OCCLUSION-AWARE DUAL-VIEW MASKED AUTOENCODING FOR FACIAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual intelligence is increasingly shifting from world-centric perception toward human-centric understanding. However, the visual evidence required to interpret human behavior and affective states remains susceptible to occlusion-induced corruption. Existing masked pretraining paradigms leave this setting underexplored, as their objectives can be satisfied by recovering missing facial content from visible context without requiring the encoder to distinguish facial evidence from occluder-induced appearance. When corrupted regions remain visible, the pretraining objective does not explicitly constrain the encoder against treating occluder-specific patterns as valid facial evidence, allowing them to persist alongside behavior-relevant facial information in the learned representation. We propose a dual-view masked autoencoding framework that learns from paired clean and synthetically occluded facial views. In the occluded branch, corrupted patches remain visible and are guided to reconstruct their corresponding clean facial content rather than the occluder appearance, directly linking corrupted observations to the facial information they obscure. This encourages the encoder to preserve facial semantics while reducing reliance on occluder-specific cues. Experiments on facial action unit detection and facial expression recognition demonstrate the effectiveness of the learned representations.

1 INTRODUCTION

Visual intelligence is increasingly shifting from world-centric perception toward human-centric understanding. Recent multimodal research has expanded beyond object and scene recognition toward fine-grained perception and interpretation of people, including facial attributes, expressions, and action units (Liu et al., 2026; Wang et al., 2025c; Zheng et al., 2025; Chaubey et al., 2026). Reliable human understanding depends on visual observations that faithfully reflect behavior and affective states. Facial occlusion weakens this correspondence because the observed appearance no longer arises solely from the face. Facial information may be hidden or altered, while appearance introduced by glasses, masks, hands, hair, or other objects remains visible in the same observation. Consequently, a facial representation must not only recover useful information from degraded inputs, but also avoid treating occluder-induced appearance as evidence of the underlying facial behavior.

Masked image modeling has become an effective paradigm for visual representation learning (He et al., 2022) and has increasingly been explored for facial representation and behavior understanding (Qiu et al., 2024; Ning et al., 2025). Existing masked pretraining paradigms primarily model missing content rather than corrupted visual evidence that remains present in the encoder input. During masked reconstruction, selected content is withheld from the encoder and inferred from the remaining visible context. The encoder is therefore supervised to recover unavailable facial information, but it is not directly constrained to associate an altered visible observation with the facial content that the alteration obscures. This distinction becomes important under occlusion because corrupted regions still enter the encoder and their object-specific appearance participates in representation formation. Without supervision connecting these observations to their underlying facial content, occluder-specific appearance may be retained together with behavior-relevant facial information. We refer to this discrepancy between missing-content reconstruction and corrupted-content representation as the masking–occlusion mismatch, as illustrated in Figure 1.

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.

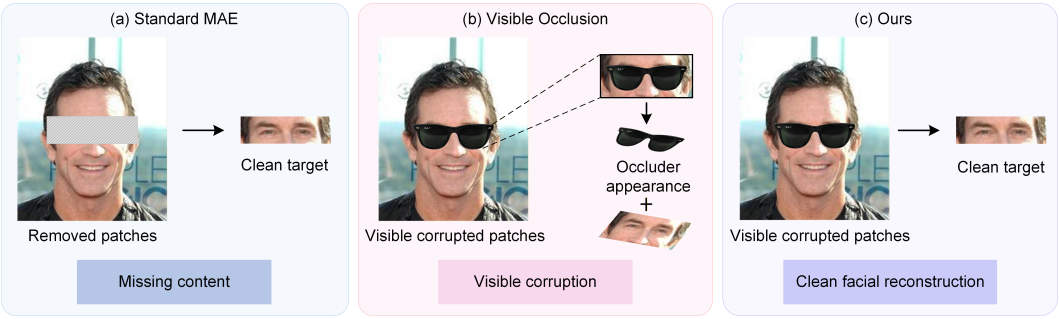


Figure 1: Conceptual comparison between standard masked autoencoding, visible facial occlusion, and our formulation. Standard MAE reconstructs facial content removed from the encoder input, whereas real occlusion leaves corrupted regions visible and introduces occluder appearance. Our method retains these visibly corrupted patches during encoding and reconstructs their corresponding clean facial content.

Recent studies address facial occlusion through representation transfer and explicit recovery. Jiang et al. (2024) mine the latent space of a pretrained masked autoencoder for occluded AU recognition. Zhang et al. (2025a) combine occlusion detection with facial expression restoration before recognition, while Wang et al. (2026) integrate occlusion detection, masking, feature-level recovery, and non-occluded guidance for FER. These approaches can suppress unreliable observations, recover affected information, or introduce guidance from cleaner facial evidence. However, task labels and feature-level alignment constrain the prediction or the overall representation, but do not identify which visual components originate from the face and which are introduced by the occluder. Occluder-specific cues may therefore remain embedded in the representation even when these downstream objectives are satisfied. Because this mixture is formed as corrupted inputs are encoded, downstream supervision can only constrain the resulting representation rather than how the corrupted evidence is initially represented. This motivates moving occlusion handling to the pretraining stage, where the encoder can learn from corrupted facial inputs before adaptation to any specific recognition task.

To address this problem, we propose an Occlusion-Aware Dual-View Masked Autoencoding framework (ODMAFace) to provide this missing supervision during pretraining. For each image, clean and synthetically occluded views are constructed and processed by a shared encoder. The clean branch retains conventional masked reconstruction, while the occluded branch keeps corrupted patches visible so that the encoder is trained on the appearance changes introduced by occlusion. A learnable occlusion embedding identifies the affected tokens during pretraining, and their reconstruction targets are derived from the corresponding clean facial content rather than the observed occluder appearance. This design encourages the encoder to preserve informative facial cues while reducing its reliance on occluder-specific appearance. After pretraining, the decoder and all occlusion-specific components are discarded, leaving only the shared encoder for downstream facial analysis without requiring occlusion information at inference time.

The main contributions are summarized as follows.

- We propose ODMAFace, a dual-view masked autoencoding framework that learns jointly from clean and synthetically occluded facial views to introduce corrupted visual evidence into facial representation pretraining.
- We introduce an occlusion-guided reconstruction mechanism that marks affected tokens with a learnable embedding and reconstructs their clean facial content, explicitly linking corrupted visual evidence to its underlying facial information.
- We establish a fixed paired object-style occlusion evaluation protocol and evaluate ODMAFace on both facial action unit detection and facial expression recognition under clean and occluded conditions.

2 Related Work

2.1 Facial Representation Learning

Recent facial representation learning has increasingly shifted toward self-supervised and transferable pretraining. LAFS (Sun et al., 2024) uses landmark-based patches and augmentations to focus representation learning on regions relevant to face recognition. This introduces a dependence on extracted landmarks and emphasizes identity discrimination. Gao & Patras (2024) extend facial representation learning beyond image-level consistency by introducing facial region awareness, jointly learning global and local representations across augmented views. Masked reconstruction provides another route for exploiting facial structure. Qiu et al. (2024) use facial MAE pretraining and AU-related adaptation for compound expression recognition, while Ning et al. (2025) further study large-scale masked pretraining with Face9M and identity adversarial training to reduce identity-related shortcut learning in AU detection. Their work highlights that representation quality depends on both the information learned during pretraining and the factors retained during downstream adaptation.

More recent work enriches facial representations through complementary objectives and semantic supervision. FSFM (Wang et al., 2025b) combines masked image modeling with instance discrimination to learn local and global facial features for face-security transfer. MF^2 (Zheng et al., 2025) incorporates structured AU and emotion descriptions to strengthen facial semantic alignment, while Face-LLaVA (Chaubey et al., 2026) extends facial analysis through multimodal instruction tuning. These approaches broaden the semantic information available to representation learning, with language-based methods additionally relying on descriptions or instruction data.

2.2 Occlusion-Aware Facial Learning

Occlusion-aware facial analysis has progressed from suppressing unreliable regions toward exploiting relationships between degraded and clean observations. MFCA-Net (Guo et al., 2024) uses multi-feature cross-attention to exploit visible facial information under occlusion. Jiang et al. (2024) further transfer information from the latent space of a pretrained masked autoencoder to occluded AU recognition, while Zhang et al. (2025b) and Chen et al. (2025) use clean-face or teacher guidance through knowledge distillation for FER. These methods improve recognition by reweighting visible evidence or transferring complementary information, but the degraded input itself remains the observation available to the recognition model.

Explicit recovery methods instead reconstruct affected information before or during recognition. Zhang et al. (2025a) combine occlusion detection with MAE-based facial restoration, while Wang et al. (2026) perform feature-level low-rank recovery with non-occluded guidance. GCA-ODN (Sadiq, 2026) further integrates global context, attentive dropout, and low-rank refinement for joint landmark and expression analysis. These approaches mainly handle occlusion within downstream recognition or recovery pipelines. We instead move the problem to pretraining by retaining occluder appearance in the encoder input and supervising affected patches with their clean facial content, while discarding all occlusion-specific components before downstream adaptation.

3 Method

3.1 Overall Framework

We propose Occlusion-Aware Dual-View Masked Autoencoding for Facial Representation Learning (ODMAFace). Given a facial image $x \in \mathbb{R}^{H\times W\times 3}$, we use it as the clean view and generate a synthetically occluded counterpart $\tilde{x}$ online during pretraining. Here, "clean" refers to the source image before synthetic corruption and does not imply that the original image is free of occlusion.

As illustrated in Figure 2, ODMAFace processes the two views through a shared ViT encoder and reconstruction decoder. The clean branch reconstructs randomly masked patches from the remaining visible tokens. The occluded branch encodes the full corrupted image, adds a learnable occlusion embedding to patches identified by the synthetic mask, and reconstructs their source-image content. Both branches use clean reconstruction targets but differ in their input conditions and supervised regions. Only the encoder is transferred to downstream tasks.

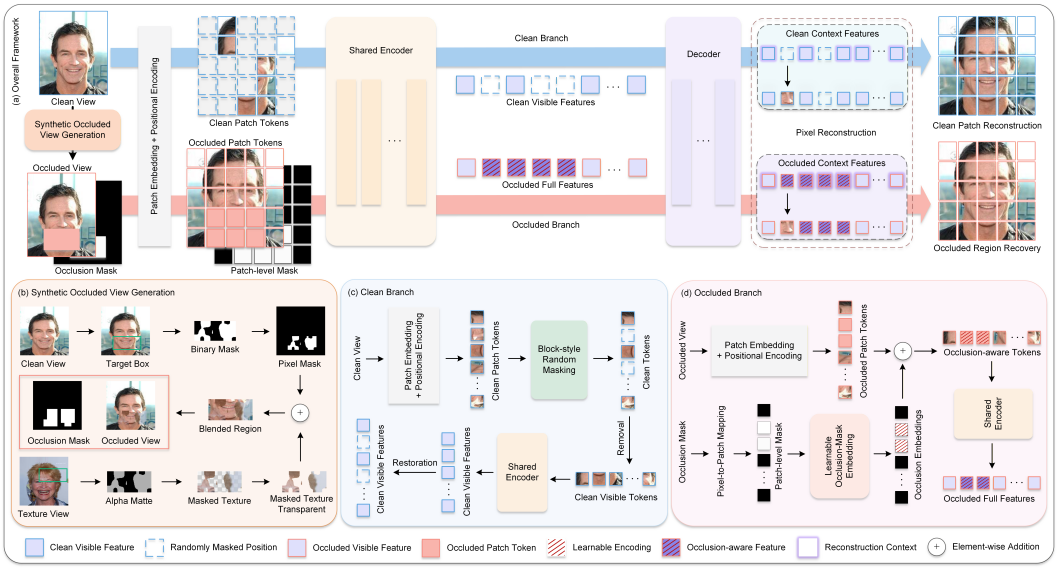


Figure 2: Overview of the proposed ODMAFace framework. (a) The overall pipeline forms a clean view and a synthetically occluded view from the same facial image, maps image patches and pixel-level occlusion masks to patch tokens and occlusion indicators, and processes both branches with shared encoder and decoder weights. (b) The synthetic occluded view is generated by injecting occluder contents into sampled facial regions using binary masks and alpha blending, while the pixel-level mask is mapped to a patch-level occlusion indicator. (c) The clean branch follows masked autoencoding: block-style random masking removes selected clean patches from the encoder input, and the decoder reconstructs missing facial content from the remaining context. (d) The occluded branch keeps corrupted patches visible to the encoder, adds learnable occlusion embeddings to affected patch tokens, and reconstructs the original clean facial content rather than the occluded appearance. After pretraining, only the shared encoder is retained for downstream tasks.

3.2 Synthetic Occluded View Generation

For each input image x, a binary pixel-level mask $M \in \{0,1\}^{H\times W}$ specifies the spatial support of the synthetic occlusion. The generator varies its shape, texture, and opacity to produce diverse local corruptions. The occluded view is formed as

$$\tilde{x} = (1 - \alpha M) \odot x + \alpha M \odot t, \tag{1}$$

where t denotes the occlusion texture, α controls the blending opacity, and $\odot$ denotes element-wise multiplication, with the spatial mask broadcast across RGB channels. The texture is obtained by cut-pasting a random region from another image or sampling random noise based on image-level color statistics. Elliptical and irregular masks provide variation in the spatial shape of the corruption.

The pixel-level mask is mapped to the patch space of the encoder. Let Ω_i denote the set of pixels in patch i. Its occlusion coverage is

$$\rho_i = \frac{1}{|\Omega_i|} \sum_{(u,v)\in\Omega_i} M_{u,v}, \tag{2}$$

and the binary patch-level indicator is

$$o_i = \mathbb{I}(\rho_i > 0), \qquad i = 1, \ldots, N, \tag{3}$$

where N is the number of image patches and $\mathbb{I}$ is the indicator function. Thus, any patch intersecting the generated occlusion is marked as affected. The continuous coverage ρ_i is used only to derive o_i and does not weight the reconstruction loss or serve as an encoder input. The indicator is obtained from the synthetic transformation during pretraining and requires no external occlusion annotation.

3.3 CLEAN BRANCH

The clean branch adopts masked autoencoding with block-style random masking. The clean view is divided into non-overlapping patches, and contiguous groups of patch tokens are removed from the encoder input. Let $m \in \{0,1\}^N$ denote the masking indicator, where $m_i = 1$ identifies a removed patch. The remaining tokens are processed by the shared ViT encoder, and a lightweight decoder D_ψ reconstructs the masked patch targets. The same decoder is used by the occluded branch.

Following normalized-pixel reconstruction, the target for patch i is

$$y_i = \frac{\mathrm{vec}(x_{\Omega_i}) - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}, \tag{4}$$

where $\mathrm{vec}(\cdot)$ flattens the RGB patch, μ_i and σ_i^2 are its pixel mean and variance, and ϵ is a small constant for numerical stability. The clean-branch loss is

$$\mathcal{L}_{\mathrm{mae}} = \frac{1}{\sum_{i=1}^N m_i} \sum_{i=1}^N m_i \left\| \hat{y}_i^{\mathrm{mae}} - y_i \right\|_2^2, \tag{5}$$

where $\hat{y}_i^{\mathrm{mae}}$ denotes the decoder prediction for patch i. This objective trains the encoder to represent facial structure from incomplete visual context.

3.4 OCCLUDED BRANCH

The occluded branch retains the full corrupted image in the encoder input. Affected patches may contain both residual facial information and appearance introduced by the synthetic occluder. Their content is preserved, while the corruption locations are indicated through a learnable embedding.

Occlusion embedding. Let W_p denote the linear patch projection, $p_i \in \mathbb{R}^d$ the positional embedding of patch i, and $e_{\mathrm{occ}} \in \mathbb{R}^d$ the learnable occlusion embedding, where d is the token dimension. The initial token for patch i is

$$\tilde{z}_i^{(0)} = W_p \,\mathrm{vec}(\tilde{x}_{\Omega_i}) + p_i + o_i e_{\mathrm{occ}}. \tag{6}$$

The occlusion embedding is an additive offset and does not replace the visual content of the patch. Unaffected patches use the standard patch embedding, while affected patches additionally receive information about their corruption status. The class token does not receive this offset.

Let F_θ denote the shared encoder. The complete token sequence is encoded as

$$\tilde{Z} = F_\theta\left([\tilde{z}_{\mathrm{cls}}^{(0)}, \tilde{z}_1^{(0)}, \ldots, \tilde{z}_N^{(0)}]\right), \tag{7}$$

where $\tilde{z}_{\mathrm{cls}}^{(0)}$ is the class token and $\tilde{Z}$ denotes the encoded sequence. The shared decoder D_ψ then produces patch reconstruction predictions.

Clean-target reconstruction. The occluded branch uses the normalized source-image targets y_i defined in Eq. (4). Its loss is restricted to patches that overlap the synthetic occlusion,

$$\mathcal{L}_{\mathrm{occ}} = \frac{1}{\sum_{i=1}^N o_i} \sum_{i=1}^N o_i \left\| \hat{y}_i^{\mathrm{occ}} - y_i \right\|_2^2, \tag{8}$$

where $\hat{y}_i^{\mathrm{occ}}$ is the prediction from the occluded branch. Supervision covers the full target of each affected patch, including partially occluded patches. The objective encourages reconstruction of the source facial content from the corrupted observation and its surrounding context.

Training objective. The two reconstruction objectives are combined as

$$\mathcal{L} = \lambda_{\mathrm{mae}} \mathcal{L}_{\mathrm{mae}} + \lambda_{\mathrm{occ}} \mathcal{L}_{\mathrm{occ}}, \tag{9}$$

where λ_{mae} and λ_{occ} control their relative contributions and are both set to one unless otherwise specified. The clean branch learns from missing inputs, while the occluded branch learns from corrupted inputs that remain visible. Their shared parameters allow both objectives to shape the transferred representation.

After pretraining, the decoder, corruption generator, and occlusion embedding are discarded. Downstream fine-tuning and inference use the encoder with a task-specific prediction head and require no occlusion indicators.

4 OBJECT-STYLE OCCLUSION EVALUATION

Everyday objects can obscure different facial regions. Glasses and hats may cover the eyes and upper face, while masks and objects held near the mouth may conceal parts of the lower face. Motivated by these common sources of facial occlusion, we construct a controlled evaluation protocol using glasses, hats, phones, cups, books, and masks. Each source image is paired with exactly one fixed object-style occluded counterpart, allowing performance changes to be assessed on the same underlying samples. The evaluation objects and placement procedures are separate from the pretraining corruption pipeline, as illustrated in Figure 3.



Figure 3: Synthetic corruptions used during pretraining and object-style occlusions used for evaluation. The latter use a separate object pool and category-specific placement procedures.

The foreground objects are collected or generated according to category. Phones, cups, and books are extracted from COCO and LVIS images (Lin et al., 2014; Gupta et al., 2019), while hats are obtained from CelebAMask-HQ (Lee et al., 2020). Masks use templates from MaskTheFace (Anwar & Raychowdhury, 2020), and glasses are generated using local parametric templates. Each object is stored in RGBA format, containing foreground RGB values f and a spatial alpha map a.

For each input image x, we generate an object-style occluded counterpart through geometric transformation, appearance adjustment, and compositing. Transparent margins are first cropped from the object, which is then scaled, rotated, and positioned according to category-specific rules. When reliable facial landmarks are available, glasses are aligned with the eyes, hats with the upper face, and the remaining categories with their designated lower-face regions. Predefined template coordinates are used otherwise. The transformed foreground subsequently undergoes brightness, contrast, and saturation adjustments, with an optional slight Gaussian blur applied jointly to its RGB and alpha channels. These operations are written as

$$(f_g, a_g) = \mathcal{T}(f, a), \qquad (f_p, a_p) = \mathcal{P}(f_g, a_g), \tag{10}$$

where $\mathcal{T}$ denotes geometric transformation and placement, and $\mathcal{P}$ denotes foreground appearance processing. The pairs (f_g, a_g) and (f_p, a_p) represent the foreground and alpha map after the respective operations.

Before compositing, the processed alpha map is always smoothed separately to soften the foreground boundary. The evaluation image is formed as

$$\bar{a} = \mathcal{B}(a_p), \qquad \tilde{x}^{\text{eval}} = (1 - \bar{a}) \odot x_{\text{bg}} + \bar{a} \odot f_p, \tag{11}$$

where $\mathcal{B}$ denotes Gaussian smoothing and $\bar{a}$ is the resulting blending map. The foreground and alpha map are expressed on the source-image canvas at their assigned positions, with zero alpha outside the placed object. The background x_{bg} is the source image after any shadow rendering and equals x when no shadow is applied. Shadows use a separately blurred alpha map with a larger blur radius. Unlike the pretraining generator, the evaluation procedure does not sample an additional global-opacity factor. All blur operations act on the foreground or alpha maps before compositing, rather than on the final image.

A binary occlusion mask is derived from the object alpha before compositing to record the affected region and its coverage. This mask is distinct from the smoothed blending map $\bar{a}$ and is not inferred

from the final composite. All generated images retain their source task labels and are stored once for reuse across methods. The masks and generation metadata support reproducibility and diagnostic analysis and are not provided as model inputs during evaluation. Representative clean–occluded pairs are presented in Appendix.

5 EXPERIMENTS

5.1 DATASETS AND EVALUATION PROTOCOLS

We evaluate the pretrained encoder on facial Action Unit (AU) detection and Facial Expression Recognition (FER), using both the original images and their fixed object-style occluded counterparts constructed following Section 4. For AU detection, we use DISFA (Mavadati et al., 2013), which contains recordings of 27 subjects with frame-level AU intensity annotations. We evaluate AU1, AU2, AU4, AU6, AU9, AU12, AU25, and AU26 under subject-independent three-fold evaluation, where an AU is considered active when its annotated intensity is greater than or equal to 2. The subject assignments are provided in Appendix, and the paired evaluation set contains 130,814 clean–occluded image pairs. For FER, we use AffectNet (Mollahosseini et al., 2019) under the eight-class setting of neutral, happy, sad, surprise, fear, disgust, anger, and contempt. The encoder is fine-tuned on the training split and evaluated on the validation split, with 4,000 clean–occluded image pairs used for paired evaluation. We report F1 and Area Under the receiver operating characteristic Curve (AUC) (Fawcett, 2006) for AU detection, averaged over the eight AUs and three folds, and classification accuracy and macro-F1 for FER.

5.2 IMPLEMENTATION DETAILS

We use ViT-Base/16 with 224×224 inputs. The model is pretrained on CelebA for 50 epochs with a batch size of 8 and gradient accumulation over 32 steps, using a learning rate of 1.5×10^{-4}, a weight decay of 0.05, 2 warm-up epochs, and an MAE mask ratio of 0.75. Synthetic occlusions occupy 4%–22% of the image, with opacity sampled from 0.55–1.0. Elliptical and irregular masks are used to vary the corruption shape, while occlusion textures are generated either by cut-pasting image content or by sampling color noise. The encoder is fine-tuned on DISFA for 20 epochs and on AffectNet-8 for 30 epochs with batch sizes of 32 and 64, respectively. We use learning rates of 5×10^{-4} and 1.25×10^{-4}, with 2 warm-up epochs for DISFA. No occlusion-related information is used during downstream training or inference. All comparison methods are reproduced or evaluated using their publicly available implementations and checkpoints under our unified evaluation protocol. For each method, the clean and pseudo results, including fold-, AU-, and class-wise analyses, are computed from the same selected checkpoint. Table 1 summarizes inference-time parameter counts, where only components retained for downstream use are included.

Table 1: Parameter counts of the compared inference models. Params denotes the number of parameters retained for downstream inference.

Method	**Params (M)** ↓	**Backbone**
EfficientFace (Zhao et al., 2021)	1.28	CNN
HSTR-Net (Song et al., 2022)	41.49	ResNet-101
ME-GraphAU (Luo et al., 2022)	93.30	ResNet-50
DAN (Wen et al., 2023)	19.72	ResNet-18
LibreFace (Chang et al., 2024)	11.24	ResNet-18
CEPrompt (Zhou et al., 2024)	127.88	CLIP ViT-B/16
FG-Net (Yin et al., 2024)	323.94	IR-SE-50 + StyleGAN2
QCS (Wang et al., 2025a)	55.17	IR50
FMAE (Ning et al., 2025)	85.81	ViT-B/16
FMAE-IAT (Ning et al., 2025)	85.81	ViT-B/16
AUNCE (Shang et al., 2026)	98.58	Swin-B
CMNet (Tian et al., 2026)	11.78	ResNet-18
ODMAFace (Ours)	85.81	ViT-B/16

5.3 Main Comparison

We compare ODMAFace with facial representation learning methods and task-specific AU and FER models under the same fixed paired evaluation protocol.

AU detection. On DISFA, ODMAFace achieves 67.08% F1 and 93.98% AUC on the clean evaluation set. Under paired object-style occlusion, the performance decreases by 4.96 and 2.23 percentage points, respectively. Among the evaluated methods, ODMAFace exhibits the smallest clean-to-pseudo degradation in both metrics while maintaining strong clean-set performance. Table 2 summarizes the fold-wise clean results and decreases under paired occlusion, while Figure 4 visualizes the Clean–Pseudo relationships. Detailed AU-wise results are provided in the Appendix.

Table 2: Clean-set comparison on DISFA. Each entry is reported as F1/AUC (%). Overall denotes the average over the three subject-independent folds. Red downward annotations beside Overall report the absolute decrease on the strictly paired pseudo set.

Method	Clean F1/AUC			
	Fold0	Fold1	Fold2	Overall†
HSTR-Net (Song et al., 2022)	60.39/89.61	50.32/88.72	49.67/83.88	53.46$^{\downarrow 9.38}$/87.40$^{\downarrow 4.31}$
ME-GraphAU (Luo et al., 2022)	**72.87/95.63**	60.20/93.93	60.83/89.16	64.63$^{\downarrow 9.41}$/92.91$^{\downarrow 3.97}$
FG-Net (Yin et al., 2024)	64.70/92.70	60.05/89.06	52.96/87.37	59.24$^{\downarrow 22.51}$/89.71$^{\downarrow 12.65}$
LibreFace (Chang et al., 2024)	58.56/89.67	59.92/92.98	52.91/81.19	57.13$^{\downarrow 11.53}$/87.95$^{\downarrow 6.01}$
FMAE (Ning et al., 2025)	64.92/91.37	65.64/94.00	54.05/88.46	61.54$^{\downarrow 9.57}$/91.28$^{\downarrow 4.17}$
FMAE-IAT (Ning et al., 2025)	56.96/88.20	51.97/87.75	55.00/86.00	54.65$^{\downarrow 8.23}$/87.32$^{\downarrow 3.81}$
AUNCE (Shang et al., 2026)	68.91/94.47	60.88/92.14	57.81/85.06	62.54$^{\downarrow 10.88}$/90.56$^{\downarrow 4.95}$
ODMAFace (Ours)	69.66/94.53	**67.75/95.51**	**63.82/91.90**	**67.08**$^{\downarrow 4.96}$/**93.98**$^{\downarrow 2.23}$

†Red annotations denote Clean minus Pseudo.

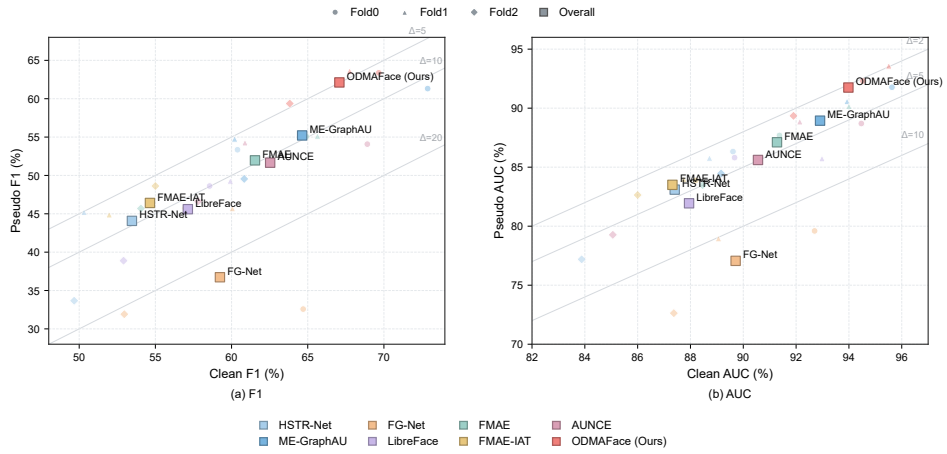


Figure 4: Clean and paired pseudo performance on DISFA. (a) Mean F1 and (b) mean AUC over the three subject-independent folds. Each labeled point represents one method's overall result; the lighter points show its fold-wise results.

Facial expression recognition. On AffectNet-8, ODMAFace achieves 62.85% accuracy and 62.87% macro-F1 on the clean set, and 54.75% accuracy and 54.60% macro-F1 on the paired pseudo set. Figure 5 compares the evaluated methods on the clean and paired pseudo sets. The results show that the learned representation transfers effectively to FER, although the larger degradation than on DISFA indicates that the influence of occlusion remains task dependent. Class-wise results are provided in the appendix.

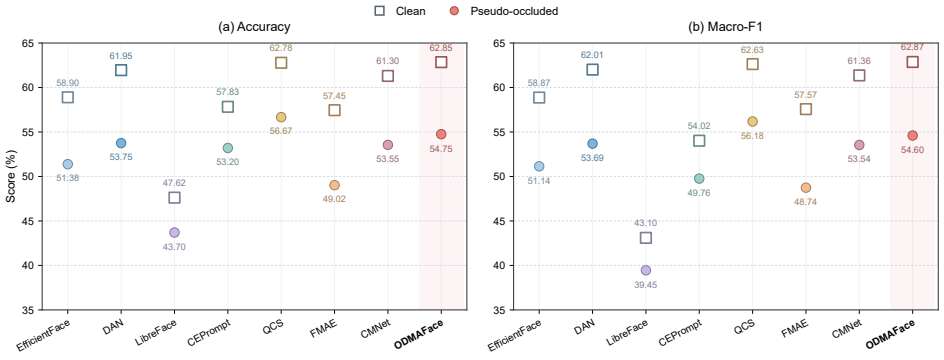


Figure 5: Recognition performance on AffectNet-8 under clean and strictly paired pseudo evaluation. (a) Classification accuracy and (b) macro-F1. Each point represents the performance of one evaluated method under the corresponding input condition.

5.4 Ablation Studies

Table 3 compares the pretraining variants under the same downstream evaluation protocol. The full model attains the highest pseudo F1 among the evaluated variants. Removing the occlusion embedding, masking corrupted patches before encoding, or reconstructing the corrupted appearance changes both clean and pseudo performance. In particular, the variants that mask corrupted patches or use corrupted reconstruction targets perform worse than ODMAFace on the paired pseudo set. These results are consistent with the intended roles of visible corrupted input, corruption indicators, and clean-target supervision.

Table 3: Ablation study of different design choices in ODMAFace on DISFA. Each entry is reported as F1/AUC (%). Masked occ. means that affected patches are removed before encoding. Δ denotes the absolute decrease from clean to pseudo-occluded evaluation.

Variant	**Input**	**Target**	**Emb.**	**Clean**	**Pseudo**	Δ
Clean-only	Clean	Clean	✗	66.76/93.54	61.55/91.14	5.21/2.40
Occluded-only	Occ.	Clean	✓	66.46/93.63	59.65/90.73	6.81/2.90
Dual-view w/o Emb.	Clean+Occ.	Clean	✗	66.43/93.29	59.88/90.77	6.55/2.52
Mask corrupted patches	Clean+Masked occ.	Clean	✗	66.55/93.41	59.72/90.66	6.83/2.75
Corrupted target	Clean+Occ.	Occ.	✓	64.32/92.48	56.71/89.21	7.61/3.27
ODMAFace (Ours)	Clean+Occ.	Clean	✓	**67.08/93.98**	**62.12/91.75**	**4.96/2.23**

6 Conclusion

We presented ODMAFace, a dual-view masked autoencoding framework for occlusion-aware facial representation learning. Existing masked pretraining primarily learns from missing-content reconstruction and does not directly constrain how corrupted visual evidence that remains visible should relate to its underlying facial content. ODMAFace addresses this gap by jointly learning from clean and synthetically occluded views, retaining corrupted patches in the encoder input and reconstructing their corresponding clean facial content with the aid of a learnable occlusion embedding. All occlusion-specific components are discarded after pretraining, allowing the learned encoder to be transferred to downstream facial analysis without requiring occlusion information at inference time.

Experiments on DISFA and AffectNet-8 demonstrate that the learned representation transfers to both facial action unit detection and facial expression recognition while maintaining competitive performance under fixed object-style occlusion. Ablation studies further support the importance of retaining corrupted patches in the encoder input and reconstructing their corresponding clean facial content. The larger performance degradation observed on FER also indicates that sensitivity to occlusion remains task dependent, motivating future work on broader facial tasks and more diverse real-world occlusion conditions.

AI USE STATEMENT

In this work, generative AI tools were used only to assist with language polishing and minor code completion and editing. All research decisions, experiments, analyses, and conclusions were made by the authors, who reviewed and verified all AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses publicly available benchmark datasets and does not involve the collection of new human-subject data. We follow the original licenses and intended research purposes of the datasets used in our experiments. The proposed method is developed for academic research, and we recognize that facial analysis technologies may raise privacy, bias, and potential misuse concerns when deployed in real-world applications. We therefore encourage responsible use of the method in compliance with applicable ethical guidelines, privacy requirements, and legal regulations.

REPRODUCIBILITY STATEMENT

We have taken several steps to support the reproducibility of our work. The model architecture, implementation details, training settings, hyperparameters, data preprocessing procedures, and evaluation protocols are described in the main manuscript. All experiments are conducted on publicly available benchmark datasets, with the dataset splits and evaluation metrics clearly specified. Additional experimental analyses and ablation studies are provided to further support the reproducibility and reliability of the reported results.

REFERENCES

Aqeel Anwar and Arijit Raychowdhury. Masked Face Recognition for Secure Authentication. *arXiv preprint arXiv:2008.11104*, 2020.

Di Chang, Yufeng Yin, Zongjian Li, Minh Tran, and Mohammad Soleymani. LibreFace: An Open-Source Toolkit for Deep Facial Expression Analysis. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8205–8215. IEEE, 2024.

Ashutosh Chaubey, Xulang Guan, and Mohammad Soleymani. Face-LLaVA: Facial Expression and Attribute Understanding Through Instruction Tuning. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2648–2660. IEEE, 2026.

Yan Chen, Kexuan Li, Feng Tian, Ganglin Wei, and Morteza Seberi. Lightweight Expression Recognition Combined Attention Fusion Network With Hybrid Knowledge Distillation for Occluded E-Learner Facial Images. *Neurocomputing*, 628:129656, 2025.

Tom Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Zheng Gao and Ioannis Patras. Self-Supervised Facial Representation Learning With Facial Region Awareness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2081–2092. IEEE, 2024.

Sheng Guo, Mian Tan, Shan Cai, Zaijun Zhang, Yihui Liang, Hongxi Feng, Xue Zou, and Lin Wang. Facial Expression Recognition Under Occlusion Conditions Based on Multi-Feature Cross-Attention. *Journal of Intelligent & Fuzzy Systems*, 46(4):9841–9856, 2024.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5356–5364. IEEE, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009. IEEE, 2022.

Minyang Jiang, Yongwei Wang, Martin J McKeown, and Z Jane Wang. Occlusion-Robust FAU Recognition by Mining Latent Space of Masked Autoencoders. *Neurocomputing*, 569:127107, 2024.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5548–5557. IEEE, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yuansen Liu, Haiming Tang, Jinlong Peng, Jiangning Zhang, Xiaozhong Ji, Qingdong He, Donghao Luo, Zhenye Gan, Junwei Zhu, Yunhang Shen, Chaoyou Fu, Chengjie Wang, Xiaobin Hu, and Shuicheng Yan. Human-MME: A Holistic Evaluation Benchmark for Human-Centric Multimodal Large Language Models. In *International Conference on Learning Representations*, 2026.

Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning Multi-Dimensional Edge Feature-Based AU Relation Graph for Facial Action Unit Recognition. In Luc De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 1239–1246. ijcai.org, 2022.

S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

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, 2019.

Mang Ning, Albert Ali Salah, and Itir Onal Ertugrul. Revisiting Representation Learning and Identity Adversarial Training for Facial Behavior Understanding. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–10. IEEE, 2025.

Feng Qiu, Heming Du, Wei Zhang, Chen Liu, Lincheng Li, Tianchen Guo, and Xin Yu. Learning Transferable Compound Expressions From Masked Autoencoder Pretraining. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4733–4741. IEEE, 2024.

Muhammad Sadiq. GCA-ODN: A Global Context-Aware Dropout Network for Joint Facial Landmark Detection and Emotion Recognition Under Occlusion. *Neural Networks*, 205:109626, 2026.

Ziqiao Shang, Bin Liu, Fengmao Lv, Fei Teng, Tianrui Li, and Lan-Zhe Guo. Learning Contrastive Feature Representations for Facial Action Unit Detection. *Pattern Recognition*, 173:112746, 2026. ISSN 0031-3203.

Wenyu Song, Shuze Shi, Yu Dong, and Gaoyun An. Heterogeneous Spatio-Temporal Relation Learning Network for Facial Action Unit Detection. *Pattern Recognition Letters*, 164:268–275, 2022.

Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. LAFS: Landmark-Based Facial Self-Supervised Learning for Face Recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1639–1649. IEEE, 2024.

Chunwei Tian, Jingyuan Xie, Qi Zhang, Chao Li, Wangmeng Zuo, and Shichao Zhang. A Cross-Modal Network for Facial Expression Recognition. *IEEE Transactions on Image Processing*, 35: 5311–5323, 2026.

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, 2025a.

Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. FSFM: A Generalizable Face Security Foundation Model via Self-Supervised Facial Representation Learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24364–24376. IEEE, 2025b.

Xiaoqin Wang, Xusen Ma, Xianxu Hou, Meidan Ding, Yudong Li, Junliang Chen, Wenting Chen, Xiaoyang Peng, and Linlin Shen. FaceBench: A Multi-View Multi-Level Facial Attribute VQA Dataset for Benchmarking Face Perception MLLMs. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9154–9164. IEEE, 2025c.

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.

Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition. *Biomimetics*, 8(2):199, 2023.

Yufeng Yin, Di Chang, Guoxian Song, Shen Sang, Tiancheng Zhi, Jing Liu, Linjie Luo, and Mohammad Soleymani. FG-Net: Facial Action Unit Detection With Generalizable Pyramidal Features. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6087–6096. IEEE, 2024.

Chaolong Zhang, Yuanping Xu, Zhijie Xu, Rongqiang Gou, Weiye Wang, Jin Jin, and Jian Huang. A Self-Supervised Facial Expression Restoration and Recognition Method Based on Improved Masked Auto-Encoder. *Neurocomputing*, 656:131495, 2025a.

Yan Zhang, Zenghui Li, Duo Shen, Ke Wang, Jia Li, and Chenxing Xia. Information Gap Based Knowledge Distillation for Occluded Facial Expression Recognition. *Image and Vision Computing*, 154:105365, 2025b.

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.

Kaiwen Zheng, Xuri Ge, Junchen Fu, Jun Peng, and Joemon M Jose. Multimodal Representation Learning Techniques for Comprehensive Facial State Analysis. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Haoliang Zhou, Shucheng Huang, Feifei Zhang, and Changsheng Xu. CEPrompt: Cross-Modal Emotion-Aware Prompting for Facial Expression Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11886–11899, 2024.

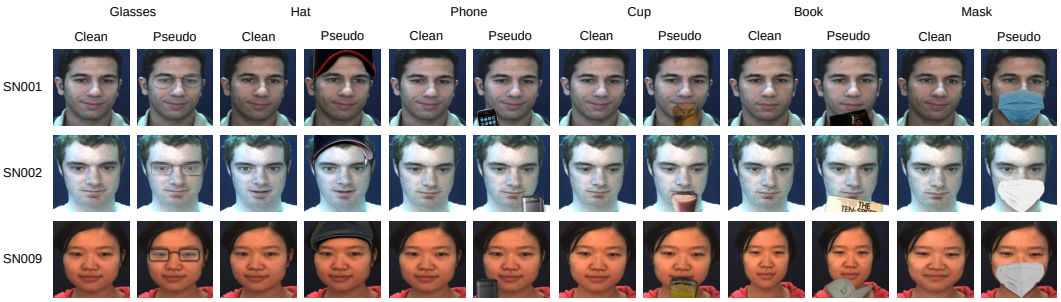
A Appendix

A.1 Evaluation Details

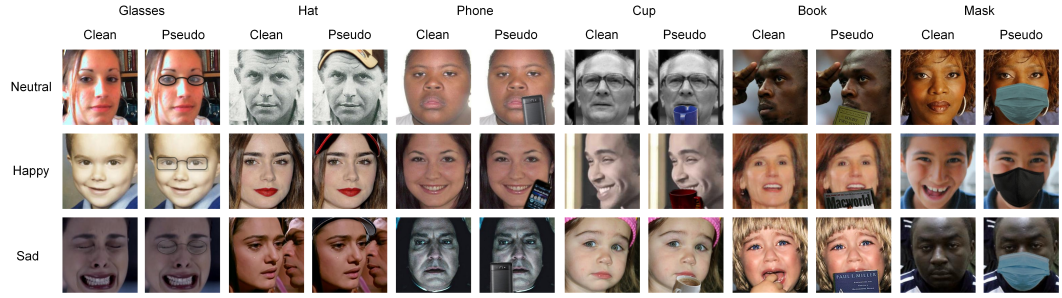
Table 4 reports the fixed subject-independent three-fold partition used for DISFA. For each fold, nine subjects are held out for evaluation and the remaining 18 subjects are used for training. Figure 6 presents representative Clean/Pseudo pairs from DISFA and AffectNet-8 across the six object-style occluder categories. Each pseudo image is generated once from its corresponding source image and reused across methods.

Table 4: Fixed subject-independent three-fold partition used for DISFA. For each fold, the listed nine subjects form the held-out test set, while the remaining 18 subjects form the training set.

Fold	Held-out test subjects
0	SN001, SN002, SN009, SN010, SN016, SN026, SN027, SN030, SN032
1	SN006, SN011, SN012, SN013, SN018, SN021, SN024, SN028, SN031
2	SN003, SN004, SN005, SN007, SN008, SN017, SN023, SN025, SN029



(a) DISFA (AU)



(b) AffectNet-8 (FER)

Figure 6: Representative clean–occluded pairs from DISFA and AffectNet-8. Each adjacent Clean/Pseudo pair shows one source image and its fixed object-style occluded counterpart.

A.2 Fine-Grained Recognition Results

We further analyze the fine-grained recognition performance on DISFA and AffectNet-8 to investigate how different facial actions and expression categories respond to occlusion. These results provide complementary evidence to the overall comparisons reported in the main manuscript.

AU-wise analysis on DISFA. Table 5 reports the F1 scores of different methods for individual action units under clean and paired pseudo-occluded conditions. ODMAFace achieves the highest pseudo-occluded F1 on seven of the eight evaluated AUs, with AU4 being the only exception. The advantage is particularly evident for AU1 and AU2, where ODMAFace reaches 59.2% and 65.0% F1, respectively, compared with 51.9% and 46.5% for the strongest competing results. Improve-

ments are also observed for AU6, AU9, AU12, AU25, and AU26, indicating that the gain under occlusion is not concentrated on a single facial action.

Table 5: Per-AU F1 scores on clean and pseudo-occluded DISFA (%). Each entry reports Clean/Pseudo F1.

Method	Per-AU F1							
	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26
HSTR-Net (Song et al., 2022)	49.4/36.6	29.3/23.2	64.5/67.6	43.1/32.1	26.9/27.1	74.8/62.8	91.9/73.1	47.8/30.1
ME-GraphAU (Luo et al., 2022)	56.6/48.2	46.8/38.4	75.0/73.1	54.2/51.1	51.1/45.2	75.6/68.9	93.6/75.2	**64.2**/41.7
FG-Net (Yin et al., 2024)	58.8/51.9	33.7/34.4	69.0/63.1	48.9/21.6	35.2/14.3	75.7/35.2	94.5/59.6	58.2/13.9
LibreFace (Chang et al., 2024)	36.3/35.8	43.4/38.9	64.3/59.8	41.0/34.7	43.1/34.4	73.6/57.7	**94.7**/69.4	60.7/34.1
FMAE (Ning et al., 2025)	47.7/42.4	51.2/46.1	66.3/63.2	50.3/42.3	55.2/49.1	75.5/68.2	91.8/74.2	54.3/30.1
FMAE-IAT (Ning et al., 2025)	46.8/41.8	46.5/42.0	67.0/61.7	46.9/41.5	30.3/20.6	74.5/65.8	86.7/70.9	38.5/27.1
AUNCE (Shang et al., 2026)	50.9/51.7	58.0/46.5	70.8/66.4	51.1/43.1	45.6/40.2	71.6/61.5	93.1/69.6	59.1/34.4
ODMAFace (Ours)	**60.7/59.2**	**65.7/65.0**	**71.1/69.8**	**56.1/51.9**	**55.1/51.5**	**76.7/72.7**	91.1/**81.6**	60.0/**45.3**

The clean-to-pseudo changes, however, remain strongly AU dependent. For ODMAFace, the decreases on AU1, AU2, and AU4 are only 1.5, 0.7, and 1.3 percentage points, whereas AU25 and AU26 decrease by 9.5 and 14.7 points, respectively. This difference is not simply explained by clean-set performance. For example, AU25 achieves a high clean F1 of 91.1% but still exhibits a noticeable decrease under occlusion, while AU2 maintains nearly unchanged performance despite its similarly strong clean result. The AU-wise comparison therefore shows that the overall improvement of ODMAFace arises from broadly improved performance across facial actions, while the remaining sensitivity to occlusion is still action dependent.

Class-wise analysis on AffectNet-8. Tables 6 and 7 report the per-class accuracy and F1 scores on AffectNet-8, respectively. Class-wise accuracy denotes the proportion of correctly classified samples within each ground-truth expression category. The class-level results reveal considerably different responses to occlusion across expression categories. For ODMAFace, the largest degradation occurs for contempt and disgust. Their class-wise accuracy decreases from 53.8% to 28.2% and from 54.4% to 36.4%, respectively, while the corresponding F1 scores decrease from 62.4% to 41.8% and from 60.4% to 47.1%. In contrast, surprise, fear, and anger exhibit considerably smaller changes, with accuracy decreases of 3.6, 1.8, and 3.4 percentage points, respectively.

Table 6: Per-class accuracy on clean and pseudo AffectNet-8 (%).

Method	Per-class Accuracy							
	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt
EfficientFace (Zhao et al., 2021)	47.4/43.6	75.0/60.6	61.8/54.0	59.0/**59.2**	60.8/60.8	52.2/39.8	61.2/61.0	**53.8**/32.0
DAN (Wen et al., 2023)	57.2/48.6	74.4/63.8	63.0/54.6	60.2/54.4	64.8/**64.4**	61.0/48.4	57.4/56.2	57.6/39.6
LibreFace (Chang et al., 2024)	**73.2/74.4**	**95.0**/80.0	57.8/52.6	36.2/34.4	44.8/43.8	20.2/12.0	53.8/52.4	0.0/0.0
CEPrompt (Zhou et al., 2024)	63.4/62.0	87.2/**82.0**	65.2/**62.4**	56.8/50.2	**65.2**/61.6	**61.8/51.8**	62.8/55.6	0.2/0.0
QCS (Wang et al., 2025a)	58.0/56.0	84.2/77.8	66.0/57.4	**61.0**/62.6	60.0/58.8	56.8/49.8	64.2/59.8	52.0/31.2
FMAE (Ning et al., 2025)	56.0/58.6	77.4/57.6	55.2/52.2	55.4/55.6	57.8/55.2	50.6/35.0	57.6/53.8	49.6/24.2
CMNet (Tian et al., 2026)	59.0/54.8	77.0/63.4	63.0/53.0	**61.0**/58.8	60.0/60.0	59.0/46.0	57.0/55.0	54.4/**37.4**
ODMAFace (Ours)	61.8/72.6	82.6/69.4	**67.4**/57.4	57.0/53.4	60.0/58.2	54.4/36.4	**65.8/62.4**	**53.8**/28.2

The accuracy results further show that occlusion does not affect all categories through a uniform reduction in recognition rate. Happy and sad exhibit moderate decreases of 13.2 and 10.0 percentage points, whereas neutral increases from 61.8% to 72.6% under pseudo occlusion. This contrast indicates that the aggregate degradation arises from different class-specific responses rather than a uniform loss of recognition ability across expressions. The increase for neutral further suggests that occlusion can alter how recognition errors are distributed among expression categories.

Table 7: Per-class F1 scores on clean and pseudo AffectNet-8 (%).

Method	Per-class F1							
	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt
EfficientFace (Zhao et al., 2021)	48.3/41.4	73.3/64.7	60.5/53.4	56.2/51.0	63.4/55.5	57.4/48.4	55.8/52.3	56.0/42.4
DAN (Wen et al., 2023)	54.0/44.8	73.9/65.3	64.1/55.3	**60.6**/51.7	66.3/58.1	**61.9**/53.5	57.1/54.9	58.3/**45.7**
LibreFace (Chang et al., 2024)	47.0/41.4	59.5/55.6	56.3/52.8	43.8/42.1	54.4/52.5	32.4/21.1	51.4/50.1	0.0/0.0
CEPrompt (Zhou et al., 2024)	52.7/46.6	69.7/68.8	64.5/**60.5**	58.4/53.0	66.2/60.6	61.8/**55.4**	58.5/53.3	0.4/0.0
QCS (Wang et al., 2025a)	**54.6/49.0**	73.3/68.9	**66.0**/60.3	60.5/**56.0**	**66.6/61.6**	60.8/53.6	**60.7/57.2**	58.6/42.9
FMAE (Ning et al., 2025)	47.4/42.9	71.9/62.0	58.7/48.2	54.3/52.7	62.8/56.6	55.1/42.0	55.3/49.8	55.1/35.8
CMNet (Tian et al., 2026)	53.4/46.0	74.3/64.6	63.3/55.0	59.5/55.4	64.3/57.9	62.0/52.4	57.0/53.8	57.1/43.2
ODMAFace (Ours)	54.1/46.9	**76.1/69.1**	64.8/59.5	59.8/55.9	65.8/61.3	60.4/47.1	59.5/55.3	**62.4**/41.8

The F1 results provide a complementary view by accounting for the balance between precision and recall. Surprise, fear, anger, and sad show limited decreases of 3.9–5.3 percentage points, while happy and neutral decrease by 7.0 and 7.2 points. Disgust and contempt remain the most affected categories, with reductions of 13.3 and 20.6 points. Neutral exhibits opposite trends in the two metrics: its accuracy increases under pseudo occlusion, whereas its F1 decreases from 54.1% to 46.9%. This discrepancy suggests that the change in recognition behavior is not fully captured by class-wise accuracy and highlights the complementary value of the F1 analysis.

A.3 Analysis by Occluder Category and Coverage

We further investigate how different object-style occlusions affect facial action unit recognition. The 130,814 paired DISFA evaluation frames are grouped according to the recorded occluder category and binary-mask coverage. Predictions are obtained from the same fold-specific checkpoints used in the main comparison, and all subgroup results follow the same evaluation settings without subgroup-specific optimization. Scores are averaged over the eight AUs and three subject-independent folds.

Analysis by occluder category. Table 8 reports ODMAFace performance under the six evaluation occluder categories. The category sizes are nearly balanced, with approximately 21.8k paired frames in each group, yet their performance degradation differs substantially. This indicates that the observed differences are not primarily driven by unequal subgroup sizes. Glasses, phones, and books produce relatively small F1 decreases of 0.79, 0.41, and 1.64 percentage points, respectively, while hats lead to a moderate decrease of 2.12 points. Cups produce a larger decrease of 6.57 points, and masks are markedly more difficult, reducing F1 by 21.56 points and AUC by 11.94 points.

Table 8: ODMAFace performance on DISFA by evaluation occluder category. Drop denotes Clean minus Pseudo in percentage points.

Occluder	Frames	F1			AUC		
		Clean	Pseudo	Drop	Clean	Pseudo	Drop
Glasses	21,804	67.28	66.49	0.79	94.02	93.48	0.54
Hat	21,800	66.78	64.66	2.12	93.81	92.70	1.11
Phone	21,800	67.10	66.70	0.41	94.03	93.96	0.07
Cup	21,801	66.65	60.08	6.57	93.92	90.16	3.76
Book	21,804	67.05	65.41	1.64	93.89	93.53	0.36
Mask	21,805	67.46	45.90	21.56	94.18	82.24	11.94

The large variation across categories also shows that the overall pseudo-occlusion result should not be interpreted as a uniform response to different forms of occlusion. In particular, mask occlusion is substantially more challenging than the other categories despite having a comparable number of evaluation frames. Under our evaluation protocol, masks primarily occupy the lower facial region,

making this result consistent with stronger interference with facial evidence in that area. However, occluder category is coupled with object shape, placement, and typical coverage, so the category comparison alone cannot isolate the contribution of any individual factor. We therefore further analyze coverage within each occluder category.

Analysis by occlusion coverage. Table 9 examines coverage within occluder categories. The effect of increasing coverage is category dependent. For books, increasing coverage from below 10% to 10–20% increases the F1 drop from 1.08 to 2.14 points. For hats, the F1 drop increases more markedly from 1.28 points at 10–20% coverage to 5.96 points above 20%. Cups show a smaller difference between the two coverage ranges, with F1 drops of 6.36 and 6.69 points.

Table 9: ODMAFace within-category performance on DISFA by occlusion coverage. Drop denotes Clean minus Pseudo in percentage points.

Occluder	Coverage	Frames	Clean F1	Pseudo F1	F1 Drop	AUC Drop
Book	$< 10\%$	9,080	66.20	65.12	1.08	0.25
Book	10–20%	12,714	67.37	65.23	2.14	0.44
Cup	$< 10\%$	18,836	66.28	59.92	6.36	3.55
Cup	10–20%	2,965	65.45	58.76	6.69	4.78
Hat	10–20%	18,062	67.07	65.79	1.28	0.72
Hat	$\geq 20\%$	3,719	60.44	54.48	5.96	2.86
Mask	10–20%	4,307	66.53	44.45	22.07	11.96
Mask	$\geq 20\%$	17,498	67.65	46.32	21.33	11.77

Mask occlusion exhibits a different pattern. Its F1 degradation remains large across the two populated coverage intervals, with drops of 22.07 points at 10–20% coverage and 21.33 points above 20%. The AUC decreases are nearly unchanged at 11.96 and 11.77 points. Thus, a larger affected area does not consistently produce a larger performance decrease, even when comparisons are restricted to the same occluder category. These results indicate that coverage alone is insufficient to explain occlusion difficulty and that the effect of the affected area depends on the type and placement of the occluding object. Coverage bins with negligible sample counts are omitted from the within-category comparison.

Coverage-stratified performance. Table 10 further pools all occluder categories into four non-overlapping coverage intervals. At the aggregate level, the F1 decrease grows from 0.63 points at 0–5% coverage to 2.56 points at 5–10%, 4.54 points at 10–20%, and 17.81 points at 20–30%. This pooled result appears to suggest a strong relationship between coverage and recognition degradation.

Table 10: ODMAFace performance on DISFA by evaluation occlusion coverage. F1 scores are three-fold means (%); Drop denotes Clean minus Pseudo in percentage points.

Coverage	Frames	Clean F1	Pseudo F1	Drop
0–5%	12,976	66.15	65.52	0.63
5–10%	58,563	66.95	64.40	2.56
10–20%	38,048	67.24	62.70	4.54
20–30%	21,227	67.25	49.44	17.81

The category-conditioned analysis suggests that this trend should be interpreted cautiously. In the 20–30% interval, 17,498 of 21,227 samples are mask occlusions, accounting for 82.4% of the group. The observed degradation therefore reflects both greater coverage and a shift in occluder composition, indicating that recognition difficulty depends on occluder characteristics as well as coverage.