the state-of-the-art results on the occlusion-RAF-DB dataset.

Method	Accuracy		Complexity	
	AffectNet	RAF-DB	Flops(G)	Parameters(M)
gACNN [26]	58.8	85.1	331	370
WGAN AE [30]	59.7	83.5	353	400
Pan's [35]	57.1	80.2	5	25
Xia's [61]	57.5	82.7	230	360
RAN [55]	52.6	86.9	390	300
OADN [9]	61.9	87.2	122	250
Wang's [53]	60.2	86.0	102	210
Latent-OFER	63.9	89.6	156	373

Table 7. Accuracy (%) and complexity comparison to the occluded FER models on the typical FER datasets

that the proposed model achieved an accuracy of 71.8%p with the default setting, a new state-of-the-art performance for this dataset, to the best of our knowledge.

Tables 5 and 6 present the FER performance for occluded images in AffectNet and RAF-DB, where the proposed Latent-OFER model achieved an accuracy of 66.5% and 84.2% with the default setting (2.5%p and 1.5%p better than OADN [9] and RAN [55]), respectively.

Table 7 compares the accuracy and complexity of the OFER models. The complexity is estimated by ourselves. As demonstrated in Table 7, OFER models exhibit poor performance on typical FER datasets. This limitation makes them unsuitable for general use unless the unoccluded image network-assisted approach described in Section 1 is employed. In contrast, the proposed model demonstrates acceptable performance on typical and occluded FER datasets. Specifically, the proposed model achieves accuracy rates of 63.9% and 89.6%, which represent improvements of 2.0%p and 2.4%p, respectively, over OADN.

5. Ablation study

This section presents the experiments investigating the effects of several modules on FER performance

Effect of self-assembly layer. We use RAF-DB test dataset and synthesize irregular masks [27] for each image to make comparison. We replaced the self-assembly layer with a conventional 3×3 layer and the CSA layer [28]. As presented in Table 8, we used standard evaluation metrics, the PSNR [56] and SSIM [56], to quantify the module per-

Method	PSNR(↑)	SSIM(↑)
Conventional	26.36	0.868
CSA	25.48	0.860
Self-assembly	26.65	0.880

Table 8. Quantitative comparisons results between conventional, CSA [28], and self-assembly.

Figure 8. Effect of the self-assembly layer. Conventional and CSA [28] are results of the proposed module which replaces the self-assembly layer with the conventional and CSA layers respectively.

	MAE [17]	CSA [28]	Hybrid reconstruction
Reconstructed-RAF-DB	72.6	71.6	77.3

Table 9. Accuracy (%) comparison of FER using irregular masked facial image reconstruction results by MAE, CSA, Hybrid reconstruction network.

formance. Moreover, as illustrated in Figure 8, the mask part fails to reconstruct reasonable content when we used conventional 3×3 layer. Although the CSA [28] can improve the performance compared to conventional. Table 9 reveals that the reconstruction results of CSA inpainting still lack facial expression attributes. Table 9 shows the average accuracy of facial expression prediction with images regenerated by MAE, CSA, and hybrid reconstruction network. For fair comparison, the same expression recognition network and the same partially masked test set were used. Compared with them, the proposed method performs better. The effect of semantic consistency loss is included in the Supplementary Material.

Effect of occlusion detection and reconstruction module. We demonstrated the effectiveness of the deocclusive auto-encoder with occluded KDEF images created through synthesis. As presented in Table 10, image reconstruction is not performed when the occlusion detector is not used, leading to low FER performance. However, even without the proposed module, the proposed model performed well compared to typical models.

Effect of latent vectors and expression-relevant feature extractor. As shown Table 10, this study evaluates the efficiency of the deocclusive autoencoder and expression-related feature extractor in FER. The results reveal that using only the ViT-latent vector extracted from occluded data results in poor performance. Although using only CNN features improved the performance, it remained unacceptable. However, training with the ViT-latent vectors and CNN features improves performance by about 4.0%p compared to only CNN features. Additionally, performance improves by 8.2%p when images are deoccluded via a hybrid reconstruction network. These results indicate that all proposed modules contribute to performance improvement,

Image reconstruction	CNN-features	Full ViT-latent vectors	Extracted ViT-latent vectors	Accuracy (%)
		✓		15.2
			✓	20.1
	✓			75.4
	✓	✓		76.5
	✓		✓	78.5
✓		✓		64.9
✓			✓	66.4
✓	✓			82.7
✓	✓	✓		84.2
✓	✓		✓	86.7

Table 10. Module evaluation in Latent-OFER on Syn-KDEF.

and the best performance is achieved when CNN features and expression-relevant ViT-latent vectors are trained cooperatively.

6. Discussion

This section discusses two topics. First, the performance of the occluded patch detector can be considered dominant. The proposed model restores images and extracts latent vectors to use for FER. In cases where occlusion patch detection fails to perform adequately, latent vectors are not extracted correctly because the image is not fully restored. However, the model uses spatial attention to extract expression-relevant features. In images with poor occlusion patch detection, the corresponding patch has lower weights As a result, the influence of poor detection on the overall performance is mitigated. The second is the potential limitations of scalability when using image reconstruction modules trained with a single dataset. These models may not be adaptable enough to handle a variety of datasets, thereby making it difficult to reconstruct a complete image. To overcome this challenge, extra training on diverse datasets or the adoption of uniform alignment across datasets is necessary. Therefore, to enhance the flexibility of image reconstruction modules, a multi-dataset training approach or standardized alignment methods can be implemented.

7. Conclusion

This paper addresses FER in real-world occlusion. Our method involves detecting the occluded parts of the face and then reconstructing them using a specially designed reconstruction network to produce images as if they were unoccluded. In addition, expression-relevant latent vectors were extracted and learned cooperatively with the CNN features. The proposed Latent-OFER model works well under occluded and unoccluded conditions. We evaluate this method on occluded FER datasets and typical FER datasets, and the proposed method achieves state-of-the-art accuracy.

Acknowledgements This work was supported by the Industrial Fundamental Technology Development Program (No. 20018699) funded by MOTIE of Korea and IITP grant funded by the Korea government (MSIT; No.2021-0-02068, RS-2023-00256629).

References

[1] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv preprint arXiv:2107.03107*, 2021.

[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016.

[3] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[4] Manuel G Calvo and Daniel Lundqvist. Facial expressions of emotion (kdef): Identification under different display-duration conditions. *Behavior research methods*, 40(1):109–115, 2008.

[5] Yuedong Chen, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. Facial motion prior networks for facial expression recognition. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2019.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 653–656, 2018.

[8] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021.

[9] Hui Ding, Peng Zhou, and Rama Chellappa. Occlusion-adaptive deep network for robust facial expression recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Qingyan Duan and Lei Zhang. Look more into occlusion: Realistic face frontalization and recognition with boostgan. *IEEE transactions on neural networks and learning systems*, 32(1):214–228, 2020.

[13] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[14] Ali Pourramezan Fard and Mohammad H Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10:26756–26768, 2022.

[15] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021.

[16] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[20] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[21] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[22] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[23] Isack Lee and Seok Bong Yoo. Latent-per: Ica-latent code editing framework for portrait emotion recognition. *Mathematics*, 10(22):4260, 2022.

[24] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[25] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, pages 10758–10768, 2022.

[26] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[27] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018.

[28] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4170–4179, 2019.

[29] Yuanyuan Liu, Jiyao Peng, Jiabei Zeng, and Shiguang Shan. Pose-adaptive hierarchical attention network for facial expression recognition. *arXiv preprint arXiv:1905.10059*, 2019.

[30] Yang Lu, Shigang Wang, Wenting Zhao, and Yan Zhao. Wgan-based robust occluded facial expression recognition. *IEEE Access*, 7:93594–93610, 2019.

[31] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE, 2017.

[32] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, pages 6927–6937. PMLR, 2020.

[33] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[34] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[35] Bowen Pan, Shangfei Wang, and Bin Xia. Occluded facial expression recognition enhanced through privileged information. In *Proceedings of the 27th ACM international conference on multimedia*, pages 566–573, 2019.

[36] Seong-Jin Park, Hyeongseok Son, Sunghyun Cho, Ki-Sang Hong, and Seungyong Lee. Srfeat: Single image super-resolution with feature discrimination. In *Proceedings of the European conference on computer vision (ECCV)*, pages 439–455, 2018.

[37] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[38] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2375, 2022.

[39] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[40] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

[41] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514:435–450, 2022.

[42] Sriparna Saha and Sanghamitra Bandyopadhyay. A symmetry based face detection technique. *Proceedings of the IEEE*, 2007.

[43] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021.

[44] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

[45] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Avinash Kumar Singh and Gora Chand Nandi. Face recognition using facial symmetry. In *Proceedings of the second international conference on computational science, engineering and information technology*, pages 550–554, 2012.

[49] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5800–5809, 2020.

[50] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5800–5809, 2020.

[51] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[52] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004.

[53] Jiahe Wang, Heyan Ding, and Shangfei Wang. Occluded facial expression recognition using self-supervised learning. In *Proceedings of the Asian Conference on Computer Vision*, pages 1077–1092, 2022.

[54] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020.

[55] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[57] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016.

[58] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021.

[59] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[60] Fangyu Wu, Chaoyi Pang, and Bailing Zhang. Facecaps for facial expression recognition. *Computer Animation and Virtual Worlds*, 32(3-4):e2021, 2021.

[61] Bin Xia and Shangfei Wang. Occluded facial expression recognition with step-wise assistance from unpaired non-occluded images. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2927–2935, 2020.

[62] Siyue Xie, Haifeng Hu, and Yongbo Wu. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern recognition*, 92:177–191, 2019.

[63] Keyu Yan, Wenming Zheng, Tong Zhang, Yuan Zong, Chuangao Tang, Cheng Lu, and Zhen Cui. Cross-domain facial expression recognition based on transductive deep transfer learning. *IEEE Access*, 7:108906–108915, 2019.

[64] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[65] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[66] Dan Zeng, Raymond Veldhuis, and Luuk Spreeuwers. A survey of face recognition techniques under occlusion. *IET biometrics*, 10(6):581–606, 2021.

[67] Liyan Zhang, Anshuman Razdan, Gerald Farin, John Femiani, Myungsoo Bae, and Charles Lockwood. 3d face authentication and recognition based on bilateral symmetry analysis. *The visual computer*, 22:43–55, 2006.

[68] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.

[69] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 418–434. Springer, 2022.

[70] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3510–3519, 2021.