Master's Thesis

Explainable Multi-class Classification Framework on Video using Interpretable Deep Learning Techniques

RENATI MULATI

Supervisor Shinobu Hasegawa

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Master of Science in Information Science)

March, 2026

Abstract

Explainable video-based classification has become increasingly important in affective computing and other human-centric artificial intelligence applications, where models are expected not only to achieve high recognition accuracy but also to provide transparent and trustworthy decision processes. In contrast to static image recognition, video understanding introduces additional challenges due to temporal variation, motion-induced noise, and frame-level ambiguity.

Although deep learning has significantly advanced video classification through spatio-temporal architectures such as 3D convolutional networks and video transformers, these models typically operate as opaque black boxes. Their lack of interpretability limits their applicability in real-world scenarios where understanding *why* a decision is made is as critical as the decision itself, particularly in domains involving human behavior and affective states.

Recent research in explainable artificial intelligence (XAI) has emphasized the importance of *intrinsic interpretability*, advocating the use of models whose internal reasoning processes are inherently transparent rather than post-hoc approximations. In image-based tasks, prototype-based and dictionary-based methods have demonstrated promising interpretability by grounding predictions in concrete visual evidence. However, extending such intrinsically interpretable models to video-based classification remains an open problem.

Existing video explainability approaches often rely on post-hoc saliency or gradient-based methods such as Grad-CAM, which may suffer from temporal inconsistency and limited semantic clarity when applied frame by frame. As a result, there is a growing need for frameworks that preserve intrinsic interpretability while addressing the unique challenges of video data.

This thesis investigates the adaptability and interpretability of *image-level interpretable deep learning models* when applied to video-based multi-class classification. Distinct from video-native architectures that jointly model space and time, we propose a unified framework that decomposes video understanding into two stages: (1) an inherently interpretable **image-level reasoner** (ILR) that performs frame-wise inference and explanation, and (2) a lightweight temporal aggregation mechanism that integrates frame-level predictions into a video-level decision. Specifically, Exponentially Weighted Moving Average (EWMA) is employed to stabilize predictions over time without introducing additional black-box temporal parameters.

This design explicitly separates explanation generation from temporal smoothing, thereby preserving frame-level interpretability while enabling video-level classification.

Within this framework, two representative interpretable paradigms are evaluated. The first is **ProtoPNet**, a prototype-based model that explains predictions by comparing local feature patches with learned prototypical parts. ProtoPNet provides case-based explanations in the form of "this looks like that," aligning model reasoning with human-understandable visual concepts. The second is **PatchSAE**, a sparse dictionary-based model inspired by sparse autoencoders and recent advances in concept decomposition. PatchSAE explains predictions through the activation of a small subset of latent dictionary atoms, offering a complementary perspective based on distributed local visual primitives rather than explicit semantic parts. By comparing these two paradigms under identical experimental conditions, this thesis aims to clarify how different intrinsic explanation mechanisms behave in dynamic video scenarios.

To systematically evaluate the proposed framework, experiments were conducted primarily on the **RAVDESS** dataset, a widely used benchmark for audio-visual emotion analysis. The task is formulated under three distinct label configurations to analyze the impact of semantic granularity and visual ambiguity. The *Original* setting employs the standard six-class facial expression taxonomy. The *V1* setting groups expressions into three coarse categories based on semantic valence, reducing classification ambiguity. The *V2* setting introduces a more challenging "Visual Interference" scenario, in which the *Sad* class is merged into the *Neutral* group, deliberately increasing intra-class visual similarity and testing model robustness against low-arousal distractors. This controlled task design allows us to investigate how interpretable models respond to varying degrees of semantic and visual complexity.

In addition to task formulation, preprocessing is treated as a first-class experimental variable. Three preprocessing configurations are examined: original frames with background, face-only representations with background removed, and reassembled facial representations emphasizing key facial regions. Prior work has shown that background context and texture bias can strongly influence convolutional networks. By explicitly controlling preprocessing, this study analyzes how different input representations affect both recognition performance and the semantic validity of generated explanations. The face-only configuration, in particular, plays a crucial role in ensuring that explanations are grounded in physiologically meaningful facial dynamics rather than spurious background cues.

Experimental results demonstrated that image-level interpretable models

can be effectively adapted to video-based classification using the proposed framework. Quantitatively, while the challenging V2 setting (Visual Interference) caused a performance drop at the image level due to semantic ambiguity (0.966 to 0.811), ProtoPNet achieved superior performance at the video level in this setting (74.3%) compared to the standard V1 setting (64.8%). This contrasts with baseline models and suggests that prototype-based reasoning aligns well with visual intensity groupings, effectively handling low-arousal expressions when they are grouped together.

Beyond quantitative metrics, qualitative analysis provides deeper insight into explanation behavior. ProtoPNet tends to produce stable, part-based explanations aligned with distinct facial organs such as the eyes and mouth, resulting in visually intuitive reasoning traces. PatchSAE, in contrast, emphasizes fine-grained local textures and distributed patterns, capturing subtle visual cues but occasionally exhibiting stronger texture bias. These differences illustrate a fundamental trade-off between semantic clarity and representational flexibility in intrinsic explanation mechanisms. A questionnaire-based human study is planned to further evaluate how prototype-based and patch-based explanations are perceived by human observers in terms of intuitiveness and trustworthiness.

Finally, preliminary experiments on supplementary datasets, including DIEM-A for body movement analysis and WLASL-10 for sign language recognition, suggest that the proposed framework generalizes beyond facial expression recognition. While these results are exploratory, they indicate the broader potential of image-level interpretable reasoning combined with temporal aggregation for multi-scale human motion understanding.

In summary, this thesis bridges the gap between static intrinsic interpretability and dynamic video analysis. By validating a unified, interpretable pipeline for video-based multi-class classification, it provides a transparent alternative to black-box video models and establishes a robust baseline for future research in explainable video understanding.

Acknowledgements

I would like to express my sincere gratitude to my supervisor at JAIST, Prof. Shinobu Hasegawa, for his continuous guidance, mentorship, and support throughout my master's program.

I would also like to express my deep appreciation to Prof. Teeradaj Racharak, Associate Professor at Tohoku University and Visiting Associate Professor at the Japan Advanced Institute of Science and Technology (JAIST), who served as my co-supervisor. His insightful feedback, rigorous academic standards, and advice on research methodology were essential to the development of this thesis. This research would not have been possible without his expertise, particularly during the critical phases of framework design and experiment analysis.

Finally, I would like to thank my family for their constant support and understanding throughout my graduate studies.

Contents

Acknowledgements **1**

1 Introduction **1**
1.1 Background 1
1.2 Explanation in Image and Video Classification 2
1.2.1 Post-hoc Explanation Techniques 2
1.2.2 Interpretable Models 3
1.3 Challenges in Video-based Expression Recognition 5
1.4 Motivation 6
1.5 Objectives 7
1.6 Contributions 8
1.7 Thesis Outline 9

2 Related Work **10**
2.1 Facial Expression Recognition (FER) 10
2.2 Video Understanding and Datasets 11
2.2.1 Deep Learning for Video Classification 11
2.2.2 Datasets for Human-Centric Video Analysis 11
2.3 Explainable AI (XAI) 12
2.3.1 Post-hoc Explanations 12
2.3.2 Evaluation of Interpretability 13
2.4 Inherently Interpretable Architectures 13
2.4.1 Prototype-based Learning (ProtoPNet) 13
2.4.2 Sparse Autoencoders and Patch-based Methods 14

3 Proposed Model **15**
3.1 Overview of the Proposed Framework 15
3.2 Image-Level Explanation Module Architecture 16
3.3 Backbone Network 17
3.4 Preprocessing Strategies 17
3.5 Interpretable Mechanisms 17

3.5.1 Prototype-based Reasoning (ProtoPNet) 18
3.5.2 Patch-based Sparse Reasoning (PatchSAE) 18
3.6 Training Procedure 19
3.6.1 Training of Prototype-based Models 19
3.6.2 Training of Patch-based Models 19
3.7 Video-level Aggregation Strategy 20
3.7.1 Frame-level Prediction Sequence 21
3.7.2 Exponentially Weighted Moving Average (EWMA) . . 21
3.7.3 Final Video-level Decision 22

4 Experimentation **23**
4.1 Datasets 23
4.1.1 RAVDESS (Primary Dataset) 24
4.1.2 DIEM-A (Supplementary Dataset) 24
4.1.3 WLASL (Supplementary Dataset) 25
4.2 Label Settings and Task Formulation 25
4.3 Preprocessing Configurations 26
4.4 Implementation Details 28
4.4.1 Common Training Settings 29
4.4.2 ProtoPNet Specifics 29
4.4.3 PatchSAE Specifics 30

5 Evaluation **32**
5.1 Quantitative Analysis on RAVDESS 32
5.1.1 Task Definition Recap 32
5.1.2 Summary of Results 33
5.2 Image-level Recognition Performance on RAVDESS 34
5.3 Video-level Recognition Performance Comparison 36
5.4 Qualitative Explainability Comparison 38
5.4.1 Visual Analysis of Temporal Dynamics 38
5.4.2 Analysis of Patch-based Explanations (Patch-SAE) . . 39
5.4.3 Analysis of Prototype-based Explanations (ProtoPNet) 41
5.4.4 Discussion 42
5.5 Supplementary Experiments on Generalization 43

6 Conclusion **46**
6.1 Summary of Contributions 46
6.2 Key Empirical Findings 46
6.3 Limitations 47
6.4 Future Work 48

A Proposed Human Study Questionnaire **54**

List of Figures

1.1 Illustration of post-hoc explanation techniques. While Saliency maps (left) capture high-resolution, fine-grained details, they often lack class discriminativity. In contrast, Grad-CAM (right) produces localized heatmaps that accurately highlight class-specific regions (e.g., distinguishing between 'Cat' and 'Dog'), despite its coarser spatial resolution. (Image adapted from Selvaraju et al. [30]) 3
1.2 Schematic comparison of explanation mechanisms 4
1.3 Illustrative examples highlighting the limitations of post-hoc explanation methods in video understanding. While saliency maps may appear reasonable in isolated frames (a), they can also become spatially diffuse or semantically misaligned in complex scenes (b), raising concerns about their temporal consistency. Figure adapted from [28]. 6

2.1 Overview of the Poster++ architecture. The model adopts a multi-stage, landmark-guided feature extraction framework with cross-attention mechanisms to progressively capture fine-grained facial deformations under occlusion and pose variation. Figure adapted from [22]. 11
2.2 Temporal instability of frame-wise post-hoc explanations. Top row shows original frames sampled from a RAVDESS video. Bottom row shows corresponding Grad-CAM visualizations generated independently for each frame. Although the facial appearance remains largely unchanged, the highlighted regions vary noticeably across frames, demonstrating saliency flickering in video settings. 13

3.1 Overall workflow of the proposed video-based emotion recognition framework. Given an input video, frames are sampled and preprocessed before being encoded by a backbone network. Interpretable representations are then extracted at the frame level and aggregated temporally to produce a video-level prediction. 16
3.2 Overview of the image-level explanation module. Two interpretable instantiations are considered: ProtoPNet and Patch-SAE. 16
3.3 Overview of the video-level aggregation strategy. Raw frame-level predictions are first smoothed using EWMA to ensure temporal consistency, followed by pooling to generate the final decision. 21

4.1 Visual comparison of the datasets. **(a) RAVDESS:** High-quality RGB studio recording. **(b) DIEM-A:** Sparse skeletal representation derived from motion capture. **(c) WLASL:** "In-the-wild" footage with varying backgrounds and lighting conditions. 23
4.2 Examples of different preprocessing configurations used in this study. **Left:** Original frames retaining background context. **Middle:** Face-only frames with background masked, forcing the model to focus on facial features. **Right:** Reassembled facial representations, constructed by stitching extracted eye and mouth regions onto a canvas to test part-based recognition. 27

5.1 Visualization of the model's high spatial discriminability. **Left column:** Input images with the yellow bounding box indicating the receptive field of the most activated prototype. **Right column:** The corresponding upsampled activation heatmaps. The results demonstrate that the learned prototypes (e.g., furrowed brows in the bottom row or closed eyes in the middle row) are highly localized and semantically meaningful, focusing on specific facial muscles rather than background noise. . . 35
5.2 Confusion Matrix Comparison. **(a) V1 Setting:** The model shows clean separation based on semantic valence. **(b) V2 Setting:** Crucially, the model successfully classifies the interference class (Sad) into the Neutral (Middle row), demonstrating its reliance on visual intensity (low arousal) rather than high-level emotion semantics. 36

5.3 Confusion Matrix of Patch-SAE on Video-level classification (V1 setting). The model shows a strong bias towards classifying Neutral/Positive videos as Negative. 37
5.4 Visual analysis of input frame sequences across different emotional classes. **(a)** The specific sequence (Neutral) analyzed in the case study, showing subtle facial movements. **(b) & (c)** Additional sequences from 'Positive' and 'Negative' classes. Validating the analysis on these multiple samples confirms that the model consistently handles varying degrees of facial intensity. 39
5.5 Image-level latent activation profile of Patch-SAE for a selected frame. The model exhibits sparsity, with only a few dominant neurons (e.g., Latent #21964) being highly activated. 40
5.6 Reference images retrieved for the dominant Latent #21964. The red bounding boxes consistently highlight the glabella (area between eyebrows) and forehead region across different subjects. This indicates that Patch-SAE utilizes this specific local texture feature as evidence for the 'Neutral' class. 40
5.7 ProtoPNet explanations using representative prototypes. (a) Prototype #17 identifies an eye-related pattern. (b) Prototype #18 identifies a nose/mid-face pattern. The heatmaps (right) show the precise location where the input matches the prototype. 41

List of Tables

4.1 Summary of label settings and task formulations. This table consolidates the mapping strategies and the rationale behind the V1 (Valence) and V2 (Visual Interference) configurations. 27
4.2 Hyperparameter settings for ProtoPNet training. 30
4.3 Hyperparameter settings for PatchSAE training. 31

5.1 Data statistics for each class under different task settings. . . . 33
5.2 Quantitative comparison of video-level classification accuracy (%) on the RAVDESS **test set**. The dataset was randomly partitioned into training (70%), validation (20%), and testing (10%) sets. Comparison between black-box baselines and interpretable models. 33
5.3 Image-level classification accuracy on the RAVDESS dataset under different label settings. 34
5.4 Comparison of recognition accuracy (%) on DIEM-A and WLASL-10. The table benchmarks our interpretable models against specialized State-of-the-Art (SOTA) black-box baselines. Note: The performance gap primarily stems from the trade-off between maximizing accuracy (black-box temporal modeling) and ensuring frame-level interpretability 44

Chapter 1

Introduction

1.1 Background

Video-based classification is a fundamental problem in computer vision and plays an important role in a wide range of applications, including action recognition, affective computing, human-computer interaction, and video understanding. Compared to image-based classification, video-based tasks require models to reason not only about spatial appearance but also about temporal dynamics across frames [29]. While temporal information provides richer context, it also introduces additional complexity for both modeling and analysis.

Recent advances in deep learning have led to strong performance improvements in visual recognition. A common paradigm is to extract frame-level features using convolutional neural networks, such as VGG [33] or ResNet [12], and to aggregate these features over time to obtain a video-level prediction. Such approaches have proven effective in practice and are widely used in contemporary video classification pipelines.

Despite their effectiveness, most video-based classifiers rely on implicit temporal representations. As a result, it is often unclear how frame-level visual evidence contributes to the final video-level decision. This lack of transparency limits the extent to which model decisions can be systematically analyzed, raising concerns about trust and reliability [26].

These challenges are particularly evident in human-centered video analysis tasks, where subtle visual cues and temporal consistency play a critical role. For example, in facial expression analysis (FER) [17], relevant information may be distributed across time and can be easily affected by background variation and motion blur. Although video-based models can capture complex temporal patterns, their decision-making processes often remain difficult

to interpret [28].

Scope of this thesis. This thesis does not aim to propose a new video-native explainability algorithm. Instead, we study how *image-level* interpretable models behave in *video-based* classification settings. Specifically, ProtoPNet [6] and PatchSAE [18] provide intrinsic explanations at the frame level, and we analyze how frame-wise predictions and explanations are accumulated to form a final video-level decision.

1.2 Explanation in Image and Video Classification

Explainability methods aim to provide insights into how deep learning models make predictions. Broadly, these methods can be categorized into two paradigms: post-hoc techniques applied to black-box models, and inherently interpretable architectures.

1.2.1 Post-hoc Explanation Techniques

Post-hoc explanation techniques are commonly used to analyze trained models without modifying their architectures, as shown in Figure 1.1. Representative approaches include saliency maps [31], gradient-based methods such as Grad-CAM [30], and perturbation-based methods. These techniques generally function by highlighting input regions (heatmaps) that are considered important for a given prediction. They are widely adopted due to their flexibility, as they can be applied to any standard CNN trained for tasks like action recognition[32].

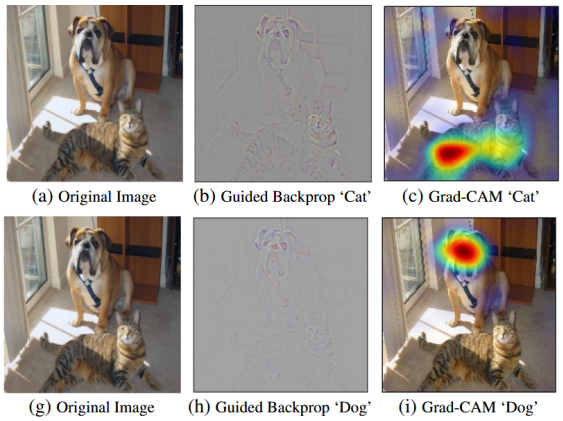


Figure 1.1: Illustration of post-hoc explanation techniques. While Saliency maps (left) capture high-resolution, fine-grained details, they often lack class discriminativity. In contrast, Grad-CAM (right) produces localized heatmaps that accurately highlight class-specific regions (e.g., distinguishing between 'Cat' and 'Dog'), despite its coarser spatial resolution. (Image adapted from Selvaraju et al. [30])

1.2.2 Interpretable Models

An alternative approach is to design models whose structures are inherently interpretable, often referred to as "glass-box" models [26]. This thesis focuses on two representative categories: prototype-based and patch-based models.

Prototype-based models, such as ProtoPNet [6], learn a set of representative prototypes during training. Each prototype corresponds to a meaningful local pattern associated with a specific class. Predictions are made by comparing input features with these learned prototypes, allowing explanations to be expressed in terms of visual similarity to concrete examples ("this looks like that"). Recent works have extended this concept to hierarchical structures [24], further enhancing interpretability.

Patch-based models rely on the concept of dictionary learning to decompose inputs into fundamental interpretable units, sharing conceptual similarities with Bag-of-Features approaches [3]. Foundational to this approach is the **Sparse Autoencoder (SAE)**, formally introduced into deep learning by **Makhzani et al. [21]**. An SAE is an unsupervised neural network that learns to reconstruct its input using a linear combination of basis functions (atoms) from a learned dictionary. Unlike standard autoencoders, it imposes a sparsity constraint (e.g., k-sparse or L_1 regularization) on the latent activations, forcing the model to represent complex data using only a small number of active elements.

Building on this modern sparse representation framework, **PatchSAE** [18] extends the methodology to interpretable image classification. Instead of processing the entire image holistically, PatchSAE treats the input feature map as a collection of local patches. It learns a dictionary of visual primitives (e.g., edges, textures). During inference, the model explains its decision by identifying which specific dictionary atoms are activated, effectively decomposing the image into a sparse set of salient local features[3]. Unlike post-hoc approximation methods, such as saliency maps or gradient-based visualizations, both prototype-based and patch-based models embed the explanation mechanism directly into the decision-making process. As illustrated in Figure 1.2, in these models, the same interpretable quantities that determine the final prediction—specifically the prototype similarity scores in ProtoPNet or sparse latent activations in PatchSAE also constitute the explanation itself. As a result, explanations are not generated retrospectively after the prediction is fixed, but are an integral and verifiable part of the predictive computation [26, 2].

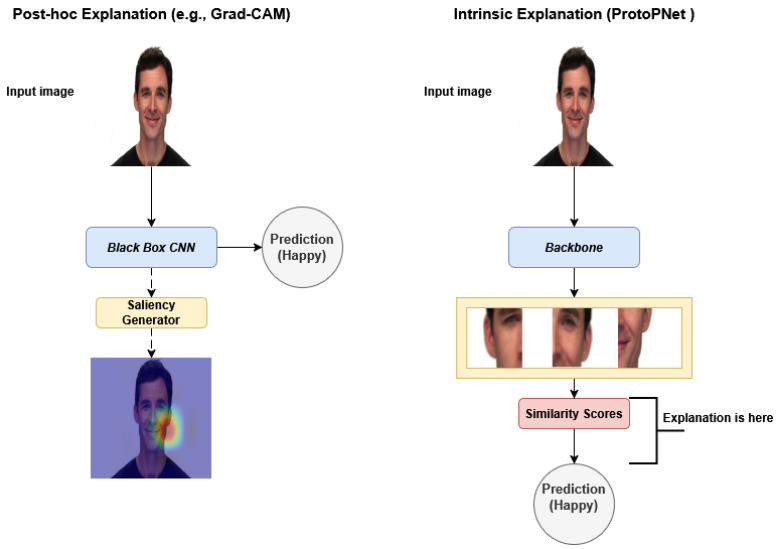


Figure 1.2: Schematic comparison between post-hoc and intrinsic explanation mechanisms. **Left:** Post-hoc methods (e.g., Grad-CAM) generate explanations via a secondary branch after the black-box prediction is made. **Right:** Intrinsic models (e.g., ProtoPNet) embed the explanation directly into the computation path, where the similarity scores to interpretable prototypes determine the final prediction.

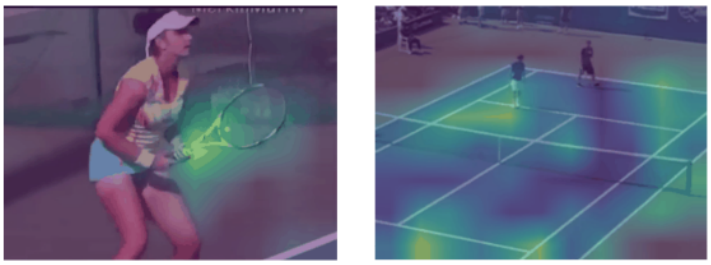
1.3 Challenges in Video-based Expression Recognition

In real-world applications, facial expressions are typically observed in video sequences rather than single images. Video-based facial expression recognition introduces additional challenges such as temporal variation, motion blur, and frame-level noise.

For instance, in the **RAVDESS** dataset, the temporal structure of expressions presents a significant challenge known as the *onset-apex-offset* evolution. A video sequence globally labeled as "Happy" or "Angry" typically begins and ends with a neutral face, with the target emotion appearing strongly only during the peak (apex) phase. Consequently, a model processing the video frame-by-frame encounters conflicting evidence: it may correctly classify the apex frames as "Happy" while tending to predict "Neutral" for onset and offset frames, creating significant temporal inconsistency. Furthermore, distinct categories such as "Calm" and "Neutral" share high visual similarity in individual frames. Without sufficient temporal context, the subtle intensity differences between these classes result in frame-level ambiguity, making it difficult for the model to distinguish them based on static appearance alone.

Limitations of Post-hoc Explanations in Video. While post-hoc techniques are widely used for interpreting static images, they face fundamental limitations when extended to video data. The primary critique is the lack of *temporal consistency*.

Since post-hoc methods are typically applied independently to each frame, even small visual perturbations can lead to large variations in the generated saliency maps across time, a phenomenon commonly referred to as *saliency flickering*. As illustrated in Figure 1.3, saliency-based explanations may become spatially unstable or semantically misaligned in complex video scenes, despite minimal changes in the underlying content. Such temporal inconsistency makes it difficult to interpret continuous motion and undermines the reliability of frame-wise explanations in video analysis [28].



(a) Positive Illustration (b) Limitation

Figure 1.3: Illustrative examples highlighting the limitations of post-hoc explanation methods in video understanding. While saliency maps may appear reasonable in isolated frames (a), they can also become spatially diffuse or semantically misaligned in complex scenes (b), raising concerns about their temporal consistency. Figure adapted from [28].

Furthermore, post-hoc explanations are often criticized for their lack of *faithfulness* [26]. Saliency maps may highlight regions that seem intuitive to humans but do not actually drive the model's prediction. In contrast, inherently interpretable models aim to provide explanations that faithfully reflect the model's internal computation.

1.4 Motivation

Recent advances in video-based classification have achieved remarkable performance by leveraging spatio-temporal architectures such as 3D convolutional networks and video transformers [34, 32]. However, these models are typically optimized for accuracy and operate as *black boxes*, making it difficult to understand which visual evidence drives their decisions. This lack of transparency limits their applicability in human-centric and high-stakes scenarios, where interpretability is as important as recognition performance [26, 2].

To address this issue, **intrinsically interpretable models** have been proposed in the image domain. Among them, **prototype-based methods**, such as ProtoPNet, explain predictions by explicitly matching input regions to learned visual prototypes that correspond to representative semantic parts [6]. These models provide intuitive, part-level explanations and have demonstrated strong interpretability on static image recognition tasks. However, existing studies primarily focus on image-level settings, and it remains unclear how prototype-based explanations behave when applied to video data,

where frame-level noise, temporal variation, and aggregation strategies may lead to unstable or inconsistent prototype activations.

In parallel, **patch-based explanation mechanisms** have emerged as an alternative interpretable paradigm. Sparse autoencoder-based approaches, such as PatchSAE, decompose visual inputs into a small set of active latent atoms, forming a learned dictionary of local visual primitives [21]. Compared to prototype-based methods, patch-based approaches do not rely on explicit semantic exemplars, and may therefore exhibit different robustness properties under temporal variations. Despite their growing importance in interpretability research, their behavior in video-based classification scenarios has not been systematically analyzed.

Another critical yet often overlooked factor in explainable video classification is **data preprocessing**. In facial expression recognition and related human-centric tasks, preprocessing choices such as background removal, region cropping, or body-part isolation can substantially alter the visual cues available to a model [17]. These transformations may not only affect recognition accuracy, but also fundamentally change the semantic meaning of learned explanations. Nevertheless, the interaction between preprocessing strategies and intrinsic interpretability mechanisms has received limited attention in prior work.

Motivated by these observations, this thesis aims to investigate how different **intrinsically interpretable image-level models** behave when extended to video-based multi-class classification. Rather than proposing new architectures, we focus on analyzing the *applicability, stability, and explanation characteristics* of prototype-based and patch-based reasoning under unified experimental settings. By systematically examining temporal aggregation, preprocessing configurations, and task formulations, this work seeks to clarify the strengths and limitations of existing interpretable models in dynamic video environments.

1.5 Objectives

Based on the motivations described above, the objectives of this thesis are summarized as follows:

- To investigate how image-level interpretable models behave when applied to video-based facial expression recognition through frame-wise inference and temporal aggregation.
- To compare prototype-based and patch-based interpretability mechanisms under a unified experimental framework.

- To analyze the influence of different preprocessing configurations on recognition performance and explanation behavior.
- To qualitatively examine the stability and semantic relevance of explanations in video-based scenarios.

1.6 Contributions

The main contributions of this thesis are summarized as follows:

- We developed and validated a unified experimental framework that adapts image-level interpretable models to video-based classification. Unlike video-native approaches, this framework decouples spatial reasoning from temporal aggregation, enabling the systematic isolation and analysis of how frame-level explanations (prototypes vs. patches) influence video-level decisions.
- We conduct a comprehensive study on video-based facial emotion classification using the **RAVDESS** dataset [20] as a primary instantiation. Multiple label settings and preprocessing configurations are considered to analyze performance and interpretability.
- We perform a direct comparison between two interpretable deep learning approaches, namely ProtoPNet [6] and PatchSAE [18], under a unified experimental protocol.
- Standard convolutional neural network models (ResNet [12] and VGG16 [33]) are included to serve as black-box baselines amenable to post-hoc analysis. This setup allows us to go beyond simple accuracy comparisons; specifically, we conduct a qualitative comparison between post-hoc explanations (derived from these baselines) and our inherently interpretable mechanisms, highlighting critical differences in temporal consistency and faithfulness.
- We qualitatively analyze explanation behaviors through visualization, and propose a questionnaire-based human study to complement the model-based analysis.
- To examine the generalizability of the proposed framework, supplementary experiments on additional video datasets, including DIEM-A [7] and WLASL [16], are conducted.

1.7 Thesis Outline

The remainder of this thesis is organized as follows:

- **Chapter 2** reviews related work on facial expression recognition, video classification datasets, and interpretable deep learning.
- **Chapter 3** introduces the proposed framework, including the image-level reasoner and video-level aggregation mechanisms.
- **Chapter 4** describes the datasets (RAVDESS, DIEM-A, WLASL), preprocessing configurations, and experimental settings.
- **Chapter 5** presents quantitative performance results and qualitative interpretability analysis, including the error analysis of confusion matrices and human study design.
- **Chapter 6** concludes the thesis and discusses future directions.

Chapter 2

Related Work

2.1 Facial Expression Recognition (FER)

Facial Expression Recognition (FER) has evolved significantly from examining static images to analyzing dynamic video sequences. **The fundamental task in FER** is to categorize facial movements into discrete emotional states. The most widely adopted taxonomy in this field is **Ekman's six basic emotions** [10]: anger, disgust, fear, happiness, sadness, and surprise, often supplemented by a 'Neutral' state. In video-based FER tasks, the objective is typically to classify a temporal sequence of frames into one of these standardized categories, requiring the model to capture the evolution of facial muscle movements over time. Early approaches primarily relied on handcrafted features extracted from static images. However, with the widespread adoption of deep learning, Convolutional Neural Networks (CNNs) such as VGG [33] and ResNet [12] have become the dominant backbones for visual feature extraction. Recent surveys [17] highlight that state-of-the-art FER methods now focus on capturing subtle facial deformations and addressing occlusions or pose variations, as demonstrated by advanced architectures, such as Poster++ [22], exemplify this trend through the use of multi-stage and region-aware processing pipelines.

Specifically, as illustrated in Figure 2.1 Poster++ introduces a multi-stage architecture that progressively integrates facial landmark information and image features at different levels. By employing landmark-guided feature extraction and window-based cross-attention mechanisms, the model explicitly focuses on localized facial regions, enabling fine-grained modeling of facial muscle movements under challenging conditions.

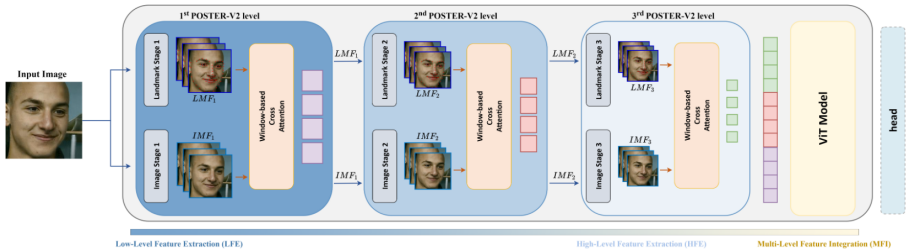


Figure 2.1: Overview of the Poster++ architecture. The model adopts a multi-stage, landmark-guided feature extraction framework with cross-attention mechanisms to progressively capture fine-grained facial deformations under occlusion and pose variation. Figure adapted from [22].

While image-based FER has achieved high accuracy, real-world expressions are inherently dynamic. Video-based FER requires modeling temporal evolution, which introduces complexities related to motion blur and temporal consistency. To address this, researchers have adapted image-based models to the video domain, often employing aggregation strategies or specialized spatio-temporal architectures.

2.2 Video Understanding and Datasets

2.2.1 Deep Learning for Video Classification

Video classification extends image recognition by incorporating the time dimension. Early works utilized 3D Convolutional Networks (C3D) [34] to jointly learn spatio-temporal features. Another influential paradigm is the Two-Stream network [32], which processes spatial (RGB) and temporal (Optical Flow) information separately. More recently, the success of Transformers in NLP has inspired Video Transformers, such as ViViT [1] and VideoMAE V2 [36], which treat video frames as sequences of tokens. Comprehensive surveys [29] indicate that while these models achieve superior performance, their decision-making processes remain opaque, necessitating the development of explainable video analysis methods [28].

2.2.2 Datasets for Human-Centric Video Analysis

Data plays a pivotal role in training and evaluating FER models. To ensure the robustness and generality of the proposed framework, this thesis utilizes three distinct video datasets, ranging from acted emotions to sign language.

RAVDESS [20]. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is the primary dataset used in this study. It contains dynamic, multimodal recordings of 24 professional actors (12 male, 12 female) from North America. The dataset is notable for its controlled environment and high-quality video, where actors perform validated expressions (e.g., calm, happy, sad, angry) with no significant background noise. This controlled setting allows for precise analysis of how interpretable models capture facial dynamics without environmental interference.

DIEM-A [7]. To test cross-cultural generalization, we employ the Diverse Intercultural E-Motion Database of Asian Performers (DIEM-A). Introduced recently, this dataset features Asian performers and emphasizes diverse emotional body movements and facial expressions. Unlike Western-centric datasets, DIEM-A provides a crucial benchmark for evaluating whether interpretability mechanisms (such as prototypes) are robust across different demographic groups.

WLASL [16]. To demonstrate the framework's applicability beyond facial expressions, we include the Word-Level American Sign Language (WLASL) dataset. This is a large-scale video dataset containing over 2,000 words performed by native signers. Video-based sign language recognition shares core challenges with FER, specifically the need to recognize fine-grained motion patterns (hand shapes vs. facial muscle movements). Using WLASL allows us to verify if the model can learn semantic parts (e.g., hands) similarly to how it learns facial organs.

2.3 Explainable AI (XAI)

The "black box" nature of deep learning has raised concerns regarding trust and reliability, especially in high-stakes decisions [26]. Explainable AI aims to make these models transparent.

2.3.1 Post-hoc Explanations

Post-hoc methods attempt to explain a model after it has been trained. Saliency maps [31] and gradient-based techniques like Grad-CAM [30] identify influential pixels in an image. While widely used, these methods are often criticized for verifying preconceived notions rather than revealing the model's true logic. Furthermore, in video settings, frame-by-frame post-hoc explanations can be temporally unstable, flickering between frames even when the

semantic content remains unchanged.This phenomenon is illustrated in Figure 2.2, where Grad-CAM explanations are applied to consecutive frames from the RAVDESS dataset.

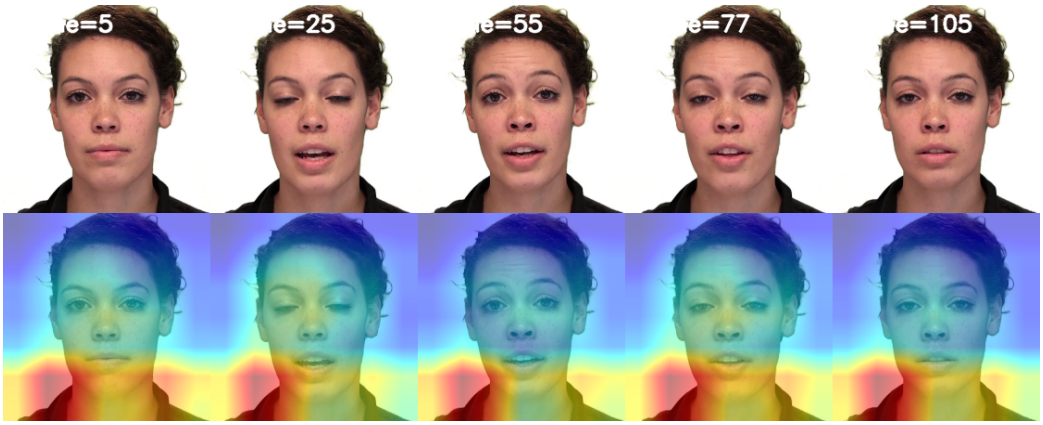


Figure 2.2: Temporal instability of frame-wise post-hoc explanations. Top row shows original frames sampled from a RAVDESS video. Bottom row shows corresponding Grad-CAM visualizations generated independently for each frame. Although the facial appearance remains largely unchanged, the highlighted regions vary noticeably across frames, demonstrating saliency flickering in video settings.

2.3.2 Evaluation of Interpretability

Evaluating explanations is challenging. Tools like Quantus [13] provide metrics to assess the faithfulness and robustness of explanations. However, quantitative metrics alone often fail to capture human-centric interpretability, motivating the need for inherent interpretability designs.

2.4 Inherently Interpretable Architectures

This thesis focuses on models that are interpretable by design, specifically Prototype-based and Patch-based approaches.

2.4.1 Prototype-based Learning (ProtoPNet)

ProtoPNet [6] introduces a case-based reasoning layer into deep neural networks. It learns a set of prototypes in the latent space, where each prototype represents a prototypical part of a class (e.g., a "smiling mouth"). Classifications are made based on the similarity (L2 distance) between the input

image patches and these learned prototypes. This embodies the reasoning process of "This looks like that" [6]. Extensions such as Neural Prototype Trees [24] further structure these prototypes hierarchically. In this thesis, we investigate how ProtoPNet's part-based reasoning adapts to video sequences.

2.4.2 Sparse Autoencoders and Patch-based Methods

An alternative to prototype matching is the use of Sparse Autoencoders (SAE)s to learn a dictionary of visual concepts [18]. These methods decompose an input into a sparse combination of basis functions (atoms). Models like Patch-SAE leverage this principle to identify salient local patches without requiring class-specific prototypes. While less semantically structured than ProtoPNet, sparse representations capture fine-grained texture details [3]. Comparing these two paradigms—explicit prototypes versus sparse dictionary atoms—in the video domain constitutes a key contribution of this work.

Extension to Video and Intrinsic Nature. A key question in applying these image-level models to video is how they fit into the temporal timeline. In this thesis, we adopt a frame-wise evidence aggregation approach. Unlike post-hoc methods that attempt to rationalise a decision after it is made by a black box, our framework uses the interpretable scores (derived from prototype or patch similarities) directly to compute the final video-level prediction. Since the decision logic is mathematically constructed from these explicit local explanations across the timeline, the framework remains inherently interpretable rather than post-hoc. This ensures that the "timeline" of the video is not just a sequence of images, but a verifiable trail of evidence leading to the classification.

Chapter 3

Proposed Model

3.1 Overview of the Proposed Framework

This thesis adopts a unified interpretable framework for video-based facial expression recognition. The overall pipeline consists of four main components: (1) frame-level preprocessing, (2) feature extraction using a backbone network, (3) image-level interpretable reasoning (incorporating both prototype-based and patch-based mechanisms), and (4) temporal aggregation for video-level prediction.

Given an input video sequence, preprocessing is first applied to each frame to normalize facial appearance and extract relevant regions. Figure 3.1 illustrates the overall workflow of the proposed method. The processed frames are then passed through the backbone and the image-level reasoner (ProtoPNet or PatchSAE) to generate transparent frame-wise explanations and class-specific prediction scores. Finally, these frame-level outputs are synthesized via the temporal aggregation module to form the final video-level decision. Different aggregation strategies are investigated to analyze their impact on recognition performance and interpretability.

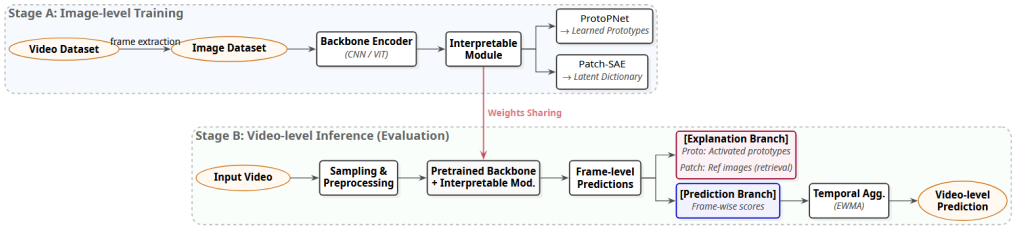


Figure 3.1: Overall workflow of the proposed video-based emotion recognition framework. Given an input video, frames are sampled and preprocessed before being encoded by a backbone network. Interpretable representations are then extracted at the frame level and aggregated temporally to produce a video-level prediction.

3.2 Image-Level Explanation Module Architecture

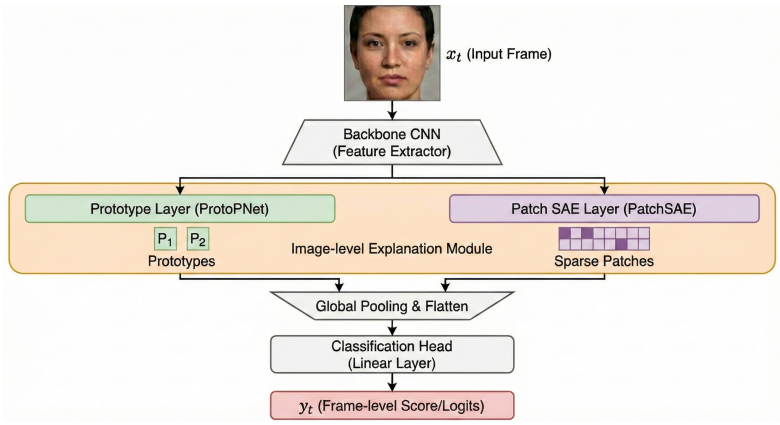


Figure 3.2: Overview of the image-level explanation module. Two interpretable instantiations are considered: ProtoPNet and PatchSAE.

This section introduces the image-level explanation module used in this study. The module receives frame-level feature representations from a shared backbone network and produces interpretable frame-level predictions. **As illustrated in Figure 3.2, the workflow proceeds in two steps:** First, the input frame is processed by the CNN backbone to extract a spatial feature map, that preserves local spatial structure and encodes high-level semantic information of facial regions. Second, this feature map is passed to the interpretable reasoner—either performing similarity matching against learned

prototypes (ProtoPNet) or sparse reconstruction via dictionary atoms (Patch-SAE). The module finally outputs both the classification logits and a visualizable activation map that highlights the relevant facial regions used in the decision.

3.3 Backbone Network

The backbone network is responsible for extracting discriminative feature representations from input images. Standard CNN architectures, specifically VGG16 and ResNet, are employed as the backbone to balance recognition performance and interpretability.

The backbone consists of multiple convolutional layers followed by non-linear activation and pooling operations. It produces a spatial feature map that preserves local facial information, which is essential for prototype-based comparison. Compared to very deep architectures, a simpler backbone allows clearer interpretation of prototype activation patterns.

In addition to the prototype-based model, the same backbone architecture is also used in baseline models for fair comparison.

3.4 Preprocessing Strategies

Preprocessing plays a critical role in facial expression recognition and directly influences both performance and interpretability. In this thesis, multiple preprocessing strategies are considered to analyze their impact on prototype-based models.

The preprocessing pipeline includes face detection and cropping, resizing, and normalization. In addition to using full-face images, region-based preprocessing is also investigated. Specifically, facial regions such as the eyes and mouth are extracted and combined to emphasize expression-related features.

For video-based data, preprocessing is applied independently to each frame. Frame sampling strategies are used to reduce redundancy while preserving expression dynamics. By comparing different preprocessing configurations, this study aims to clarify how input representations affect prototype learning and activation behavior.

3.5 Interpretable Mechanisms

Standard convolutional features extracted by the backbone are high-dimensional and abstract, making them difficult for humans to interpret directly. To

bridge this semantic gap, explicit calculation mechanisms are necessary to map these abstract representations into human-understandable concepts. Without this explicit transformation, the model decision process remains a black box. Therefore, this section details the mathematical formulation of two distinct mechanisms—prototype-based matching and patch-based sparse coding—that explicitly convert latent features into verifiable visual evidence.

3.5.1 Prototype-based Reasoning (ProtoPNet)

The ProtoPNet module explains predictions by comparing the extracted feature map with a set of learned class-specific prototypes [6]. Each prototype represents a prototypical local pattern in the feature space. During inference, the model computes similarity scores between input features and prototypes, and aggregates these scores to produce class-specific prediction scores together with prototype-based explanations. Let $\mathbf{z} \in R^{H \times W \times D}$ denote the spatial feature map extracted by the backbone network. Let $\mathbf{P} = \{\mathbf{p}_k\}_{k=1}^{K}$ denote the set of K learnable prototypes, where $\mathbf{p}_k \in R^D$.

The similarity between a local feature patch $\mathbf{z}_{i,j}$ and the k-th prototype is computed using the negative squared Euclidean distance:

$$s_k(i,j) = -\mathbf{z}_{i,j} - \mathbf{p}_k^2. \tag{3.1}$$

The global activation score for prototype k is obtained by max-pooling over the spatial dimensions:

$$S_k = \max_{i,j} s_k(i,j). \tag{3.2}$$

These scores represent how strongly a specific prototype is present in the input image. The final logits are computed via a fully connected layer connected to these similarity scores.

3.5.2 Patch-based Sparse Reasoning (PatchSAE)

In contrast to fixed prototypes, PatchSAE employs a sparse dictionary learning approach. It consists of a dictionary of atoms (reference patches) and an autoencoder that enforces sparsity in the latent space.

Given the feature map $\mathbf{z}$, PatchSAE aims to reconstruct it using a linear combination of atoms from a learned dictionary $\mathbf{D}_{dict}$. The encoding process involves projecting the features into a latent space $\mathbf{h}$ subject to a sparsity constraint (e.g., L_1 regularization or Top-k activation). The learned dictionary atoms capture recurring visual primitives in the feature space, such as localized textures or facial components [18]. During inference, the activation strengths in the sparse code $\mathbf{h}$ serve as interpretable signals that indicate

which atoms are responsible for representing the input. These atom activations can be aggregated to compute class-specific prediction scores, while simultaneously providing patch-level explanations.

$$\mathbf{h} = Encoder(\mathbf{z}), \quad s.t. \quad \|\mathbf{h}\|_0 \leq k \tag{3.3}$$

Unlike ProtoPNet, where prototypes are class-specific, PatchSAE's dictionary atoms capture generic local textures (as analyzed in Chapter 5). The interpretability arises from retrieving the most activated dictionary atoms and visualizing the corresponding image patches from the training set.

3.6 Training Procedure

Since the two interpretable mechanisms utilize different learning paradigms, their training procedures differ accordingly.

3.6.1 Training of Prototype-based Models

The prototype-based model (ProtoPNet) is trained in multiple stages to ensure semantically meaningful prototypes. First, the backbone network and prototype layer are jointly optimized using a classification loss (typically Cross-Entropy) combined with clustering and separation losses to structure the latent space.

After the initial training stage, a projection (push) operation is applied. The push operation updates each prototype by projecting it onto the nearest feature patch from the training data that belongs to the same class:

$$\mathbf{p}_k \leftarrow \arg\min_{\mathbf{z} \in \mathcal{Z}_{class}} \|\mathbf{z} - \mathbf{p}_k\|_2 \tag{3.4}$$

This procedure enforces prototypes to correspond to actual training samples, ensuring that explanations are grounded in real visual data. Unless otherwise specified, we adopt push-based models as the standard configuration.

3.6.2 Training of Patch-based Models

In contrast, the PatchSAE model is trained essentially as an autoencoder, focusing on faithful reconstruction with sparsity constraints. The training objective $\mathcal{L}_{patch}$ consists of a reconstruction loss and a sparsity penalty:

$$\mathcal{L}_{patch} = \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 + \lambda \|\mathbf{h}\|_1 \tag{3.5}$$

where $\mathbf{z}$ is the input feature map, $\hat{\mathbf{z}}$ is the reconstructed feature map derived from the dictionary atoms, and λ controls the degree of sparsity. This optimization forces the model to learn a compact set of "atomic" visual patterns (edges, textures) in the dictionary $\mathbf{D}_{dict}$, rather than class-specific prototypes. Once trained, the dictionary is fixed, and the sparse activations $\mathbf{h}$ serve as the interpretable representation for classification.

3.7 Video-level Aggregation Strategy

For video-based facial expression recognition, relying solely on independent frame-level predictions often leads to temporal inconsistency due to motion blur, occlusion, or frame-level noise. To address this, we employ **Exponentially Weighted Moving Average (EWMA)** as the primary temporal aggregation mechanism. Unlike simple averaging which treats all frames equally, EWMA applies a recursive smoothing filter that enforces temporal continuity while retaining sensitivity to recent changes in facial dynamics.

Figure 3.3 provides a schematic overview of this process. The system takes a sequence of raw frame-level prediction scores (depicted as the noisy input signal) and processes them sequentially. Instead of treating frames in isolation, the EWMA filter recursively updates its internal state, producing a smoothed score trajectory that suppresses high-frequency flickering. Finally, the state at the last time step is extracted to serve as the consensus prediction for the entire video.

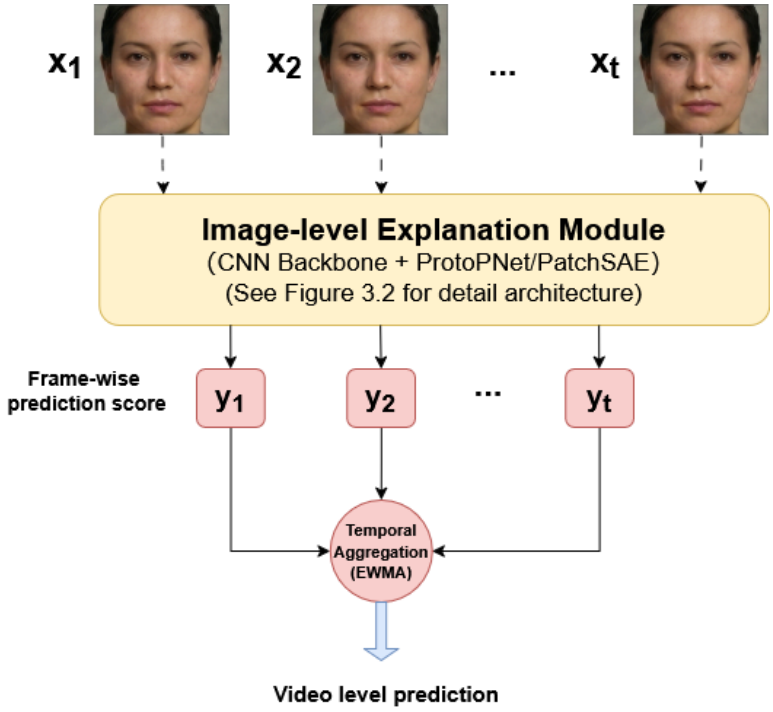


Figure 3.3: Overview of the video-level aggregation strategy. Raw frame-level predictions are first smoothed using EWMA to ensure temporal consistency, followed by pooling to generate the final decision.

3.7.1 Frame-level Prediction Sequence

Let $V = \{I_1, I_2, \ldots, I_T\}$ be a video sequence consisting of T frames. The backbone network first processes each frame I_t independently to produce a raw prediction score vector (logits or probabilities) $\mathbf{y}_t \in R^C$, where C is the number of emotion classes. This results in a raw score sequence $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$.

3.7.2 Exponentially Weighted Moving Average (EWMA)

To mitigate prediction fluctuation between consecutive frames, we apply EWMA to the raw score sequence. Let $\tilde{\mathbf{y}}_t$ denote the smoothed prediction vector at time step t. The recursive update rule is defined as:

$$\tilde{\mathbf{y}}_t = \alpha \cdot \mathbf{y}_t + (1 - \alpha) \cdot \tilde{\mathbf{y}}_{t-1} \tag{3.6}$$

where:

- $\mathbf{y}_t$ is the raw output from the model at frame t.

- $\tilde{\mathbf{y}}_{t-1}$ is the smoothed output from the previous time step (with $\tilde{\mathbf{y}}_0$ initialized as $\mathbf{y}_1$ or a zero vector).
- $\alpha \in [0, 1]$ is the smoothing factor (or decay rate).

Role of α. The parameter α controls the degree of temporal smoothing. A smaller α gives more weight to historical information, resulting in smoother trajectories but higher latency. A larger α makes the model more responsive to the current frame but more susceptible to noise. In this thesis, EWMA serves as a low-pass filter, effectively suppressing high-frequency noise (e.g., prediction flickering) caused by minor facial variations.

3.7.3 Final Video-level Decision

Since the EWMA update rule recursively accumulates historical information, the smoothed prediction at the final time step T effectively represents a weighted summary of the entire video sequence. Therefore, we define the video-level prediction $\mathbf{Y}_{video}$ as the smoothed score of the last frame:

$$\mathbf{Y}_{video} = \tilde{\mathbf{y}}_T \tag{3.7}$$

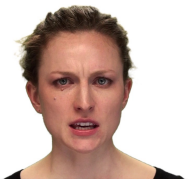
where $\tilde{\mathbf{y}}_T$ denotes the **final smoothed prediction vector** obtained from the EWMA recursion at the last frame index T. The tilde symbol ($\sim$) signifies that this vector is not the raw output of the backbone network, but rather the result of the temporal smoothing process defined in Eq. (3.6). By utilizing the final state of the EWMA sequence, the model captures the cumulative emotional evolution, giving higher importance to the most recent evidence while maintaining robustness against earlier noise.

Chapter 4

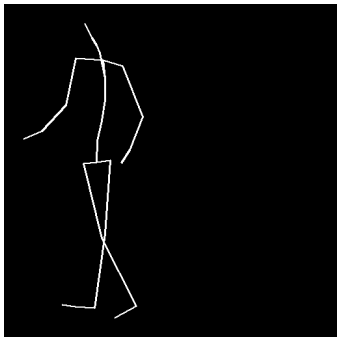
Experimentation

4.1 Datasets

To evaluate the effectiveness and robustness of the proposed interpretable framework, we utilize three distinct video datasets: RAVDESS (primary), DIEM-A (cross-cultural generalization), and WLASL (cross-domain generalization). These datasets differ significantly in terms of recording environment, video quality, and subject diversity. Representative visual samples from these datasets are illustrated in Figure 4.1, highlighting the domain shift from dense RGB textures to sparse skeletal structures.



(a) RAVDESS



(b) DIEM-A

(c) WLASL

Figure 4.1: Visual comparison of the datasets. **(a) RAVDESS:** High-quality RGB studio recording. **(b) DIEM-A:** Sparse skeletal representation derived from motion capture. **(c) WLASL:** "In-the-wild" footage with varying backgrounds and lighting conditions.

4.1.1 RAVDESS (Primary Dataset)

The **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** [20] serves as the primary benchmark for this study due to its high-quality annotations and controlled environment.

Data Quantity. The dataset contains recordings from **24 professional actors** (12 male, 12 female). For the Speech portion used in this thesis, there are a total of **1,440 video files**. Each actor performs 60 distinct trials, covering 8 emotional categories: *neutral, calm, happy, sad, angry, fearful, disgust, and surprised*. Each video clip has an average duration of approximately 3 to 5 seconds, capturing the full onset, apex, and offset of the expression.

Video Quality. RAVDESS is characterized by **high-fidelity studio quality**. Videos are recorded at **720p resolution (1280×720)** with a standard frame rate of 30 fps. The recording environment features professional three-point lighting and a consistent neutral background (green screen or plain wall), ensuring no background clutter. This high signal-to-noise ratio makes it an ideal testbed for analyzing whether the model correctly focuses on facial muscle movements rather than environmental artifacts.

4.1.2 DIEM-A (Supplementary Dataset)

To assess the model's robustness on non-Western subjects, we employ the **Diverse Intercultural E-Motion Database of Asian Performers (DIEM-A)** [7].

Data Quantity. This dataset includes **60 Asian performers** from diverse ethnic backgrounds and comprises approximately **2,200 motion sequences** labeled with basic emotions similar to Ekman's categories. Unlike RAVDESS, the actors perform both posed and spontaneous expressions, providing a larger variance in expression intensity.

Video Quality. Although DIEM-A does not consist of camera-captured RGB videos, the motion capture data represent **time-continuous human body animations**. These animations can be rendered and temporally sampled into **frame-based skeletal visualizations** (stick figures), allowing them to be treated as *video-like temporal sequences* for frame-wise analysis. This representation provides fine-grained temporal continuity and enables

evaluation of temporal aggregation mechanisms independently of visual appearance, texture, or background cues.

4.1.3 WLASL (Supplementary Dataset)

To verify if the proposed framework can generalize beyond facial expressions to body/hand motion, we utilize the **Word-Level American Sign Language (WLASL)** dataset [16].

Data Quantity. WLASL is a large-scale dataset containing over **2,000 distinct sign language words** performed by more than 100 signers. For our experiments, we select a subset of 10 common classes (e.g., 'book', 'drink', 'help') to maintain comparability with the classification complexity of FER tasks, totaling approximately **500 video samples**.

Video Quality ("In-the-Wild"). Unlike the studio-grade RAVDESS and DIEM-A, WLASL consists of videos aggregated from YouTube and other web sources. Therefore, the video quality is **highly variable**: resolutions range from low-quality (360p) to high-definition (720p+), and frame rates vary between 24 and 30 fps. Crucially, these videos feature **uncontrolled environments**, including cluttered backgrounds, varying lighting conditions, and partial occlusions. This serves as a robustness stress test for our preprocessing and attention mechanisms.

4.2 Label Settings and Task Formulation

To investigate how task granularity and label ambiguity affect interpretability, we formulate the facial expression recognition task under different label settings. While the image-level datasets were provided with pre-defined labels, we constructed the video-level tasks by mapping the original RAVDESS categories into simplified groupings.

The RAVDESS dataset originally contains 8 classes. In this thesis, we define our experimental scope by selecting a subset of 6 categories: Neutral, Calm, Happy, Sad, Angry, and Fearful. 'Surprise' and 'Disgust' are excluded to streamline the classification task and focus the analysis on these primary emotional states. Furthermore, for the simplified 3-class settings, we also exclude "Fearful" to remove excessive ambiguity, resulting in a core subset.

The specific label mappings are defined as follows:

Original (6-class). This setting uses the standard emotion categories: Neutral, Calm, Happy, Sad, Angry, and Fearful. This serves as the fine-grained baseline.

V1 (3-class): Semantic Valence Grouping (Clean). This setting represents a standard, semantically clean classification task based on valence (Positive vs. Negative). The boundaries between classes are distinct.

- **Positive:** Happy
- **Neutral:** Neutral, Calm
- **Negative:** Sad, Angry

V2 (3-class): Visual Interference Grouping (Robustness Test). This setting is designed to test the model's robustness to **visual ambiguity** and "semantic interference". Unlike V1, we intentionally introduce the "Sad" category into the **Neutral** group.

- **Positive:** Happy
- **Neutral (Expanded):** Neutral, Calm, **Sad** (Interference added)
- **Negative:** Angry

Rationale for V2: While 'Sad' is semantically negative, it shares high visual similarity with standard Neutral expressions (low facial muscle activation). By grouping them into the **Neutral** class, we create a "Hard" category containing semantic conflict. This setting evaluates whether the interpretable model relies on *high-level semantic labels* (which would separate Sad from Neutral) or *low-level visual intensity* (which would correctly group them together as static faces).

4.3 Preprocessing Configurations

Table 4.1: Summary of label settings and task formulations. This table consolidates the mapping strategies and the rationale behind the V1 (Valence) and V2 (Visual Interference) configurations.

Setting	**Target Class**	**Source Emotions**	**Rationale & Hypothesis**
Original (6-class)	(6 Distinct Classes)	Neutral, Calm, Happy, Sad, Angry, Fearful	**Baseline:** Standard fine-grained emotion recognition.
V1 (3-class)	Positive	{Happy}	**Semantic Valence Grouping:** Groups emotions by psychological meaning (Good vs. Bad).
	Neutral	{Neutral, Calm}	
	Negative	{Sad, Angry}	*Challenge:* Combines visually distinct emotions (Static Sad vs. Dynamic Angry).
V2 (3-class)	Positive	{Happy}	**Visual Intensity (Arousal) Grouping:**
	Neutral*	{Neutral, Calm, Sad}	**Robustness Test:** Neutral expanded by low visual intensity.
	Negative	{Angry}	*Hypothesis:* Grouping driven by visual motion rather than semantic valence.

Data preprocessing plays a pivotal role in determining the features available to the model. To systematically analyze the impact of background noise and isolate the contribution of specific facial regions, we design three distinct preprocessing configurations. All input frames are resized to a fixed resolution of 224×224 pixels and normalized using standard ImageNet mean and standard deviation values.

Figure 4.2 illustrates visual examples of these three configurations.

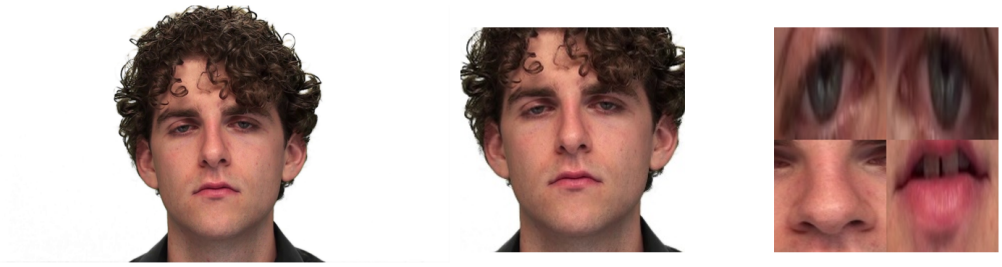


Figure 4.2: Examples of different preprocessing configurations used in this study. **Left:** Original frames retaining background context. **Middle:** Face-only frames with background masked, forcing the model to focus on facial features. **Right:** Reassembled facial representations, constructed by stitching extracted eye and mouth regions onto a canvas to test part-based recognition.

1. Original Frames (Baseline). As shown in the **Left panel** of Figure 4.2, the raw video frames are used without explicit region filtering. The entire image is resized to the target resolution using standard interpolation. This configuration serves as a baseline to evaluate whether the model relies on non-facial cues (e.g., hair, clothing, or background wall patterns) for classification.

2. Face-only Frames (Primary Setting). Illustrated in the **Middle panel**, this setting minimizes environmental bias by employing a standard face extraction pipeline. First, a face detector identifies the facial landmarks to determine the boundaries of the face. Based on these coordinates, a tight "bounding box" is computed, and the image is "cropped" to this region. Although a minimal amount of the original uniform background remains at the periphery, this process effectively removes high-variance background factors (such as clothing and body posture). **This is the default configuration** for the quantitative experiments presented in Chapter 5, ensuring that the model's prototypes are driven primarily by facial muscle dynamics.

3. Reassembled Face (Qualitative Analysis). Finally, the **Right panel** shows the "Reassembled" representation, designed to force part-based reasoning. We utilized the "dlib" library [15] to detect 68-point facial landmarks. From these, four specific ROIs (left eye, right eye, nose, and mouth) are extracted and reorganized into a "2×2 grid layout" on a blank canvas. Specifically, the eyes are placed in the top row, while the nose and mouth are placed in the bottom row. This artificial setting is primarily used for qualitative analysis to verify whether the model can recognize expressions based purely on local semantic parts, completely independent of global face shape and skin texture.

Video Processing. For video-based experiments, these preprocessing steps are applied independently to each frame in the sequence. This frame-wise independence ensures that the subsequent temporal aggregation (Section 3.8) operates on consistent visual features.

4.4 Implementation Details

All models are implemented using the **PyTorch** deep learning framework. We build our models on top of standard PyTorch modules and utilities, with `torchvision` used for backbone architectures and image preprocessing. The

backbone network for all experiments is initialized with weights pre-trained on ImageNet [8] to ensure stable convergence.

4.4.1 Common Training Settings

We use Stochastic Gradient Descent (SGD) with a momentum of 0.9. The initial learning rate is set to 0.001 for the backbone and 0.003 for the explanation modules. A step-based learning rate scheduler is applied. The batch size is fixed at 32 for all experiments.

4.4.2 ProtoPNet Specifics

The ProtoPNet model is implemented using a VGG-16 with Batch Normalization backbone, pretrained on ImageNet. The input images are resized to 224×224. The prototype layer is configured to have a depth of 512, matching the output feature dimension of the backbone.

Prototype Allocation. To ensure fair comparison across different label settings, we allocate a fixed number of "10 prototypes per class". Consequently, the total number of prototypes varies with the task formulation: 60 prototypes are used for the Original (6-class) setting, and 30 prototypes are used for the V1/V2 (3-class) settings. Each prototype is a 1×1 patch in the latent feature space.

Loss Function and Optimization. The model is optimized using a weighted sum of four loss terms: Cross-Entropy Loss (L_{ce}), Cluster Cost (L_{clst}), Separation Cost (L_{sep}), and L_1 regularization. We adopt a multi-stage training strategy:

1. **Warm-up Phase (5 epochs):** Only the add-on layers and prototype vectors are trained (LR $= 3 \times 10^{-3}$), while the backbone is frozen.

2. **Joint Training Phase:** The backbone (LR $= 1 \times 10^{-4}$), add-on layers, and prototypes are trained jointly. The learning rate decays every 5 epochs.

3. **Prototype Projection (Push):** Starting from epoch 30, a projection operation is performed every 10 epochs. This step replaces each prototype with the nearest latent training patch to ensure interpretability.

The detailed hyperparameters are summarized in Table 4.2.

Table 4.2: Hyperparameter settings for ProtoPNet training.

Parameter	**Value**
Architecture & Training	
Backbone Network	VGG-16 (Batch Norm)
Input Resolution	224×224
Batch Size	16
Total Epochs	100
Warm-up Epochs	5
Push Start Epoch	30
Push Frequency	Every 10 epochs
Learning Rates (Joint)	
Backbone Features	1×10^{-4}
Prototype Vectors	3×10^{-3}
Add-on Layers	3×10^{-3}
Step Size (LR Decay)	5 epochs
Loss Coefficients	
Cross-Entropy (L_{ce})	1.0
Cluster Cost (L_{clst})	0.8
Separation Cost (L_{sep})	-0.08
L_1 Regularization	1×10^{-4}

4.4.3 PatchSAE Specifics

For the PatchSAE model, we employ a Sparse Autoencoder architecture to learn a dictionary of local visual primitives. The implementation is based on a standard encoder-decoder structure with sparsity constraints.

Architecture and Dictionary. We set the dictionary size to 512 atoms. The sparsity is enforced via an L_1 regularization penalty on the latent activations. This encourages the model to reconstruct the input feature maps using only a small subset of active dictionary atoms, thereby isolating distinct visual textures.

Training Strategy. The model is trained to minimize a composite loss function, balancing feature reconstruction (MSE) and classification performance. We employ a "Constant with Warmup" learning rate scheduler to stabilize the early training phase. The key hyperparameters used in our experiments are summarized in Table 4.3.

Table 4.3: Hyperparameter settings for PatchSAE training.

Hyperparameter	**Value**
Learning Rate (LR)	4×10^{-4}
L_1 Coefficient (Sparsity Penalty)	8×10^{-5}
LR Scheduler	Constant with Warmup
Warm-up Steps	500
Batch Size	16
MSE/CLS Coefficient	1.0
Total Training Tokens	20,000

Chapter 5

Evaluation

5.1 Quantitative Analysis on RAVDESS

In this section, we present the quantitative evaluation of the proposed framework on the RAVDESS dataset. All experiments follow the implementation protocols detailed in Chapter 4.

5.1.1 Task Definition Recap

To ensure clarity in interpreting the results, we briefly recapitulate the specific class compositions for the three label settings used in this analysis. The number of samples for each class under different settings is summarized in Table 5.1.

Table 5.1: Data statistics for each class under different task settings.

Setting	**Class**	**# Samples**
Original (6-class)	Neutral	743
	Calm	1619
	Happy	1490
	Sad	1580
	Angry	1472
	Fearful	1408
V1 (3-class)	Positive	2703
	Neutral	3701
	Negative	5487
V2 (3-class)	Positive	1490
	Neutral*	3942
	Negative	1472

5.1.2 Summary of Results

Table 5.2 summarizes the classification accuracy and F1-scores comparing our interpretable models (ProtoPNet, PatchSAE) against the black-box baselines (ResNet, VGG) across different settings using the face-only preprocessing configuration.

Table 5.2: Quantitative comparison of video-level classification accuracy (%) on the RAVDESS **test set**. The dataset was randomly partitioned into training (70%), validation (20%), and testing (10%) sets. Comparison between black-box baselines and interpretable models.

Model	**Original (6-class)**	**V1 (3-class Clean)**	**V2 (3-class Robust)**
Black-box Baselines			
ResNet-18	27.8	33.4	33.4
VGG-16	57.7	**76.6**	62.1
Interpretable Models			
ProtoPNet	42.8	64.8	**74.3**
PatchSAE	16.9	57.2	49.7

The results indicate that while VGG-16 achieves the highest accuracy in the standard V1 setting, our interpretable **ProtoPNet** remains highly competitive, especially in the challenging V2 setting (74.3%), significantly

outperforming the ResNet baseline. This demonstrates that incorporating interpretable prototypes does not necessarily lead to a drastic drop in performance, and in some cases (like V2) may offer better regularization against noise.

5.2 Image-level Recognition Performance on RAVDESS

To provide a controlled reference for video-based recognition, we first evaluate all models under an image-level setting. In this configuration, individual frames extracted from RAVDESS videos are treated as independent samples, without modeling temporal relationships. This design ensures that any performance gap observed later can be attributed to the challenges of temporal aggregation rather than spatial feature extraction.

Table 5.3: Image-level classification accuracy on the RAVDESS dataset under different label settings.

Model	Original (6-class)	V1 (3-class)	V2 (3-class)
ProtoPNet	0.917	0.966	0.811
PatchSAE	0.792	0.887	0.906
ResNet	0.967	0.977	0.999
VGG16	0.991	0.997	0.995

From Table 5.3, several key observations can be made:

High Spatial Discriminability. Quantitatively, all models achieve high recognition accuracy at the image level. Standard CNN baselines reach near-saturated performance (97%+), indicating that the static facial appearance in RAVDESS provides sufficient discriminative cues. ProtoPNet also demonstrates strong performance, particularly in the V1 setting (96.6%), suggesting that the learned prototypes effectively capture the static semantics of facial expressions. **Crucially, beyond these numerical metrics, the model demonstrates the ability to localize fine-grained facial features.** As illustrated in **Figure 5.1**, the learned prototypes do not activate globally but instead focus on distinct, localized regions. The **Left column** shows the receptive fields (yellow boxes), while the **Right column** displays the corresponding heatmaps. The visualizations confirm that the model attends to specific semantic parts—ranging from the "forehead region" (Top row) and

"closed eyes" (Middle row) to the "furrowed brows" (Bottom row) rather than relying on background noise.

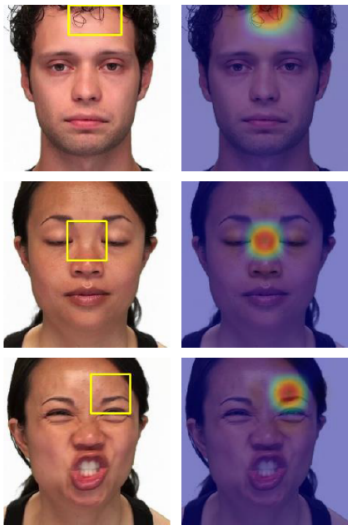


Figure 5.1: Visualization of the model's high spatial discriminability. **Left column:** Input images with the yellow bounding box indicating the receptive field of the most activated prototype. **Right column:** The corresponding upsampled activation heatmaps. The results demonstrate that the learned prototypes (e.g., furrowed brows in the bottom row or closed eyes in the middle row) are highly localized and semantically meaningful, focusing on specific facial muscles rather than background noise.

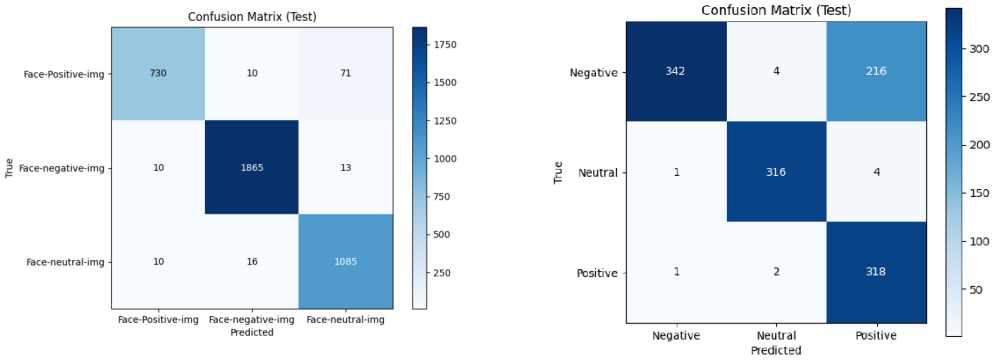
Confusion Matrix Analysis. To validate the discriminative mechanism, we compare the V1 and V2 settings in Figure 5.2.

- **Hypothesis Validation (Target Class):** Most importantly, the **Neutral class (Middle row)**, which contains the interfering 'Sad' samples, achieves near-perfect isolation with a **98.4% recall rate** (316/321). This is the core evidence confirming our hypothesis: the model successfully groups 'Sad' with 'Neutral' based on shared low-intensity features (Passive), ignoring the semantic conflict.

- **Arousal-based Confusion (The Reason for Lower Acc):** Although the overall accuracy drops to ∼81.0% compared to V1, this is driven entirely by the confusion between high-arousal classes. As seen in the top-right cell, a significant portion of **Negative** samples (216) are misclassified as **Positive**. This is an expected outcome of the V2 design: without the semantic anchor of 'Sad', the model struggles to distinguish between 'Angry' (Negative) and 'Happy' (Positive) because

both share **High Arousal** visual attributes (e.g., open mouth, widened eyes).

- **Conclusion:** This pattern confirms that the model's decision boundary is dominated by **visual intensity** rather than emotional valence.



(a) V1 Setting

(b) V2 Setting

Figure 5.2: Confusion Matrix Comparison. **(a) V1 Setting:** The model shows clean separation based on semantic valence. **(b) V2 Setting:** Crucially, the model successfully classifies the interference class (Sad) into the Neutral (Middle row), demonstrating its reliance on visual intensity (low arousal) rather than high-level emotion semantics.

5.3 Video-level Recognition Performance Comparison

This section analyzes the video-level recognition performance, where frame-level predictions are aggregated to form a final decision. The accuracy data corresponds to Table 5.2 presented in Section 5.1.

Overall Performance and Baseline Trends. Standard CNN-based models exhibit divergent behaviors in the video setting. VGG16 acts as a high-performance upper bound (0.766 in V1). **However, ResNet performs significantly worse (0.334 in V1).** We hypothesize that ResNet's deeper residual architecture makes it highly sensitive to frame-level noise (e.g., motion blur during expression onset) in this specific dataset. When aggregated

via simple EWMA without attention mechanisms, this frame-level instability degrades the final video-level score. In contrast, VGG's simpler features appear more robust to temporal fluctuation here.

Interpretable Models: ProtoPNet vs. PatchSAE. A consistent performance gap is observed between the two interpretable approaches: **ProtoPNet consistently outperforms PatchSAE** across all configurations (e.g., 0.648 vs 0.572 in V1). This suggests that the explicit prototype-matching mechanism—which enforces the learning of semantic facial organs like eyes and mouth—is more robust for video classification than PatchSAE's sparse dictionary learning, which relies on local texture primitives.

Error Analysis of Patch-SAE. To investigate *why* PatchSAE underperforms, we visualize its confusion matrix for the V1 setting in Figure 5.3.

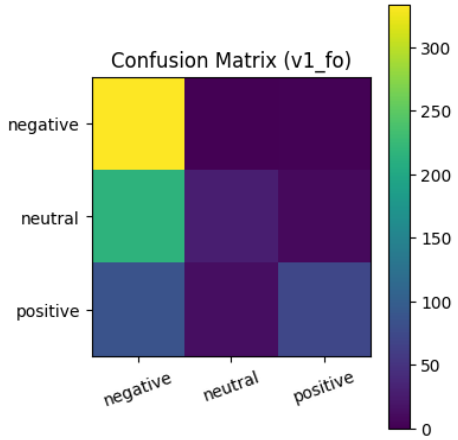


Figure 5.3: Confusion Matrix of Patch-SAE on Video-level classification (V1 setting). The model shows a strong bias towards classifying Neutral/Positive videos as Negative.

The confusion matrix reveals a significant **class bias** (Left column). The model frequently misclassifies *Neutral* and *Positive* samples as *Negative*. **We hypothesize that this is due to the inherent nature of sparse dictionary learning.** Unlike ProtoPNet which learns semantic object parts, PatchSAE focuses on **low-level local patterns (e.g., edges and textures)**. In dynamic videos, minor pixel variations or video noise in a neutral face might be misinterpreted by the model as significant texture features associated with negative expressions, leading to false positive predictions for the Negative class.

Impact of Preprocessing Strategies. Finally, to address the research question regarding the necessity of background removal, we compared the performance of models trained on *Original* frames (retaining background) versus the *Face-only* configuration.

Our experiments reveal a significant performance drop when background information is retained. For instance, ProtoPNet's accuracy in the V1 setting drops from **64.8% (Face-only)** to approximately **42% (Original frames)**. Qualitative inspection confirms that without masking, the model's prototypes often mistakenly latch onto high-contrast background edges or clothing patterns rather than facial muscles. **This quantitative gap justifies our decision to adopt the Face-only configuration as the standard setting for all main analyses presented in Table 5.2.**

5.4 Qualitative Explainability Comparison

To validate the interpretability advantages of our framework, we contrast the proposed models with standard post-hoc explanations. As widely reported in literature [26, 28], post-hoc saliency maps derived from baseline CNNs (e.g., ResNet) often exhibit significant noise and temporal flickering, failing to provide stable visual evidence. In contrast, our framework employs inherently interpretable mechanisms specifically **Patch-SAE** and **ProtoPNet** to generate structured and semantically consistent explanations. In the following subsections, we first present a visual analysis of the input processing dynamics across multiple samples, followed by a detailed examination of the internal reasoning mechanisms (Latents vs. Prototypes) on a representative case.

5.4.1 Visual Analysis of Temporal Dynamics

To systematically analyze how the model processes dynamic expressions, we visualize the input frame sequences. We primarily focus on the representative sequence **(ID: 02-02-01-01-02-01-15, labeled as 'Neutral')**, as requested, to illustrate the inference capability on low-intensity expressions. The frame-wise evolution of this sequence is visualized in the **top row** of Figure 5.4.

Robustness Validation (3 Videos per Class). To ensure the robustness of our observations and avoid cherry-picking, we extended this visual analysis to include **three randomly selected videos per emotion class**. Figure 5.4 presents representative samples from three distinct categories: Neu-

tral (Top), Positive (Middle), and Negative (Bottom). We consistently observed that the model successfully captures temporal intensity changes—detecting the high-arousal onset in 'positive' (wide smile) and the subtle mouth depression in 'negative', while maintaining stability for 'Neutral'. This confirms that the interpretable mechanisms are not limited to a single sample but generalize across different emotional intensities.

(a) Representative Sequence (ID: 02-02-01-01-02-01-15) - Class: Neutral

(b) Additional Analysis Sample - Class: Positive

(c) Additional Analysis Sample - Class: Negative

Figure 5.4: Visual analysis of input frame sequences across different emotional classes. **(a)** The specific sequence (Neutral) analyzed in the case study, showing subtle facial movements. **(b) & (c)** Additional sequences from 'Positive' and 'Negative' classes. Validating the analysis on these multiple samples confirms that the model consistently handles varying degrees of facial intensity.

5.4.2 Analysis of Patch-based Explanations (Patch-SAE)

Having validated the input processing dynamics across multiple classes, we now examine the internal decision mechanism of Patch-SAE on the primary representative sequence(ID: 02-02-01-01-02-01-15). Unlike prototype-based models that match inputs to fixed exemplars, Patch-SAE provides explanations through a sparse dictionary learning mechanism. Its decision process can be traced by analyzing the *latent activation profile* and the corresponding *reference images*.

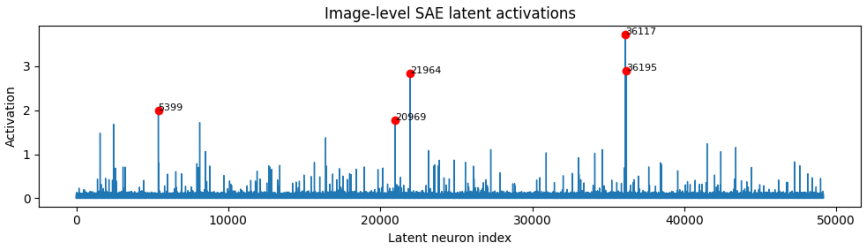


Figure 5.5: Image-level latent activation profile of Patch-SAE for a selected frame. The model exhibits sparsity, with only a few dominant neurons (e.g., Latent #21964) being highly activated.

Figure 5.5 visualizes the sparse latent activation profile for a representative frame. The x-axis corresponds to the index of latent neurons in the Patch-SAE dictionary, while the y-axis denotes their activation magnitude. Most latent units exhibit near-zero activation, reflecting the sparsity of the learned representation. The highlighted red dots indicate the most strongly activated latent neurons for this input. We observe that specific latent neurons, such as **Latent #21964**, show strong activations. To interpret the semantic meaning of this latent neuron, we retrieve the top-activating reference images from the training set, as shown in Figure 5.6.

Figure 5.6: Reference images retrieved for the dominant Latent #21964. The red bounding boxes consistently highlight the glabella (area between eyebrows) and forehead region across different subjects. This indicates that Patch-SAE utilizes this specific local texture feature as evidence for the 'Neutral' class.

The retrieved reference images in Figure 5.6 consistently focus on the

glabella and forehead regions. This suggests that Patch-SAE explains the prediction by identifying specific local texture primitives (e.g., the relaxed state of the forehead) that are statistically associated with the 'Neutral' expression. Crucially, this allows us to interpret Patch-SAE not just through activation maps, but through its **learned visual dictionary**. However, since Patch-SAE explanations are latent-driven and class-agnostic, the semantic meaning of each latent requires additional interpretation through reference retrieval. Here, *latent-driven* indicates that explanations are derived from the activation patterns of latent units learned by the Patch-SAE, rather than from class-specific prototypes or labels. Meanwhile, *class-agnostic* means that individual latent units are not explicitly associated with any particular emotion category, and the same latent feature may be activated across different classes. As a result, such explanations may be less immediately intuitive than prototype-based explanations, which directly associate exemplars with class labels. Furthermore, it is important to note that interpreting the exact physical meaning of these small retrieved patches (e.g., definitively identifying them as 'skin tension' versus mere illumination changes) solely from visual bounding boxes inherently involves subjectivity.

5.4.3 Analysis of Prototype-based Explanations (ProtoPNet)

For comparison, we examine the reasoning process of ProtoPNet on the same input. ProtoPNet explains predictions by matching the input frame with learned visual prototypes. Figure 5.7 displays the most frequently activated prototypes for this video.

(a) Prototype #17 (Eye region)

(b) Prototype #18 (Nose/Mid-face region)

Figure 5.7: ProtoPNet explanations using representative prototypes. (a) Prototype #17 identifies an eye-related pattern. (b) Prototype #18 identifies a nose/mid-face pattern. The heatmaps (right) show the precise location where the input matches the prototype.

As observed in Figure 5.7, ProtoPNet utilizes semantic object parts. **Prototype #17** focuses on the eye region, while **Prototype #18** highlights the nose and mid-face area. Unlike Patch-SAE, which explains predictions by activating recurring local texture patterns, ProtoPNet provides more object-centric explanations that are aligned with distinct facial organs. As shown in Figure 5.7, ProtoPNet explanations are localized on specific semantic facial parts, such as the eye region (Prototype #17) or the nose/mid-face region (Prototype #18), where the input closely matches a learned prototype. In contrast, Patch-SAE explanations in Figure 5.6 consistently highlight fine-grained texture cues on the glabella and forehead region across different subjects, without explicitly binding these cues to a predefined facial organ. This contrast illustrates that ProtoPNet relies on part-level exemplar matching, whereas Patch-SAE captures distributed texture-level evidence through its learned visual dictionary.

5.4.4 Discussion

The comparison reveals complementary strengths of the two interpretable approaches, specifically in their fundamental explanatory units:

- **Patch-SAE (Latent-based):** Explains decisions by activating specific **latent neurons** (e.g., Latent #21964). As shown in our analysis, these latents correspond to *fine-grained local textures* (like the forehead skin) and are interpreted by retrieving a dictionary of reference patches.
- **ProtoPNet (Prototype-based):** Explains decisions by matching the input to learned **visual prototypes** (e.g., Prototype #17). These prototypes correspond to *semantic object parts* (like eyes or mouth) and align well with human intuition of facial anatomy.

This analysis clarifies the distinction: while ProtoPNet performs *part-level matching*, Patch-SAE functions as a *dictionary of local primitives*. Both mechanisms provide transparency into the features driving the video-level decision, fulfilling the interpretability requirements for video analysis. These differences suggest that prototype-based explanations may be more suitable for human-facing video explanation, while patch-based explanations may better capture subtle appearance variations.

5.5 Supplementary Experiments on Generalization

To examine the performance of the models beyond the specific setting of RAVDESS (facial landmarks), we extended the evaluation to two additional datasets with different motion characteristics:

- **DIEM-A (Body Skeletons):** A body-movement-based emotion dataset represented by skeletal motion data. The original dataset contains 13 emotion categories. In this study, we focus on a subset of six classes (*anger, fear, joy, sadness, surprise, and disgust*) to maintain consistency with the Original (6-class) setting used in our image-based experiments and to simplify the experimental setup. The dataset is publicly available athttps://www.cr-ict.riec.tohoku.ac.jp/diem-a/.

- **WLASL-10 (Hand-Level Subset):** A subset of the World Level American Sign Language (WLASL) dataset consisting of the 10 most frequent gloss classes in the training split of the processed version used in this study. The selected classes are:{*accident, before, cool, cousin, drink, environment, go, shirt, trade, and who*}.Only classes that appear in the training, validation, and test splits are considered.The dataset is obtained from a publicly available processed version hosted on Kaggle:https://www.kaggle.com/datasets/risangbaskoro/wlasl-processed.

Table 5.4 presents the video-level recognition accuracy obtained from these experiments.

Internal Comparison. As shown in Table 5.4, we first observe the recognition performance of both interpretable models on non-facial domains.

- On **DIEM-A**, PatchSAE achieves an accuracy of 25.0%, compared to 21.0% for ProtoPNet.

- On **WLASL-10**, PatchSAE reaches an accuracy of 63.0%, showing a noticeable improvement over ProtoPNet (47.0%).

These results indicate that the proposed PatchSAE framework maintains a performance advantage over ProtoPNet when applied to sparse skeleton data, demonstrating better adaptability to non-facial structures.

Table 5.4: Comparison of recognition accuracy (%) on DIEM-A and WLASL-10. The table benchmarks our interpretable models against specialized State-of-the-Art (SOTA) black-box baselines. Note: The performance gap primarily stems from the trade-off between maximizing accuracy (black-box temporal modeling) and ensuring frame-level interpretability .

Method	**DIEM-A (Acc)**	**WLASL-10 (Acc)**
Ours (Frame-based Interpretable)		
ProtoPNet	21.0	47.0
PatchSAE	**25.0**	**63.0**
SOTA Baselines (Temporal & Black-box)		
ST-GCN (Skeleton) [37]	∼65.0	–
Pose-TGCN (Skeleton) [16]	–	85.3
I3D (RGB Video) [5]	–	94.0

Benchmarking against SOTA and Performance Analysis. However, when compared to specialized black-box baselines (SOTA), a significant performance gap is observed. We attribute this discrepancy to two fundamental structural limitations, which represent the necessary trade-off for achieving model transparency:

- **1. Lack of Learnable Temporal Dynamics:** SOTA methods such as ST-GCN [37] and I3D [5] explicitly learn the temporal evolution of actions using recurrent or 3D convolution modules. In contrast, our ProtoPNet and PatchSAE frameworks are strictly **frame-based**. Although we apply **Exponentially Weighted Moving Average (EWMA)** to smooth the predictions, the model itself lacks the capacity to capture temporal dependencies (i.e., how an action evolves over time). This limitation is evident in the facial domain (RAVDESS), where accuracy drops from **96.6% (Image-level)** to **∼65% (Video-level)**. The EWMA appears insufficient to fully compensate for the model's inability to distinguish complex dynamic transitions, leading to performance degradation in continuous video tasks.

- **2. Modality Mismatch (Vector vs. Pixel):** Crucially, standard CNNs are heavily biased towards **dense textures** rather than sparse shapes [11]. While CNNs excel at facial textures (RAVDESS), they struggle to extract features from the sparse, texture-less binary skeleton images used in DIEM-A, leading to lower efficiency compared to graph-based methods that process coordinates directly.

Why WLASL outperforms DIEM-A? Despite these limitations, PatchSAE achieves significantly higher accuracy on **WLASL-10 (63.0%)** compared to DIEM-A (21.0%). This highlights that sign language is largely driven by **static hand configurations** (Pose-dominant), which our model captures effectively. In contrast, DIEM-A relies on continuous trajectories (Motion-dominant), making frame-based classification inherently more challenging.

Chapter 6

Conclusion

6.1 Summary of Contributions

This thesis presented a systematic study on explainable video-based multi-class classification, with a primary focus on interpreting facial expression recognition (FER) models. The core contribution of this work is the development and validation of a **unified interpretable pipeline** that demonstrates how image-level intrinsic interpretability can be effectively extended to dynamic video sequences.

Specifically, we established a framework that bridges the gap between static prototype-based learning and temporal data aggregation. By evaluating two distinct paradigms—**ProtoPNet** (global-based) and **PatchSAE** (patch-based)—across three granularities of human movement (facial, body, and hand dynamics), this research proves that complex spatio-temporal tasks can be made transparent without sacrificing competitive recognition performance. This work serves as a foundational step toward more reliable and accountable human-centric AI systems.

6.2 Key Empirical Findings

Based on the experiments conducted on the RAVDESS dataset and supplementary benchmarks (DIEM-A and WLASL-10), several key conclusions are drawn:

- **Feasibility of Temporal-Intrinsic Interpretability:** The results confirm that image-level interpretable models can be successfully adapted to the video domain via frame-wise inference and temporal smoothing

(e.g., EWMA). This approach retains frame-level granularity in explanations while achieving video-level classification, providing a viable alternative to "black-box" 3D convolutional networks.

- **Trade-offs in Model Behavior and Adaptability:** Comparative analysis revealed distinct characteristics between the two frameworks. In the FER task (RAVDESS), **ProtoPNet** exhibited higher temporal stability and provided intuitive explanations based on prototypical parts, a concept formalized in interpretable deep learning [**?**], where the prediction is derived from local semantic similarities (e.g., focusing on eyes or mouth) rather than global holistic features. However, in non-facial domains such as body skeletons (DIEM-A) and hand trajectories (WLASL-10), **PatchSAE** demonstrated superior adaptability and higher recognition accuracy. This suggests that patch-based sparse representations may be more resilient to the diverse structural variations found in broader human-centric tasks.
- **Critical Role of Targeted Preprocessing:** Our experiments reinforce that for human-centric video understanding, face-only or joint-only preprocessing is essential for "faithful" explanations. By removing background noise and environmental distractors, the models are forced to derive evidence solely from relevant physiological dynamics, thereby improving the reliability of the generated prototypes.
- **Impact of Task Formulation and Visual Interference (V2):** The introduction of the V2 experimental setting (incorporating 'Sad' as a distractor within 'Neutral' groups) highlighted the sensitivity of interpretable models to visual intensity (arousal). This finding underscores the challenge of disentangling semantic emotional meaning from low-level visual cues when class boundaries are ambiguous.

6.3 Limitations

Despite the promising results, this study has several limitations:

1. **Simplified Temporal Modeling:** The current aggregation strategy relies primarily on statistical pooling and smoothing. While effective for feasibility verification, it does not explicitly model complex causal relationships or long-term temporal dependencies between frames.
2. **Scalability and Hyperparameter Sensitivity:** The optimal patch size and the number of prototypes or atoms currently require dataset-specific tuning. While this setting is sufficient for controlled analysis,

the scalability of the proposed framework to larger vocabularies and long-tail class distributions remains an important direction for future work.

3. **Subjectivity in Patch Interpretation:** While PatchSAE successfully identifies salient local regions (e.g., the forehead), definitively concluding what specific micro-feature (e.g., skin tension vs. illumination) the model focuses on requires further quantitative verification, as visual inspection alone can be ambiguous.

6.4 Future Work

Building on the findings of this thesis, several directions are identified for future research:

- **Learnable Temporal Interpretability:** Future work could incorporate interpretable temporal modules, such as "temporal prototypes," to automatically identify which specific frames or segments are most influential for the final video-level decision.

- **Multi-modal Explainability:** Since human communication is inherently multi-modal, integrating acoustic signals (available in datasets like RAVDESS) to study the interaction between visual and audio-based prototypes would provide a more holistic understanding of human behavior.

- **Human-in-the-loop Evaluation:** Conducting user studies with domain experts (e.g., psychologists or sign language specialists) to verify the practical utility and "trustworthiness" of the generated explanations in real-world diagnostic or educational scenarios. To validate the interpretability from a human perspective, we propose a user study framework to quantitatively assess whether the patch-based explanations (PatchSAE) align better with human intuition compared to prototype-based ones. The study follows a within-subject design with two key metrics:

  - *Comparative Clarity:* Participants choose which visualization better explains the prediction (A/B testing).
  - *Trustworthiness:* Participants rate their trust in the explanation on a 5-point Likert scale.

The detailed questionnaire structure is provided in **Appendix A**. Data collection is planned for the next phase of this research, targeting approximately 20 participants.

In conclusion, this thesis confirms that applying interpretable deep learning to multi-scale human motion data is a promising and necessary direction. The proposed framework provides a robust baseline for future research into transparent video understanding systems.

Bibliography

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021.

[2] Alejandro Barredo Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019.

[4] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Kundu, Carson Denison, Danny Hernandez, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[7] Miao Cheng, Chia-Huei Tseng, Ken Fujiwara, Victor Schneider, and Yoshifumi Kitamura. Asian emotional body movement database: Diverse intercultural e-motion database of asian performers (diem-a). In *2025 13th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2025.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[10] Paul Ekman. An argument for basic emotions. *Cognition and emotion*, 6(3-4):169–200, 1992.

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

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

[13] Anna Hedström, Leander anders Weber, Daniel Krakowczyk, Daria Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations. In *Journal of Machine Learning Research*, volume 24, pages 1–11, 2023.

[14] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *NeurIPS*, volume 33, pages 19000–19012, 2020.

[15] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[16] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1469, 2020.

[17] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(4):1195–1215, 2022.

[18] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations (ICLR)*, 2025.

[19] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022.

[20] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196327, 2018.

[21] Alireza Makhzani and Brendan Frey. k-sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2014.

[22] Jiawei Mao, Rui Li, Xinyu Cai, Runhua Huang, and Shiguang Li. Poster++: A cleaner and stronger facial expression recognition network. 2023.

[23] Meeke Nauta et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[24] Meeke Nauta, Ron Van Breugel, Christin Seifert, and Sezer Jemaa. Neural prototype trees for interpretable fine-grained image recognition. In *CVPR*, pages 14933–14943, 2021.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[26] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[27] Dawid Rymarczyk, Lukasz Struski, Jacek Tabor, and Bartosz Zielinski. Protopshare: A parameter-efficient prototypical network for interpretable image recognition. In *arXiv preprint arXiv:2011.14340*, 2021.

[28] Avinab Saha, Shashank Gupta, Sravan Kumar Ankireddy, Karl Chahine, and Joydeep Ghosh. Exploring explainability in video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3698–3708, 2024.

[29] Javiera Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Gholamreza Anbarjafari. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10), 2023.

[30] Ramprasad R Selvaraju, Michael Cogswell, Abhishek Das, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2016.

[31] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2013.

[32] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014.

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

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[35] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6897–6906, 2020.

[36] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023.

[37] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Appendix A

Proposed Human Study Questionnaire

This appendix outlines the preliminary design of the questionnaire for the proposed human study on interpretability. The study aims to evaluate user trust and the clarity of the explanations generated by the proposed framework (Patch-SAE and ProtoPNet).

Study Overview

- **Objective:** To compare the human-perceived interpretability of patch-based (texture) vs. prototype-based (part) explanations.
- **Target Audience:** Approximately 20 participants (university students and researchers).
- **Methodology:** Online survey containing comparative questions (A/B testing) and Likert-scale rating questions.

Draft Questions

Part 1: Clarity Comparison (A/B Testing) *Instruction: Participants are shown a video frame along with two visual explanations: Image A (ProtoPNet) and Image B (Patch-SAE).*

- **Q1:** Which visualization helps you better understand why the model predicted the emotion as 'Happy'?

  Image A is clearer

Image B is clearer

Both are similar

- **Q2:** Which visualization highlights facial regions that you consider more relevant (e.g., eyes, mouth)?

  Image A

  Image B

Part 2: Trustworthiness (Likert Scale) *Instruction: Please rate your agreement with the following statements on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).*

- **Q3:** The visual highlights generated by **Method A (ProtoPNet)** align with my own judgment of the emotion.
  (1) – (2) – (3) – (4) – (5)

- **Q4:** The visual highlights generated by **Method B (Patch-SAE)** align with my own judgment of the emotion.
  (1) – (2) – (3) – (4) – (5)

Part 3: Qualitative Feedback

- **Q5:** (Optional) Briefly describe which features (e.g., specific body parts or textures) you found most helpful for identifying the emotion.