# Revisiting the Volumetric Data of 4DME: Compression, Extension and Benchmarking for Micro-Expression Analysis

Samuel Boccara[1] ⋆, Amar Tious[2] ⋆, Guoying Zhao[31]⋆⋆, Yante Li[4], Toinon Vigier[2], and Vincent Ricordel[2]

[1] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
{name}.{surname}@oulu.fi
[2] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, France
{name}.{surname}@univ-nantes.fr
[3] Ellis Institute Finland
[4] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, China liyante{at}caas.cn

**Abstract.** Micro-expressions are brief, involuntary facial movements that reveal subtle affective states. While most computational work focuses on micro-expression recognition, which classifies clips into discrete emotion categories, Micro-Expression Action Unit (ME-AU) detection offers a more theory-consistent formulation grounded in the Facial Action Coding System. Despite its finer granularity, ME-AU detection remains comparatively under-explored due to the limited availability of high-quality data for efficient automatic models. Recent volumetric micro-expression datasets provide temporally coherent 3D facial recordings with reliable AU labels, creating new opportunities for fine-grained analysis. However, their adoption remains limited due to large data volumes, heterogeneous formats, and the absence of standardized benchmarks, which together hinder reproducibility and cross-study comparison. Moreover, because of the small spatio-temporal spanning of the AUs, designing models that effectively combine volumetric representations with temporal dynamics remains a non-trivial research challenge. To address these barriers, we introduce an extended and standardized release of the 4DME dataset with cleaned data and normalized formatting, accompanied by a compressed version to reduce storage overhead. We further design VoluME, a family of dual-stream optical-flow-based architecture for volumetric ME-AU detection. To determine appropriate settings for compression, we combine state-of-the-art compression baselines, perceptual evaluation, and downstream task-specific benchmarking. As the first work dealing with volumetric ME-AU detection under compression, our results show that a compression ratio of approximately $3000\times$ can be achieved with $\approx 1\%$ F1 degradation. The extended release additionally introduces explicit two-way cultural balancing, supporting future cross-cultural micro-expression studies.

⋆ Equal contribution
⋆⋆ Corresponding author
Project page: https://github.com/1nO1SB/4DMEplus

**Keywords:** Micro-expression · Computer Vision · 3D Data · Multimedia Coding

## 1 Introduction

Micro-expressions (MEs) are brief, involuntary facial movements occurring under emotional suppression. Lasting only 40–500 milliseconds and involving subtle activations of small facial muscle groups, they are inherently sparse in both space and time, making computational modeling particularly challenging.

ME research comprises four main tasks: spotting, recognition (MER), Action Unit detection (ME-AU), and generation. MER has received the most attention but relies on coarse emotion categories that limit fine-grained analysis and may suffer from inconsistent labels derived from elicitation protocols, self-reports, or AU annotations [24]. In contrast, ME-AU detection [10] operates at the level of facial muscle activations defined by the FACS (Facial Action Coding System) [2], offering a more objective and compositional representation of facial behavior. Despite these advantages, ME-AU detection remains under-explored due to its reliance on high-quality spatio-temporal data capable of capturing subtle muscle deformations.

Historically, 2D ME datasets have been limited by small sample sizes, low spatial resolution, and insufficient temporal sampling. Cross-dataset protocols [21,24] partially mitigated data scarcity but introduced heterogeneity and quality imbalance. Recent volumetric (3D+t) datasets, notably 4DME [9] and CASME$^3$ [8], address these limitations by providing temporally coherent 3D facial reconstructions with improved geometric fidelity and more consistent AU annotations. Beyond explicit depth cues, volumetric data enable modeling of out-of-plane facial deformations that are ambiguous in 2D. Human perception studies further suggest that 3D reconstructions improve AU recognition accuracy compared to 2D video [8].

Despite these advantages, 3D data remain underutilized in ME research due to large storage requirements, heterogeneous formats, and the absence of standardized benchmarks. 3D data coding standards are developed to address the cost in storage and transmission of volumetric data. But because micro-expressions are spatio-temporally sparse, global fidelity of the 3D data after decoding might not ensure the preservation of subtle deformations that defines AUs. In addition, 3D compression must therefore be considered and evaluated with adapted methods. Consequently, two key challenges emerge: establishing a principled benchmark for volumetric ME-AU detection and developing efficient compression strategies to reduce the size of 3D MER datasets while preserving subtle AU-related signals.

Modeling volumetric ME-AU detection introduces additional complexity. While features based on optical flow (OF) have long been effective in 2D ME analysis [5, 11, 15, 17, 25], extending motion modeling to 3D is non-trivial, as multiple volumetric motion representations exist without consensus for facial analysis. Moreover, high-resolution 3D inputs impose severe computational constraints,

limiting the feasibility of deep attention-based architectures. Even with suitable models, large-scale 4D data must be compressed to enable reproducible and scalable research.

To address the lack of standardized volumetric benchmarks, the modeling challenges of subtle 3D facial dynamics, and the practical barriers posed by large-scale 4D data, we introduce a unified framework for 3D ME-AU analysis spanning dataset consolidation, dual-stream model design, and task-aware compression.

Our contributions are threefold:

- We consolidate the 4DME dataset, designing a three-stage de-noising procedure to add previously unreleased volumetric data, establishing the first ethnically balanced micro-expression dataset with 3D modality.
- We propose VoluME (Volumetric Optical flow Learning for action Unit ME detection), a family of dual-stream architectures that fuse optical and depth flow to jointly model 3D spatial structure and temporal dynamics, establishing, to the best of our knowledge, the first cross-dataset volumetric ME-AU detection baseline.
- We systematically evaluate 3D compression under the micro-expression framework, combining standard-compliant codecs, perceptual assessment, and downstream AU detection performance to identify an optimal strategy that substantially reduces data size while preserving AU-level fidelity.

## 2 Related work

**Micro-expression Action Unit Detection Models.** Micro-expressions involve low-intensity, short-duration facial deformations that become even subtler at the Action Unit (AU) level. A central challenge is integrating spatial appearance and fine-grained temporal dynamics under weak motion signals. Existing methods largely follow two paradigms.

The first relies on hand-crafted or motion-driven descriptors with explicit temporal modelling. LBP-TOP [30] encodes motion across three orthogonal planes, while optical-flow-based methods such as MDMO [17], HOOF [5], STST-Net [11], SSSNet [24], and Off-ApexNet [15] model onset-to-apex dynamics to capture subtle facial displacements. By explicitly encoding motion priors, these approaches remain effective despite the low amplitude of ME movements.

The second paradigm employs attention-based architectures. SCA [12] introduces joint spatial-temporal attention, whereas AUFormer [28] uses a transformer architecture to model long-range dependencies. Although competitive, these methods generally require larger training sets and do not explicitly encode the onset-to-apex dynamics that characterise micro-expressions.

These limitations are amplified in volumetric settings. Extending AU detection to 3D representations significantly increases computational cost, as attention complexity scales cubically with spatial resolution. Consequently, end-to-end 2D foundation models or directly adapted transformer architectures are ill-suited to

3D micro-expression analysis. This motivates architectures that retain motion-aware advantages while remaining computationally tractable in low-data volumetric contexts.

**Volumetric Video Coding and Quality.** In multimedia coding, consistency and compatibility across devices and use cases is very important. This is ensured through standardization efforts led by MPEG (Moving Pictures Expert Group) for volumetric media. In the case of 4D volumetric videos, two standards based on point clouds, geometry-based (G-PCC) and video-based point cloud compression (V-PCC) have been established [20]. Both approaches greatly reduce the storage and transmission cost for 4D content. With V-PCC, which codes 4D videos with 2D video codecs, lossless coding is not possible as the data is always transformed through projections.

For visual media, lossy coding can allow optimized visually lossless coding, maintaining high quality for human observers [33]. Objective quality metrics are also developed to measure this perceived quality for streaming and user-generated visual content [14]. However, such metrics are not sufficient in other applicative contexts. Related work investigates the impact of lossy coding on radiology images [7] where good visual quality might still affect important medical cues. For immersive media, user experience and behavior must be considered to represent perceived quality [23, 32]. In Micro-Expression Recognition, it is important to preserve the fine AU information, as explored in previous studies for motion magnification [31], but never for lossy coding, especially not for 4D data.

## 3 Data Extension and Cleaning

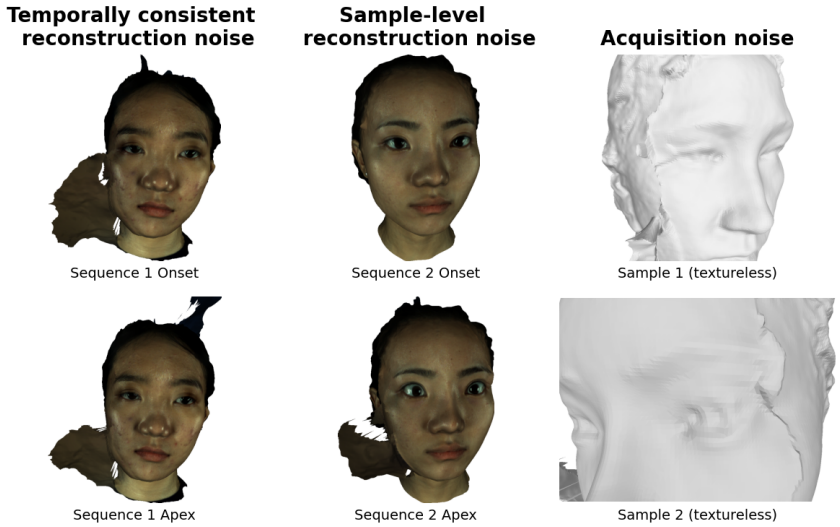


**Fig. 1:** The three identified noise types in the volumetric set of 4DME.

The initial release of the 4DME dataset comprises 130 annotated micro-expression sequences collected from 24 subjects. Each sequence consists of temporally ordered high-resolution triangular meshes reconstructed via synchronized multi-camera stereo capture [9]. Annotators subsequently restricted the temporal extent of each sequence to the micro-expression interval. The resulting clips provide dense spatio-temporal measurements, with an average geometric resolution of approximately 0.5 $mm^2$ per triangle (150k triangles per frame on average) at 60 fps.

Beyond the 130 released volumetric micro-expression clips, approximately 130 additional sequences were collected but not included in the volumetric subset due to severe acquisition or reconstruction artifacts. In the following section, we provide a structured characterization of these noises, their sources and their implications for volumetric micro-expression analysis.

### 3.1 Noise Characterization and Downstream Impact

We identify three primary categories in the volumetric sequences, distinguished by their origin and structural impact on the reconstructed meshes :

1. **Temporally consistent reconstruction noise.** This noise arises from errors in stereo reconstruction following multi-camera capture [9]. It manifests as large spurious surfaces that blend with the facial texture and persist across multiple consecutive frames, and in some cases entire sequences.
2. **Sample-specific reconstruction noise.** This form of degradation is caused by parasitic objects entering the reconstruction volume. It produces peripheral geometric artifacts that appear in isolated frames and may partially overlap with the facial region.
3. **Statistical acquisition noise.** Sensor-level perturbations introduce isolated erroneous points in the reconstructed geometry. Although individually small, the mesh topology enforces connectivity to these outliers, resulting in extended planar artifacts that distort local surface structure.

Figure 1 gives a visual example of the three types of noise that we have identified. The reconstruction noises are shown with samples from the same sequence to compare the temporal spanning of the noise. The acquisition noise samples are shown without texture to highlight the impact on local geometry.

Sequence-wide artifacts induce scale and geometry inconsistencies over time, preventing reliable alignment. Heuristic alignment strategies based on bounding boxes become unstable due to the presence of large peripheral artifacts, while learned registration approaches degrade when substantial portions of the facial surface are occluded or geometrically distorted. As a consequence, methods that rely on temporal geometric consistency, such as point-to-point distance estimation, optical or volumetric flow computation, and volumetric discretization strategies (e.g., voxelization or occupancy mapping) become unreliable, since their formulations implicitly assume aligned and topologically coherent sequences.

### 3.2 Denoising Methods

A quantitative inspection of the 130 released sequences reveals substantial geometric degradation: 40% of sequences exhibit temporally consistent reconstruction noise, 55% contain frame-specific artefacts, and 45% are affected by acquisition-level outliers. Overall, only 33% of sequences remain free from facial-surface corruption. These statistics confirm that geometric inconsistency affects the majority of the original volumetric release and directly compromises the structural assumptions required for automated micro-expression analysis. Removing these noise sources is challenging due to their heterogeneity and modality-dependent manifestation. In particular, sequence-wide and sample-specific reconstruction artefacts originate from the stereo pipeline and remain statistically consistent with the facial surface in terms of texture mapping, point density, and mesh resolution, making semantic or object-aware filtering unreliable. We therefore propose a three-stage denoising pipeline targeting each noise category:

**Statistical Denoising.** To suppress acquisition-level noise, we analyse mesh edge-length distributions within each sample. Isolated outliers produce abnormally long edges inconsistent with local geometric resolution. We model these distributions using sample-specific Gaussian Mixture Models and remove statistical outliers, followed by deletion of associated vertices. This step removes low-level perturbations while preserving high-curvature facial regions.

**Sequence-wide Density-Aware Denoising.** To mitigate reconstruction artefacts persisting across frames, we aggregate each sequence into a unified occupancy mask. Valid facial regions form dense, temporally consistent clusters proportional to sequence length, whereas reconstruction artefacts exhibit lower spatial consistency. We apply neighborhood-based statistical outlier removal on the aggregated representation to isolate the stable facial region, and use the resulting mask to filter each frame.

**Manual Mask Refinement.** Residual dense artefacts with statistics comparable to facial regions are removed through manual refinement of the sequence masks. This final step ensures that only facial and peri-facial geometry is retained throughout the sequence.

The denoised sequences exhibit substantially improved consistency, with an **86%** reduction in normal-distribution divergence, **52%** lower occupancy instability, and **47%** improved curvature stability. Raw metrics are reported in the supplementary materials.

### 3.3 Data extension

The proposed denoising pipeline enables the recovery and integration of previously excluded volumetric sequences from the 4DME collection, as well as correction of previously unusable volumetric samples in the released set. As a result, the number of usable micro-expression clips is substantially increased, along with the prevalence of several benchmark Action Units (AUs). Tab. 1 quantifies the increase in sequence count for the most commonly evaluated AUs, including

**Table 1:** Comparison table between the current release and extension of the 4DME volumetric dataset, based on most used AUs, total sequences, and Cross-culture split.

| Dataset | AU1 | AU2 | AU4 | AU6 | AU7 | AU12 | AU17 | AU45 | Sequences | Culture Split |
|---|---|---|---|---|---|---|---|---|---|---|
| Released 4DME | 29 | 31 | 66 | 9 | 56 | 25 | 6 | 15 | 130 | (17/8) |
| Extended 4DME | 46 | 54 | 117 | 30 | 93 | 82 | 11 | 35 | 265 | (23/20) |
| Relative Increase | +59% | +74% | +78% | +233% | +67% | +224% | +84% | +127% | +102% | - |

those defined in the original 4DME benchmark protocol. For several frequently studied AUs, the number of available samples more than doubles. A comprehensive comparison across all annotated AUs is provided in the supplementary material. Beyond the increase in volume, the extended subset improves subject-level and cultural representation. The resulting dataset achieves approximate two-way Asian/European cultural balance, making 4DME, to our knowledge, the first 3D micro-expression dataset with explicit cross-cultural representation. This extension provides a foundation for culturally aware ME analysis.

## 4 Compression of 4DME

### 4.1 Compression Methods

We select two open-source codecs representing the main standardized approaches to volumetric compression. As a geometry-based method, we use Draco [4], a widely adopted codec for meshes and point clouds. While it supports real-time transmission of 4D content [1], its coding efficiency is limited compared to more advanced schemes [27, 29].

As a video-based alternative, we employ tmc2 [18], the MPEG reference implementation of V-PCC, which achieves state-of-the-art dynamic point cloud compression [27,29], enabling visually lossless coding beyond 1000× compression, albeit without real-time capability. In Draco, the quantization parameter (QP) determines the effective bits per vertex; higher QP therefore implies lower compression, unlike conventional codecs. Each mesh frame of the extended 4DME dataset is encoded independently at four level. For V-PCC, meshes are converted to colored point clouds voxelized on $1024^3$ grids, and geometry and color are jointly encoded in all-intra mode at three rates recommended by the MPEG common test conditions [6], defined by $(QP_{geom}; QP_{color})$ pairs. All QPs are reported in Tab. 2, and their effect on a 3D frame is illustrated in Fig. 2.

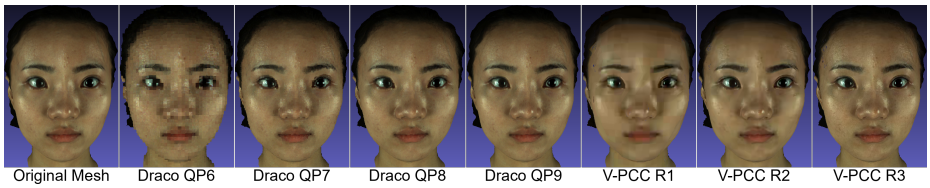


**Fig. 2:** Example of decoded 3D frame at all levels of compression.

### 4.2 Quality and Coding Efficiency

To compare the two compression schemes, we evaluate coding efficiency and visual quality. Visual fidelity is measured using objective metrics on 2D depth maps, colored projections, and original 3D vertex data, enabling joint analysis of geometric and appearance degradation across compression levels.

On depth maps and projections, we compute PSNR and SSIM [26], comparing Draco-compressed meshes to their originals and V-PCC reconstructions to voxelized point clouds. Geometric fidelity in 3D is further assessed using PSNR-D1 [22], defined in the V-PCC common test conditions [6], which extends PSNR to point-to-point geometry evaluation. Metrics are computed per frame and averaged per sequence. Table 3 summarises results at each distortion level, with $PSNR_{depth}$ and $SSIM_{depth}$ evaluated on depth maps, and $PSNR_{color}$ and $SSIM_{color}$ on colored projections.

Results show comparable quality at the highest settings (QP9 and R3), while V-PCC better preserves geometry and appearance under stronger compression (R1). Severe Draco levels (QP6, QP7) introduce noticeable texture-mapping artefacts, whereas V-PCC produces more uniform color degradation.

Coding efficiency is measured via the geometric compression rate ($CR_{geom}$), defined as the ratio between bits per vertex in the original meshes and the encoded bitstream. For V-PCC, we also report the total compression rate ($CR_{total}$), including color (Tab. 2). V-PCC significantly outperforms Draco, achieving over 1500× reduction at near-lossless quality and up to 5780× under moderate degradation. However, these results do not consider the loss of AU-related information, which is critical for downstream analysis.

### 4.3 Perception-Based Compression Validation

To assess the impact of 3D compression on subtle AU signals, we conducted a subjective perceptual study in which human observers evaluated the decoded sequences and their consistency with the original 4DME annotations.

**Experimental Setup.** In this experiment, naive observers viewed denoised and compressed short sequences from 4DME in random order using a Unity 6-based interface. For each sequence, they completed a checkbox questionnaire listing all annotated AUs and could replay sequences as needed. All sessions were conducted under identical conditions. Participants were seated in a lit room at a distance of three screen heights from a 27-inch Lenovo P27q-10 monitor, using a mouse and keyboard. The display was connected to a workstation equipped with

**Table 2:** Quantization Parameters ($QP_{geom}$; $QP_{color}$ pairs for V-PCC) and Compression Ratios for all levels of compression.

| | Draco | | | | V-PCC | | |
|---|---|---|---|---|---|---|---|
| | QP6 | QP7 | QP8 | QP9 | R1 | R2 | R3 |
| $QP$ | 6 | 7 | 8 | 9 | 36;47 | 28;37 | 20;27 |
| $CR_{geom}$ | 16.0 | 13.7 | 12.0 | 10.7 | 2955 | 1665 | 640 |
| $CR_{total}$ | - | - | - | - | 5780 | 3155 | 1575 |

**Table 3:** Objective quality metrics for the compression method and levels.

| | **Draco** | | | | **V-PCC** | | |
|---|---|---|---|---|---|---|---|
| Metric | QP6 | QP7 | QP8 | QP9 | R1 | R2 | R3 |
| PSNR-D1 | 61.4 | 64.4 | 67.4 | 70.5 | 66.1 | 69.4 | 71.4 |
| $PSNR_{depth}$ | 32.7 | 37.5 | 40.7 | 42.2 | 36.4 | 39.1 | 40.1 |
| $PSNR_{color}$ | 21.1 | 25.3 | 28.6 | 30.2 | 26.1 | 28.9 | 29.7 |
| $SSIM_{depth}$ | 0.952 | 0.976 | 0.984 | 0.986 | 0.975 | 0.982 | 0.984 |
| $SSIM_{color}$ | 0.783 | 0.899 | 0.940 | 0.955 | 0.877 | 0.923 | 0.938 |

an Intel Xeon W-2235 CPU and an NVIDIA Quadro RTX 6000 GPU, running the compiled Unity application and loading the 3D sequences from an external USB-C SSD. A capture of the setup and interface is provided in the supplementary material. Following Section 4.2, only the three V-PCC compression levels were evaluated. The test included all short 4DME sequences, each presented in three compressed versions. Observers could pause and resume the experiment without losing progress. Evaluation continued until 44 distinct micro-expression sequences had been assessed across all three compression levels. **Subjective Experimental Results.** A total of 10 observers (6 men, 4 women; mean age 26.3) participated in the study. Each evaluated approximately 50 sequences, spending on average 45 seconds per sequence. We compared their responses with the original 4DME annotations and computed an AU perception score for each V-PCC compression level, defined as the proportion of AUs still perceived after compression. Table 4 reports the mean AU perception score and per-AU binary F1-score across V-PCC levels. The relatively low perceptual F1-scores reflect the inherent difficulty of AU recognition. Prior studies show that human AU perception remains substantially below automated performance, even for FACS-trained observers, with cross-observer agreement below 80% F1-score [13, 19]. Overall, AU visibility is only marginally affected by compression, with noticeable degradation observed only at the lowest quality level. Although further validation is required, these findings suggest that V-PCC could reduce the size of 4DME by up to 3000× while largely preserving perceptual quality (Table 2). A more conservative strategy using V-PCC R1 would still achieve a 1500× reduction.

## 5 Volumetric AU Detection Framework

As stated earlier, the main challenges of ME-AU detection lie in the subtlety and sparsity of the movements in the spatial and temporal domains. When dealing with volumetric data, recent architectures and techniques become ill-fitted to the increased sample size and relatively limited data for large scale model training.

**Table 4:** Perceptual Experiment Results on AU detection.

| Compression level | AU Perception | F1-score | AU1 | AU2 | AU4 | AU6 | AU7 | AU12 | AU17 | AU45 |
|---|---|---|---|---|---|---|---|---|---|---|
| V-PCC R1 | 0.949 | 0.55 | 0.593 | 0.424 | 0.522 | 0.105 | 0.303 | 0.333 | 0.222 | 0.944 |
| V-PCC R2 | 0.956 | 0.64 | 0.807 | 0.610 | 0.769 | 0.370 | 0.715 | 0.554 | 0.800 | 0.944 |
| V-PCC R3 | 0.957 | 0.66 | 0.784 | 0.610 | 0.767 | 0.686 | 0.733 | 0.732 | 0.909 | 0.943 |

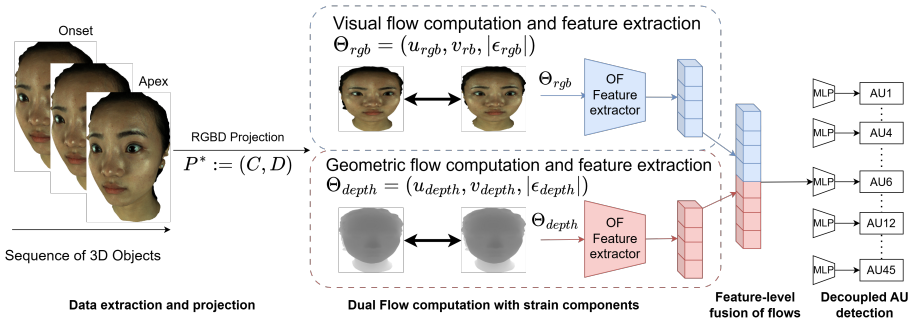


**Fig. 3:** Feature and model design within VoluME.

Starting from these challenges, we propose VoluME, a family of models that use lightweight CNN-based architecture and built-in fusion of the OF features for efficient feature extraction across two separate flow streams. The proposed data pipeline of VoluME is shown in figure Fig. 3

### 5.1 Volumetric Spatio-Temporal Feature Design

Recent volumetric approaches to micro-expression recognition have explored the incorporation of depth information by passing an RGB-D image of the apex frame through modern vision backbones [8]. While this enables efficient fusion of spatial features, these single-frame strategies discard temporal dynamics entirely. For micro-expression AU detection, where movements are subtle and fast, temporal information is critical for capturing the fine-grained dynamics of facial deformations. As a result, relying solely on a single frame is insufficient for detailed AU recognition, as the temporal evolution of micro-movements is completely ignored. A second methodological stream attempts to extend the traditional optical flow formulation to volumetric data by defining the voxel intensity function $I(x, y, z, t)$ and applying the optical flow hypotheses [8]: $\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial z}w = 0$, where $(u, v, w)$ is the volumetric flow vector. The full volumetric flow equations can be found in the supplementary materials.

However, this formulation has key shortcomings. By treating planar and depth displacements on the same scale, it mixes visual motion and depth variation in a single vector, which limits the ability to exploit complementary cues from appearance and geometry. Furthermore, the volumetric flow does not explicitly include optical strain, despite strong evidence from the literature that strain features improve AU detection by capturing subtle surface deformations that are characteristic of micro-expressions [11,16,25]. As a result, this approach can fail to fully represent the fine-grained, surface-specific dynamics that are essential for accurate ME analysis.

To address these limitations, we reformulate flow computation under the assumption that the 3D object represents a single visible facial surface. As such, our 3D object $P(x, y, z, t)$ is fully defined by the RGB-D projected surface $P^*$,

which encodes the color $C$ and the depth $D$:

$$P^*(x, y, t) := \{C(x, y, t), D(x, y, t)\}, \tag{1}$$

where $C$ is the RGB color and $D$ the depth value. The OF assumptions can be applied separately to the color and depth functions :

$$\begin{cases} \dfrac{\partial C}{\partial t} = \dfrac{\partial C}{\partial x} u_{\text{rgb}} + \dfrac{\partial C}{\partial y} v_{\text{rgb}} = 0, \\ \dfrac{\partial D}{\partial t} = \dfrac{\partial D}{\partial x} u_{\text{depth}} + \dfrac{\partial D}{\partial y} v_{\text{depth}} = 0. \end{cases} \tag{2}$$

Here, $of_{\text{rgb}} = (u_{\text{rgb}}, v_{\text{rgb}})$ and $of_{\text{depth}} = (u_{\text{depth}}, v_{\text{depth}})$ are the OF components computed on the XY projection of the 3D object and the depth map, respectively. Our proposed volumetric feature design alleviates the limitations of volumetric flow while being specifically tailored for micro-expression analysis. Indeed, the dual flow streams allow full exploitation of surface-specific dynamics, enabling optical strain computation for both appearance and geometry, which have been shown to significantly increase detection robustness in the 2D domain. Using our two flow streams, we also decouple visual and depth signals to better model AUs that manifest predominantly in one of the two streams.

### 5.2 Model design

**Dual Stream Feature Extraction and Fusion.** Early optical-flow-based methods for micro-expression analysis, such as MDMO and HOOF [5,17], relied on handcrafted features with conventional classifiers. With deep learning, performance improved as convolutional networks learned data-driven representations directly from OF maps. However, micro-expression datasets remain limited in size, which has favoured lightweight architectures that reduce overfitting while retaining discriminative capacity. Accordingly, strong OF-based models such as SSSNet, Off-ApexNet, and STSTNet [11,15,25] employ shallow CNNs to process the OF components $(u, v, |\epsilon_{uv}|)$, balancing representational power and data efficiency for small-scale MER datasets. Following this principle, we extend these OF models within a volumetric dual-stream framework. Prior work [16,25] shows that fusing the three OF components is critical for robust AU detection, as it jointly encodes motion and deformation cues. We therefore improve STSTNet and Off-ApexNet by performing input-level fusion, representing $(u, v, |\epsilon_{uv}|)$ as a three-channel image processed directly by convolutional layers. To preserve the separation between appearance-driven and geometry-driven motion, we keep visual and geometric OF streams independent and perform fusion at the feature level rather than at the input level. This design allows complementary modeling of AUs that manifest differently in texture and geometric displacement before integration.

**AU-Specific Classifier.** Due to the spatial proximity of many facial AUs, simultaneous activations often produce highly correlated motion patterns. Such

co-occurring movements can interfere with each other in a shared classification setting, leading to ambiguity and degraded performance in traditional multi-label classifiers. Prior work has shown that decoupled classification strategies, such as one-vs-rest SVMs [9] or independent AU-specific heads [12], are more robust in this context, as they reduce inter-AU interference and better handle overlapping facial activations. Following this principle, we employ one dedicated multi-layer perceptron per Action Unit. This design limits competition between AUs during training and alleviates confusion in regions where multiple movements are spatially concentrated. By decoupling the classification process, the model is better equipped to perform fine-grained discrimination in high-density AU regions of the face.

## 6 Experimental Validation

### 6.1 Experimental setting

To establish a benchmark for volumetric ME-AU detection, we propose two validation procedures using the 3D data available in the extended 4DME dataset, and in CASME$^3$.

**Validation Within the 4DME Dataset.** The within-dataset evaluation adopts a Leave-One-Subject-Out (LOSO) protocol on the extended and cleaned 4DME dataset. Each subject defines one fold, ensuring strict subject-independent assessment consistent with established practice in micro-expression analysis. VoluME is evaluated alongside the two existing 3D ME-AU detection approaches, CCDN [9] and RGBD AlexNet [8]. Evaluation is restricted to a predefined subset of commonly studied Action Units to ensure comparability with prior work. The selected AUs are listed in Tab. 1.

**Cross-Dataset Validation.** Cross-dataset evaluation follows a Leave-One-Dataset-Out (LODO) protocol. Models are trained on one dataset and evaluated on the other, thereby assessing robustness to inter-dataset domain shift. For CASME$^3$, we use Part C of the dataset, collected under the mock crime paradigm, due to its improved ecological validity and reliability of elicitation [8]. The AU subset matches the within-dataset protocol, with the exception of AU6, which is excluded due to missing annotations.

**Task-Aware Compression Validation.** Task-aware compression evaluation extends the 4DME LOSO protocol to compressed data. VoluME is assessed across four compression levels using Draco [4] and three levels using V-PCC [3], as described in Sec. 4. This allows direct measurement of the effect of geometric and color distortion on downstream AU detection.

**Validation Metrics.** Performance is primarily evaluated using the pooled binary F1-score (BF1), computed over the complete set of predictions under the LOSO and LODO protocols. As shown in [24], pooling predictions across folds provides a more reliable estimate of model performance by reducing sensitivity to fold-level variance. To assess the impact of compression on AU information, we report the AU-wise macro average of the F1 (macro-F1). While macro-F1

is less representative of overall model performance due to its equal weighting of classes, it is better suited for evaluating data quality in this context. By treating all AUs equally, it provides a clearer indication of whether AU signals remain detectable after compression, independently of class imbalance.

### 6.2 Results

**Within-Dataset Results.** In Tab. 5 are reported the results of the models using the LOSO validation protocol. Across folds, VoluME consistently outperforms the benchmarked 3D baselines, highlighting the effectiveness of input-level spatio-temporal fusion through optical flow and the complementary feature-level fusion of geometric and visual streams. Among the proposed variants, SSSNet achieves the highest overall performance, while STSTNet yields competitive results and surpasses SSSNet on selected AUs. Using the SSSNet feature extractor, VoluME reaches a mean binary F1-score of 0.475 across folds, establishing a reference baseline for volumetric ME-AU detection under subject-independent evaluation. Performance remains heterogeneous across AUs. In particular, AU17 exhibits near-zero F1-scores, which can be attributed to its limited prevalence in the dataset: only six subjects contain positive AU17 samples, typically in one or two sequences. This extreme class sparsity makes subject-invariant modeling of AU17 especially challenging in the LOSO setting.

**Cross-Dataset Results.** In Tab. 6 are reported the results of the models using the LODO validation protocol. Overall, VoluME (SSSNet) remains the strongest performer, though the performance gap between VoluME and other models is reduced compared to the LOSO experiments. This convergence reflects the inherent challenges of domain adaptation, as the CASME$^3$ sequences exhibit lower geometric fidelity and more reconstruction artifacts, which increase variability during training. Notably, VoluME(STSTNet) and VoluME(Off-ApexNet) show a larger drop in performance under these conditions, highlighting their sensitivity to noisy or lower-quality data. Interestingly, the RGBD-AlexNet baseline slightly outperforms these two VoluME variants in this cross-dataset setting. This can be attributed to its input-level fusion of RGB and depth channels, which partially compensates for the degraded or incomplete geometry in CASME$^3$ samples. In contrast, our feature-level fusion strategy is designed to extract finer-grained spatio-temporal representations and is therefore better suited to cleaner, higher-fidelity datasets, where subtle facial deformations can be captured more effectively. These findings underscore the complementary strengths of input- and

**Table 5:** LOSO F1 score on the extended 4DME volumetric data. Best result per-AU is underlined, best result overall is **bolded**.

| Model | AU1 | AU2 | AU4 | AU6 | AU7 | AU12 | AU17 | AU45 | BF1 |
|---|---|---|---|---|---|---|---|---|---|
| RGBD AlexNet [8] | 0.111 | 0.182 | 0.264 | 0.200 | 0.371 | 0.452 | 0.000 | 0.000 | 0.323 |
| CCDN [9] | 0.130 | 0.255 | 0.361 | 0.000 | 0.323 | 0.304 | 0.000 | 0.085 | 0.336 |
| VoluME (Off-ApexNet) | 0.283 | 0.291 | 0.247 | 0.071 | 0.398 | 0.457 | 0.039 | 0.210 | 0.397 |
| VoluME (STSTNet) | 0.279 | 0.311 | 0.511 | 0.117 | 0.402 | 0.474 | 0.073 | 0.145 | 0.424 |
| VoluME (SSSNet) | 0.344 | 0.327 | 0.467 | 0.189 | 0.383 | 0.424 | 0.000 | 0.452 | **0.475** |

**Table 6:** LODO F1 score on our two volumetric datasets. Best result per-AU is underlined, best result overall is **bolded**.

| Model | AU1 | AU2 | AU4 | AU7 | AU12 | AU17 | AU45 | BF1 |
|---|---|---|---|---|---|---|---|---|
| RGBD AlexNet [8] | 0.223 | 0.367 | 0.600 | 0.088 | 0.119 | 0.000 | 0.167 | 0.318 |
| CCDN [9] | 0.033 | 0.338 | 0.530 | 0.080 | 0.105 | 0.125 | 0.187 | 0.274 |
| VoluME (STSTNet) | 0.294 | 0.309 | 0.420 | 0.378 | 0.000 | 0.058 | 0.068 | 0.282 |
| VoluME (Off-ApexNet) | 0.222 | 0.305 | 0.506 | 0.000 | 0.362 | 0.070 | 0.227 | 0.308 |
| VoluME (SSSNet) | 0.197 | 0.386 | 0.425 | 0.483 | 0.465 | 0.066 | 0.165 | **0.387** |

feature-level fusion: the former provides robustness against noisy inputs, while the latter enhances discriminative power when high-quality 3D data is available.

**Task-focused Compression Evaluation.** To assess the impact of compression on downstream performance, we evaluate the VoluME models on Draco and V-PCC decoded sequences across multiple compression levels (Table 7). As expected, detection performance decreases with increasing geometric and visual degradation. In particular, V-PCC R1 and Draco QP6–QP7 introduce distortions that substantially affect ME-AU detection accuracy. At higher fidelity settings (Draco QP9 and V-PCC R2/R3), performance remains stable, with F1-score variations limited to approximately 1% relative to the original data. Notably, despite achieving significantly higher compression ratios, V-PCC maintains and even improves detection performance compared to Draco. These results indicate that V-PCC offers a more favourable trade-off between compression efficiency and task preservation for 3D micro-expression analysis.

**Ablation study of VoluME components** To validate the main design choices of VoluME under both LOSO and LODO evaluation protocols. We evaluate the single-stream RGB and depth variants of VoluME, input-level instead of feature-level fusion, and removal of optical strain. Results are reported in table 8. Combining RGB and depth motion consistently improves performance over single-stream variants, while feature-level fusion outperforms direct input fusion, particularly under cross-dataset evaluation. Removing optical strain produces the largest degradation, confirming the importance of our dual optical flow formulation compared to volume flow.

**Table 7:** LOSO macro-F1 comparison between uncompressed and compressed sets of the volumetric 4DME data. Best result per codec is **bolded**

.

| Compression | VoluME (SSSNet) | VoluME (STSTNet) | VoluME (Off-ApexNet) | Average |
|---|---|---|---|---|
| Uncompressed | 0.386 | 0.375 | 0.305 | 0.356 |
| VPCC R3 | 0.384 | 0.371 | 0.313 | **0.356** |
| VPCC R2 | 0.382 | 0.346 | 0.304 | 0.344 |
| VPCC R1 | 0.314 | 0.328 | 0.276 | 0.306 |
| Draco QP9 | 0.376 | 0.336 | 0.345 | **0.352** |
| Draco QP8 | 0.371 | 0.293 | 0.298 | 0.321 |
| Draco QP7 | 0.367 | 0.293 | 0.272 | 0.308 |
| Draco QP6 | 0.330 | 0.257 | 0.261 | 0.283 |

**Table 8:** Ablation study of the VoluME design choices under within- and cross-dataset evaluation. Best results are **bolded**.

| | Our Method (VoluME-SSSNet) | RGB-only (Single stream) | Depth-only (Single stream) | Input Fusion (RGB+Depth) | No Strain (Dual-stream) |
|---|---|---|---|---|---|
| LOSO | **0.485** | 0.405 | 0.364 | 0.396 | 0.364 |
| LODO | **0.389** | 0.300 | 0.312 | 0.217 | 0.175 |

## 7 Conclusion

This work revisits the volumetric component of 4DME and establishes a solid foundation for 3D micro-expression analysis. Through a systematic characterization of geometric noise, we demonstrate that a substantial portion of the original release does not satisfy the structural assumptions underlying automated spatio-temporal modeling. By addressing these limitations, we recover previously unusable sequences and extend the effective size and reliability of the volumetric benchmark. Building upon this cleanFed and consolidated dataset, we introduce the first cross-dataset evaluation protocol for volumetric ME-AU detection and propose the VoluME family of models, achieving state-of-the-art detection results using our spatio-temporal feature formulation, providing standardized baselines for future research. Finally, we present a task- and perception-aware validation of modern point cloud compression schemes, releasing a compressed version of the 4DME data that achieves a 3000× reduction in storage with negligible downstream ($\approx 1\%$) and perceptual ($< 1\%$) performance degradation. Together, these contributions transform the 4DME data from a partially constrained resource into a scalable and reproducible benchmark, enabling reliable and storage-efficient volumetric micro-expression research.

## Acknowledgements

The authors gratefully acknowledge support from the Marie Skłodowska-Curie Actions (MSCA) through the ACMOD project (Grant No. 101130271), which enabled this collaboration. The authors also acknowledge the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 359894), the University of Oulu and Research Council of Finland Profi 7 (grant 352788), and the computational resources provided by the Mahti supercomputer and CSC – IT Center for Science, Finland, and the GLiCID Computing Facility (https://doi.org/10.60487/glicid, Pays de la Loire, France) which supported the experiments conducted in this study. Y. Li acknowledges support from the CAAS-ASTIP-2026-AII project.

## References

1. De Fré, M., van der Hooft, J., Wauters, T., De Turck, F.: Demonstrating adaptive many-to-many immersive teleconferencing for volumetric video. In: Proceedings of the 15th ACM Multimedia Systems Conference. pp. 453–458 (2024)

2. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology **17**, 124–129 (1971). https://doi.org/10.1037/h0030377
3. Fréneau, L., Gautier, G., Mercat, A., Vanne, J.: uvgvpccenc: Practical open-source encoder for fast v-pcc compression. In: Proceedings of the 16th ACM Multimedia Systems Conference. pp. 235–241 (2025)
4. Google: Draco. https://google.github.io/draco/ (2025), [SW]
5. Happy, S.L., Routray, A.: Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Transactions on Affective Computing **PP**, 1–1 (07 2017). https://doi.org/10.1109/TAFFC.2017.2723386
6. ISO/IEC JTC1/SC29/WG11: Common test conditions for V3C and V-PCC. Tech. Rep. N19518, MPEG (Jul 2020), online
7. Koff, D.A., Shulman, H.: An overview of digital compression of medical images: can we use lossy image compression in radiology? Canadian association of radiologists journal **57**(4), 211 (2006)
8. Li, J., Dong, Z., Lu, S., Wang, S.J., Yan, W.J., Ma, Y., Liu, Y., Huang, C., Fu, X.: CAS(ME)3: A Third Generation Facial Spontaneous Micro-Expression Database With Depth Information and High Ecological Validity. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 2782–2800 (Mar 2023). https://doi.org/10.1109/TPAMI.2022.3174895, https://ieeexplore.ieee.org/document/9774929
9. Li, X., Cheng, S., Li, Y., Behzad, M., Shen, J., Zafeiriou, S., Pantic, M., Zhao, G.: 4DME: A Spontaneous 4D Micro-Expression Dataset With Multimodalities. IEEE Transactions on Affective Computing **14**(4), 3031–3047 (Oct 2023). https://doi.org/10.1109/TAFFC.2022.3182342, https://ieeexplore.ieee.org/abstract/document/9796028, conference Name: IEEE Transactions on Affective Computing
10. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikäinen, M.: Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods. IEEE Transactions on Affective Computing **9**(4), 563–577 (Oct 2018). https://doi.org/10.1109/TAFFC.2017.2667642, https://ieeexplore.ieee.org/document/7851001
11. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikäinen, M.: Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8. IEEE (2019)
12. Li, Y., Huang, X., Zhao, G.: Micro-expression action unit detection with spatial and channel attention. Neurocomputing **436**, 221–231 (May 2021). https://doi.org/10.1016/j.neucom.2021.01.032, https://www.sciencedirect.com/science/article/pii/S0925231221000539
13. Li, Y., Wei, J., Liu, Y., Kauttonen, J., Zhao, G.: Deep learning for micro-expression recognition: A survey (2022), https://arxiv.org/abs/2107.02823
14. Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., Manohara, M., et al.: Toward a practical perceptual video quality metric. The Netflix Tech Blog **6**(2), 2 (2016)
15. Lionel, S.B., Wong, Y.K., See, J.: Off-apexnet on micro-expression recognition system (2018), https://arxiv.org/abs/1805.08699, arXiv:1805.08699
16. Liong, S.T., See, J., Phan, R.C.W., Oh, Y.H., Cat Le Ngo, A., Wong, K., Tan, S.W.: Spontaneous subtle expression detection and recognition based on facial strain. Signal Processing: Image Communication **47**, 170–182 (2016). https://doi.org/https://doi.org/10.1016/j.image.2016.06.004, https://www.sciencedirect.com/science/article/pii/S092359651630090X

17. Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X.: A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. IEEE Trans. Affect. Comput. **7**(4), 299–310 (Oct 2016). https://doi.org/10.1109/TAFFC.2015.2485205, https://doi.org/10.1109/TAFFC.2015.2485205
18. MPEG: mpeg-pcc-tmc2. https://github.com/MPEGGroup/mpeg-pcc-tmc2 (2025)
19. Qu, F., Wang, S.J., Yan, W.J., Fu, X.: CAS(ME)2: A Database of Spontaneous Macro-expressions and Micro-expressions. In: Kurosu, M. (ed.) Human-Computer Interaction. Novel User Experiences. pp. 48–59. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-39513-5_5
20. Schwarz, S., Preda, M., Baroncini, V., Budagavi, M., Cesar, P., Chou, P.A., Cohen, R.A., Krivokuća, M., Lasserre, S., Li, Z., Llach, J., Mammou, K., Mekuria, R., Nakagami, O., Siahaan, E., Tabatabai, A., Tourapis, A.M., Zakharchenko, V.: Emerging MPEG Standards for Point Cloud Compression. IEEE Journal on Emerging and Selected Topics in Circuits and Systems **9**(1), 133–148 (Mar 2019). https://doi.org/10.1109/JETCAS.2018.2885981, https://ieeexplore.ieee.org/abstract/document/8571288
21. See, J., Yap, M.H., Li, J., Hong, X., Wang, S.J.: MEGC 2019 – The Second Facial Micro-Expressions Grand Challenge. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–5 (May 2019). https://doi.org/10.1109/FG.2019.8756611, https://ieeexplore.ieee.org/document/8756611
22. Tian, D., Ochimizu, H., Feng, C., Cohen, R., Vetro, A.: Geometric distortion metrics for point cloud compression. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3460–3464 (Sep 2017). https://doi.org/10.1109/ICIP.2017.8296925, https://ieeexplore.ieee.org/abstract/document/8296925?casa_token=IAJYelpWpNOAAAAA:zsbC1U7QLI_3smZAZxmcpDy82aBcYIlnSV6IBnTKDX_J3UAUYvIe7WMhasdGrJfYKAEXlqglbSA
23. Tious, A., Vigier, T., Ricordel, V.: New challenges in point cloud visual quality assessment: a systematic review. Frontiers in Signal Processing **4**, 1420060 (2024)
24. Varanka, T., Li, Y., Peng, W., Zhao, G.: Data leakage and evaluation issues in micro-expression analysis. IEEE Transactions on Affective Computing **15**(1), 186–197 (Jan 2024). https://doi.org/10.1109/taffc.2023.3265063, http://dx.doi.org/10.1109/TAFFC.2023.3265063
25. Varanka, T., Peng, W., Zhao, G., Peng, W., Zhao, G.: Micro-Expression Recognition with Noisy Labels. Electronic Imaging **33**, 1–8 (Jan 2021). https://doi.org/10.2352/ISSN.2470-1173.2021.11.HVEI-157, https://library.imaging.org/ei/articles/33/11/art00009, publisher: Society for Imaging Science and Technology
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
27. Wu, C.H., Hsu, C.F., Hung, T.K., Griwodz, C., Ooi, W.T., Hsu, C.H.: Quantitative Comparison of Point Cloud Compression Algorithms With PCC Arena. IEEE Transactions on Multimedia **25**, 3073–3088 (2023). https://doi.org/10.1109/TMM.2022.3154927, https://ieeexplore.ieee.org/document/9722998
28. Yuan, K., Yu, Z., Liu, X., Xie, W., Yue, H., Yang, J.: Auformer: Vision transformers are parameter-efficient facial action unit detectors. arXiv preprint arXiv:2403.04697 (2024)
29. Zerman, E., Ozcinar, C., Gao, P., Smolic, A.: Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression. In: 2020

Twelfth International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–6 (May 2020). https://doi.org/10.1109/QoMEX48832.2020.9123137, https://ieeexplore.ieee.org/abstract/document/9123137?casa_token=f9fZkRVe2C4AAAAA:HBfHalRxUtDBo66fTyKljdLCyWb1hLnHZ2CQPlcFks3AXu6_3WtK__ta2GpX-SbdQQWKww5k

30. Zhao, G., Pietikainen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(6), 915–928 (Jun 2007). https://doi.org/10.1109/TPAMI.2007.1110, https://ieeexplore.ieee.org/document/4160945
31. Zhou, L., Wang, M., Huang, X., Zheng, W., Mao, Q., Zhao, G.: An empirical study of super-resolution on low-resolution micro-expression recognition. IEEE Transactions on Affective Computing (2025)
32. Zhou, X., Viola, I., Rossi, S., Cesar, P.: Comparison of visual saliency for dynamic point clouds: Task-free vs. task-dependent. IEEE Transactions on Visualization and Computer Graphics (2025)
33. Zhu, J.: High-end Video Streaming Qualityin the Wild: measuring and Predicting Satisfied User Ratio. Ph.D. thesis, Nantes Université (2024)