




Article

Emotion Recognition beyond Pixels: Leveraging Facial Point Landmark Meshes

Herag Arabian ¹, Tamer Abdulbaki Alshirbaji ^{1,2}, J. Geoffrey Chase ³ and Knut Moeller ^{1,*}

¹ Institute of Technical Medicine (ITeM), Furtwangen University, 78054 Villingen-Schwenningen, Germany; h.arabian@hs-furtwangen.de (H.A.)

² Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, 04103 Leipzig, Germany

³ Department of Mechanical Engineering, University of Canterbury, Christchurch 8041, New Zealand

* Correspondence: moe@hs-furtwangen.de

Featured Application: This work is being developed as part of a digital health system designed to assist in the therapeutic treatment of people with autism spectrum disorder.

Abstract: Digital health apps have become a staple in daily life, promoting awareness and providing motivation for a healthier lifestyle. With an already overwhelmed healthcare system, digital therapies offer relief to both patient and physician alike. One such planned digital therapy application is the incorporation of an emotion recognition model as a tool for therapeutic interventions for people with autism spectrum disorder (ASD). Diagnoses of ASD have increased relatively rapidly in recent years. To ensure effective recognition of expressions, a system is designed to analyze and classify different emotions from facial landmarks. Facial landmarks combined with a corresponding mesh have the potential of bypassing hurdles of model robustness commonly affecting emotion recognition from images. Landmarks are extracted from facial images using the Mediapipe framework, after which a custom mesh is constructed from the detected landmarks and used as input to a graph convolution network (GCN) model for emotion classification. The GCN makes use of the relations formed from the mesh along with the special distance features extracted. A weighted loss approach is also utilized to reduce the effects of an imbalanced dataset. The model was trained and evaluated with the Aff-Wild2 database. The results yielded a 58.76% mean accuracy on the selected validation set. The proposed approach shows the potential and limitations of using GCNs for emotion recognition in real-world scenarios.

Keywords: deep learning; digital health; emotion recognition; facial point landmarks; graph convolution network; mental well-being; mesh analysis; pattern recognition



Citation: Arabian, H.; Abdulbaki Alshirbaji, T.; Chase, J.G.; Moeller, K. Emotion Recognition beyond Pixels: Leveraging Facial Point Landmark Meshes. *Appl. Sci.* **2024**, *14*, 3358. <https://doi.org/10.3390/app14083358>

Academic Editors: Hyeonjoon Moon and Lien Minh Dang

Received: 28 February 2024

Revised: 11 April 2024

Accepted: 15 April 2024

Published: 16 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Digital health apps have seen rapid growth over recent years as they promote physical and mental well-being and provide motivation for a more health-conscious lifestyle. As many healthcare systems are increasingly overwhelmed by demand, these health apps coupled with digital therapies offer the potential to benefit patients and clinicians through digital patient-led care [1,2]. Digital health apps can also offer a more individualized patient-centered approach to care. By supplying vital information collected from these intelligent systems over a longer duration than the few minutes physicians spend with patients in one-on-one sessions, more efficient diagnosis and treatment can be achieved. Intelligent learning systems may offer further opportunity to enhance the quality and productivity of patient care, and equity of access to it, e.g., [3].

A digital emotion recognition tool to assist in the therapeutic intervention of people with autism spectrum disorder (ASD) is being developed. ASD is estimated to affect 1~2% of the general population, which constitutes roughly 1 out of every 59 people [4,5]. ASD is defined as a neuro-developmental condition impairing a person's social skills, such as their interaction,

communication, behaviors, and interests [4,6,7]. This condition may often lead to more severe health problems attributed to isolation and unemployment (or reduced/under employment), which can result in depression and anxiety [4]. To counteract these problems, group or single therapies are devised for each level of ASD. A closed-loop emotional feedback system is in development to immerse subjects in a virtual world and have them take part in different gamified scenarios and tasks designed to stimulate emotional responses as part of therapy. This system also aims to assist clinicians by providing vital information to monitor progress and adapt the therapy level, assuring better quality and more personalized care. Such digital therapeutic intervention systems have the potential to assist individuals with ASD to cope better in different social environments [8–10].

Defining an emotion is a topic of great debate [11]. In this work, an emotion is considered as defined by the framework of a component process model, a sequence of triggers interrelated to changes in organism states in response to an external or internal stimulus [11]. Over the years, the understanding of emotional stimuli, in particular their influence and role in clinical settings, has evolved. Emotional stimuli, more specifically those identified as threats, have significant influence on selective attention, processing prioritization, and Pavlovian response [12,13]. However, an interesting perspective has emerged in recent studies, that reactions to such stimuli are aligned with personal objectives. This perspective suggests a context-dependent response of facial expressions [14,15].

To ensure efficient emotional feedback of the proposed system, an effective model for the recognition of expressions must be developed. Identifying emotions is often a difficult task given the numerous ways a person can express themselves [16]. However, emotions are often perceived through either facial expressions (55%); speech and voice patterns (35%); or physiological signals (10%) [17]. This study concentrates on the use of facial expressions as a basis for emotion recognition. It is also worth noting that previous research [18,19] has shed light on the importance of whole-body expressions in relating to the emotional states of others, such as by recognizing emotions when the face is occluded and conveying action intentions.

The prevalence of studies [20–25] on non-verbal emotional cues highlights the growing interest in the evolving field of recognizing emotions to bridge the gap in human–machine interactions. Extracting information from facial expressions is achieved via two methods: (1) image-based; and (2) geometric-based. This study uses a geometric-based approach, where facial point landmarks are extracted and used in combination with a graph convolution network to provide a robust representation of emotion state data.

The use of facial landmarks provides a more interrelated and holistic approach to the identification of emotions from facial expressions. In [16], the use of facial landmark locations incorporated in the classification process with a unique loss function revealed promising results when tested on distinct emotional datasets. The work of [26] highlighted the use of facial landmarks extracted from a Kinect 3D device to identify action units (AUs) based on the representation from the facial action coding system (FACS) [27]. The use of the AUs allowed the classification model performance to reach 96% accuracy on the Karolinska directed emotional faces (KDEF) [28] dataset. A combination of facial landmark localization, with 68 facial landmarks, and physiological signals was studied in [29]. The implemented model was able to effectively classify six emotional classes with an accuracy of 86.94% on the gathered dataset. In [30], key facial landmarks were selected and used for geometric analysis of facial gestures on three different datasets using machine learning models for classification. The models achieved good performance, reaching 97% accuracy on the extended Cohn–Kanade (CK+) dataset with a k-nearest neighbor (KNN) classifier and real-time processing time of 250 Hz.

This study's proposed approach relies on a graph convolution network (GCN) as the classification model. The GCN adopts a spatially unconstrained methodology, allowing points to exist freely in three-dimensional space. Unlike methods constrained to a specific plane, the GCN leverages the interconnections among linkages and establishes relations with point anchors, resulting in a more resilient and robust outcome [31]. Graph networks

can be used in a wide array of applications from generative models and traffic network predictions [32] to text sequence labeling [33].

To evaluate the proposed approach, the Aff-Wild2 [34–44] database was employed. This database is considered large and is composed of in-the-wild, e.g., non-posed, captured image frames. The images were first passed through the point landmark feature extractor of Mediapipe [45], where 478 facial landmarks were extracted and depicted in a 3D coordinate plane. A subset of Aff-Wild2 was used for the analysis of the proposed approach. The subset was split according to an 80% training and 20% validation partition.

As Aff-Wild2 is a challenge database, different techniques and approaches have been studied for emotion recognition. In [46], a semi-supervised approach to improve facial expression recognition was implemented using unlabeled data and a dynamic threshold module, achieving an F1-score of 0.3075 [44]. An expression-related self-supervised learning method was developed in [47] to classify facial expressions, achieving an F1-score of 0.3218 [44]. A multi-layered perceptron ensemble was studied in [48] with a pre-trained EmotiEffNet architecture for feature extraction from frames. This approach achieved an F1-score of 0.3292 [44]. A fused transformer encoder model using audio-visual input with an affine module was implemented in [49], reaching an F1-score of 0.3337 [44]. In [50], a multi-modal fusion model was developed. This approach leverages a temporal convolutional network and transformer models to enhance performance, reaching an F1-score of 0.3532 [44].

Specifically, the study in this work leverages the unique characteristics of facial meshes and the interrelations among facial landmarks, which is an approach which, to the best of the authors' knowledge, has not been extensively explored in prior research. In particular, while previous studies touched on the use of a fraction of the 478 facial landmarks with a distinct relational geometry, there is an opportunity in this work to delve deeper into capturing subtle changes in facial expressions. These subtle changes are often overlooked in pursuit of broader emotional generalizations but may contain key elements to improve the overall emotion recognition capability.

This study's primary aim is to show that combining facial landmark points with a graph convolution network provides better efficacy in identifying emotions. Section 2 describes the methods used for the network architecture, feature selection, and analysis criteria. Section 3 provides the results, followed by their respective discussions in Section 4, and concluding with a summary of the main findings in Section 5.

2. Materials and Methods

2.1. System Methodology

The proposed system workflow is displayed in Figure 1. The input to the system is an image that is passed through the pre-processing stage, where the point landmark feature extraction algorithm of Mediapipe [45] is implemented. The algorithm extracts the relevant facial landmarks and projects the data in a 3D coordinate plane. This information is then processed with the features for each node, and an adjacency matrix is created from the generated mesh. The data then pass through a graph convolution network (GCN) used as the emotion classification model.

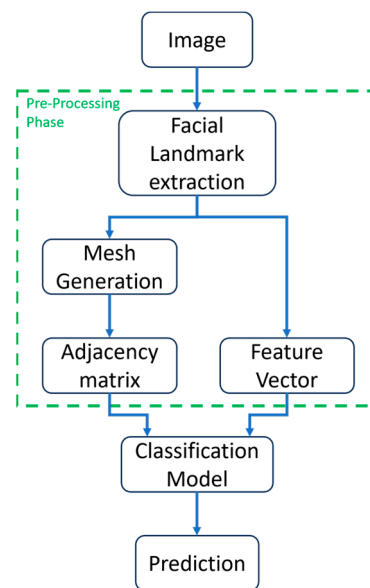


Figure 1. Proposed system workflow. Green dashed lines represent the pre-processing phase.

2.2. Pre-Processing Stage

To develop an efficient emotion recognition model that is robust to changes in light and subject demographic, a point landmark detection algorithm was chosen for the estimation of facial landmarks. The Mediapipe [45] framework of facial landmark detection was selected, as it is robust and provides landmarks in a 3D space. The algorithm adjusts to facial orientation and distance relative to the camera and provides an output of 478 facial point landmarks encompassing the face and key features, including eyes, nose, mouth, and eyebrows. The algorithm also accounts for dynamic obstructions, such as the placement of a hand in front of the face.

After extracting the facial landmarks, a custom mesh was created to highlight the relation of each node with its corresponding neighbor as the linkages. The mesh is symmetrical on the horizontal axis and depicted in Figure 2. The adjacency matrix is later obtained from the given mesh.

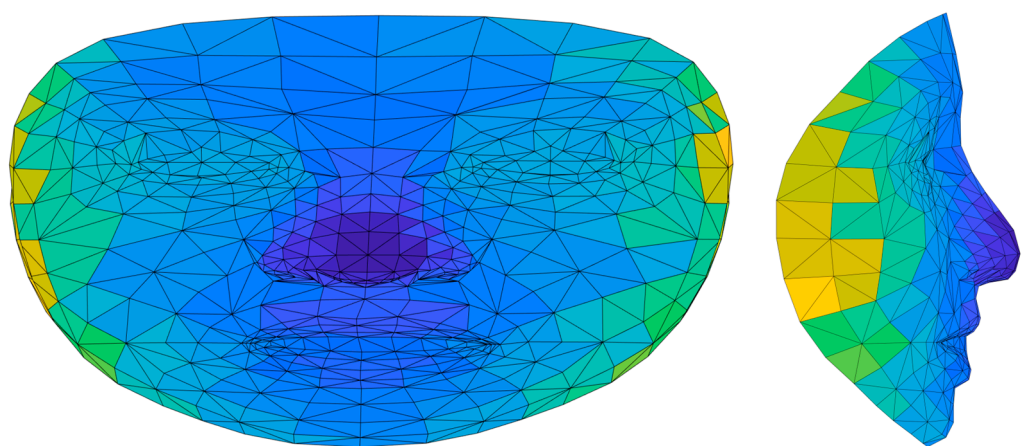


Figure 2. The generated mesh from the 478 facial point landmarks extracted: front view (left) and side view (right). The image is shown in a 3D plane, and the color variations depict the depth of the corresponding surface.

The overall impact of the facial mesh, with all 478 facial landmark features being given equal significance, was set to achieve a holistic representation of facial dynamics, emphasizing its collective influence on the proposed model's performance.

2.3. Feature Extraction

To achieve accurate classification of the point landmark collection into a specific single class, it is crucial to identify distinguishing features. To maintain objectivity across different demographics, four key points were selected as anchors within the landmark points. Subsequently, the Euclidean distance function was employed to calculate the distances between these anchors and corresponding points in the collection. The anchor points chosen were the landmarks of 5, 11, 94, and 324, which represent the tip of the nose, the top of the forehead, and the outermost left and right points of the face.

2.4. Classification Model

2.4.1. Network Architecture

The graphical convolution network (GCN) [31] was chosen as the classification model. The GCN employs layer-wise propagation, enabling a first-order approximation of spectral convolutions on graphs [31]. This approach facilitates encoding both the graph structure and node features, leading to improved modeling of relational data. By effectively capturing dependencies between nodes and linkages, GCNs extract robust, hierarchical features from graph-structured data.

The proposed model takes two inputs, the adjacency matrix, and the feature vector. The model was designed with four multiplication layers. These layers capture and refine the information from both the graph structure and node features. The first layer (32-feature output) aggregates the data from neighboring nodes while the second (32-feature output) and third (478-feature output) layers update the node representation based on the aggregated information, refining the features to capture high-level representations. The last multiplication layer (8-feature output) is followed by the SoftMax activation function, which provides the output of the architecture and is equal to the number of classes. Figure 3 represents the model architecture.

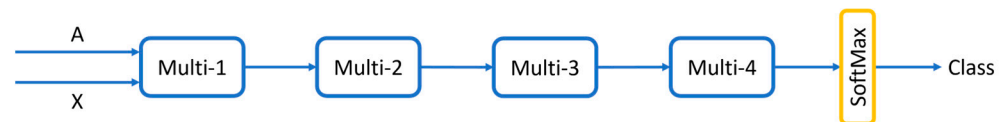


Figure 3. Proposed model architecture. A is the adjacency matrix; X is the feature vector. Multi stands for the multiplication layer.

2.4.2. Weighted Loss

To ensure a fair representation of all classes during the network training, a weighted cross-entropy function was used. Weights were calculated by

$$W_c = \frac{\max(S)}{S_c} \quad (1)$$

where W_c is the weight of the corresponding class c , S is a vector representing each class's data count, and S_c is the amount of data for a certain class c .

2.4.3. Training Options

The model was trained on 1500 epochs with a fixed learning rate of 0.01 using the adaptive moment estimation (Adam) optimization function. The model was run in a MATLAB 2023b environment (The MathWorks, Natick, MA, USA) on a desktop with 512.00 GB memory (RAM) and an NVIDIA graphics card RTX A6000 (NVIDIA Corporation, Santa Clara, CA, USA).

2.5. Performance Measures

To evaluate the performance of the proposed model, three techniques were implemented. The assessment was based on both the mean true positive (TP) accuracy and

mean F1-score across all classes, as well as Cohen’s Kappa coefficient. The F1-score was calculated by

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c} \quad (2)$$

where $F1_c$ is the F1-score of the corresponding class c . TP_c , FP_c , and FN_c are the TP , false positive (FP), and false negative (FN) for the given class c .

In order to further highlight the model’s abilities of robust emotion recognition, a custom heat map was created using the visualization technique of class activation mapping (CAM) for model explain-ability. Since CAM is most commonly associated with image-based approaches and not graphical node representations, a custom visual was developed. The explain-ability method used highlights the point landmarks that are showing a strong impact on the decision-making process of the model with a sphere that varies in size and color based on the intensity of the impact. A color scale was used, and the feature maps were extracted from the last layer of the model architecture.

2.6. Database Description

To train and assess the proposed model in close to real-world situations, an in-the-wild-collected database was selected. Aff-Wild2 is a relatively large dataset composed of 564 videos with ~2.8 million image frames. The database is annotated for different tasks of valence–arousal, expression, and AU classification. For this study, the expression classification annotations of 8 unique classes were selected, where the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) along with a neutral state and an “other” state are conveyed. This subset is composed of 546 videos with 2.6 million frames gathered from a total of 554 subjects (326 male and 228 female) of diverse demographics and environments [34–44].

For this study, a total of 30 subjects were randomly selected from the database for the training and evaluation. Of the 30, 12 were from the original database validation set. The model training was conducted agnostically to individual subjects, with the data split into 80% training and 20% testing.

3. Results

3.1. Dataset Distribution

Table 1 represents the selected dataset’s class distribution. As observed, there is an imbalanced distribution in the emotion classes with a particular bias towards the neutral class and the weakest representation of the fear class. The selected dataset recorded a high imbalance ratio of 37.61, calculated as the ratio between the sample numbers of the majority class and the minority class [51]. This distribution presents a challenge, emphasizing the proposed model’s capacity for generalization in prediction.

Table 1. Class distribution of the selected dataset.

Emotion Class	Selected	Training	Validation
Anger	3239	2581	658
Disgust	1129	903	226
Fear	990	797	193
Happiness	12,022	9647	2375
Neutral	37,233	29,852	7381
Other	19,364	15,417	3947
Sadness	16,567	13,271	3296
Surprise	9267	7381	1886
Total	99,811	79,849	19,962

3.2. Model Performance

The TP performance of the proposed model is represented in Figure 4 for each class for both the training and validation sets. The weakest performance was attributed to the

happiness class with 49.92% and 49.98% for the training and validation sets, respectively. The best results were for the fear class with 86.07% and anger with 83.59% for both the training and validation sets, respectively. The mean TP accuracy yielded 58.76% on the validation set.

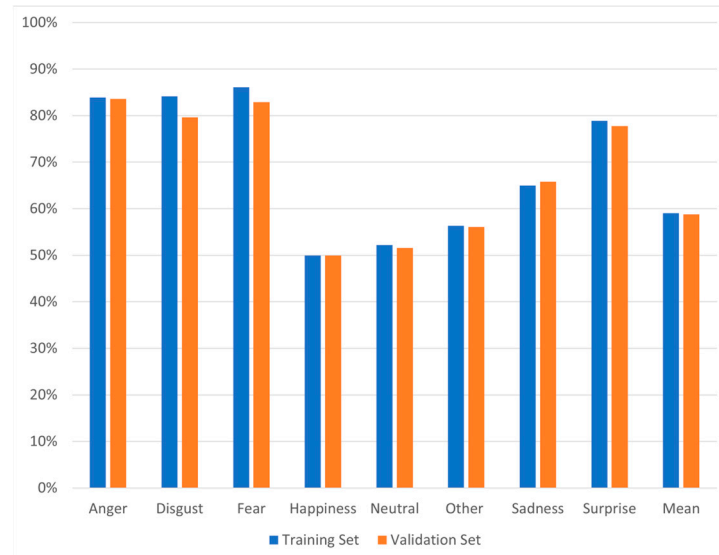


Figure 4. Evaluation results of the true positive (TP) accuracy for each class and its mean for both the validation and training sets.

In Figure 5, the results from the F1-score are depicted for each class for both the training and validation sets. The outcomes show weak performance on the fear class with an F1-score of 28.39% and 27.00% for both the training and validation sets, respectively. The strongest performance was observed for the surprise class with 71.75% and 71.13% for both the training and validation sets, respectively. The mean F1-score was $53.07\% \pm 14.87\%$ on the validation set.

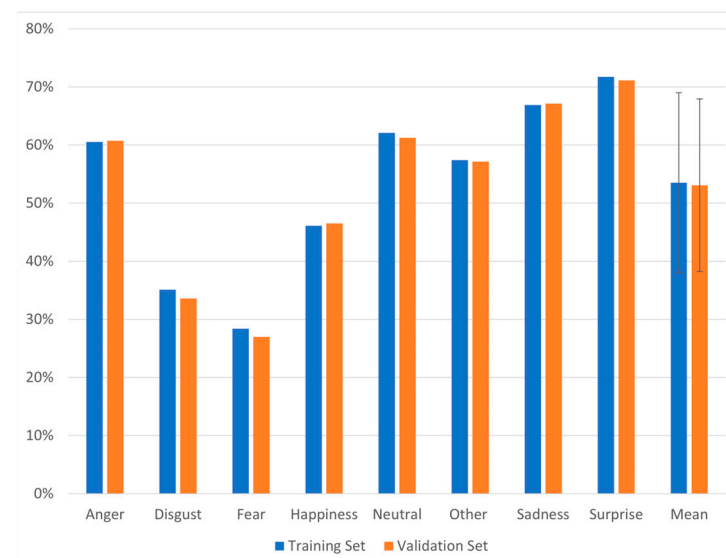


Figure 5. Results of the F1-score for each class and its mean for both the validation and training sets. The error bars represent the standard deviation.

The confusion matrices of both the training and validation sets are shown in Figures 6 and 7. A strong misclassification and confusion can be observed between the neutral and “other” classes with 14.42% and 13.87%, respectively, for the validation set.

Another strong confusion is observed between the neutral and happiness classes with 14.25% and 15.20% for the validation and training sets, respectively.

True Class	Other	2213	569	206	254	188	102	285	130
	Neutral	1024	3805	553	1052	189	308	297	153
	Sadness	263	239	2168	162	114	270	33	47
	Happiness	192	340	166	1187	99	74	81	236
	Disgust	4	6	4	8	180	1	21	2
	Surprise	82	58	61	32	51	1467	100	35
	Fear	5	6			18	4	160	
	Anger	12	23	4	35	6	13	15	550
		Other	Neutral	Sadness	Happiness	Disgust	Surprise	Fear	Anger
		Predicted Class							

Figure 6. Confusion matrix results of the validation set for all classes. The blue color represents the true positive (TP) predictions.

True Class	Other	8681	2287	744	955	815	332	1036	567
	Neutral	3833	15591	2206	4538	703	1238	1243	500
	Sadness	1062	900	8626	647	502	1124	197	213
	Happiness	870	1332	627	4816	387	264	383	968
	Disgust	20	26	10	28	760	6	52	1
	Surprise	277	147	289	119	188	5821	380	160
	Fear	29	10		3	41	25	686	3
	Anger	64	61	20	147	30	34	59	2166
		Other	Neutral	Sadness	Happiness	Disgust	Surprise	Fear	Anger
		Predicted Class							

Figure 7. Confusion matrix results of the training set for all classes. The blue color represents the true positive (TP) predictions.

To evaluate the agreement between the observed and expected predictions, the Kappa value was calculated. A Kappa of 0.49 was achieved for the proposed model.

4. Discussion

Table 1 shows a strong bias towards the four classes of neutral, other, happiness, and sadness. To mitigate their impact on the model's learning process and ensure a fair representation of the classes, the weighted loss method described in Section 2.4.2 was implemented. The efficacy of this strategy is evident in Figures 4 and 5, showcasing the model's ability to balance feature representation. It demonstrated good performance across both the underrepresented and well-represented classes.

Figure 4 shows that the proposed approach was able to achieve good performance on the selected dataset. The results also demonstrated that the model was not overfitting the training data, since the difference in the measures of the validation and training sets had a margin of less than 5%. The model's ability to achieve strong performance on the underrepresented classes was expected, as there were limited data to learn from, coupled with the adopted weighted loss strategy. The happiness class's weak performance was due to it being mistaken as neutral with an error of 14.32%, as revealed by the confusion matrix. The other and neutral classes were also often confused, leading to their low performance scores. This phenomenon can be attributed to the minute differences in the features and linkages obtained from the facial mesh that make it difficult to distinguish between these classes. These outcomes also suggest emotion recognition via image alone might combine these emotions for simplicity and improved performance, or, in contrast, require extra inputs for full classification, such as voice stress analysis, to better segregate these otherwise less distinguishable emotions.

While the model's overall performance did not reach a high standard, achieving a mean TP accuracy of 58.76% is notable. This outcome is particularly significant given the complexity of the dataset, which encompasses a wide range of facial expressions. The dataset, collected "in-the-wild", mirrors real-world scenarios as closely as possible. In such settings, distinguishing between facial expressions becomes inherently challenging, as there are no standardized poses or specific facial reactions to replicate. Instead, the expressions captured reflect visceral reactions and perceptions, making accurate predictions more complex. The difficulties encountered in distinguishing emotional classes also highlight the intricate nature of both data annotation and facial expression decoding.

The results of the F1-score reveal that the proposed approach has some hurdles to overcome. The approach's limitations were noticed in the fear class, where the model's ability to correctly identify instances belonging to the positive class and to distinguish TP predictions from FP was notably lacking. Such a score suggests significant challenges in accurately classifying the fear class and highlights potential limitations in the model's ability to generalize to unseen data. This outcome was expected as the fear class was the lowest representation, and although the weighted loss was adopted for performance boost and imbalance stabilization, it was not able to improve on the generalizability of unseen data due to the weak representation of this class.

Achieving a 53.07% F1-score, while not considered high, is still acceptable and reasonable in this type of application [44,46–50]. The complexity and inherent ambiguity of emotion recognition, coupled with the dataset's representation, make it challenging to achieve notable performance. However, the proposed approach was able to demonstrate the ability to capture the key patterns from within the data and thereby make informed predictions within the given constraints. This outcome is strengthened by the Kappa score of 0.49, which falls within the moderate range and demonstrates the model's predictive capability that is not equivalent to chance but rather a moderate agreement between the predicted and true outcome. This demonstrates the model's ability to capture complex facial expressions. Additionally, it provides insights into the model's effectiveness and potential areas for improvement.

Table 2 shows the comparison of the proposed approach's results to other methods used during the Aff-Wild2 competition as highlighted in [44]. The proposed approach was able to achieve a higher F1-score compared to other works. This notable performance improvement underscores the efficacy of the method implemented in effectively recognizing and interpreting facial expressions. The margin between the proposed approach and existing methodologies highlights the robustness and potential of the model in using the GCN in emotion recognition. Notably, the model successfully identifies relevant patterns within the data, leading to improved efficacy in emotion recognition. While using a fraction of the Aff-Wild2 dataset, the proposed approach was able to achieve better performance compared to the other works mentioned.

Table 2. Comparison of the F1-score results from different methods on the Aff-Wild2 dataset.

Method	F1-Score
Yu et al. [46]	0.3075
Xue et al. [47]	0.3218
Savchenko [48]	0.3292
Zhang et al. [49]	0.3337
Zhou et al. [50]	0.3532
Proposed approach	0.5307

* Value in bold represents the best performance.

In Figure 8, the models explain-ability is displayed, which highlights the areas of strong impact on the class prediction. As observed, the model concentrates on the forehead of the face for the classification of the other class. The focus on the forehead in emotion recognition holds significance for identifying emotions that deviate from the traditional emotional classes. The forehead is a crucial area for detecting subtle and culturally specific emotional cues. Unique expressions may manifest through slight movements or muscle contractions in the forehead, particularly in instances where emotions are complex or socially influenced. Moreover, cultural variations in facial expression lead to differing associations between emotions and facial regions, with some cultures emphasizing forehead movements for emotional expression. Therefore, by focusing on the forehead, the model was able to achieve improved accuracy and comprehensiveness, enabling a better interpretation of emotional cues beyond the constraints of traditional emotional categories.

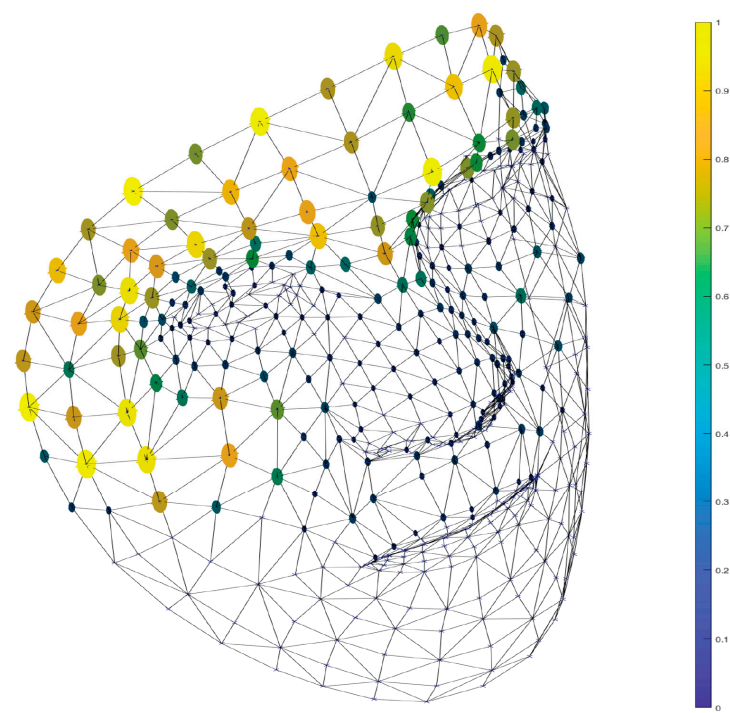


Figure 8. Custom model explainability visualization of a face representing the other class. The color bar represents the intensity of the impact of the particular point landmark, with dark blue representing no influence and yellow a strong influence. The radius of the circle also varies according to the intensity of the impact.

In Figure 9, the model’s focus on the forehead, complemented by its focus on the lips, eyes, and cheeks, underscores its proficiency in distinguishing different emotional expressions. By analyzing multiple facial features, including the lips, eyes, and cheeks, the model achieves enhanced precision and depth in deciphering emotional cues. This refined

approach overcomes the constraints of traditional emotional classifications, allowing for a more detailed and comprehensive understanding of emotional expressions.

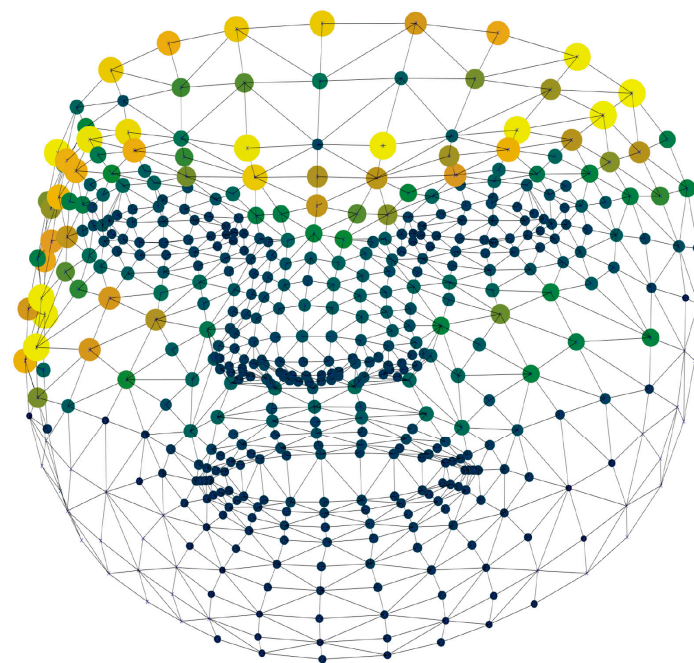


Figure 9. Custom model explainability visualization of a face representing the happiness class. The color bar represents the intensity of the impact of the particular point landmark, with dark blue representing no influence and yellow a strong influence. The radius of the circle also varies according to the intensity of the impact.

The limitations of this approach include no optimization on the hidden layers, partial use of the dataset, no hyperparameter tuning, and the absence of data normalization. The hidden layers and hyperparameters were not fine-tuned in this study, potentially leaving room for further performance enhancement. The absence of data normalization may have had an impact on the model's ability to generalize well to unseen data. The use of a small portion of the database, which restricted the representation of some classes, had a significant impact on the performance. Another limitation is that in this study equal significance was given to all facial point landmarks.

To address these drawbacks, future work will focus on incorporating normalization functions coupled with the fine-tuning of the hidden layers and hyperparameters. The use of a larger portion of the dataset will also be considered by optimizing the model training approach so that the system can effectively handle the associated computational demands. A study that focuses on facial muscle movements, where certain features play a more prominent role in capturing subtle expressions, will also be undertaken.

5. Conclusions

In this study, facial landmarks were used with a graph convolution network (GCN) for facial emotion recognition. The proposed approach showcased the potential of using GCNs for emotion recognition in real-world scenarios. By leveraging the graph-based representation, the model was able to capture intricate relationships between facial expressions, leading to a mean TP classification accuracy of 58.76% and mean F1-score of 53.07% on the selected validation set. Given the inherent challenges of classifying non-posed emotional expressions and the constraints of limited data, the proposed approach yields compelling results, particularly when compared to previous research efforts. Further enhancements are planned for future work, including cross-dataset evaluations for model generalizability assessments.

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

Funding: This research was partially funded by the German Federal Ministry of Research and Education (BMBF) under grant LESSON FKZ: 3FH5E10IA, a grant from KOMPASS funded by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) of Baden-Wuerttemberg Germany, a grant from the ERAPERMED2022-276—ETAP BMG FKZ 2523FSB110, a grant from the New Zealand Ministry of Business, Innovation and Employment (MBIE) under the Catalyst Leaders Grant (UOCX2201), and a grant from the Education New Zealand Programme for Project Related Personal Exchange (PPP): New Zealand–German Academic Exchange (DAAD) Programme under grant AIDE-ASD FKZ 57656657.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The database used in this study was the Aff-Wild2 database. The Aff-Wild2 database (<https://ibug.doc.ic.ac.uk/resources/aff-wild2/>, accessed on 21 August 2023) is available from the publisher upon request.

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

References

1. Dorsey, E.R.; Venkataraman, V.; Grana, M.J.; Bull, M.T.; George, B.P.; Boyd, C.M.; Beck, C.A.; Rajan, B.; Seidmann, A.; Biglan, K.M. Randomized Controlled Clinical Trial of “Virtual House Calls” for Parkinson Disease. *JAMA Neurol.* **2013**, *70*, 565–570. [CrossRef] [PubMed]
2. Campbell, S. From Face-to-Face to FaceTime. *IEEE Pulse* **2020**, *11*, 7–11. [CrossRef] [PubMed]
3. Holder-Pearson, L.; Chase, J.G. Socio-Economic Inequity: Diabetes in New Zealand. *Front. Med.* **2022**, *9*, 756223. [CrossRef] [PubMed]
4. Tebartz van Elst, L.; Fangmeier, T.; Schaller, U.M.; Hennig, O.; Kieser, M.; Koelkebeck, K.; Kuepper, C.; Roessner, V.; Wildgruber, D.; Dziobek, I. FASTER and SCOTT&EVA Trainings for Adults with High-Functioning Autism Spectrum Disorder (ASD): Study Protocol for a Randomized Controlled Trial. *Trials* **2021**, *22*, 261. [PubMed]
5. Rylaarsdam, L.; Guemez-Gamboa, A. Genetic Causes and Modifiers of Autism Spectrum Disorder. *Front. Cell. Neurosci.* **2019**, *13*, 385. [CrossRef] [PubMed]
6. Grifantini, K. Detecting Faces, Saving Lives. *IEEE Pulse* **2020**, *11*, 2–7. [CrossRef]
7. Sandler, A.D.; Brazdziunas, D.; Cooley, W.C.; de Pijem, L.G.; Hirsch, D.; Kastner, T.A.; Kummer, M.E.; Quint, R.D.; Ruppert, E.S. The Pediatrician’s Role in the Diagnosis and Management of Autistic Spectrum Disorder in Children. *Pediatrics* **2001**, *107*, 1221–1226.
8. Golan, O.; Ashwin, E.; Granader, Y.; McClintock, S.; Day, K.; Leggett, V.; Baron-Cohen, S. Enhancing Emotion Recognition in Children with Autism Spectrum Conditions: An Intervention Using Animated Vehicles with Real Emotional Faces. *J. Autism Dev. Disord.* **2010**, *40*, 269–279. [CrossRef] [PubMed]
9. Yuan, S.N.V.; Ip, H.H.S. Using Virtual Reality to Train Emotional and Social Skills in Children with Autism Spectrum Disorder. *Lond. J. Prim. Care* **2018**, *10*, 110–112. [CrossRef] [PubMed]
10. Ravindran, V.; Osgood, M.; Sazawal, V.; Solorzano, R.; Turnacioglu, S. Virtual Reality Support for Joint Attention Using the Floreo Joint Attention Module: Usability and Feasibility Pilot Study. *JMIR Pediatr. Parent.* **2019**, *2*, e14429. [CrossRef] [PubMed]
11. Scherer, K.R. What Are Emotions? And How Can They Be Measured? *Soc. Sci. Inf.* **2005**, *44*, 695–729. [CrossRef]
12. Lang, P.J.; Bradley, M.M. Emotion and the Motivational Brain. *Biol. Psychol.* **2010**, *84*, 437–450. [CrossRef]
13. Vuilleumier, P. How Brains Beware: Neural Mechanisms of Emotional Attention. *Trends Cogn. Sci.* **2005**, *9*, 585–594. [CrossRef] [PubMed]
14. Mancini, C.; Falciani, L.; Maioli, C.; Mirabella, G. Happy Facial Expressions Impair Inhibitory Control with Respect to Fearful Facial Expressions but Only When Task-Relevant. *Emotion* **2022**, *22*, 142. [CrossRef] [PubMed]
15. Mirabella, G.; Grassi, M.; Mezzarobba, S.; Bernardis, P. Angry and Happy Expressions Affect Forward Gait Initiation Only When Task Relevant. *Emotion* **2023**, *23*, 387. [CrossRef] [PubMed]
16. Tautkute, I.; Trzcinski, T.; Bielski, A. I Know How You Feel: Emotion Recognition with Facial Landmarks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1878–1880.
17. Mehrabian, A. Communication without Words. In *Communication Theory*; Mortensen, C.D., Ed.; Routledge: London, UK, 2017; pp. 193–200. ISBN 978-1-315-08091-8.

18. De Gelder, B. Why Bodies? Twelve Reasons for Including Bodily Expressions in Affective Neuroscience. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 3475–3484. [\[CrossRef\]](#) [\[PubMed\]](#)
19. De Gelder, B.; Van den Stock, J.; Meeren, H.K.; Sinke, C.B.; Kret, M.E.; Tamietto, M. Standing up for the Body. Recent Progress in Uncovering the Networks Involved in the Perception of Bodies and Bodily Expressions. *Neurosci. Biobehav. Rev.* **2010**, *34*, 513–527. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Arabian, H.; Battistel, A.; Chase, J.G.; Moeller, K. Attention-Guided Network Model for Image-Based Emotion Recognition. *Appl. Sci.* **2023**, *13*, 10179. [\[CrossRef\]](#)
21. Sepas-Moghaddam, A.; Etemad, A.; Pereira, F.; Correia, P.L. Facial Emotion Recognition Using Light Field Images with Deep Attention-Based Bidirectional LSTM. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3367–3371.
22. Khaireddin, Y.; Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. *arXiv* **2021**, arXiv:2105.03588.
23. Mehendale, N. Facial Emotion Recognition Using Convolutional Neural Networks (FERC). *SN Appl. Sci.* **2020**, *2*, 446. [\[CrossRef\]](#)
24. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-Piloted Deep Network for Facial Expression Recognition. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14. Springer: Berlin/Heidelberg, Germany, 2016; pp. 425–442.
25. Li, Y.; Lu, G.; Li, J.; Zhang, Z.; Zhang, D. Facial Expression Recognition in the Wild Using Multi-Level Features and Attention Mechanisms. *IEEE Trans. Affect. Comput.* **2020**, *14*, 451–462. [\[CrossRef\]](#)
26. Tarnowski, P.; Kołodziej, M.; Majkowski, A.; Rak, R.J. Emotion Recognition Using Facial Expressions. *Procedia Comput. Sci.* **2017**, *108*, 1175–1184. [\[CrossRef\]](#)
27. Ekman, P.; Friesen, W.V. Facial Action Coding System. *Environ. Psychol. Nonverbal Behav.* **1978**. [\[CrossRef\]](#)
28. Goeleven, E.; De Raedt, R.; Leyman, L.; Verschuere, B. The Karolinska Directed Emotional Faces: A Validation Study. *Cogn. Emot.* **2008**, *22*, 1094–1118. [\[CrossRef\]](#)
29. Li, D.; Wang, Z.; Gao, Q.; Song, Y.; Yu, X.; Wang, C. Facial Expression Recognition Based on Electroencephalogram and Facial Landmark Localization. *Technol. Health Care* **2019**, *27*, 373–387. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Siam, A.I.; Soliman, N.F.; Algarni, A.D.; El-Samie, A.; Fathi, E.; Sedik, A. Deploying Machine Learning Techniques for Human Emotion Detection. *Comput. Intell. Neurosci.* **2022**, *2022*, 8032673. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
32. Derrow-Pinion, A.; She, J.; Wong, D.; Lange, O.; Hester, T.; Perez, L.; Nunkesser, M.; Lee, S.; Guo, X.; Wiltshire, B. Eta Prediction with Graph Neural Networks in Google Maps. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, Australia, 1–5 November 2021; pp. 3767–3776.
33. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph Neural Networks: A Review of Methods and Applications. *AI Open* **2020**, *1*, 57–81. [\[CrossRef\]](#)
34. Zafeiriou, S.; Kollias, D.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Kotsia, I. Aff-Wild: Valence and Arousal ‘In-the-Wild’ challenge. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 34–41.
35. Kollias, D.; Tzirakis, P.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Schuller, B.; Kotsia, I.; Zafeiriou, S. Deep Affect Prediction In-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *Int. J. Comput. Vis.* **2019**, *127*, 907–929. [\[CrossRef\]](#)
36. Kollias, D.; Sharmanska, V.; Zafeiriou, S. Face Behavior a La Carte: Expressions, Affect and Action Units in a Single Network. *arXiv* **2019**, arXiv:1910.11111.
37. Kollias, D.; Zafeiriou, S. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and Arcface. *arXiv* **2019**, arXiv:1910.04855.
38. Kollias, D.; Zafeiriou, S. Affect Analysis In-the-Wild: Valence-Arousal, Expressions, Action Units and a Unified Framework. *arXiv* **2021**, arXiv:2103.15792.
39. Kollias, D.; Sharmanska, V.; Zafeiriou, S. Distribution Matching for Heterogeneous Multi-Task Learning: A Large-Scale Face Study. *arXiv* **2021**, arXiv:2105.03790.
40. Kollias, D.; Schulc, A.; Hajiyeve, E.; Zafeiriou, S. Analysing Affective Behavior in the First Abaw 2020 Competition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 637–643.
41. Kollias, D.; Zafeiriou, S. Analysing Affective Behavior in the Second Abaw2 Competition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3652–3660.
42. Kollias, D. Abaw: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Multi-Task Learning Challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2328–2336.
43. Kollias, D. ABAW: Learning from Synthetic Data & Multi-Task Learning Challenges. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 157–172.
44. Kollias, D.; Tzirakis, P.; Baird, A.; Cowen, A.; Zafeiriou, S. Abaw: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5888–5897.

45. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M.; Lee, J. Mediapipe: A Framework for Perceiving and Processing Reality. In Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; Volume 2019.
46. Yu, J.; Cai, Z.; Li, R.; Zhao, G.; Xie, G.; Zhu, J.; Zhu, W.; Ling, Q.; Wang, L.; Wang, C. Exploring Large-Scale Unlabeled Faces to Enhance Facial Expression Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5802–5809.
47. Xue, F.; Sun, Y.; Yang, Y. Exploring Expression-Related Self-Supervised Learning for Affective Behaviour Analysis. *arXiv* **2023**, arXiv:2303.10511.
48. Savchenko, A.V. Emotiefnet Facial Features in Uni-Task Emotion Recognition in Video at Abaw-5 Competition. *arXiv* **2023**, arXiv:2303.09162.
49. Zhang, Z.; An, L.; Cui, Z.; Dong, T. Facial Affect Recognition Based on Transformer Encoder and Audiovisual Fusion for the ABAW5 Challenge. *arXiv* **2023**, arXiv:2303.09158.
50. Zhou, W.; Lu, J.; Xiong, Z.; Wang, W. Leveraging TCN and Transformer for Effective Visual-Audio Fusion in Continuous Emotion Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5755–5762.
51. Zhu, R.; Guo, Y.; Xue, J.-H. Adjusting the Imbalance Ratio by the Dimensionality of Imbalanced Data. *Pattern Recognit. Lett.* **2020**, *133*, 217–223. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.