OCA3D: OCCLUSION-CAUSE-AWARE 3D FACIAL EVIDENCE COMPLETION FOR FACIAL EXPRESSION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world facial occlusion arises from two fundamentally different causes: external objects, such as masks, glasses, hair, and hands, and pose-induced self-occlusion. Existing methods typically address external occlusion and self-occlusion either independently or as generic appearance corruption, without explicitly distinguishing their underlying visibility mechanisms or exploiting cause-specific 3D facial correspondences for evidence recovery. We propose **OCA3D**, an occlusion-cause-aware 3D facial evidence completion framework that explicitly models these two sources of missing facial evidence. From each input image, OCA3D predicts an external-occlusion mask and regresses FLAME parameters, based on which a mesh-bound 3D Gaussian representation is constructed over the fixed facial topology. Self-occluded regions are completed through topology-defined bilateral symmetry, while Gaussian attributes for externally occluded regions are recovered from reliable visible mesh faces through nearest-neighbor correspondence within fine-grained semantic regions. The completed representation is rendered from both the input and frontal views, and the resulting evidence is fused with original-image features for FER. We further introduce **Occlu-FER-Mask**, a manually annotated dataset providing pixel-level supervision for external-occlusion segmentation in real-world FER images. OCA3D achieves Top-1 accuracies of 93.06%, 92.15%, 90.19%, and 89.73% on RAF-DB, FERPlus, Occlusion-RAF-DB, and Occlusion-FERPlus, respectively, and remains robust under large-pose self-occlusion on Pose30/45-RAF-DB and Pose30/45-FERPlus (93.50%, 93.01%, 92.85%, 92.58%) as well as real-world external occlusions. These results demonstrate the effectiveness of modeling occlusion according to its cause and recovering expression evidence directly on the 3D facial surface.

1 INTRODUCTION

Facial expression recognition (FER) aims to infer human emotional states from facial visual signals and is widely used in human-computer interaction, driver monitoring, and assisted health analysis (Zhang et al., 2018; Lee et al., 2023; Wang et al., 2026). Recent deep FER models have achieved strong performance on standard benchmarks, but their success often assumes that expression-related facial evidence is sufficiently observable. This assumption frequently breaks in real-world images: masks, glasses, hair, and hands may cover informative facial regions, while large head poses can make parts of the facial surface invisible from the camera. Such missing observations remove discriminative expression cues and may introduce irrelevant visual content, substantially degrading recognition performance. Robust occluded FER therefore requires not only suppressing unreliable observations, but also recovering the missing expression evidence needed for expression recognition.

Existing occluded FER methods mainly suppress corrupted observations (Li et al., 2019; Wang et al., 2020; Ding et al., 2020; Wang et al., 2021; Huang et al., 2025; Zhai et al., 2025) or reconstruct missing evidence (Xing et al., 2022; Lee et al., 2023; Wang et al., 2026). Although effective, external occlusion and pose-induced self-occlusion are typically handled either independently or as generic sources of appearance degradation. This abstraction overlooks an important fact: facial evidence can be missing for fundamentally different physical reasons, and these causes determine what in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

formation remains available for recovery. We therefore argue that robust facial evidence completion should depend not only on *where* information is missing, but also on *why* it is invisible.

Specifically, real-world facial invisibility mainly arises from two distinct mechanisms. *External occlusion* occurs when a facial surface is geometrically visible but its appearance is corrupted by a foreground object. Pose-induced *self-occlusion*, by contrast, occurs when the facial surface itself is invisible from the current viewpoint because of 3D orientation and depth ordering. Although both appear as "missing content" after image projection, they provide different recovery conditions. Externally occluded regions still correspond to geometrically visible facial surfaces and can exploit appearance cues from structurally related visible regions. Self-occluded regions, however, have no direct observation in the current view and require correspondences defined by the underlying facial geometry and topology. This distinction motivates a cause-aware recovery strategy in 3D rather than a shared completion mechanism in the image plane.

Based on this observation, we propose **OCA3D**, an occlusion-cause-aware 3D facial evidence completion framework for robust FER. Given an input image, OCA3D first predicts an external-occlusion mask and regresses input-aligned FLAME (Li et al., 2017b) parameters. The predicted mask and geometry-derived visibility jointly partition the facial surface into reliable, externally occluded, and self-occluded regions. We then bind 3D Gaussians (Kerbl et al., 2023) to the fixed FLAME topology, providing a structured representation for transferring appearance evidence across the facial surface. For externally occluded regions, Gaussian attributes are recovered from reliable visible mesh faces through Euclidean nearest-neighbor correspondence constrained within fine-grained semantic regions. For self-occluded regions, attributes are propagated through topology-defined bilateral correspondence. The completed representation is finally rendered from both the input and frontal views, and their features are fused with those extracted from the original image to provide complementary evidence for FER.

Explicitly modeling external occlusion further requires accurate pixel-level supervision, which is largely absent from existing FER benchmarks. We therefore introduce **Occlu-FER-Mask**, a manually annotated dataset providing pixel-level masks for external facial occlusions in real-world FER images. Experiments on nine benchmarks show that OCA3D is effective across both standard and occluded FER settings and remains robust under large-pose self-occlusion and real-world external occlusions, supporting the effectiveness of modeling missing facial evidence according to its underlying cause. Our contributions are summarized as follows:

- **Occlusion-cause-aware facial evidence recovery.** We distinguish external occlusion from pose-induced self-occlusion according to their underlying visibility mechanisms and formulate robust FER as cause-specific evidence recovery on the 3D facial surface.
- **Structured 3D completion with mesh-bound Gaussians.** We introduce a FLAME-constrained, mesh-bound Gaussian representation that uses semantic-region-constrained nearest-neighbor correspondence for external occlusion and bilateral topology correspondence for self-occlusion. Dual input-view and frontal-view renderings further provides complementary evidence for FER.
- **Pixel-level supervision for real-world facial occlusions.** We construct **Occlu-FER-Mask**, a manually annotated dataset with pixel-level external-occlusion masks for real-world FER images, enabling explicit occlusion localization and structured 3D evidence recovery.

2 Related Work

2.1 Occluded Facial Expression Recognition

Occluded FER aims to recognize expressions when facial observations are partially missing or corrupted. Existing approaches can be broadly grouped into two directions: suppressing unreliable observations and recovering missing expression evidence.

Occlusion suppression. Early methods reduce the influence of occluded regions through local modeling or attention. PG-CNN (Li et al., 2018) performs weighted prediction over landmark-based facial patches, while ACNN (Li et al., 2019) combines global and local representations. MAPNet (Ju & Zhao, 2022) generates expression-related masks from landmarks, and RAN (Wang et al., 2020)

adaptively aggregates region features to improve robustness to both occlusion and pose-indeced self-occlusion. Although these methods can suppress local occlusion noise, their performance is often constrained by landmark errors and predefined region partitioning. Subsequent approaches increasingly incorporate geometric or semantic information as auxiliary priors. POSTER++ (Mao et al., 2025) jointly models appearance and geometry information, and ORSANet (Zhai et al., 2025) combines facial semantic maps with landmarks. Data augmentation methods, including Cutout (DeVries & Taylor, 2017), Mixup (Zhang et al., 2017), CutMix (Yun et al., 2019), and FERMix (Huang et al., 2025), further improve robustness by exposing models to diverse appearance corruption. These approaches remain dependent on visible evidence, and become less effective when highly informative regions are fully occluded.

Expression evidence recovery. A complementary line of work attempts to reconstruct missing information. WGAN (Lu et al., 2019) jointly optimize face reconstruction and expression supervision. Latent-OFER (Lee et al., 2023) adopts a detection–masking–reconstruction pipeline to restore regions corrupted by external objects. Because visually plausible restoration may not necessarily preserve expression semantics, Co-Completion (Xing et al., 2022) instead completes discriminative representations, and subsequent low-rank recovery (Wang et al., 2026) exploits correlations among facial regions to infer missing expression features. Within Occluded FER, however, such recovery is predominantly performed in the 2D image or feature space and is mainly designed for corrupted observations caused by external occlusion. Pose-induced self-occlusion has also been studied jointly with occlusion, e.g., by RAN (Wang et al., 2020), but is typically handled as another source of appearance degradation rather than as a geometrically distinct visibility mechanism.

Our work builds on a different abstraction. External occlusion and pose-induced self-occlusion may both remove expression evidence, yet they provide fundamentally different recovery conditions: the former covers facial surfaces that remain geometrically visible, whereas the latter makes facial surfaces unobservable from the current viewpoint. OCA3D explicitly distinguishes these causes and performs cause-specific evidence recovery on the 3D facial surface, using semantic correspondence for external occlusion and topology-defined bilateral correspondence for self-occlusion.

2.2 3D Face Reconstruction and Gaussian Representations

Monocular 3D face reconstruction estimates facial geometry from a single image, commonly using parametric priors such as 3DMM (Blanz & Vetter, 2023) and FLAME (Li et al., 2017b) to constrain the inherently ill-posed problem. RingNet (Sanyal et al., 2019) regresses FLAME parameters with identity-consistency constraints, while DECA (Feng et al., 2021) decomposes facial geometry into coarse shape and expression-related details. EMOCA (Daněček et al., 2022) introduces emotion-related supervision to improve expression fidelity, and SMIRK (Retsinas et al., 2024) uses analysis-by-neural-synthesis to better capture subtle and extreme expressions. These advances make parametric face models particularly useful for our setting: beyond reconstructing facial shape, they provide a canonical topology from which camera visibility and structural correspondence can be defined.

3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) provides an explicit representation that combines efficient differentiable rendering with flexible appearance modeling. For facial avatars, GaussianAvatars (Qian et al., 2024) rigs Gaussian primitives to FLAME triangles so that their positions and orientations follow the underlying facial surface. SplattingAvatar (Shao et al., 2024) similarly embeds Gaussians on a deformable mesh, establishing persistent correspondence between surface geometry and appearance primitives. More recent methods (Chu & Harada, 2024; Gerogiannis et al., 2025; Lyu et al., 2025; Xia et al., 2026; Ren et al., 2026) extend Gaussian facial modeling toward generalizable or single-image reconstruction. Arc2Avatar (Gerogiannis et al., 2025) uses generative priors to infer expressive 3D avatars from a single image, while OMG-Avatar (Ren et al., 2026) directly predicts multi-level Gaussian head representations using global features, projected local features, and depth-guided fusion.

These methods primarily optimize for avatar reconstruction, animation, or novel-view synthesis, where photorealism, identity preservation, and cross-view consistency are central objectives. OCA3D instead uses mesh-bound Gaussians as a structured carrier of expression evidence. The fixed FLAME topology establishes persistent geometric and semantic correspondence across facial regions, while the Gaussian attributes retain appearance information that can be propagated between

surface elements. This combination allows us to reason explicitly about visibility, distinguish external occlusion from self-occlusion, and recover missing evidence according to its underlying cause rather than generating a universally completed face.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERALL FRAMEWORK

Given a facial image $I \in [0,1]^{3\times H\times W}$ and its expression label y, our goal is to recover missing expression-relevant evidence that becomes unavailable under occlusion and use the recovered evidence to improve FER. Rather than treating all missing observations uniformly in the image plane, we distinguish them according to their underlying visibility causes.

$$\begin{aligned} \mathcal{D}_{\text{obs}} &= \{f \mid \nu_{\text{src}}(f) = 1, M_{\text{ext}}(f) = 0\}, \\ \mathcal{O}_{\text{ext}} &= \{f \mid \nu_{\text{src}}(f) = 1, M_{\text{ext}}(f) = 1\}, \\ \mathcal{O}_{\text{self}} &= \{f \mid \nu_{\text{src}}(f) = 0, \nu_{\text{fr}}(f) = 1\}, \end{aligned} \tag{1}$$

where $\nu_{\text{src}}(f)$ and $\nu_{\text{fr}}(f) \in \{0,1\}$ are geometric visibility indicators of face f under the source and frontal views, determined from the FLAME geometry and camera configuration. $M_{\text{ext}}(f)$ denotes the external-occlusion state obtained by mapping the pixel-level mask M_{ext} to the projected face f.

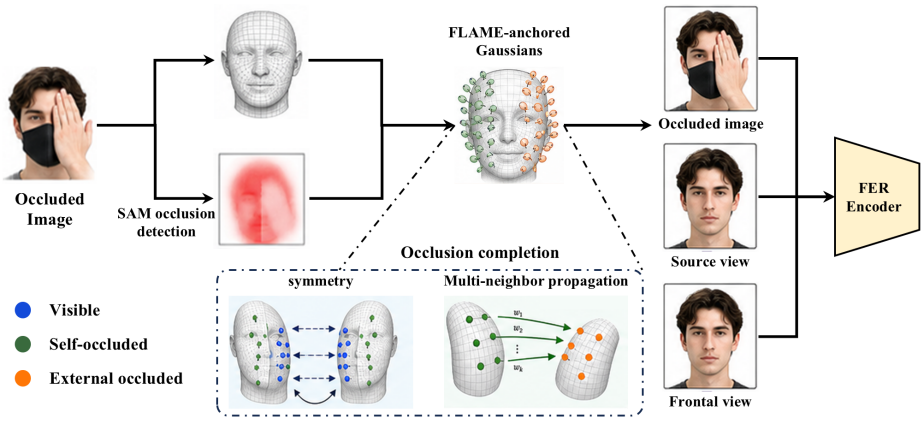


Figure 1: Overall framework of the proposed OCA3D. The framework consists of the occlusion detector, the FLAME parameter estimation module, and the FER classifier.

As shown in Figure 1, OCA3D contains three stages. First, an external-occlusion detector predicts a binary mask M_{ext}, while a FLAME regressor estimates the input-aligned 3D facial geometry. We then bind one Gaussian primitive to each mesh face, establishing a topology-aligned appearance representation. Finally, the completed representation is rendered from the input and frontal views to obtain R_{src} and R_{fr}, which are combined with the original image to provide complementary evidence for FER.

3.2 EXTERNAL OCCLUSION DETECTION

Reliable separation of external occlusion and self-occlusion requires accurate localization of foreground objects covering the facial surface. Although pretrained SAM3 Carion et al. (2025) provides strong general-purpose segmentation, directly applying it to occluded FER is non-trivial. SAM3 typically requires point, box, mask, or text prompts, whereas the location and shape of facial occluders are unknown at inference time. Moreover, facial occlusion is not a fixed object category, but a spatial relation between the face and diverse foreground objects such as hands, masks, and glasses. We therefore formulate external occlusion as a fixed-concept binary segmentation task and use the same task prompt during training and inference.

Training such a model requires pixel-level supervision that is largely unavailable in existing FER benchmarks, which typically provide only expression labels. Existing facial-occlusion segmenta-

tion data such as FaceOcc (Yin & Chen, 2022) provide useful supervision, but their limited scale and scene diversity can leave a distribution gap relative to unconstrained FER images containing diverse occluders, imaging conditions, and backgrounds. Although synthetic occlusion data such as Webface-OCC (Huang et al., 2021) can provide large-scale training data, it differs from real-world occlusion patterns. Both types of data limit model generalization to unseen occlusion scenarios. To provide supervision that better matches the target setting, we construct **Occlu-FER-Mask**, a manually annotated dataset containing pixel-level external-occlusion masks for real-world FER images. The annotations cover diverse foreground occluders, including hands, masks, glasses, hair, and other objects appearing within the facial region. Detailed annotation protocols and representative examples are provided in the Sup. A.1.

After training, the external-occlusion detector is frozen. The predicted M_{ext} is projected onto the reconstructed facial surface and combined with geometry-derived visibility to determine whether a mesh-bound Gaussian belongs to $\mathcal{D}_{\text{obs}}$, $\mathcal{O}_{\text{ext}}$, or $\mathcal{O}_{\text{self}}$. Consequently, the segmentation mask is not used merely to suppress occluded pixels; it determines the physical cause of missing facial evidence and selects the corresponding 3D completion strategy.

3.3 FLAME-Constrained 3D Gaussian Representation

In parallel with external-occlusion detection, we adopt the FLAME estimator E_{F} used in OMG-Avatar (Ren et al., 2026) to regress FLAME parameters from the input image.

$$\Theta = E_{\text{F}}(I) = (\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}_{\text{pose}}, \boldsymbol{\kappa}), \tag{2}$$

$$\mathcal{V}, \mathcal{F} = \mathcal{M}_{\text{F}}(\Theta), \tag{3}$$

where $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, $\boldsymbol{\theta}_{\text{pose}}$, and $\boldsymbol{\kappa}$ denote the identity shape, expression, pose, and camera parameters, respectively, $\mathcal{V}$ and $\mathcal{F}$ are the vertex and triangular-face sets.

For each mesh face $f = (i, j, k) \in \mathcal{F}$, we bind a 3D Gaussian primitive $\mathcal{G}_f$ whose center $\boldsymbol{\mu}_f$ is fixed at the triangle centroid. Unlike free Gaussian optimization, no learnable positional offset is introduced, so each Gaussian deforms consistently with the FLAME geometry and remains anchored to a fixed facial topology. This prevents geometric drift and provides stable correspondences for topology-aware attribute propagation. Each Gaussian is represented as,

$$\mathcal{G}_f = (\boldsymbol{\mu}_f, \underbrace{\boldsymbol{\Sigma}_f, \alpha_f, \boldsymbol{\chi}_f}_{\boldsymbol{\xi}_f}), \tag{4}$$

where $\boldsymbol{\Sigma}_f$, α_f, and $\boldsymbol{\chi}_f$ denote covariance, opacity, and spherical-harmonic color coefficients. We collectively denote these transferable Gaussian attributes by $\boldsymbol{\xi}_f$. The Gaussian center is determined by the reconstructed FLAME, while $\boldsymbol{\xi}_f$ is predicted from global image context, local observations, and face-level geometric and semantic information.

Not all local image features are reliable. Features from $\mathcal{D}_{\text{obs}}$ are treated as valid observations, whereas those from externally or self-occluded faces are masked out. For these regions, the attribute predictor produces an initial estimate $\boldsymbol{\xi}_f^0$ using global context, geometry, and semantic priors. Such estimates are not used as appearance donors for other mesh faces. Instead, they are refined in the following section according to the occlusion cause, using bilateral facial correspondence for self-occlusion and semantic topology neighborhoods for external occlusion.

3.4 Occlusion-Cause-Aware Gaussian Attribute Completion

The mesh-bound Gaussian representation establishes a stable correspondence between 2D observations and 3D facial attributes on a fixed FLAME topology. Based on the surface partition defined above, attributes associated with $f \in \mathcal{D}_{\text{obs}}$ are supported by reliable image observations, whereas those in $\mathcal{O}_{\text{ext}}$ and $\mathcal{O}_{\text{self}}$ lack trustworthy local evidence for different reasons. We therefore recover their attributes using two cause-specific correspondence mechanisms: symmetric completion and nearest-neighbor completion.

3.4.1 Symmetry-Based Completion for Self-Occlusion

Pose-induced self-occlusion often hides one facial side while leaving its contralateral counterpart visible. Since FLAME has a fixed topology, we define a bilateral face correspondence in canonical space $\mathcal{S} : \mathcal{F} \to \mathcal{F}$, $\mathcal{S}(\mathcal{S}(f)) = f$.

For each face f, $\mathcal{S}(f)$ is determined by reflected centroid distance across the sagittal plane together with facial semantic labels. For a self-occluded face $f \in \mathcal{O}_{\text{self}}$, if its bilateral counterpart $\mathcal{S}(f) \in \mathcal{D}_{\text{obs}}$, $\boldsymbol{\xi}_f \leftarrow \boldsymbol{\xi}_{\mathcal{S}(f)}$. Importantly, only appearance-related Gaussian attributes are propagated, the target Gaussian remains anchored to its own FLAME face and therefore retains its center $\boldsymbol{\mu}_f$. If $\mathcal{S}(f) \notin \mathcal{D}_{\text{obs}}$, we instead use the semantic-constrained neighborhood strategy described below. Since faces in $\mathcal{O}_{\text{self}}$ are invisible from the source view, their completed attributes mainly contribute to frontal-view rendering.

3.4.2 Semantic Completion for External Occlusion

External occlusion differs fundamentally from self-occlusion: the corresponding facial surface remains geometrically visible, but its projected appearance is corrupted by a foreground object. We therefore recover its attributes from nearby reliable surface elements that share the same facial semantics. To avoid cross-part propagation, e.g., from lips to surrounding skin, we assign each FLAME face a fixed fine-grained semantic label $r(f)$ in canonical space and restrict donors to reliable faces within the same region. Let $d(f, g)$ denote the Euclidean distance between faces f and g on the canonical FLAME surface. For $f \in \mathcal{O}_{\text{ext}}$, we select the K nearest reliable faces satisfying

$$C_f = \{g \in \mathcal{D}_{\text{obs}} \mid r(g) = r(f)\}, \quad N_K(f) = \text{KNN}_d\left(f, C_f\right). \tag{5}$$

The completed attribute is obtained by distance-weighted aggregation,

$$\widetilde{\boldsymbol{\xi}}_f = \mathcal{A}\left(\left\{\boldsymbol{\xi}_g^0, \frac{e^{-d(f,g)}}{\sum_{h \in N_K(f)} e^{-d(f,h)}}\right\}_{g \in N_K(f)}\right), \tag{6}$$

where $\mathcal{A}$ applies attribute-specific aggregation to Gaussian covariance, opacity, and spherical-harmonic coefficients.

The two completion mechanisms therefore operate on the same mesh-bound representation but exploit different evidence according to the cause of invisibility. During training, pixel-level reconstruction loss is applied only to reliably observed regions to avoid supervision from occluded pixels, while geometric and expression constraints stabilize FLAME estimation.

$$\mathcal{L}_{\text{3D}} = \mathcal{L}_{\text{render}} + \lambda_{\text{lmk}}\mathcal{L}_{\text{lmk}} + \lambda_{\text{expr}}\mathcal{L}_{\text{expr}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{7}$$

where,

$$\begin{aligned} \mathcal{L}_{\text{render}} &= \|M_{ext} \odot (R_{src} - I)\|_1, \quad \mathcal{L}_{\text{lmk}} = \|\boldsymbol{\ell}(I) - \boldsymbol{\ell}(R_{src})\|_2^2, \\ \mathcal{L}_{\text{expr}} &= 1 - \cos(P_{\text{expr}}(R_{src}), P_{\text{expr}}(I)), \quad \mathcal{L}_{\text{reg}} = \|\boldsymbol{\psi} - \boldsymbol{\psi}^0\|_2^2, \end{aligned} \tag{8}$$

where $\boldsymbol{\psi}^0$ denote the expression parameters predicted by the pretrained FLAME regressor. Following expression-consistency objectives in EMOCA (Daněček et al., 2022) and SMIRK (Retsinas et al., 2024), we employ a pretrained frozen QCS (Wang et al., 2025) network as the expression feature extractor P_{expr}.

Training proceeds in two stages. We first freeze the FLAME estimator and optimize only the Gaussian predictor. We then jointly fine-tune both modules with a smaller learning rate, while keeping the identity, pose, and camera branches frozen. The completed representation is finally rendered under the source and frontal views. R_{src} restores evidence occluded by external objects under the original pose, whereas R_{fr} reveals structures hidden by pose-induced self-occlusion. Both renderings are fused with the original image for final FER.

3.5 3D-Structure-Assisted Expression Recognition

The dual-view renderings recover expression evidence missing from the occluded input, but may still contain errors from segmentation, geometry estimation, or attribute completion. We therefore use them as auxiliary cues rather than replacing the original image. Given a shared FER encoder E_{FER} Wang et al. (2025), we extract

$$\mathbf{f}_o = E_{\text{FER}}(I), \quad \mathbf{f}_{src} = E_{\text{FER}}(R_{src}), \quad \mathbf{f}_{fr} = E_{\text{FER}}(R_{fr}), \tag{9}$$

where $\mathbf{f}_o$ preserves reliable appearance from the original image, while $\mathbf{f}_{\mathrm{src}}$ and $\mathbf{f}_{\mathrm{fr}}$ provide complementary evidence recovered from external occlusion and self-occlusion, respectively. These features are fused residually,

$$\begin{aligned} \triangle src = \mathbf{W}_{src}\,\mathrm{LN}(\mathbf{f}_{src} - \mathbf{f}_o), \quad \triangle fr = \mathbf{W}_{fr}\,\mathrm{LN}(\mathbf{f}_{fr} - \mathbf{f}_o), \\ \mathbf{z} = \mathbf{W}_c\,\mathrm{LN}(\mathbf{f}_o + \alpha \triangle src + \beta \triangle fr), \quad \mathcal{L}_{\mathrm{FER}} = \mathrm{CE}(\mathbf{z}, y), \end{aligned} \tag{10}$$

where $\mathbf{W}_{[*]}$ denotes a fully connected projection weight, LN is Layer Normalization, and α and β control the contributions of the source-view and frontal-view features.

This residual design keeps the original image as the primary recognition pathway while using 3D renderings only to supplement missing evidence. After training the reconstruction modules, we freeze the occlusion detector and the entire 3D pipeline, and optimize only E_{FER}, $\mathbf{W}_{src}$, $\mathbf{W}_{fr}$, and the classification head. This decouples recognition from reconstruction and prevents the classification objective from distorting the rendered facial appearance.

4 Experiments

4.1 Dataset and Implementation Details

(1) RAF-DB and FERPlus. RAF-DB (Li et al., 2017a) contains about 30K in-the-wild images. We use the standard seven-class split with 12,271 training and 3,068 test images. FERPlus (Barsoum et al., 2016) re-annotates FER2013 with crowd-sourced label distributions and contains 35,887 images. We follow the official split and the standard eight-class protocol including Contempt. **(2) Occlu-FER and Occlu-FER-Mask.** Occlu-FER (Zhai et al., 2025) contains eight expression categories under real-world occlusion. Since it provides only expression labels, we manually annotate pixel-level external-occlusion masks to construct **Occlu-FER-Mask**, which directly supervises our occlusion detector and supports fine-grained occlusion analysis. **(3) External-occlusion benchmarks.** Occlusion-RAF-DB and Occlusion-FERPlus (Wang et al., 2020) contain 735 and 605 real-occluded test images selected from RAF-DB and FERPlus. **(4) Pose benchmarks.** Pose-RAF-DB and Pose-FERPlus (Wang et al., 2020) evaluate pose-induced self-occlusion. Pose30 subsets contain samples with pose angles above 30°, while Pose45 uses a 45° threshold and presents stronger unilateral facial invisibility.

All images are resized to 224×224, using the official dataset splits. For 3D representation learning, we set $\lambda_{\mathrm{lmk}} = 2$, $\lambda_{\mathrm{expr}} = 0.2$, and $\lambda_{\mathrm{reg}} = 0.05$. FER training uses random rotation, translation, scaling, and horizontal flipping. AdamW is used with batch size 28, weight decay $1e^{-4}$, and 15 training epochs. The initial learning rate is $3e^{-5}$ for RAF-DB and $1e^{-5}$ for FERPlus and Occlu-FER, followed by cosine decay to $1e^{-6}$. α and β are 1.165 and 1.565 for RAF-DB, 0.15 and 0.075 for FERPlus, and 1.45 and 0.215 for Occlu-FER. Top-1 accuracy is reported for all datasets.

4.2 Comparison with State-of-the-Art Methods

We compare OCA3D with representative general and occluded FER methods across seven benchmarks, covering standard FER, real external occlusion, and pose-induced self-occlusion.

(1) Conventional FER and external occlusion. As shown in Table 1, OCA3D achieves 93.06%, 92.15%, 90.19%, and 89.73% Top-1 accuracy on RAF-DB, FERPlus, Occlusion-RAF-DB, and Occlusion-FERPlus, respectively, obtaining the best results across all four benchmarks. Compared with QCS, the strongest general FER baseline, the gains are 0.56% and 0.74% on RAF-DB and FERPlus, increasing to 1.63% and 1.72% on their occluded subsets. The larger improvements under occlusion indicate that OCA3D particularly benefits recognition when expression evidence is missing. Compared with occlusion-specific methods, OCA3D improves over ORSANet from 87.75% to 90.19% on Occlusion-RAF-DB. While ORSANet enhances visible-region representations using landmarks and semantic segmentation, OCA3D additionally recovers missing evidence through explicit 3D topology. Unlike 2D reconstruction methods such as Latent-OFER, our completion is constrained by FLAME geometry and fixed structural correspondences, reducing unconstrained propagation in image space.

(2) Pose-Induced Self-Occlusion. Table 2 evaluates large-pose scenarios, where parts of the facial surface become geometrically invisible. OCA3D achieves 93.50%, 93.01%, 92.85%, and 92.58%

Table 1: Top-1 Accuracy on RAF-DB, FERPlus, Occlusion-RAF-DB, and Occlusion-FERPlus.

Method	RAF-DB	FERPlus	Occlusion-RAF-DB	Occlusion-FERPlus
gACNN (Li et al., 2019)	85.07	–	–	–
PG-CNN (Li et al., 2018)	83.27	–	–	–
WGAN (Lu et al., 2019)	83.49	–	–	–
RAN (Wang et al., 2020)	86.90	88.55	82.72	83.63
OADN (Ding et al., 2020)	89.83	–	–	84.57
VTFF (Ma et al., 2021)	88.14	64.80	83.95	–
EfficientFace (Zhao et al., 2021)	88.36	–	83.24	–
EAC (Zhang et al., 2022)	89.99	89.64	85.46	–
Wang et al. (2022)	85.95	–	82.45	–
MAPNet (Ju & Zhao, 2022)	87.26	–	–	–
POSTER (Zheng et al., 2023)	92.05	91.62	86.82	–
Latent-OFER (Lee et al., 2023)	89.60	–	84.20	–
FERMixNet (Huang et al., 2025)	89.86	89.58	86.67	85.95
POSTER++ (Mao et al., 2025)	92.21	–	87.02	–
ORSANet (Zhai et al., 2025)	92.28	–	87.75	–
QWTR (Zhou et al., 2025)	91.31	–	85.35	–
QCS (Wang et al., 2025)	<u>92.50</u>	<u>91.41</u>	<u>88.56</u>	<u>88.01</u>
Wang et al. (2026)	91.75	–	86.80	–
Ours	**93.06**	**92.15**	**90.19**	**89.73**

on Pose30-RAF-DB, Pose45-RAF-DB, Pose30-FERPlus, and Pose45-FERPlus, respectively. These results exceed the best previous results by 4.24%, 4.08%, 4.30%, and 5.08%, with an average gain of 4.43%. The advantage remains pronounced at 45°, where stronger unilateral facial invisibility occurs. Unlike approaches that rely mainly on attention or visible-region modeling, OCA3D explicitly identifies geometrically invisible surfaces, completes them through bilateral topological correspondence, and exposes the recovered structures through frontal-view rendering. The larger gains under extreme poses are consistent with the intended role of self-occlusion modeling.

Table 2: Comparison of Top-1 Accuracy on Pose-RAF-DB and Pose-FERPlus.

Method	Pose30-RAF-DB	Pose45-RAF-DB	Pose30-FERPlus	Pose45-FERPlus
RAN (Wang et al., 2020)	86.74	85.20	82.23	80.40
OADN (Ding et al., 2020)	–	–	88.52	<u>87.50</u>
VTFF (Ma et al., 2021)	87.97	88.35	88.29	87.20
EfficientFace (Zhao et al., 2021)	88.13	86.92	–	–
FERMixNet (Huang et al., 2025)	89.01	87.63	<u>88.55</u>	86.73
QWTR (Zhou et al., 2025)	<u>89.26</u>	<u>88.93</u>	–	–
Ours	**93.50**	**93.01**	**92.85**	**92.58**

(3) Real-world complex occlusion. On Occlu-FER, OCA3D achieves 70.91% accuracy, outperforming QCS (68.52%), ORSANet (68.07%), and POSTER (66.82%) by 2.39%, 2.84%, and 4.09%. It also surpasses EfficientFace (65.96%), EAC (66.36%), and POSTER++ (66.09%). Since Occlu-FER contains more diverse occluders and spatial configurations than the curated occlusion subsets, this result suggests that OCA3D generalizes beyond a narrow set of predefined occlusion patterns.

Overall, local-region and attention-based methods (Li et al., 2019; Wang et al., 2020; Ding et al., 2020; Ju & Zhao, 2022; Zhai et al., 2025) mainly suppress corrupted evidence, while general FER methods (Mao et al., 2025; Zhou et al., 2025; Wang et al., 2025) strengthen feature representation without explicitly recovering missing information. Reconstruction-based approaches (Lu et al., 2019; Lee et al., 2023) recover missing content but largely operate in 2D space. OCA3D instead distinguishes external occlusion from pose-induced self-occlusion through 3D geometry and applies cause-specific completion to each. Its consistent gains across standard, external-occlusion, and

large-pose settings support the benefit of occlusion-cause-aware 3D evidence recovery for FER. For more experimental results, such as ablation studies and visualizations, please refer to the Sup. A.2.

4.3 Robustness under Different Occlusion Ratios

To evaluate robustness to different levels of occlusion, we construct a controlled protocol on RAF-DB by covering approximately 10%, 20%, and 30% of the image region with black patches. By fixing the occluder appearance while varying its spatial extent, this setting isolates the effect of progressively missing facial evidence. This setting fixes the occlusion type while varying its extent, providing a direct measure of sensitivity to missing expression evidence. As shown in Table 3, OCA3D achieves 90.84% and 88.01% accuracy at 10% and 20% occlusion, ranking first in both settings. It outperforms ORSANet by 0.33% and 0.14%, and QCS by 0.59% and 0.76%, respectively. At 30% occlusion, it reaches 83.83%, only 0.19% below ORSANet, while remaining ahead of QCS, POSTER++, and other baselines. The performance degrades gradually as the occluded area increases, reflecting the reduced amount of reliable facial evidence available for recovery. Nevertheless, OCA3D remains competitive across all occlusion levels. This robustness is consistent with its cause-aware completion mechanism: externally occluded surface elements are not inferred solely from their corrupted image observations, but recover Gaussian attributes from reliable mesh faces within the same semantic region. The fixed FLAME topology further provides a stable 3D reference for locating and transferring expression-related evidence as the visible facial area decreases.

Table 3: Evaluation results of artificially simulated occlusion scenarios ($10\% - 30\%$) on RAF-DB.

Method	Original	10%	20%	30%
EfficientFace (Zhao et al., 2021)	88.36	87.93	84.79	80.93
POSTER (Zheng et al., 2023)	92.05	89.76	86.86	82.53
FERMixNet (Huang et al., 2025)	89.86	84.70	82.54	78.91
POSTER++ (Mao et al., 2025)	92.21	89.76	86.86	82.53
ORSANet (Zhai et al., 2025)	92.28	<u>90.51</u>	<u>87.87</u>	**84.02**
QCS (Wang et al., 2025)	<u>92.50</u>	90.25	87.25	82.49
Wang's (Wang et al., 2026)	91.75	86.23	84.38	80.61
Ours	**93.06**	**90.84**	**88.01**	<u>83.83</u>

4.4 Discussion

Our results show that distinguishing external occlusion from self-occlusion enables effective cause-specific evidence recovery, but several limitations remain. First, as occlusion increases, fewer reliable faces are available for Gaussian attribute propagation, limiting recovery when most expression-relevant surfaces are missing. This partly explains the reduced advantage at (30%) occlusion. Second, external-occlusion segmentation, FLAME estimation, Gaussian prediction, completion, and dual-view rendering introduce greater computational cost than most 2D FER methods. Lightweight 3D representations, shared encoders, and more efficient rendering could improve practical deployment. Finally, bilateral correspondence may be less reliable for strongly asymmetric expressions. More flexible cross-region propagation could better handle such cases and extreme occlusions.

5 Conclusion

We presented OCA3D, an occlusion-cause-aware framework for robust FER that distinguishes external occlusion from pose-induced self-occlusion and recovers missing expression evidence on a FLAME-constrained 3D Gaussian representation. OCA3D uses semantic-region-constrained correspondence for external occlusion and topology-defined bilateral correspondence for self-occlusion, with dual-view rendering providing complementary evidence for recognition. We also introduced Occlu-FER-Mask, a manually annotated dataset with pixel-level external-occlusion masks for real-world FER. Experiments across standard, real-world occlusion, and large-pose benchmarks demonstrate the effectiveness of cause-aware 3D completion, highlighting the importance of modeling why facial evidence is unavailable rather than only where it is missing.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring mandatory disclosure, and such tasks are not applicable to this work. We used generative AI tools solely for language editing, including improving grammar, clarity, and academic writing style. All AI-assisted revisions were carefully reviewed by the authors to ensure that they faithfully preserved the original technical content, claims, and intended meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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*, pp. 279–283, 2016.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 157–164. 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *Advances in Neural Information Processing Systems*, volume 37, pp. 57642–57670, 2024. doi: 10.52202/079017-1838.

Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). In *International Conference on Learning Representations*, volume 2024, pp. 32573–32592, 2024.

Radek Daněček, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322, 2022.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Hui Ding, Peng Zhou, and Rama Chellappa. Occlusion-adaptive deep network for robust facial expression recognition. In *2020 IEEE International Joint Conference on Biometrics*, pp. 1–9, 2020. doi: 10.1109/IJCB48548.2020.9304923.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Dimitrios Gerogiannis, Foivos Paraperas Papantoniou, Rolandos Alexandros Potamias, Alexandros Lattas, and Stefanos Zafeiriou. Arc2Avatar: Generating expressive 3D avatars from a single image via ID guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10782, 2025.

Baojin Huang, Zhongyuan Wang, Guangcheng Wang, Kui Jiang, Kangli Zeng, Zhen Han, Xin Tian, and Yuhong Yang. When face recognition meets occlusion: A new benchmark. In *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4240–4244, 2021. doi: 10.1109/ICASSP39728.2021.9413893.

Yansong Huang, Junjie Peng, Wenqiang Zhang, Tong Zhao, Gan Chen, Shuhua Tan, Fen Yi, and Lu Wang. FERMixNet: An occlusion robust facial expression recognition model with facial mixing augmentation and mid-level representation learning. *IEEE Transactions on Affective Computing*, 16(2):639–654, 2025. doi: 10.1109/TAFFC.2024.3454102.

Lingzhao Ju and Xu Zhao. Mask-based attention parallel network for in-the-wild facial expression recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2410–2414, 2022. doi: 10.1109/ICASSP43922.2022.9747717.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Isack Lee, Eungi Lee, and Seok Bong Yoo. Latent-OFER: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1536–1546, 2023. doi: 10.1109/ICCV51070.2023.00148.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2584–2593, 2017a. doi: 10.1109/CVPR.2017.277.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, 2017b. doi: 10.1145/3130800.3130813.

Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated CNN for occlusion-aware facial expression recognition. In *2018 24th International Conference on Pattern Recognition*, pp. 2209–2214, 2018. doi: 10.1109/ICPR.2018.8545853.

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, 2019. doi: 10.1109/TIP.2018.2886767.

Yang Liu, Fei Wang, Jiankang Deng, Zhipeng Zhou, Baigui Sun, and Hao Li. MogFace: Towards a Deeper Appreciation on Face Detection . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4083–4092, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00406. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00406.

Yang Lu, Shigang Wang, Wenting Zhao, and Yan Zhao. WGAN-based robust occluded facial expression recognition. *IEEE Access*, 7:93594–93610, 2019. doi: 10.1109/ACCESS.2019.2928125.

Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. FaceLift: Learning generalizable single image 3D face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12691–12701, 2025.

Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 14(2):1236–1248, 2021. doi: 10.1109/TAFFC.2021.3122146.

Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, and Yigang Wang. POSTER++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 157:110951, 2025. doi: 10.1016/j.patcog.2024.110951.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20299–20309, 2024.

Jianqiang Ren, Lin Liu, and Steven Hoi. OMG-Avatar: One-shot multi-LOD gaussian head avatar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11017–11028, 2026.

George Retsinas, Panagiotis P. Filntisis, Radek Daněček, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3D facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2490–2501, 2024.

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7755–7764. IEEE, 2019.

Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1606–1616, 2024.

Chengpeng Wang, Li Chen, Lili Wang, Zhaofan Li, and Xuebin Lv. QCS: Feature refining from quadruplet cross similarity for facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7563–7572, 2025. doi: 10.1609/aaai.v39i7.32814.

Cong Wang, Jian Xue, Ke Lu, and Yanfu Yan. Light attention embedding for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1834–1847, 2021. doi: 10.1109/TCSVT.2021.3083326.

Jiahe Wang, Heyan Ding, and Shangfei Wang. Occluded facial expression recognition using self-supervised learning. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1077–1092, 2022. URL https://openaccess.thecvf.com/content/ACCV2022/html/Wang_Occluded_Facial_Expression_Recognition_using_Self-supervised_Learning_ACCV_2022_paper.html.

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. doi: 10.1109/TIP.2019.2956143.

Wei Wang. X-anylabeling: A unified desktop platform for ai-assisted data annotation. https://github.com/CVHub520/X-AnyLabeling, 2023.

Yanzhong Wang, Daming Shi, and Muhammad Sadiq. Occlusion-aware low-rank learning framework for facial expression recognition. *Computer Vision and Image Understanding*, 269:104820, 2026. doi: 10.1016/j.cviu.2026.104820.

Zehao Xia, Yiqun Wang, Zhengda Lu, Kai Liu, Jun Xiao, and Peter Wonka. OMEGA-Avatar: One-shot modeling of 360° gaussian avatars. *arXiv preprint arXiv:2602.11693*, 2026. URL https://arxiv.org/abs/2602.11693.

Zhen Xing, Weimin Tan, Ruian He, Yangle Lin, and Bo Yan. Co-completion for occluded facial expression recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 130–140. Association for Computing Machinery, 2022. doi: 10.1145/3503161.3548183.

Xiangnan Yin and Liming Chen. FaceOcc: A diverse, high-quality face occlusion dataset for human face extraction. *CoRR*, abs/2201.08425, 2022. URL https://arxiv.org/abs/2201.08425.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Huiyu Zhai, Xingxing Yang, Yalan Ye, Chenyang Li, Bin Fan, and Changze Li. Rethinking occlusion in FER: A semantic-aware perspective and go beyond. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5567–5576, 2025. doi: 10.1145/3746027.3754978.

Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi S Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. 2023.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Ligang Zhang, Brijesh Verma, Dian Tjondronegoro, and Vinod Chandran. Facial expression analysis under partial occlusion: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–49, 2018.

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*, volume 13686 of *Lecture Notes in Computer Science*, pp. 418–434. Springer, 2022. doi: 10.1007/978-3-031-19809-0_24.

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*, volume 35, pp. 3510–3519, 2021. doi: 10.1609/aaai.v35i4.16465.

Ce Zheng, Matias Mendieta, and Chen Chen. POSTER: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3146–3155, 2023.

Yu Zhou, Jialun Pei, Weixin Si, Jing Qin, and Pheng-Ann Heng. Delving into quaternion wavelet transformer for facial expression recognition in the wild. *IEEE Transactions on Multimedia*, 27: 3895–3909, 2025. doi: 10.1109/TMM.2025.3535361.

A Appendix

A.1 Occlu-FER-Mask

Figure 2 shows several manually annotated examples. The first row contains real occluded faces cropped using detected face bounding boxes, while the second row shows the corresponding pixel-level external-occlusion masks. Specifically, we first use MogFace (Liu et al., 2022) to detect the facial bounding box and then employ X-AnyLabeling (Wang, 2023) to manually annotate, at the pixel level, external occlusions within the facial region caused by hands, masks, glasses, hair, and other foreground objects. This procedure produces occlusion supervision that is spatially aligned with the facial regions used as input to the FER model.



Figure 2: Examples from Occlu-FER-Mask. The first row shows the original occluded facial images, and the second row shows the manually annotated pixel-level occlusion masks, where white pixels indicate external occlusions.

Existing datasets such as FaceOcc (Yin & Chen, 2022) provide useful supervision for facial occlusion segmentation, but remain limited in both scale and scene diversity. FaceOcc contains approximately 2,200 images, most of which exhibit relatively clear facial appearances, while their resolution, imaging conditions, and background variations still differ from those encountered in unconstrained in-the-wild FER scenarios. Consequently, occlusion segmentation models trained solely on such data may have limited generalization to complex real-world conditions. In contrast, our Occlu-FER-Mask contains 7,718 manually annotated images, approximately 3.5 times the size of FaceOcc, and covers a substantially broader range of real occlusion categories, spatial locations, scales, and challenging imaging conditions. More importantly, all samples are drawn directly from real-world occluded FER scenarios, making their data distribution better aligned with the downstream task considered in this work. This provides more targeted supervision for learning

external-occlusion segmentation models with stronger generalization to in-the-wild FER. Therefore, the value of Occlu-FER-Mask lies not only in expanding the scale of pixel-level occlusion annotations, but also in reducing the distribution gap between generic facial occlusion datasets and real-world occluded FER scenarios.

A.2 Ablation Studies

A.2.1 Hyperparameter Sensitivity

We further analyze the sensitivity of four key hyperparameters: the landmark constraint weight λ_{lmk}, the expression-consistency weight λ_{expr}, the FLAME regularization weight λ_{reg}, and the number of nearest neighbors K used for semantic neighborhood completion.

As shown in Table 4, the default configuration achieves the highest accuracy of 70.9%, indicating that an appropriate balance among these constraints is important. For the neighborhood size K, reducing it from 4 to 2 and 1 decreases accuracy to 69.6% and 68.5%, respectively. This suggests that relying on only a few neighboring faces provides insufficient evidence for stable local attribute estimation, whereas a moderately larger set of reliable donors within the same semantic region offers more informative support for completion. The geometry- and expression-related losses exhibit similar sensitivity. When λ_{lmk} is changed from 2 to 1 or 4, performance drops by 2.2% and 1.9%, respectively. A weak landmark constraint may lead to insufficient alignment between FLAME geometry and the input face, whereas an overly strong constraint can limit the flexibility needed to adapt expression-related geometry. Likewise, $\lambda_{\mathrm{expr}} = 0.2$ outperforms 0.1 and 0.4 by 1.8% and 2.1%, respectively. This indicates that a moderate expression-consistency constraint helps preserve discriminative expression information in the rendered faces, while either insufficient or excessive supervision can degrade the final recognition performance. The regularization term shows a comparable trend. Increasing λ_{reg} from 0.05 to 0.1 reduces accuracy to 69.3%, while decreasing it to 0.025 further lowers performance to 67.9%. These results suggest that appropriate regularization is important for stabilizing FLAME expression parameters during joint optimization: overly strong regularization restricts adaptation to the expression of the current input, whereas overly weak regularization allows the geometry parameters to deviate excessively from the pretrained initialization. Overall, the results indicate that the adopted default configuration provides a favorable balance among geometric alignment, expression preservation, parameter stability, and the spatial extent of local attribute propagation.

Table 4: Classification accuracy on Occlu-FER under different hyperparameter configurations.

λ_{lmk}	2	2	2	2	2	2	2	1	4
λ_{expr}	0.2	0.2	0.2	0.2	0.2	0.1	0.4	0.2	0.2
λ_{reg}	0.05	0.05	0.05	0.1	0.025	0.05	0.05	0.05	0.05
K	4	2	1	4	4	4	4	4	4
Accuracy (%)	70.9	69.6	68.5	69.3	67.9	69.1	68.8	68.7	69.0

A.2.2 Qualitative and Quantitative Comparison of Facial Evidence Recovery

To verify whether the proposed occlusion-cause-aeare 3D completion mechanism can recover facial evidence that is truly useful for expression recognition, we compare it with representative 2D restoration and 3D face reconstruction methods from both qualitative and quantitative perspectives. The 2D baselines include CoPaint (Zhang et al., 2023) and Latent-OFER Lee et al. (2023). CoPaint restores missing regions using a diffusion model, whereas Latent-OFER reconstructs occluded facial content with a generative adversarial network. For 3D methods, we use the same estimated FLAME geometry and obtain facial appearance with SMIRK (Retsinas et al., 2024), DECA (Feng et al., 2021), and GPAvatar (Chu et al., 2024), respectively, and compare them with our mesh-bound 3D Gaussian completion. Specifically, SMIRK adopts sampling-based appearance rendering, DECA employs neural rendering, and GPAvatar represents facial appearance with a Neural Radiance Field Mildenhall et al. (2020). To minimize the influence of classifier differences, all quantitative experiments use the same QCS (Wang et al., 2025) model as the FER classifier.

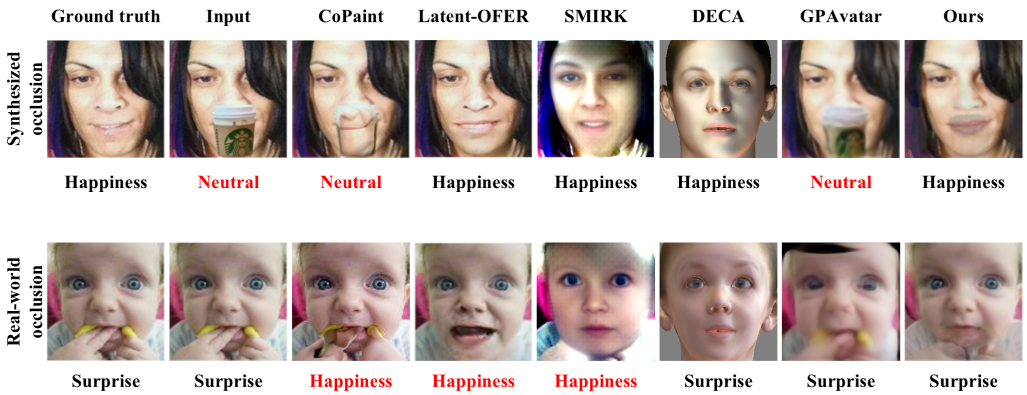


Figure 3: Qualitative comparison of synthetic and real-world occlusion recovery. Predicted labels are obtained from the recovered images, with incorrect predictions highlighted in red.

As shown in Figure 3, the compared recovery methods exhibit clear differences between visual plausibility and expression consistency. In the synthetic occlusion example, the occluder covers the highly discriminative mouth region, causing the occluded image to be incorrectly classified as Neutral. Although CoPaint fills in the missing region, the recovered mouth structure still fails to provide the correct expression cue, and the prediction remains Neutral. GPAvatar produces an incorrect prediction, due to residual occlusion or appearance reconstruction errors. Latent-OFER, SMIRK, DECA, and our method recover sufficient facial evidence to support the correct Happiness prediction. However, the outputs of SMIRK and DECA exhibit more noticeable changes in identity, texture, or rendering domain. In contrast, our method builds on the original reliable observations and performs 3D structure-constrained attribute propagation only for missing regions, allowing it to recover local evidence consistent with the ground-truth expression while preserving the visible appearance of the input as much as possible.

This distinction is even more evident in the real-world occlusion example. Although CoPaint, Latent-OFER, and SMIRK all generate visually complete or plausible faces, their recovery of the mouth shape alters the original expression evidence, causing QCS to misclassify all three results as Happiness. DECA, GPAvatar, and our method preserve the correct Surprise prediction. In particular, our method recovers the occluded region while retaining the originally visible facial content. This example highlights an important property of occluded FER: visually plausible completion is not equivalent to expression-consistent completion. A useful recovery process should not only produce a credible facial appearance, but more importantly avoid altering the local structures that determine the underlying expression category.

Table 5: Quantitative comparison on RAF-DB and Occlusion-RAF-DB.

Method	RAF-DB	Occlusion-RAF-DB
QCS	92.50	88.56
QCS (masked image)	92.76	89.78
QCS (CoPaint)	92.57	90.07
QCS (DECA)	92.89	89.78
QCS (SMIRK)	92.56	88.86
QCS (GPAvatar)	92.73	89.65
Ours	93.06	90.19

The quantitative results further support this observation. As shown in Table 5, the original QCS achieves 92.50% and 88.56% accuracy on RAF-DB and Occlusion-RAF-DB, respectively. Simply masking out the occluded regions improves performance on Occlusion-RAF-DB to 89.78%, indicating that removing occlusion-related noise alone is already beneficial. However, suppressing unreliable observations cannot recover expression information that has become unavailable. CoPaint further improves accuracy to 90.07%, while DECA, SMIRK, and GPAvatar achieve 89.78%,

88.86%, and 89.65%, respectively. In comparison, our complete 3D Gaussian completion reaches 93.06% on RAF-DB and 90.19% on Occlusion-RAF-DB, achieving the highest accuracy among all compared settings.

Notably, on Occlusion-RAF-DB, our method improves over the original QCS by 1.63% and over the masked-QCS baseline by 0.41%. This difference indicates that the gain does not arise solely from removing occluder interference, but also from effectively recovering missing expression evidence. Compared with the strongest image-restoration baseline, CoPaint, our method still provides a further 0.12% improvement. Meanwhile, accuracy on the original RAF-DB benchmark increases from 92.50% to 93.06%, suggesting that introducing 3D auxiliary evidence does not come at the cost of a noticeable degradation in conventional FER performance.

Overall, the qualitative and quantitative results reveal an important principle for FER-oriented completion: the objective should not be visual realism alone, but the preservation and recovery of discriminative expression information. 2D generative models may synthesize visually plausible content whose expression semantics drift from the original input, while general-purpose 3D face reconstruction methods are primarily designed for geometric reconstruction or novel-view rendering rather than the recovery of missing expression evidence. By anchoring Gaussian attributes to the FLAME topology and propagating them only from reliable observations according to the underlying cause of occlusion, our method constrains recovery through stable 3D correspondences, yielding a better balance between visual restoration and downstream FER.