electronics

MDPI

Article

SlowR50-SA: A Self-Attention Enhanced Dynamic Facial Expression Recognition Model for Tactile Internet Applications

Nikolay Neshov *,†, Nicole Christoff †, Teodora Sechkova †, Krasimir Tonchev † and Agata Manolova †

Faculty of Telecommunications, Technical University of Sofia, 8 Kliment Ohridski Blvd., 1000 Sofia, Bulgaria; nicole.christoff@tu-sofia.bg (N.C.); teodora.sechkova@gmail.com (T.S.); k_tonchev@tu-sofia.bg (K.T.); amanolova@tu-sofia.bg (A.M.)

* Correspondence: nneshov@tu-sofia.bg

† These authors contributed equally to this work.

Abstract: Emotion recognition from facial expressions is a challenging task due to the subtle and nuanced nature of facial expressions. Within the framework of Tactile Internet (TI), the integration of this technology has the capacity to completely transform real-time user interactions, by delivering customized emotional input. The influence of this technology is far-reaching, as it may be used in immersive virtual reality interactions and remote tele-care applications to identify emotional states in patients. In this paper, a novel emotion recognition algorithm is presented that integrates a Self-Attention (SA) module into the SlowR50 backbone (SlowR50-SA). The experiments on the DFEW and FERV39K datasets demonstrate that the proposed model achieves good performance in terms of both Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) metrics, achieving a UAR (WAR) of 57.09% (69.87%) on the DFEW dataset, and UAR (WAR) of 39.48% (49.34%) on the FERV39K dataset. Notably, SlowR50-SA operates with only eight frames of input at low temporal resolution, highlighting its efficiency. Furthermore, the algorithm has the potential to be integrated into Tactile Internet applications, where it can be used to enhance the user experience by providing real-time emotion feedback. SlowR50-SA can also be used to enhance virtual reality experiences by providing personalized haptic feedback based on the user's emotional state. It can also be used in remote tele-care applications to detect signs of stress, anxiety, or depression in patients.

Keywords: emotion recognition; facial expression; deep learning; SlowFast networks; SlowR50; self-attention; DFEW; FERV39K; Tactile Internet

Citation: Neshov, N.; Christoff, N.; Sechkova, T.; Tonchev, K.; Manolova, A. SlowR50-SA: A Self-Attention Enhanced Dynamic Facial Expression Recognition Model for Tactile Internet Applications. *Electronics* **2024**, *13*, 1606. https://doi.org/10.3390/electronics13091606

Academic Editor: Beiwen Li

Received: 20 February 2024
Revised: 12 April 2024
Accepted: 20 April 2024
Published: 23 April 2024

1. Introduction

The Tactile Internet (TI) is a fundamental aspect of 6G technology that predicts real-time haptic connection and offers revolutionary possibilities for distant work and immersive interactions. TI is a very low-latency, ultra-high-reliability communication system [1]. It is designed to allow remote access, monitoring, manipulation, or control of physical or virtual objects or processes that are perceived as happening in real time, either by humans or automated systems [2]. The development of TI presents obstacles, particularly in the field of tactile cognition, which refers to the understanding and processing of tactile interactions. An essential aspect of delivering superior haptic feedback and facilitating immediate performance assessment is a comprehensive grasp of context. In order to address the unavoidable delays in remote work, the development of artificial intelligence approaches is necessary. However, to enhance motion prediction and feedback, it is crucial to gather more information about the context and terminal interactions [3]. Key features include ultra-low latency, reliability, and a human-centric approach through projects such as Tactile Internet with Human-in-the-Loop (TaHiL) [4].

TI incorporates tactile and kinesthetic content to complement the visual and aural aspects of augmented reality (AR) and virtual reality (VR) experiences [5]. It embodies

a concept of the Internet that integrates the sensation of touch with conventional communication methods, with the goal of facilitating remote operation of systems without requiring physical proximity. TI applications enable tactile communications, especially in the context of haptic-enabled VR. These applications require extremely low latency, less than 50 ms, making them suitable for remote phobia treatment via VR [6]. A critical challenge in implementing TI is the "1 ms challenge", which requires the system reaction time to be less than 1ms in order to prevent users from being able to differentiate between local and remote control [7].

Another paradigm emerges as TI enables multisensory experiences in the Metaverse, integrating tactile and kinesthetic components with visual and audio content. As users interact within the Metaverse, the TI enhances the sensation of immersion by adding the additional sense of touch and pressure. However, it presents challenges related to reliability, low latency, and sensitivity to network jitter, making its integration into the Metaverse a complex but promising endeavor. In addition, it is essential to address issues related to the emotional and mental well-being of users, ethical norms, and the preservation of a safe and healthy environment within this virtual realm [8]. As the Metaverse enhances digital experiences with tactile elements, the integration of semantic compression refines the efficient communication of emotions, thus augmenting the immersive quality of virtual interactions. Semantic compression also addresses high-latency challenges in affective computing [9], playing a pivotal role in the TI's pursuit of diverse experiences, including emotional, behavioral, and cognitive dimensions, in the 6G era [1]. Unlike remote vision and hearing, haptic sensing remains an uncharted area that the TI seeks to explore [10]. In the realm of 6G TI, the convergence of semantic compression and emotions, exemplified by the work of Akinyoade and Eluwole [11], initiates further exploration of applications and research directions, recognizing its transformative role in shaping the era of 6G technology.

The convergence of the field of emotional intelligence with artificial intelligence (AI) and machine learning (ML) enables the development of robotic systems that can accurately perceive and react to human emotions [12]. Emotion, a complex and subjective feeling encompassing physiological, psychological, and behavioral aspects, arises from internal or external stimuli in the human being. It can be expressed in various forms, such as joy, sadness, surprise, or fear, observable in facial expressions or heard in a person's voice, or even through touch. Not only does touch affect emotion, but emotional expressions also affect touch perception. So emphasis is placed on enhancing the Quality of Experience (QoE) through machine learning and QoE models [13,14]. The potential of the Tactile Internet extends to applications like Exergames, where users interact with emotions through speech recognition. Exergames allows users to express feelings verbally during exercise, recording and analyzing emotions to generate tactile vibrations corresponding to users' feelings. This real-time feedback enhances users' overall experience and satisfaction during physical activities [15].

Facial expression recognition is a crucial component in the larger context of emotional recognition. It entails the analysis of the facial expressions of persons, which is achieved by both humans and computer systems using image processing and AI technologies.

The motivation behind this work stems from the burgeoning field of TI and its potential to revolutionize real-time user experiences through the integration of emotion recognition technology. With the advent of 6G technology, there is a growing emphasis on ultra-low-latency communication systems like TI, which enable remote interactions with physical or virtual objects in real time. However, the full realization of TI's potential hinges on the incorporation of emotional intelligence, particularly in applications such as immersive virtual reality interactions and remote tele-care. Recognizing the crucial role of facial expressions in conveying emotions, the objective of this study is to develop a novel deep learning architecture specifically tailored for Dynamic Facial Expression Recognition (DFER) within the TI framework. By integrating an SA module into the SlowR50 backbone (SlowR50-SA), this research aims to enhance the accuracy and efficiency of emotion recognition, thus paving the way for personalized emotion feedback in TI applications. Through

rigorous experimentation and benchmarking against state-of-the-art methods, this work seeks to demonstrate the effectiveness and computational efficiency of the proposed model, thereby addressing a critical need in the evolving landscape of TI-enabled interactions.

The main contributions of the presented work are as follows:

- Presenting a novel deep learning architecture for DFER, the model effectively extracts spatiotemporal features using the SlowR50 (8×8) model. This architecture integrates a slow pathway with low temporal resolution for capturing long-range temporal information and identifying subtle changes in facial expressions. The inclusion of an SA module further refines the feature vector, dynamically attending to relevant spatial and temporal details, enhancing the representation of nuanced facial expressions.
- The proposed algorithm achieves superior performance on benchmark datasets (DFEW and FERV39K) compared to state-of-the-art methods, demonstrating its effectiveness in Dynamic Facial Expression Recognition. The model outperforms competitors in terms of both UAR and WAR, showcasing its capability to accurately classify emotions.
- The model demonstrates computational efficiency by achieving state-of-the-art results with only eight frames of input. This efficiency, combined with its high performance, positions the algorithm as a promising candidate for real-world applications, especially in TI scenarios, where it can effectively recognize and respond to facial expressions with reduced computational cost.

The rest of this paper is structured as follows. Section 2 provides an overview of related works in the field of DFER. Section 3 details the proposed model and its implementation. In Section 4, we present the datasets used, experimental setups, comparisons with state-of-the-art methods, and an ablation study. This section also includes visualizations of 2D t-SNE features and confusion matrices. Finally, Section 5 concludes the paper, summarizing key findings and contributions.

2. Related Work

Dynamic facial expression recognition is a complex task in computer vision and affective computing. Its goal is to classify a facial video clip, rather than a still image, into one of the basic emotions. The field of DFER has attracted considerable attention from researchers [16–22]. These studies share a common goal of addressing challenges within environmental scenarios, such as occlusion, pose variation, and noisy frames. Despite the progress made by these methods, it is evident that they still fall short in extracting comprehensive temporal features that encompass both short-term and long-term aspects. So a prevailing trend in the recently published works is the adoption of the transformer architecture for modeling spatiotemporal relationships in facial expressions. This architectural choice, as highlighted by [16–21], underscores the importance of capturing complex dependencies within dynamic facial expressions. Evaluation of these DFER methodologies extends to commonly used datasets such as DFEW, FERV39K, AFEW, and BU-3DFE, reflecting the comparative analysis of DFER models by [16–21].

Examining the differences between these DFER methods reveals different approaches to spatiotemporal modeling. One notable study by Liu et al. (2023) [16] proposes an Expression Snippet Transformer (EST) that decomposes videos into expression fragments and predicts the order of scrambled fragments. This approach emphasizes the importance of unifying video lengths through interpolation and clipping, achieving high accuracy across multiple datasets. However, EST focuses on fragment-based analysis, leaving room for improvement in capturing long-range temporal dependencies. Zhao et al. (2021) [17] introduced the Former-DFER by using a combination of a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former) to train spatial and temporal features. While Former-DFER effectively captures spatiotemporal relationships, its performance may be limited by the complexity of the transformer architecture and the computational resources required. Lee et al. (2023) [18] present Frame-Level Emotion-Driven Dynamic Facial Expression Recognition featuring an Affectivity Extraction Network (AEN) with frame-level emotion-driven loss features. This method incorporates emotion-driven loss

functions to enhance recognition accuracy, but it may lack robustness in handling diverse environmental scenarios. Li et al. (2023) [22] contributed to intensity-adaptive loss for dynamic facial expression recognition by integrating a global attentional bias (GCA) block and intensity-adaptive loss (IAL) to handle different expression intensities. While effective in addressing intensity variations, this approach may require additional computational overhead. Li et al. (2022) [19] propose NR-DFERNet, addressing noisy frames using a dynamic–static fusion module (DSF) and a fragment-based filter (SF) to mitigate the impact of neutral frames. These different methodological approaches also involve variations in the training paradigm. Wang et al. (2023) [20] reimagined the learning paradigm for DFER by treating it as a weakly supervised problem and introduced the multi-3D dynamic facial expression learning (M3DFEL) framework with multi-instance learning (MIL). Additionally, variations in loss functions are investigated with Li et al. (2023) [22], introducing Intensity-Aware Loss to distinguish samples with low expression intensity. While the Intensity-Aware Loss effectively handles expression intensity variations, it may introduce additional computational overhead during training, potentially limiting scalability to larger datasets. Attention mechanisms are also a focal point, as seen in Ma et al.'s Logo-Former (2022) [21], which proposes a local–global spatiotemporal transformer (LOGO-Former) with attention mechanisms to capture local and global dependencies. However, this work may face challenges in capturing long-range dependencies and subtle temporal changes in dynamic facial expressions, potentially impacting its effectiveness in recognizing nuanced expressions. Addressing the challenge of modeling noisy frames, Li et al. (2022) [19] propose NR-DFERNet, introducing a dynamic class tag (DCT) and an SF to process noisy frames in the decision stage. However, this work may have limited effectiveness in handling complex noise patterns in dynamic video sequences, potentially leading to misclassification of facial expressions in challenging environmental conditions. In the ever-evolving field of Facial Expression Recognition, researchers have developed creative approaches to address the complexities of analyzing facial expressions in dynamic video sequences. The EST method [16] takes a distinctive approach by unifying video lengths through interpolation and clipping, using face detection, and randomly selecting frames to create expression snippets. With an implementation in PyTorch, EST achieves an average FER accuracy of 88.17% across datasets such as BU-3DFE, MMI, AFEW, and DFEW. Noteworthy is its real-time speed and computational efficiency. Since the EST method relies on random frame selection and interpolation to create expression snippets, it may introduce biases or artifacts in the extracted snippets, potentially leading to inaccuracies in recognition. The Frame-Level Emotion-Guided Dynamic Facial Expression Recognition with Emotion Grouping method [18] introduces the AEN architecture, incorporating temporal transformers and pre-processing involving face region detection. Trained on a PyTorch platform with an NVIDIA RTX 3090 GPU, it utilizes pre-trained networks and introduces fusion parameters for dynamic emotion grouping. This method emphasizes the efficacy of proposed loss functions and fusion parameters. Addressing the nuances of expression intensity, the Intensity-Aware Loss for Dynamic Facial Expression Recognition in the Wild method [22] employs a GCA Block, Dynamic–Static Fusion Module, and Temporal Transformer for feature extraction. Trained on PyTorch-GPU and Tesla V100 GPUs, it achieves performance on dynamic facial expression recognition tasks. As mentioned, this architecture relies heavily on pre-trained networks and fusion parameters, which may limit its adaptability to novel datasets or dynamic environmental conditions.

In conclusion of this section, the DFER field has witnessed substantial progress with various methodological approaches, all aiming to address challenges posed by environmental scenarios in videos, such as occlusion, pose variation, and noisy frames. The adoption of the transformer architecture, as evident in several studies, emphasizes the importance of capturing complex spatiotemporal relationships within dynamic facial expressions. Methodological variations include innovative techniques like EST, Former-DFER, Frame-Level Emotion-Driven Dynamic Facial Expression Recognition, NR-DFERNet, and M3DFEL, each proposing unique solutions to the challenges at hand. These methods em-

ploy diverse strategies, including attention mechanisms, intensity-adaptive loss, and novel training paradigms like weakly supervised learning and multi-instance learning. Evaluations on benchmark datasets reveal competitive performance, showcasing advancements in addressing expression intensity, noisy frames, and long-term temporal relationships. While these methodologies exhibit promising results, ongoing research and exploration of new techniques remain crucial for further advancements in the dynamic facial expression recognition domain.

3. Algorithm Description

3.1. Proposed Model

The proposed video classification model for DFER is illustrated in Figure 1.

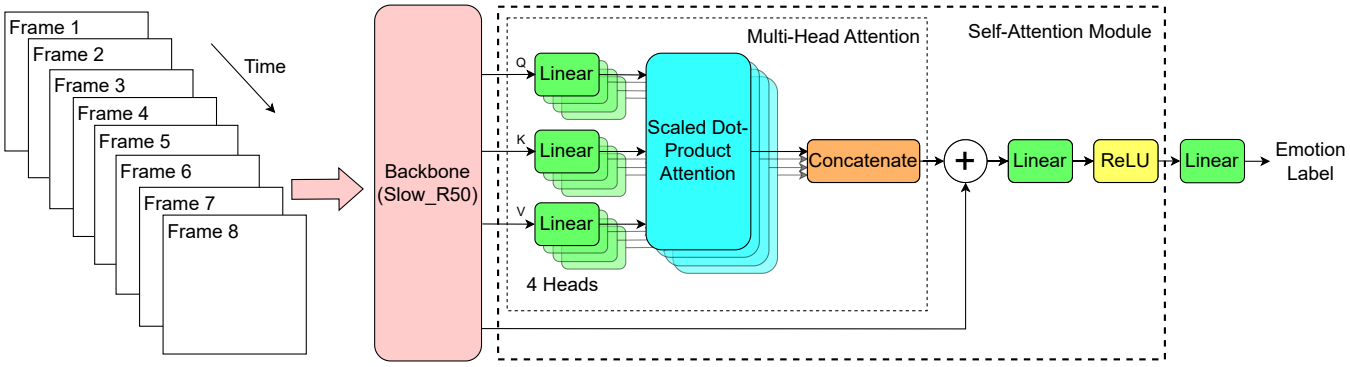

Figure 1. A pipeline of the proposed model.

For the extraction of spatiotemporal features, the SlowFast architecture's slow path model is utilized. In this case, we are specifically using the SlowR50 (8×8) model, as introduced in [23]. There are several reasons why it is encouraged to use a slow pathway with low temporal resolution. Initially, it can be used to extract long-range temporal information. When recognizing emotions in videos, it is necessary to identify the overall emotional state of the person, taking into account the temporal dynamics of their facial expressions. This can be effectively achieved by processing the video at a low temporal frame resolution, as it allows the model to focus on the subtle changes in facial expressions that occur over time. The slow pathway's ability to capture long-range temporal information allows it to identify long-range relationships, which is crucial for accurate expression recognition. Furthermore, it can reduce the computational cost of DFER algorithms. Because the slow pathway operates at a lower temporal resolution than the fast pathway, it requires fewer computations and less memory, making it more efficient to train and use. For the proposed model, the videos are processed by segmenting each video into C frames of resolution $M \times N$, resulting in an input tensor of a shape $M \times N \times C$. This approach allows us to efficiently capture the temporal dynamics of facial expressions while still maintaining a manageable input size. The feature vector extracted by the SlowR50 backbone is further refined by the SA module, which consists of multiple blocks: a Multi-Head Attention, a Summation block, a Linear (FC layer) and ReLU activation function (see Figure 1). The working flow of SA is further described in the details below.

Multi-Head Attention: Initially, the Multi-Head Attention mechanism with four heads is applied to the single feature vector, allowing the model to dynamically attend to the most relevant spatial and temporal information encoded within the vector. This attention mechanism enables the model to capture the subtle nuances of facial expressions, even in low temporal resolutions. In the context of the Multi-Head Attention operation, where the objective is to capture relationships within the same vector, all the query (Q), key (K), and value (V) vectors are set equal to the feature vector extracted by the backbone. This

simplification allows focusing on the self-attention mechanism applied to the feature vector. In general, the Multi-Head Attention operation is expressed as follows:

$$\text{MultiheadAttention}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \text{head}_3, \text{head}_4) \times \text{W}^O, \quad (1)$$

where each attention head (head_i) is computed as:

$$\text{head}_i = \text{Attention}(Q \times \text{W}_i^Q, K \times \text{W}_i^K, V \times \text{W}_i^V), \quad (2)$$

where W_i^Q, W_i^K, W_i^V, and W^O are weight matrices associated with the query, key, value, and output transformations, respectively, for each attention head i. The attention function (Attention) is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \times K^\text{T}}{\sqrt{d_k}}\right) \times V. \quad (3)$$

The variable K^T denotes the transposed matrix of the key vectors, ensuring compatibility with the query vectors for the attention calculation, and d_k represents the dimensionality of the key vectors.

Summation block: After the Multi-Head Attention mechanism has identified and weighted the most relevant spatial and temporal information within the feature vector, the attended features are added back to the original feature vector. This summation operation effectively integrates the attention mechanism's insights into the feature representation, weighting the features according to their importance and enriching the representation with additional information. This enhanced feature representation allows the model to better capture the subtle details in facial expressions.

Linear (FC layer) and ReLU activation function: The model further refines the feature representation by passing it through a Fully Connected (FC) layer and ReLU activation function. This combination of layers serves to normalize the feature representation, enhancing its complexity, and improving the model's ability to generalize to unseen data. This refined feature representation provides the model with a more accurate and informative basis for making predictions about the underlying emotion in the video frames. After the SA module, the final Linear (FC) layer is the classification layer, which produces the Emotion Label. During the training phase, the backbone is fine-tuned, while the SA module and classification layer are fully trained.

3.2. Implementation Details

The algorithm is implemented in PyTorch-GPU (v1.12.1) [24] and trained on an NVIDIA GeForce RTX 2080 Ti GPU (Graphics Processing Unit). It uses a SlowR50 (8×8) model [23] for feature extraction, which is fine-tuned during training. The models are trained for 100 epochs with an AdamW optimizer, learning rate of 5×10^{-4} , and weight decay of 0.05. After finding the best model, it is fine-tuned for another 30 epochs with a learning rate of 5×10^{-5}. Each video is input to the algorithm as eight frames of 196×196 pixels each. Horizontal flipping, random cropping, and color jitter are applied to augment the data.

4. Experiments

4.1. Datasets

Dynamic Facial Expression in-the-Wild (DFEW) [25] is a comprehensive dataset captured in real-world settings, introduced in 2020. Comprising over 16,000 video clips featuring dynamic facial expressions, these clips are collected from a broad range of over 1500 global movies, presenting diverse and real-world scenarios with challenges such as extreme illuminations, self-occlusions, and unpredictable pose changes. Each video clip is carefully annotated by ten well-trained experts under professional guidance. The anno-

tations classify expressions into seven categories: Happy, Sad, Neutral, Angry, Surprise, Disgust, and Fear.

FERV39K [26] encompasses 38,935 video clips sourced from four scenarios, further categorized into 22 fine-grained scenes. Distinguished by its unprecedented scale of 39K clips, scenario–scene division, and cross-domain supportability, FERV39K marks a milestone in DFER datasets. Each video clip within FERV39K undergoes meticulous annotation by 30 professional annotators, ensuring the provision of high-quality labels. These annotations classify expressions into the same seven primary categories as in DFEW.

4.2. Experimental Protocol

In this study, UAR and WAR are employed as primary evaluation metrics, aligning with established practices in the field of dynamic facial expression recognition. These metrics are widely used in previous studies for their effectiveness in evaluating model performance across various domains, including facial expression recognition [17–20,22,27]. UAR, computed as the average recall across all classes, provides an unbiased assessment of the model's ability to accurately classify facial expressions without favoring any specific class. It can be defined as:

$$UAR = \frac{1}{N}\sum_{i=1}^{N} R_i, \tag{4}$$

where N is the number of classes and R_i is the recall for class i. Similarly, WAR extends the evaluation beyond UAR by considering the distribution of samples across different classes. By weighting the recall of each class based on its sample size, WAR offers a more nuanced evaluation that accounts for class imbalances commonly encountered in real-world datasets. It can be expressed as:

$$WAR = \frac{\sum_{i=1}^{N}(R_i \times S_i)}{\sum_{i=1}^{N} S_i}, \tag{5}$$

where S_i is the number of samples for class i. Given the widespread use of UAR and WAR in existing literature, their adoption in this study enables direct comparisons with prior research outcomes. This ensures the consistency and reliability of the findings while facilitating a deeper understanding of the proposed model's performance relative to state-of-the-art approaches.

To ensure fair and consistent comparisons, we adopted a 5-fold cross-validation setup as suggested by DFEW [25] for evaluating various methods. For the FERV39K dataset, we followed the recommended approach from [26] by partitioning the data into 80% training and 20% testing sets.

4.3. Comparison of the Proposed Method with the State-of-the-Art Methods

The comparative analysis of the proposed SlowR50-SA algorithm with other state-of-the-art methods on both DFEW and FERV39K datasets is presented in Table 1. The research works included in the comparison analysis were chosen based on their use of the identical experimental protocol employed in this study. The table's results demonstrate that SlowR50-SA outperforms all other approaches in terms of both UAR and WAR metrics.

It surpasses the AEN model [18] with a difference of 0.43% (0.5%) for UAR (WAR) on DFEW. Additionally, SlowR50-SA outperforms the M3DFEL model [20] by UAR (WAR) of 0.99% (0.62%) on DFEW and 3.54% (1.67%) on FERV39K, despite using only eight frames as input compared to M3DFEL's sixteen frames. In addition, SlowR50-SA outperformed the second-best model in terms of UAR for the FERW39K dataset, surpassing ResNet18-ViT by 0.13%. Similarly, the proposed model outperformed IAL, the second-best model for FERW39K in terms of WAR, by 0.8%. This demonstrates the effectiveness of SlowR50-SA, which achieves superior performance using fewer frames and outperforms other methods.

Table 1. Comparison of proposed SlowR50-SA model with the state-of-the-art methods on DFEW and FERV39K datasets (bold indicates the best result, while underline indicates the second-best result). The evaluation metrics UAR and WAR for the methods compared with the SlowR50-SA algorithm are derived from corresponding literature data.

	DFEW		FERV39K	
Method	**UAR (%)**	**WAR (%)**	**UAR (%)**	**WAR (%)**
R(2+1)D-18 [20,28]	42.79	53.22	31.55	41.28
C3D [20,29]	42.74	53.54	22.68	31.69
I3D [20,30]	43.4	54.27	30.17	38.78
P3D [20,31]	43.97	54.47	23.2	33.39
EC-STFL [20,25]	45.35	56.51	-	-
3D-ResNet18 [20,32]	46.52	58.27	26.67	37.57
ResNet18-LSTM [22,33,34]	51.32	63.85	30.92	42.95
Former-DFER [17,22]	53.69	65.7	37.2	46.85
EST [16,27]	53.94	65.85	-	-
Logo-Form [21,27]	54.21	66.98	38.22	48.13
ResNet18-ViT [27,33,35]	55.76	67.56	<u>38.35</u>	48.43
NR-DFERNet [19,27]	54.21	68.19	33.99	45.97
IAL [22]	55.71	69.24	35.82	<u>48.54</u>
M3DFEL [20]	56.1	69.25	35.94	47.67
AEN [18]	<u>56.66</u>	<u>69.37</u>	38.18	47.88
SlowR50-SA (proposed)	**57.09**	**69.87**	**39.48**	**49.34**

4.4. Ablation Study on Self-Attention Module

Adding a Self-Attention Module after the SlowR50 backbone on the DFEW dataset resulted in an improvement of over 0.3% in the UAR metric and almost 0.5% in the WAR metric (see Table 2). This improvement came with an increase of 17.82M parameters and a slight increase of 40 M FLOPs (Floating Point Operations Per second). Recall the results shown in Table 1, in which the SlowR50 backbone exhibits impressive performance on the DFEW dataset, even outperforming the state-of-the-art AEN method [18]. However, it is important to note that the integration of the Self-Attention (SA) module further enhances the model's ability to capture subtle spatiotemporal dependencies within facial expression sequences. Despite the notable performance of the SlowR50 backbone alone, the additional complexity introduced by the SA module contributes to further enhancing the model's performance in DFER tasks.

Table 2. Ablation study on the effect of SA module added after the SlowR50 backbone on DFEW database.

Method	**UAR (%)**	**WAR (%)**	**Params (M)**	**FLOPs (G)**
SlowR50	56.78	69.40	31.65	71.88
SlowR50-SA	57.09	69.87	49.47	71.92

4.5. Detailed Results

In this section, a visual representation of the data using t-SNE [36] is provided. Specifically, two-dimensional t-SNE plots are employed to visualize samples from both the DFEW and FERV39K datasets, aiding in the comprehension of their distribution. Additionally, the confusion matrices of both datasets are presented for further analysis.

Two-dimensional t-SNE feature visualization: Figure 2a,b show the distribution of features in different colors and example image samples for each emotion in the DFEW and FERV39K datasets, respectively. For the DFEW dataset, it is evident that the features for the neutral, happy, sad, and angry emotions are more clearly separated into clusters, whereas the features for fear and surprise are more dispersed. The samples belonging to the disgust class do not form a cluster, likely due to the low proportion of disgust videos (1.22%) in the dataset. The model's inability to form a distinct cluster for the expressions of disgust indicates that it has

difficulty accurately classifying these emotions. Regarding the FERV39K dataset (Figure 2b), it is apparent that the clusters exhibit a more diffuse distribution than DFEW. Similar to the DFEW dataset, clusters representing neutral, happy, sad, and angry emotions appear more tightly grouped, whereas fear, surprise, and disgust exhibit a more dispersed arrangement. The two figures depicting t-SNE visually reinforce the findings presented in Table 1.

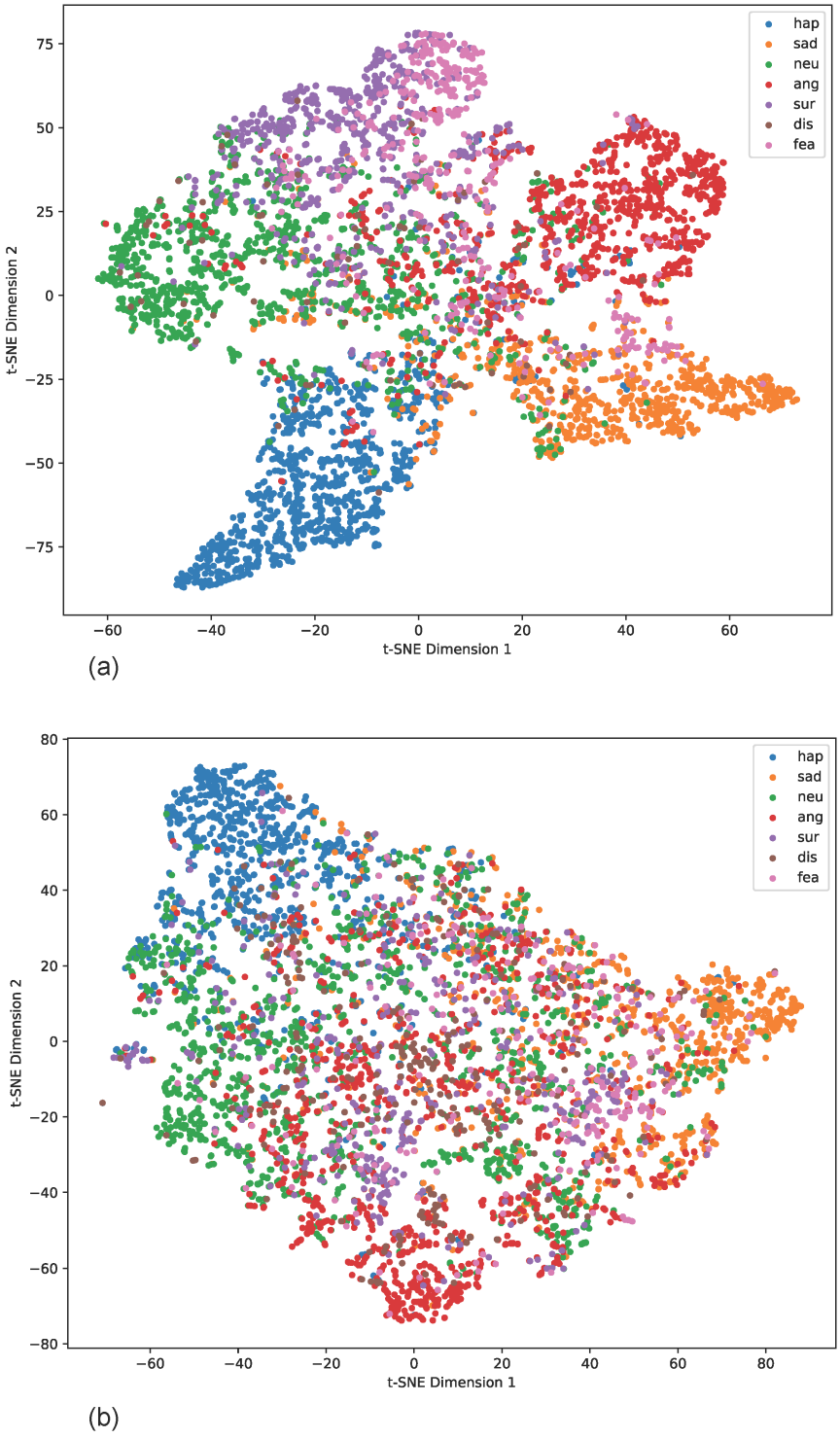

Figure 2. Two-dimensional t-SNE visualization [36] of facial expression features obtained by the proposed algorithm on videos from DFEW [25] dataset (**a**) and from FERV39K [26] dataset (**b**) .

Confusion matrices: The proposed SlowR50-SA algorithm is tested for its effectiveness on the DFEW dataset by examining confusion matrices generated across all five folds (Figure 3). These matrices reveal that the model struggles to accurately predict both the expressions of disgust and fear. The model performs particularly poorly with expressions of disgust, as observed in the earlier t-SNE visualization. While the model performs better with the expressions of fear, it still struggles to achieve an accurate classification rate due to the fact that videos with these emotions are also rarely presented in the dataset (only 8.14%). This suggests that the task of distinguishing between disgust and fear among the other expressions is particularly challenging. Additionally, the model tends to classify samples as neutral expressions in an attempt to minimize the risk of misclassification. Figure 4 depicts the confusion matrix for the FERV39K dataset. It reveals that happy, sad, and neutral emotions are identified more frequently, with rates exceeding 50%, while the remaining four emotions exhibit lower recognition rates.

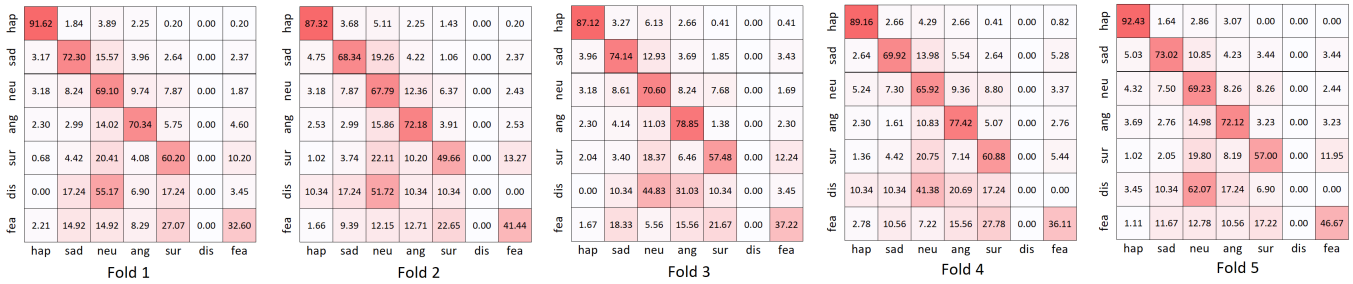

Figure 3. The confusion matrices obtained by the proposed SlowR50-SA algorithm on DFEW dataset.

Figure 4. The confusion matrix obtained by the proposed SlowR50-SA algorithm on FERV39K dataset.

The comparison of the confusion matrices between the DFEW and FERV39K datasets indicates notable differences in recognition performance. Analysis reveals that the DFEW dataset demonstrates superior classification accuracy compared to FERV39K. This discrepancy is particularly evident in the recognition of various emotions, where DFEW exhibits more robust performance across multiple emotion categories. These findings underscore the importance of dataset selection in training emotion recognition models and suggest the need for further investigation into the factors contributing to the variance in performance between datasets.

Variation of training and testing WAR across epochs: Figure 5a,b depict the evolution of training and testing WARs across epochs for the DFEW (fold 2) and FERV39K datasets, respectively. The learning rate for the DFEW dataset is reduced from 5×10^{-4} to 5×10^{-5} at epoch 44, while for the FERV39K dataset, it is reduced at epoch 51, as outlined in the training process detailed in Section 3.2. It is evident that the optimal testing WAR is achieved at epoch 50 (67.62%) for the DFEW dataset and at epoch 64 (49.34%) for the FERV39K dataset. The total time required for training and testing one epoch is approximately 450 s for the DFEW dataset and 1500 s for the FERV39K dataset.

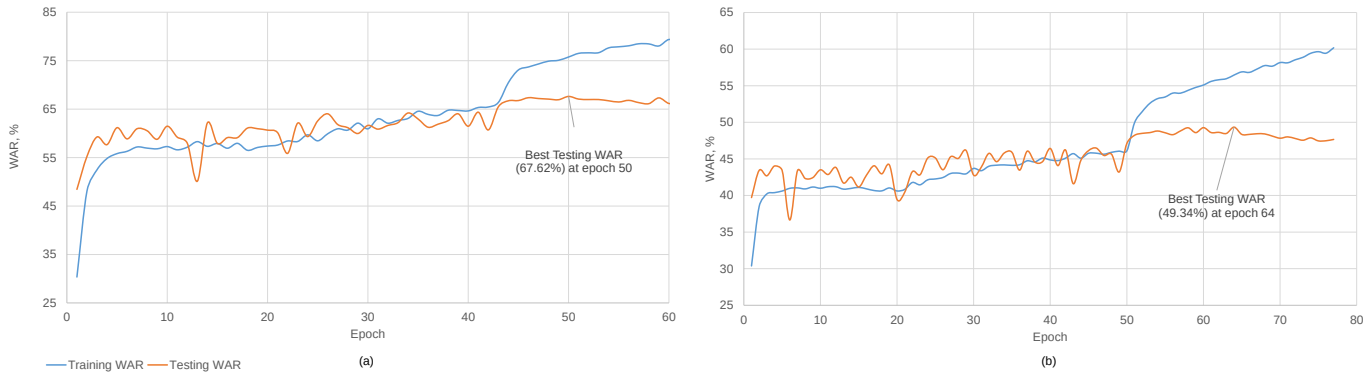

Figure 5. The variation of WAR across epochs for DFEW (fold 2) dataset (**a**) and FERV39K dataset (**b**).

GradCAM [37] visualizations: Figure 6 illustrates GradCAM visualizations from the final layers of the SlowR50 backbone across the seven emotions. It is evident that the activations primarily occur in facial regions that are characteristic of distinct emotions. This observation holds especially true for emotions depicted in the first row of Figure 6, including happy, sad, neutral, and angry. Nevertheless, when considering the emotions of disgust and fear, it is noticeable that the model does not focus on the relevant facial regions associated with these emotions. Consequently, the performance is not satisfactory for these emotions, as evidenced by the confusion matrices and t-SNE visualization, depicted above.

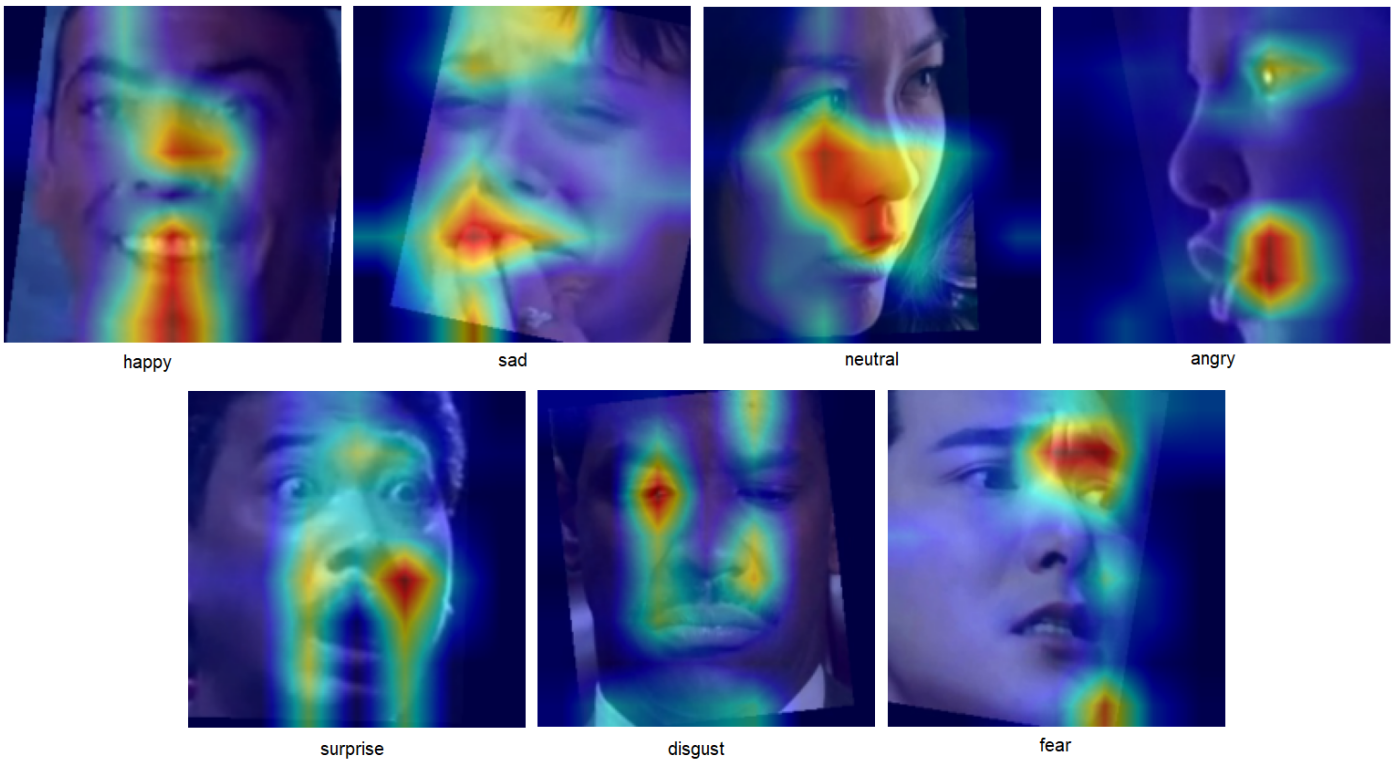

Figure 6. GradCAM [37] visualization of different emotions, utilizing weight gradients from the last SlowR50 backbone layer.

4.6. Limitations of the Presented Work

The following are considered limitations of the present work:

- While the proposed SlowR50-SA algorithm demonstrates superior performance on the DFEW and FERV39K datasets, its property to generalize to other datasets or real-world scenarios remains untested. The datasets used may not fully represent the

diversity of facial expressions encountered in real-world settings, potentially limiting the algorithm's applicability in practical situations.

- Both DFEW and FERV39K datasets may suffer from class imbalance issues, which can affect the model's performance, especially for minority classes such as disgust and fear. Imbalanced datasets may lead to biased models that prioritize majority classes, potentially resulting in lower accuracy for minority classes.
- The ablation study focuses solely on the addition of the Self-Attention module to the SlowR50 backbone. Further analyses, such as investigating the impact of different hyperparameters or architectural variations, could provide deeper insights into the algorithm's performance and help optimize its design.
- Although the proposed algorithm achieves good performance with only eight frames of input, its computational efficiency in real-world applications, especially on resource-constrained devices or in real-time systems, remains unclear. Assessing the algorithm's efficiency in practical deployment scenarios is essential for its feasibility in TI applications.

To address the limitations highlighted above, future research efforts could focus on the following areas:

- Generalization to diverse datasets: We acknowledge the importance of evaluating the algorithm's performance on a wider range of datasets, including those with more diverse facial expressions and real-world scenarios. Future work could involve testing the SlowR50-SA algorithm on additional datasets and assessing its robustness across various settings.
- Mitigating class imbalance issues: To mitigate the impact of class imbalance on model performance, future studies could explore techniques such as data augmentation, oversampling of minority classes, or using advanced loss functions tailored to handle imbalanced datasets. Additionally, efforts could be made to collect or curate datasets that better represent the distribution of facial expressions in real-world scenarios.
- Extended scope of ablation study: Further analysis could extend beyond the addition of the Self-Attention module to explore the effects of different hyperparameters, architectural variations, or alternative model components. Conducting comprehensive experiments would provide deeper insights into the algorithm's behavior and aid in optimizing its performance.
- Evaluation of computational efficiency: Future research should prioritize assessing the algorithm's computational efficiency in practical deployment scenarios. This could involve benchmarking the algorithm on resource-constrained devices, evaluating its runtime performance, and optimizing its implementation for real-time applications.

5. Conclusions

This paper presents SlowR50-SA, a novel emotion recognition algorithm that appends a Self-Attention module to the SlowR50 backbone. The experimental results on two benchmark datasets, DFEW and FERV39K, indicate that SlowR50-SA performs favorably compared to other algorithms, demonstrating good or better performance in terms of both UAR and WAR. Additionally, the model uses only eight frames of input, indicating its efficiency. The ablation study in Table 2 further highlights the positive impact of the Self-Attention module, which significantly improves the model's performance. These findings demonstrate the potential of SlowR50-SA as a powerful tool for emotion recognition. Its state-of-the-art performance, computational efficiency, and ability to operate with fewer input frames make it a promising candidate for real-world TI applications. Based on the promising outcomes of this study, future research could explore further enhancements to SlowR50-SA, such as experimenting with different variations of the Self-Attention module, integrating multimodal data sources for more robust emotion recognition, and conducting experiments with different backbone architectures and hyperparameters. Additionally, evaluating SlowR50-SA in real-world TI scenarios and exploring transfer learning techniques could accelerate its deployment and improve its effectiveness across diverse

applications. These avenues offer opportunities to advance emotion recognition technology and its integration into real-world settings.

Author Contributions: Conceptualization, N.N., K.T. and T.S.; methodology, K.T. and N.N.; software, N.N. and T.S.; validation, N.N., N.C. and T.S.; formal analysis, A.M. and N.C.; investigation, A.M., K.T. and N.C; resources, N.N.; data curation, N.N.; writing—original draft preparation, A.M., K.T., N.C., N.N. and T.S.; writing—review and editing, A.M. and N.C.; visualization, N.N.; supervision, A.M.; project administration, A.M.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research is financed by the European Union-Next Generation EU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0005: "Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia" (IDEAS).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AR	Augmented Reality
AEN	Affectivity Extraction Network
CS-Former	convolutional spatial transformer
DCT	dynamic class tag
DFER	Dynamic Facial Expression Recognition
DFEW	Dynamic Facial Expression in-the-Wild
DLIAM	Dynamic Long-term Instance Aggregation Module
DSF	dynamic–static fusion module
EST	Expression Snippet Transformer
FC	Fully Connected
FLOPs	Floating Point Operations Per second
GCA	global attentional bias
GPU	Graphics Processing Unit
IAL	intensity-adaptive loss
LOGO-Former	local–global spatiotemporal transformer
MIL	multi-instance learning
M3DFEL	multi-3D dynamic facial expression learning
QoE	quality of experience
SA	Self-Attention
SF	fragment-based filter
TaHiL	Tactile Internet with Human-in-the-Loop
TI	Tactile Internet
T-Former	temporal transformer
UAR	Unweighted Average Recall
VR	Virtual Reality
WAR	Weighted Average Recall

References

1. Fanibhare, V.; Sarkar, N.I.; Al-Anbuky, A. A survey of the tactile internet: Design issues and challenges, applications, and future directions. *Electronics* **2021**, *10*, 2171. [CrossRef]
2. Holland, O.; Steinbach, E.; Prasad, R.V.; Liu, Q.; Dawy, Z.; Aijaz, A.; Pappas, N.; Chandra, K.; Rao, V.S.; Oteafy, S.; et al. The IEEE 1918.1 "tactile internet" standards working group and its standards. *Proc. IEEE* **2019**, *107*, 256–279. [CrossRef]
3. Oteafy, S.M.; Hassanein, H.S. Leveraging tactile internet cognizance and operation via IoT and edge technologies. *Proc. IEEE* **2018**, *107*, 364–375. [CrossRef]
4. Ali-Yahiya, T.; Monnet, W. *The Tactile Internet*; John Wiley & Sons: Hoboken, NJ, USA, 2022.
5. Xu, M.; Ng, W.C.; Lim, W.Y.B.; Kang, J.; Xiong, Z.; Niyato, D.; Yang, Q.; Shen, X.S.; Miao, C. A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges. *IEEE Commun. Surv. Tutor.* **2022**, *25*, 656–700. [CrossRef]

6. Rasouli, F. A Framework for Prediction in a Fog-Based Tactile Internet Architecture for Remote Phobia Treatment. Ph.D. Thesis, Concordia University, Montreal, QC, Canada, 2020.
7. Van Den Berg, D.; Glans, R.; De Koning, D.; Kuipers, F.A.; Lugtenburg, J.; Polachan, K.; Venkata, P.T.; Singh, C.; Turkovic, B.; Van Wijk, B. Challenges in haptic communications over the tactile internet. *IEEE Access* **2017**, *5*, 23502–23518. [CrossRef]
8. Tychola, K.A.; Voulgaridis, K.; Lagkas, T. Tactile IoT and 5G & beyond schemes as key enabling technologies for the future metaverse. *Telecommun. Syst.* **2023**, *84*, 363–385.
9. Amer, I.M.; Oteafy, S.M.; Hassanein, H.S. Affective Communication of Sensorimotor Emotion Synthesis over URLLC. In Proceedings of the 2023 IEEE 48th Conference on Local Computer Networks (LCN), Daytona Beach, FL, USA, 2–5 October 2023; pp. 1–4.
10. Dar, S.A. International conference on digital libraries (ICDL)-2016 Report, Teri, New Delhi. *Libr. Tech News* **2017**, *34*, 8. [CrossRef]
11. Akinyoade, A.J.; Eluwole, O.T. The internet of things: Definition, tactile-oriented vision, challenges and future research directions. In *Proceedings of the Third International Congress on Information and Communication Technology: ICICT 2018, London*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 639–653.
12. Gupta, M.; Jha, R.K.; Jain, S. Tactile based intelligence touch technology in IoT configured WCN in B5G/6G-A survey. *IEEE Access* **2022**, *11*, 30639–30689. [CrossRef]
13. Steinbach, E.; Strese, M.; Eid, M.; Liu, X.; Bhardwaj, A.; Liu, Q.; Al-Ja'afreh, M.; Mahmoodi, T.; Hassen, R.; El Saddik, A.; et al. Haptic codecs for the tactile internet. *Proc. IEEE* **2018**, *107*, 447–470. [CrossRef]
14. Alja'Afreh, M. A QoE Model for Digital Twin Systems in the Era of the Tactile Internet. Ph.D. Thesis, Université d'Ottawa/University of Ottawa, Ottawa, ON, Canada, 2021.
15. Shamim Hossain, M.; Muhammad, G.; Al-Qurishi, M.; Masud, M.; Almogren, A.; Abdul, W.; Alamri, A. Cloud-oriented emotion feedback-based Exergames framework. *Multimed. Tools Appl.* **2018**, *77*, 21861–21877. [CrossRef]
16. Liu, Y.; Wang, W.; Feng, C.; Zhang, H.; Chen, Z.; Zhan, Y. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognit.* **2023**, *138*, 109368. [CrossRef]
17. Zhao, Z.; Liu, Q. Former-dfer: Dynamic facial expression recognition transformer. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 1553–1561.
18. Lee, B.; Shin, H.; Ku, B.; Ko, H. Frame Level Emotion Guided Dynamic Facial Expression Recognition With Emotion Grouping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5680–5690.
19. Li, H.; Sui, M.; Zhu, Z. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv* **2022**, arXiv:2206.04975.
20. Wang, H.; Li, B.; Wu, S.; Shen, S.; Liu, F.; Ding, S.; Zhou, A. Rethinking the Learning Paradigm for Dynamic Facial Expression Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 17958–17968.
21. Ma, F.; Sun, B.; Li, S. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv* **2022**, arXiv:2205.04749.
22. Li, H.; Niu, H.; Zhu, Z.; Zhao, F. Intensity-aware loss for dynamic facial expression recognition in the wild. *Proc. Aaai Conf. Artif. Intell.* **2023**, *37*, 67–75. [CrossRef]
23. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 6202–6211.
24. Pytorch.org, Instalation of Pytorch v1.12.1. Available online: https://pytorch.org/get-started/previous-versions/ (accessed on 25 March 2024).
25. Jiang, X.; Zong, Y.; Zheng, W.; Tang, C.; Xia, W.; Lu, C.; Liu, J. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2881–2889.
26. Wang, Y.; Sun, Y.; Huang, Y.; Liu, Z.; Gao, S.; Zhang, W.; Ge, W.; Zhang, W. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20922–20931.
27. Awesome Dynamic Facial Expression Recognition. Available online: https://github.com/zengqunzhao/Awesome-Dynamic-Facial-Expression-Recognition (accessed on 25 March 2024).
28. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
29. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
30. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 6299–6308.
31. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3D residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
32. Hara, K.; Kataoka, H.; Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6546–6555.

33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
36. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
37. Gildenblat, J. Contributors. PyTorch Library for CAM Methods. 2021. Available online: https://github.com/jacobgil/pytorch-grad-cam (accessed on 12 April 2024).