sensors

Article

Can Hierarchical Transformers Learn Facial Geometry?

Paul Young [1], Nima Ebadi [2], Arun Das [2,3], Mazal Bethany [4], Kevin Desai [1,*] and Peyman Najafirad [1,4]

[1] Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249, USA
[2] Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA
[3] Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15260, USA
[4] Department of Information Systems, University of Texas at San Antonio, San Antonio, TX 78249, USA
* Correspondence: kevin.desai@utsa.edu

Abstract: Human faces are a core part of our identity and expression, and thus, understanding facial geometry is key to capturing this information. Automated systems that seek to make use of this information must have a way of modeling facial features in a way that makes them accessible. Hierarchical, multi-level architectures have the capability of capturing the different resolutions of representation involved. In this work, we propose using a hierarchical transformer architecture as a means of capturing a robust representation of facial geometry. We further demonstrate the versatility of our approach by using this transformer as a backbone to support three facial representation problems: face anti-spoofing, facial expression representation, and deepfake detection. The combination of effective fine-grained details alongside global attention representations makes this architecture an excellent candidate for these facial representation problems. We conduct numerous experiments first showcasing the ability of our approach to address common issues in facial modeling (pose, occlusions, and background variation) and capture facial symmetry, then demonstrating its effectiveness on three supplemental tasks.

Keywords: face geometry; hierarchical transformers; anti-spoofing; facial expression recognition; deepfakes

check for updates

Citation: Young, P.; Ebadi, N.; Das, A.; Bethany, M.; Desai, K.; Najafirad, P. Can Hierarchical Transformers Learn Facial Geometry? *Sensors* **2023**, *23*, 929. https://doi.org/10.3390/s23020929

Academic Editors: Roberto Vezzani, Mohamed Daoudi, Guido Borghi, Marcella Cornia, Claudio Ferrari, Federico Becattini and Andrea Pilzer

Received: 1 December 2022
Revised: 10 January 2023
Accepted: 11 January 2023
Published: 13 January 2023

1. Introduction

Facial representation has always been a core part of computer vision. One of its applications, face detection, was one of the most successful image analysis tasks in the early years of computer vision [1]. Since then, facial representation applications have grown to include many security [2,3], forensics [4], and even medical [5] applications. Facial representation efforts continue to motivate innovations in the computer vision field such as federated learning advancements [6–8] or the introduction of angular softmax loss [9]. The face is a core element of identity and expression. People are identified primarily through their faces and communicate verbally and non-verbally through facial movements. Correspondingly, it is critical for machines attempting to interact with people to have an effective way of representing their faces.

Unexpected changes to the environment often have an adverse effect on computer vision tasks. Implementations that have vastly different background conditions from training data often generalize poorly [10]. These conditions can include variations in occlusion, lighting, background, pose, and more. Mitigating this by collecting target domain data is often costly or impractical. As a result, many efforts have been made to look for ways to mitigate these conditions without additional data. There have been more generalized attempts to address this domain shift [11] as well as specific models and methods tailored to specific tasks.

As a result of these challenges, many authors have worked to overcome these difficulties in various facial tasks [12–16]. Facial geometry is a representation of the physical layout of the face. This is a complex 3D structure that has some unique properties and can be simplified and represented in a variety of ways. Facial 'keypoint' detection is the process of detecting the location of various important parts of the face. The layout of these keypoints can be used to modify, reconstruct, or identify a face [17–19], whereas keypoint-based modeling can be useful for many tasks such as facial expression recognition [20], the lack of fine-grained pixel information makes it unsuitable for such tasks as spoofing or deepfake detection. Another approach uses 3D morphable models to construct a facial geometry representation, where a given face is represented by a combination of various template models [17,21–23]. These models can be further modified for desired deformations. However, morphable models have difficulty when encountering occluded or angled views of the face [24]. Although not fully symmetric, most faces have a certain degree of symmetry that can be exploited for facial representation tasks. This can be used to compensate for occluded information [25] or even to perform recognition with half faces [26]. We seek to capture facial geometry to create a consistent representation irrespective of these changes. Three use cases that can be used to further evaluate the capability of our facial geometry learning are face anti-spoofing, facial expression recognition, and deepfake detection.

The identity component of facial representation corresponds to face recognition and re-identification tasks. These tasks are extensively integrated into biometric security systems [27]. These systems represent an effective method of limiting access based on identity without the need or vulnerability of physical or password-based systems, whereas current face recognition methods generally operate at a high level of accuracy [28], these systems present a vulnerability to spoofing attacks. Spoofing attacks come in many forms, but the most common are photo presentation, video replay, and mask attacks. The reason for this vulnerability is that face recognition systems are trained to differentiate between different identities, not to identify false presentations. If an adversary can use a presentation attack to fool systems, the security of face-recognition-dependent systems may be compromised. The threat of these attacks has motivated many works of research into the field of face anti-spoofing (FAS); whereas the facial domain is the same, the focus for FAS shifts from a global identity to looking for subtle liveness cues such as fine-grained texture [29] or style information [12].

In addition to recognizing the person's identity, facial representations are important for understanding the state of a person. People communicate in many more ways than just the words that are spoken. The expressions we present while interacting shape the context and meaning of the words we speak. To understand communication better, computer systems must learn to capture human emotions through expression. In addition, facial expression recognition (FER) can be used for certain medical diagnosis purposes [30,31]. The understanding of human emotion is heavily tied to multiple areas of the face: eyes, mouth, and between the eyes [32]. A facial representation that understands these expressions must be able to capture these small details while having a larger structure to localize and relate these smaller features.

Human perception can be attacked with facial manipulation techniques. This incorporates many techniques through procedures such as face swapping and expression manipulation. We will refer to this category of attacks as deepfakes. Deepfakes present a substantial threat to various facets of our lives. False criminal or civil accusations can be made with fabricated video as proof. Conversely, the utility of video evidence deteriorates if false videos cannot be detected. Similarly, elections can be swayed by conjured footage of politicians engaged in incriminating activity. Detection of these attacks can be done through the examination of regions of abnormal distortion as well as inconsistencies in the face. When facial manipulation algorithms replace or modify facial identities or expressions, there are regions that span the gap between the modified content and the original content. Deepfake algorithms are trained to make these regions as realistic as possible, but such images are still artificial, generated content with the potential for an unnatural distribution of pixels. Examining the image with sufficient resolution makes it possible to detect the artifacts left by deepfake manipulations.

We propose a deep-learning facial representation architecture that uses a multi-level transformer model to capture fine-grained and global facial characteristics. To capture the physical characteristics of the target in each sample, our proposed method uses a Swin transformer [33] which achieves state-of-the-art results in image classification tasks. The model yields facial embeddings which are used to perform various face-related tasks. Transformer architectures use attention mechanisms to represent relationships between model inputs. These architectures are effective modeling tools, but they suffer from resolution trade-offs due to their large size. The shifted window approach of our selected backbone addresses this problem allowing for better fine-grained modeling which is key for face representation tasks. Figure 1 shows a high-level representation of the proposed hierarchical architecture compared to a standard ViT transformer model. To further validate this capability, we apply this solution to three facial representation tasks: face anti-spoofing, facial expression recognition, and deepfake detection. Our performance studies show that we are able to robustly detect spoofing, deepfake attacks, and human facial expressions.

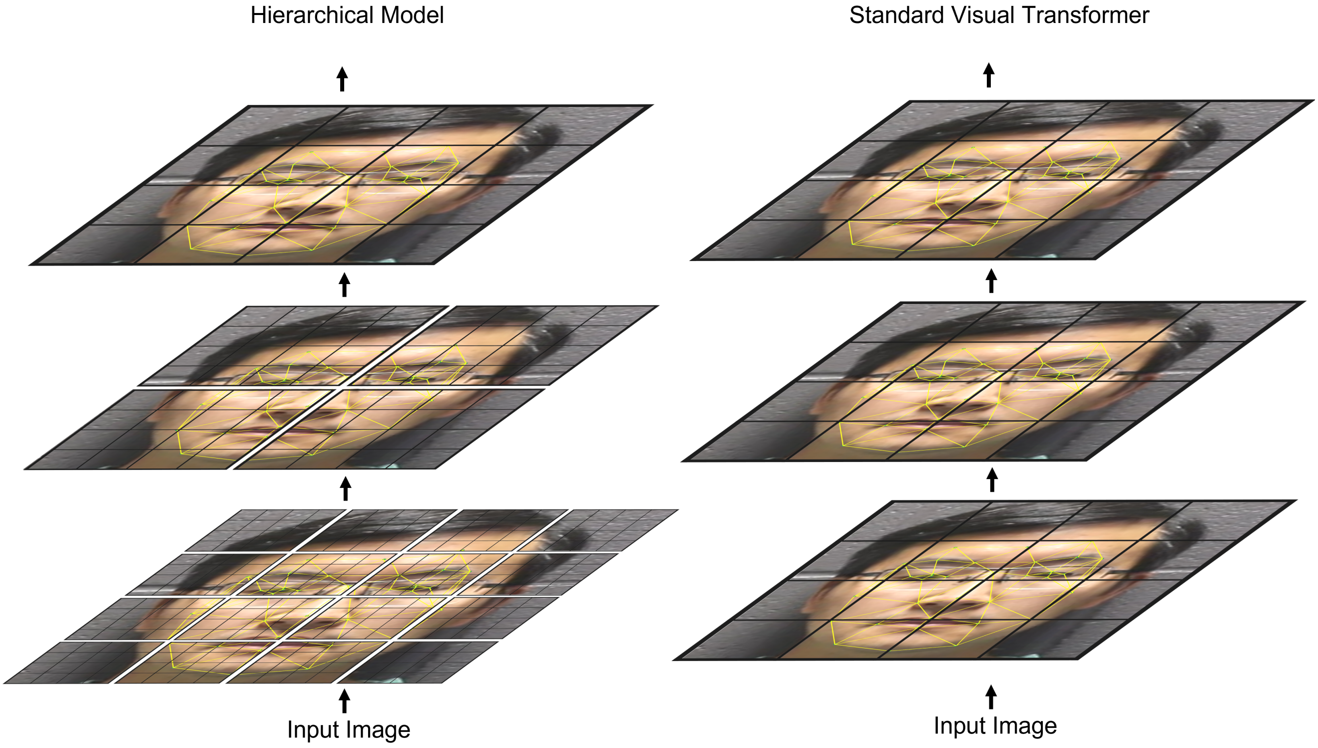

Figure 1. Comparison of the hierarchical model (**left**) to standard visual transformer such as ViT (**right**). In the hierarchical model, the lower layers are embedded as a greater number of smaller patches. Attention for these patches is only calculated within the local patch windows. As the image embedding progresses through the model, patches are gradually merged, and patch windows are expanded to allow for global representations.

Our contributions are summarized as follows:

- We propose a hierarchical transformer model to learn multi-scale details of face geometry. We structure this model to capture both fine-grained details and global relationships.
- We validate the geometry learning capability of our facial representation approach with various tests. This demonstrates its resilience to rotation, occlusion, pose, and background variation.
- We apply and validate the model on three facial representation use cases. The three tasks, face anti-spoofing, facial expression recognition, and deepfake detection, highlight the robustness and versatility of facial geometry representation.

2. Related Work

Various early works involving facial representation used Haar-like features for face detection. The Haar sequence, proposed by Alfred Haar [34], forms the basis for Haar wavelet functions. These functions, which take the shape of a square wave over the unit interval, have been used to detect transitions such as tool failures [35]. Papageorgiou et al. and others [36–38] use this property on images to find boundaries for object detection. Viola and Jones [39] use a similar technique with a cascaded model for face detection. Various adaptions have been made on this such as by Mita et al. [40] which addresses correlated features and Pham et al. [41] which uses statistics for feature selection.

As convolutional neural network (CNN) models permeated computer vision, various CNN solutions for facial representation emerged [42–45]. The increased depth and number of parameters along with CNN strengths of locality and translation invariance have facilitated more sophisticated tasks such as re-identification and face recognition. The increased depth of CNNs allowed for more robust and sophisticated tasks, but tracing gradients through a large number of layers created an exploding/vanishing gradient effect. This effect limited gradient transmission to deeper layers, obstructing training and convergence of models. The introduction of residual connections [46] between layers alleviated this issue and allowed for deeper, more sophisticated ResNet architectures. However, even with this improvement, CNNs still suffer from some drawbacks. Their pooling layers can be lossy, and the convolutional architecture makes relationships between distant image regions more difficult to model.

In addition to their success in NLP, transformers have shown promising results for image processing. The capability to model image relationships with attention allows the model to focus on key regions and relationships in the image. It also better facilitates the modeling of relationships between more distant image regions. The visual transformer (ViT) [47] achieved state-of-the-art performance on image classification tasks. Due to the quadratic growth of attention parameters based on the input size, Dosovitskiy et al. structured ViT to accept relatively large patch sizes of 16 × 16. Various modifications have been made to utilize visual transformers for tasks such as object detection [48–50]. Specifically, two problems have emerged to expand the capability of visual transformers. The first is how to create different-scale feature maps to facilitate tasks such as segmentation and detection. The second is how to attend to fine-grained details without losing global relationships or overly increasing the number of parameters. The solution to both of these problems has appeared in multi-level models. Three such models [33,51,52] have appeared and each performs the following tasks in different ways. They attend to fine-grained information locally, while attending to global relationships on a coarser scale.

Structure learning problems in images require learning information from granular-level data. Transformer models provide fine-grained representation learning with attention mechanisms. The quadratic cost of attention parameters inspires many solutions to address the large size and unique challenges of images [53–57]. The problem distills to effectively modeling fine-grained details while preserving global relationships.

One of the simplest and most common adaptations has been the ViT model [47]. This model splits an image into 256 patches and feeds each patch as an input into a transformer encoder. The large patch sizes needed to limit the number of attention parameters deteriorate the model's effectiveness on finer tasks such as small object detection or FAS. The Swin transformer [33] is able to shrink the size of these patches by limiting attention to local areas. It then achieves global representation by shifting the boundaries of these areas and gradually merging patches together. These smaller patch sizes make it ideal for more fine-grained tasks as it allows the model's parameters to attend to smaller details.

Much of the recent literature on FAS has focused on building models that achieve domain generalization, attack type generalization, or both. Jia et al. [13] use an asymmetric triplet loss to group live embeddings from different domains to facilitate consistent live embeddings in new domains. Wang et al. [58] use CNN autoencoders with decoders to separate the embedding of live and spoof features to different modules. Wang et al. [12]

also use CNN networks to separate style features and perform contrastive learning to suppress domain-specific features; whereas CNN architectures can capture locality information well, their global representations are limited. Similarly, certain methods such as PatchNet [29] forgo global information entirely and opt to focus anti-spoofing efforts on texture information from small patches of an image. On the other hand, there are a couple of anti-spoofing methods that make use of transformer architectures for FAS tasks. George et al. [59] use pretrained vision transformer models for zero-shot FAS capabilities. Similarly, Wang et al. [60] use a transformer model modified to capture temporal information for spoof detection. The large patch sizes of ViT [47] limit the fine-grained representation of these ViT-based models.

The challenge of facial expression recognition (FER) comes from two directions. First, models must ensure expression recognition is not tied to the identity of the individual with the expression. Second, models must learn to navigate intra-expression variance and inter-expression similarities. Zhang et al. [61] separate the identity of the face representation from the facial expression by using a ResNet-based deviation module to subtract out identity features. Ruan et al. [62] use a feature decomposition network and feature relation modules to model the relationships between expressions and mitigate the challenges related to intra-expression and inter-expression appearance. Similar to previous models, these models lack global connections and are therefore limited in their corresponding modeling ability. Xue et al. [32] use a CNN network to extract features and relevant patches and then feed them through a transformer encoder to model relations between the features and classify the image. Hong et al. [63] use a video-modified Swin transformer augmented with optical flow analysis to detect facial microexpressions. The success of this approach on that subset of expression recognition is promising for the broader use of multi-level transformer models in modeling facial expressions.

Zhao et al. [64] look at the deepfake detection problem as a fine-grained texture classification problem. They use attention with texture enhancement to improve detection. Their work sheds light on the utility of effectively modeling fine-grained image details when performing deepfake detection. This emphasis aligns with multiple other approaches which rely on fine-grained artifacts for detection [65–67]. In contrast, Dong et al. [68] compare identities of inner and outer face regions for their detection. These competing needs highlight the value of a combined representation of fine-grained and global relationships. There has also been some research that relates deepfake detection to expression detection. Mazaheri and Roy-Chowdhury [69] use FER to localize expression manipulation, whereas Hosler et al. [70] use discrepancies between detected audio and visual expressions to detect deepfakes. These correlations indicate that a similar model and approach may be appropriate for both problems; whereas Ilyas et al. [71] performed a combined evaluation on the audio and visual elements of a video with a Swin transformer network, it remains to be seen if the task can be performed effectively solely on image elements.

3. Proposed Method

Facial geometry consists of both smaller local details and larger shape and positioning. To model these well, we implement a multi-level transformer architecture that captures fine-grained details at lower layers and regional relationships at higher layers. These layers are connected hierarchically to combine these features into a complete representation. The structure of the hierarchy is determined by two factors: model depth and window size. The model depth is composed of multiple transformer blocks divided by patch merging layers as shown in Figure 2. Each transformer block consists of a multi-headed self-attention (MSA) layer and a multi-layer perceptron each with layer normalization and a skip connection. The increasing scale of these layers captures different levels of geometry features throughout the model. The window size is the number of inputs along one dimension which are attended to by each input. For example, with a window size of 7, each input would attend to a 7 × 7 window of inputs. In addition, the window size hyperparameter also determines the number of windows at each layer by virtue of it dividing up the number of inputs at each layer. Thus, if a layer with window size 7 receives a 56 × 56 input, it would contain 64 attention windows arranged in an 8 × 8 pattern. This change in window count affects how quickly the patch merging process achieves global attention since the number of windows is quartered at every patch merging stage. Figure 3 visualizes the feature scales using a patch size of 7.

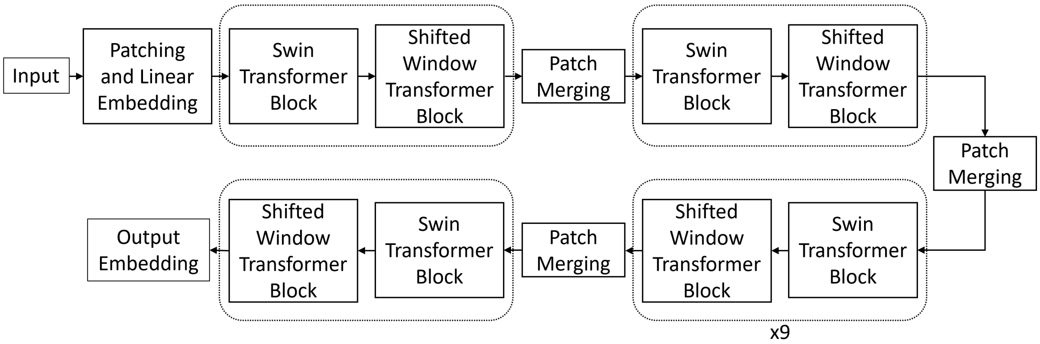
Input
Patching and Linear Embedding
Swin Transformer Block
Shifted Window Transformer Block
Patch Merging
Swin Transformer Block
Shifted Window Transformer Block
Patch Merging
Swin Transformer Block
Shifted Window Transformer Block
x9
Patch Merging
Swin Transformer Block
Shifted Window Transformer Block
Output Embedding

Figure 2. Diagram of the architecture of transformer backbone. Each transformer block has its attention parameters partitioned by windows as shown in Figure 4.

As described by Vaswani et al. [72], MSA can be defined by the expression

$$\mathrm{MSA}(Q, K, V) = \mathrm{Concat}(head_1, \ldots, head_h)W^O \tag{1}$$

where

$$head_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

Q, K, and V refer to the query, key, and value vectors derived from each input by matrix multiplication. Each W variable represents a different set of weights learned during training.

For each of the facial geometry tasks, the initial embeddings are created by dividing the input image into 4 × 4 patches. All three channels of the pixel values for these patches are concatenated into a 48-length vector. Following Dosovitskiy et al. [47], this vector is embedded linearly and given a positional encoding according to this equation:

$$z_0 = [x_1\mathbf{E}; x_2\mathbf{E}; \ldots; x_N\mathbf{E}] + \mathbf{E}_{pos} \tag{3}$$

Each x term represents one element of the patch pixel value vector, $\mathbf{E}$ is the linear embedding matrix, $\mathbf{E}_{pos}$ is the positional encoding term, and z_0 represents the final patch embedding. After the linear embedding and positional encoding, these patches are input into the transformer. The attention is localized to a $M \times M$ window of patches, where M is the window size. These windows are connected through a process in which window borders are shifted in subsequent layers. This places previously separated nearby patches into the same window, allowing for representation across patches, as shown in Figure 4. As the features move through the layers, these patches are gradually merged, reducing the resolution of the features while broadening the scope of the local windows. This continues until the entire representation fits within one window. Algorithm 1 breaks down this process in a step-by-step manner. Finally, the embedding is sent through a classification head according to the task required.

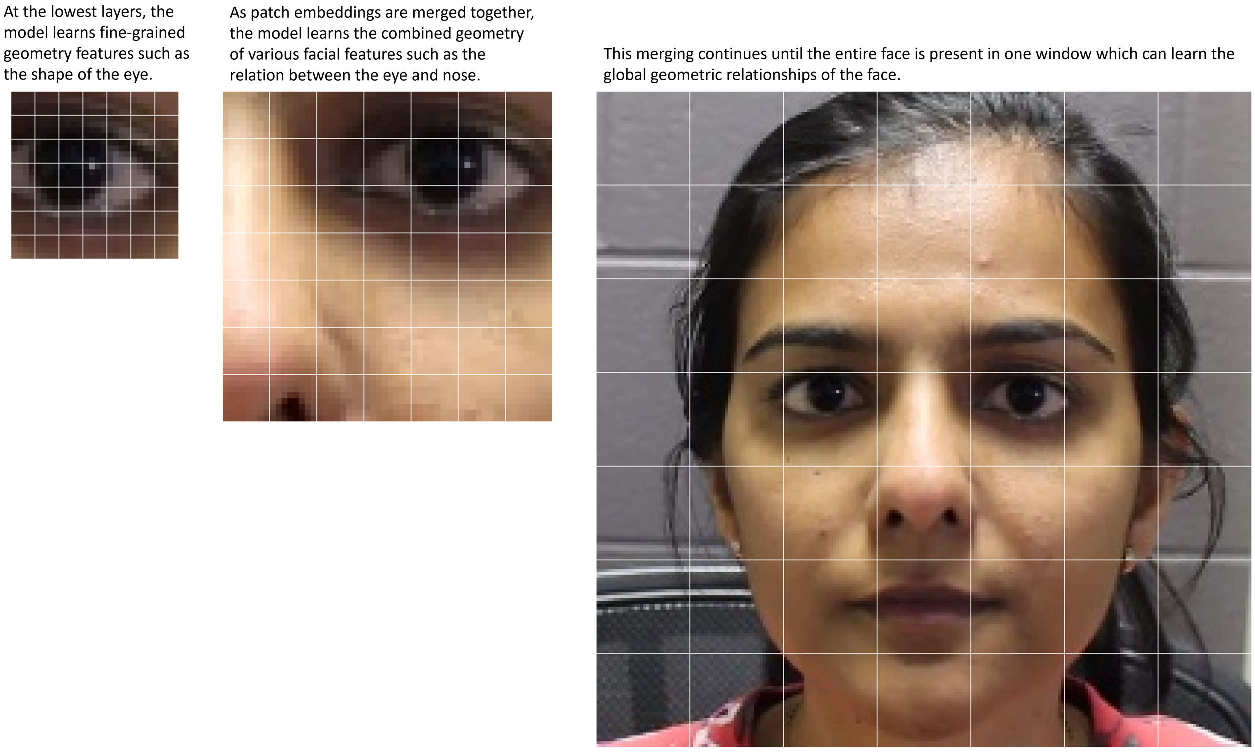
At the lowest layers, the model learns fine-grained geometry features such as the shape of the eye.
As patch embeddings are merged together, the model learns the combined geometry of various facial features such as the relation between the eye and nose.
This merging continues until the entire face is present in one window which can learn the global geometric relationships of the face.

Figure 3. The hierarchical architecture allows the different layers to learn different parts of the facial geometry. The illustration is based on a window size of 7.

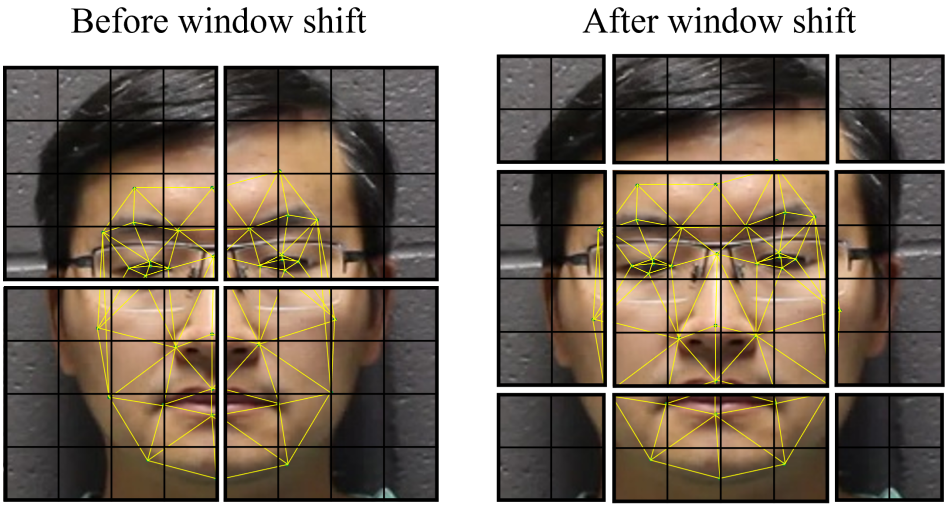

Figure 4. Because of localized attention windows, a mechanism is needed to represent fine-grained relationships across window borders. Window shifting provides this mechanism. By shifting the attention windows borders every other layer, we can model the relationships between each image patch and all nearby patches.

3.1. Use Cases

For the face anti-spoofing problem, we hypothesize that spoof-related features can be found in both the fine-grained details and the global representations. The fine-grain details include unnatural texture, reflection, or lighting. Global cues involve unnatural bowing or skew. We use our hierarchical architecture as a backbone for binary classification. The detailed representation layers give us the capability to detect based on fine cues, whereas the coarser layers enable the discovery of global spoofing cues. For training and inference, live and spoof images are sampled from their corresponding videos and classified through the model.

Algorithm 1 Face geometry representation using hierarchical shifted windows architecture

```
Input: $P = \{p_x^1, p_x^2, \ldots, p_x^n\}$ where $p_x^n$ is the $n$th $4 \times 4$ patch of image $x$.
    $\mathbf{E}$ = learned linear embedding matrix.
    $\mathbf{E}_{pos}$ = positional encoding matrix.
    MSA, SW-MSA = multiheaded self-attention and shifted window MSA.
    MLP = multi-layer perceptron
    LN = layer normalization
Output: Face Representation Classification
for $p \in P$ do                                                    ▷ For each patch
    $p \leftarrow \text{flatten}(p)$
    $p \leftarrow [p_f^1\mathbf{E}; p_f^2\mathbf{E}; \ldots; p_f^{48}\mathbf{E}]$
    $p \leftarrow p + \mathbf{E}_{pos}$
end for
$X \leftarrow P$
for block pair in transformer blocks do
    $X \leftarrow \text{MSA}(\text{LN}(X)) + X$
    $X \leftarrow \text{MLP}(\text{LN}(X)) + X$
    $X \leftarrow \text{SW-MSA}(\text{LN}(X)) + X$
    $X \leftarrow \text{MLP}(\text{LN}(X)) + X$
If at block pair 1, 2, 11:
    for $x_{1,1} \ldots x_{M,N} \in X$ do
        $x_{m,n} \leftarrow \text{merge}(x_{2m,2n}, x_{2m+1,2n}, x_{2m,2n+1}, x_{2m+1,2n+1})$
    end for
end for
```

Similarly, deepfake detection also makes use of these diverse layers. Deepfake cues can be found both in fine-grained textures as demonstrated by Zhao et al. [64] or in larger representations, as shown by the regional identity discrepancy work of Dong et al. [68]. As with the FAS problem, we use a classification of image frames with our hierarchical architecture to detect deepfake attacks.

Facial expression recognition is somewhat different as most of the clues come from certain key regions (eyes, mouth, brow) [32]. However, these regions are not always located exactly in the same location, and thus localizing representations are needed. Furthermore, the combination of regional expressions is needed as different expressions can exhibit similar facial movements [62]. By using the aforementioned hierarchical transformer in conjunction with a multi-label classifier, we can use the various layers of features together to address these detection challenges.

3.2. Training

For the FAS, FER, and deepfake experiments, we fine-tuned the Swin transformer using binary and multi-class cross-entropy loss with one additional fully connected layer. Cross-entropy is an entropy-based measurement of the difference between two inputs. Specifically, it refers to the amount of information required to encode one input using the other. The cross-entropy loss for a given class n can be found by the equation

$$l_n = -\sum_{c=1}^{C} w_c \log \frac{\exp x_{n,c}}{\sum_{i=1}^{C} \exp x_{n,i}} \tag{4}$$

where x, y, w, and C represent the input, target, weight, and number of classes, respectively. For binary problems, this simplifies to

$$l_n = -w_n[y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)] \tag{5}$$

We selected the AdamW [73] optimizer with β_1 = 0.9, β_2 = 0.999 and weight decay = 0.01 for all training purposes.

4. Experiments

We test the symmetry and robustness capabilities of our facial representations with multiple experiments. The effect of pose and occlusion variance is tested by comparing the embeddings using one quantitative and one qualitative experiment. We also perform one separate experiment measuring the effect of background variation. Then, we examine the capability of our facial geometry modeling on three use cases, FAS, FER, and deepfake detection. Finally, we further explore and test the limits of the symmetry modeling capability through two occlusion-based experiments.

4.1. Machine Specification and Parameters

Experiments were performed on a Tesla V100-SXM2 GPU with the assistance of up to 16 Intel(R) Xeon(R) Gold 6152 CPU @ 2.10 GHz processors. For embedding experiments, and pretraining for additional experiments, we used a model pretrained on facial recognition from the FaceX-Zoo suite [74]. This locked the layer count to 2, 2, 18, 2, and the window size to 7. When training, we varied the number of frozen pretrained layers, the learning rate, and the number of training epochs. Testing was performed using train/test splits either built into the dataset (SiW) or created. Deepfake detection experiments were performed with 150 frozen layers, with a learning rate of 0.0001, and a training time of 40 epochs. Facial expression recognition was performed with 200 frozen layers, a learning rate of 0.001, and 30 training epochs. All SiW protocols were performed with a learning rate of 0.0001. Protocol 1 used 200 frozen layers and 10 epochs, protocol 2 used 200 frozen layers and 30 epochs, and protocol 3 used 200 frozen layers and 20 epochs.

4.2. Data Preparation

Pose and occlusion variance experiments were performed using selected images sampled from the SiW dataset [75] because of the accessibility of varied facial poses. Details of the selected frames are available in the code. Background variation experiments were performed using selected images sampled from the OULU-NPU datset [76], due to the variation in the image backgrounds.

For the use case of the facial representation capabilities of our approach, we tested our approach on three additional face-related tasks: face anti-spoofing, deepfake detection, and facial expression recognition. For the face anti-spoofing task, we used the SiW dataset [75]. SiW consists of 4478 videos of 165 subjects divided between live and spoofing videos. For the deepfake detection task, we tested the FaceForensics++ (FF++) dataset [77]. FF++ has 9431 videos consisting of 8068 attack videos as well as 1363 benign videos. The attack videos are split into five categories based on the technique used to generate them: DeepFake, Face2Face, FaceShifter, FaceSwap, and NeuralTextures. Finally, for the facial expression recognition task, we used CK+ dataset [78]. The CK+ dataset contains 593 emotion frame sequences, but we only use the 327 sequences which have labels associated with them. These images exhibit 8 emotion categories: neutral, anger, contempt, disgust, fear, happiness, sadness, and surprise. Details of these three datasets can be found in Table 1.

Table 1. Dataset descriptions: * Recorded actors only. YouTube video count not included.

Dataset	Subjects	Dataset Type	# Samples	Distribution
SiW	165	FAS	4620	1320 Live 3300 Spoof
FF++	26 *	DeepFake Detection	9431	1363 Live 8068 Manipulated
CK+	123	FER	10,735	8 Expression Categories

For testing on the video datasets, four frames were selected at random from each video. Each sequence in the CK+ dataset progresses from a neutral expression to the most expressive. The final three frames of each sequence were selected for the corresponding emotion category, and the first frame was used as the neutral category. The dataset was divided into train and test sets with a roughly 70/30 split. The frames were cropped to each subject's face using the facial detection and cropping code in the FaceX-Zoo [74] suite. Detecting and cropping the face with this method narrows the scope of the problem to images with a single centered face image. This helps reduce the effect of multiple objects in the input image. The resulting images contained 224 × 224 pixels.

4.3. Results

To evaluate the pose and occlusion variance quantitatively, we compare the cosine similarity of four embeddings. The first embedding is one generated from an ordinary frontal image of a person. The second embedding comes from an askance image of the same person's face. The third is from the original face with the occlusion mask placed over one eye. For comparison, we select the fourth image from a different person. The similarity among these embeddings is presented in Table 2. Note the large numbers relating images of the same person compared to that of another individual; whereas the askance embeddings show some variance from the originals, it is far less than the comparison to another individual.

Table 2. Correlation matrix comparing average embedding distance between individuals, including occlusions and varying pose. Note the high correlation between representations of the same person and the low correlation between different people.

	Subject	**Askance**	**Occluded**	**Second Subject**
Subject	1	0.6957	0.9342	0.0995
Askance	0.6957	1	0.6694	0.109
Occluded	0.9342	0.6694	1	0.0944
Second Subject	0.0995	0.109	0.0944	1

For a qualitative measure of pose and occlusion variance, Figure 5 presents some examples from these occluded and askance samples. Figure 6 gives a tSNE visualization of the closeness of the intra-person embeddings for these images compared to the inter-person distances. Here you can see groupings of points representing original and alternative images of the same person labeled as the same color. The alternatives are produced either by occluding the original image (represented by the +) or selecting an image with a different pose (represented by ×). The grouping of intra-person embeddings and the separation of inter-person embeddings demonstrate the robustness of our approach to occlusion and pose.

Table 3 shows a comparison between embeddings across individuals and backgrounds. The high correlation between embeddings generated from the same person with different backgrounds shows the robustness of the proposed approach to background variation. The vastly lower correlation to embeddings from other people of the same background confirms its ability to filter out background information when performing facial geometry modeling.

Table 3. Testing robustness to changes in the background. Compares the cosine similarity of embeddings from each image to itself, an image from the same person with different background, an image from a different person with the same background, and an image from a different person and background. Used a selection of 60 images from 20 different actors in 3 background scenes. Each score represents the average of the comparisons taken. Images were taken from the OULU-NPU dataset [76].

	Cosine Similarity
Self	1
Different Background	0.76
Different Person	0.12
Different Person and Background	0.11

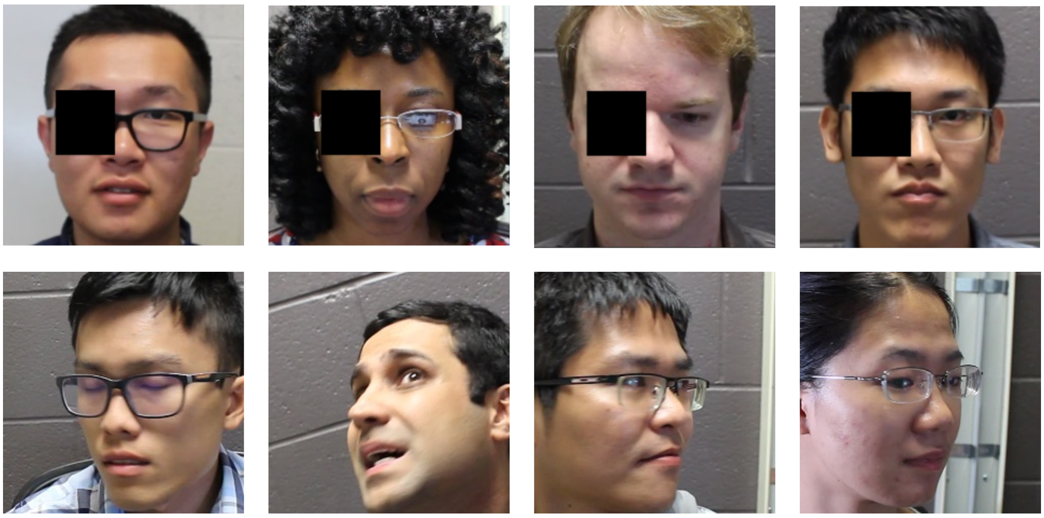

Figure 5. We test facial representations and symmetry capability with various modifications such as occlusions and varying facial poses. Here are examples of individuals both with eye occlusions as well as askance facial poses.

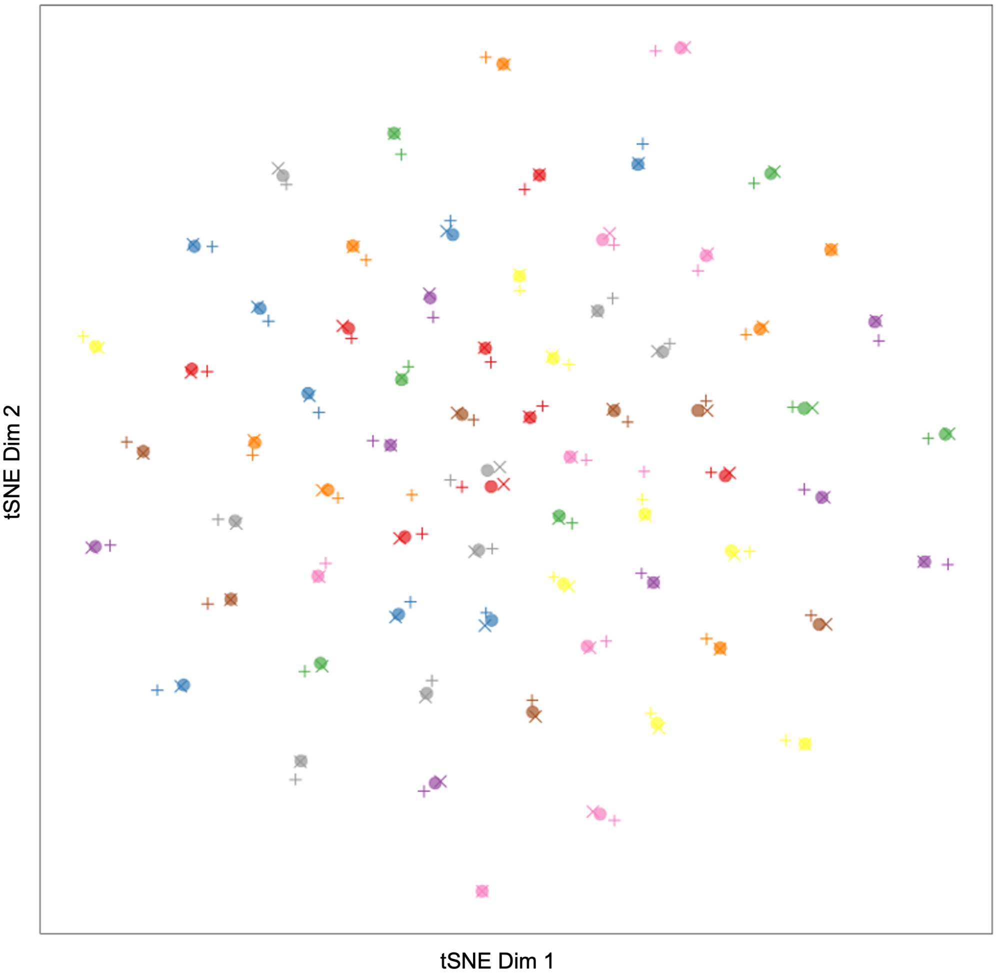

Figure 6. tSNE of facial embeddings from 69 individuals to show robustness to pose and occlusion. Frontal face images are represented with dots, askance images with +, and occluded images with ×. Images from the same individual are shown as the same color (with some color reuse due to a limited color palette). Note the clustering of points relating to single identities regardless of occlusion or pose.

Table 4 compares our results with existing works on deepfake and expression recognition. Table 5 compares anti-spoofing capability on the three protocols of the SiW dataset. Protocol 1 tests pose invariance, training on frontal views, and testing on a variety of poses. Protocol 2 performs a four-part leave-one-out strategy for the replay device. Protocol 3 tests the capability of unseen attack types by training on either print or video attacks and testing on the other.

Table 4. Comparing AUC results (left) on FF++ [77] dataset and accuracy results (right) on CK+ [78] dataset against existing methods. Our approach shows its effectiveness in both deepfake detection and facial expression recognition.

Method	AUC	Method	Accuracy
MADD [64]	0.998	Ruan et al. [62]	0.995
Nirkin et al. [79]	0.997	PPDN [80]	0.973
Face X-ray [81]	0.984	IPA2LT [82]	0.924
Chen et al. [83]	0.984	DeRL [84]	0.974
SPSL [85]	0.969	FN2EN [86]	0.986
SMIL [87]	0.968	DDL [88]	0.992
Ours	0.943	Ours	0.957

Table 5. Comparison on SiW protocols for face anti-spoofing task. Protocol 1 tests anti-spoofing with unseen poses, protocol 2 tests it with varying replay mediums, and protocol 3 tests on unseen attack types (print vs. video). Our facial geometry representation is sensitive enough to transfer to fine tasks such as anti-spoofing.

Protocol	Method	APCER (%)	BPCER (%)	ACER (%)
1	ResNet(CeFA) [89]	1.03	0.83	0.93
	Yang et al. [90]	-	-	0.30
	Wang et al. [91]	0.64	0.17	0.40
	Wang et al. [58]	0.00	0.00	0.00
	PatchNet [29]	0.00	0.00	0.00
	Wang et al. [60]	0.00	0.00	0.00
	Ours	0.96	0.67	0.82
2	ResNet(CeFA) [89]	0.20 ± 0.11	0.25 ± .022	0.23 ± 0.15
	Yang et al. [90]	-	-	0.15 ± 0.05
	Wang et al. [91]	0.00 ± 0.00	0.04 ± 0.08	0.02 ± 0.04
	Wang et al. [58]	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	PatchNet [29]	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	Wang et al. [60]	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	Ours	1.73 ± 1.29	1.60 ± 0.86	1.67 ± 0.81
3	ResNet(CeFA) [89]	6.35 ± 3.67	6.72 ± 3.75	6.57 ± 3.46
	Yang et al. [90]	-	-	5.85 ± 0.85
	Wang et al. [91]	2.63 ± 3.72	2.92 ± 3.42	2.78 ± 3.57
	Wang et al. [58]	4.77 ± 5.04	2.44 ± 2.74	3.58 ± 3.93
	PatchNet [29]	3.06 ± 1.10	1.83 ± 0.83	2.45 ± 0.45
	Wang et al. [60]	2.69 ± 2.05	2.67 ± 2.00	2.68 ± 2.03
	Ours	16.81 ± 1.66	5.03 ± 4.24	10.92 ± 1.29

5. Discussion

5.1. Strengths

The experiments have shown that a hierarchical transformer architecture learns a robust facial geometry representation. As shown in Figure 7, we prepare our approach to give a consistent performance with different poses and occlusions. The pose and occlusion experiments demonstrate that our model is robust against missing information and that it can extrapolate information using facial geometry representations. Similarly, the occlusion experiments demonstrate that it can use facial symmetry to work around missing information to form a consistent representation. To explore the symmetry and occlusion robustness of our approach, we performed a gradual horizontal occlusion of an image and captured the embedding outputs. Figure 8 graphs the cosine similarity between the embedding of the original and the occluded image as occlusion increases. The ability to effectively use symmetry to model facial geometry is shown by the stark contrast before and after the 50% occlusion mark. This is validated by a similar experiment performed with vertical occlusion in Figure 9. When symmetry is not present, the similarity drops far more rapidly than in the previous experiment. This shows the role that symmetry plays in accounting for missing facial geometry information.

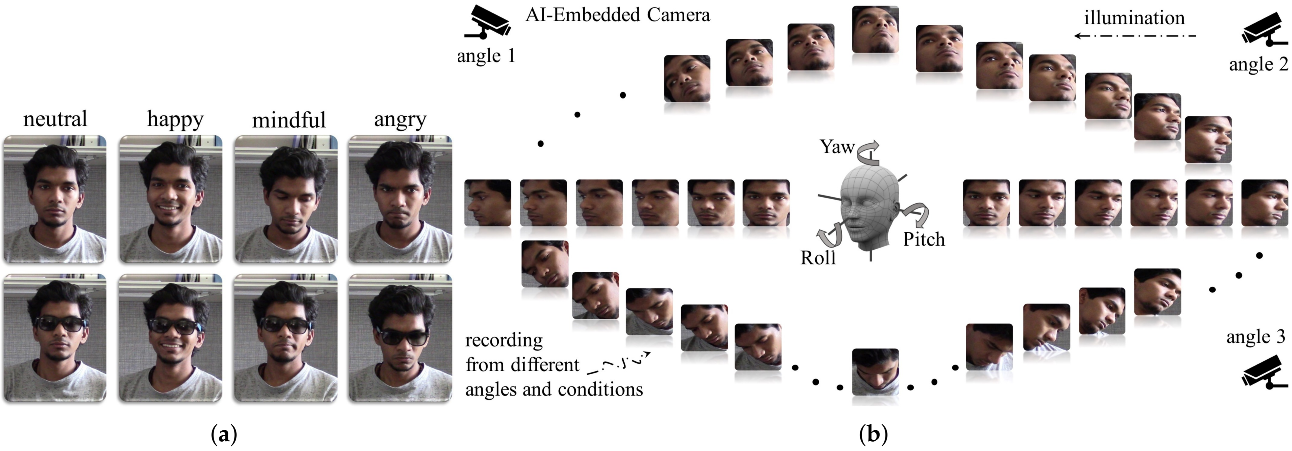

Figure 7. We structure and train our model to be robust to variance in pose, lighting, and occlusion. Occlusions and expressions are illustrated in (**a**), and different Yaw, Pitch, and Roll variations are illustrated in (**b**).

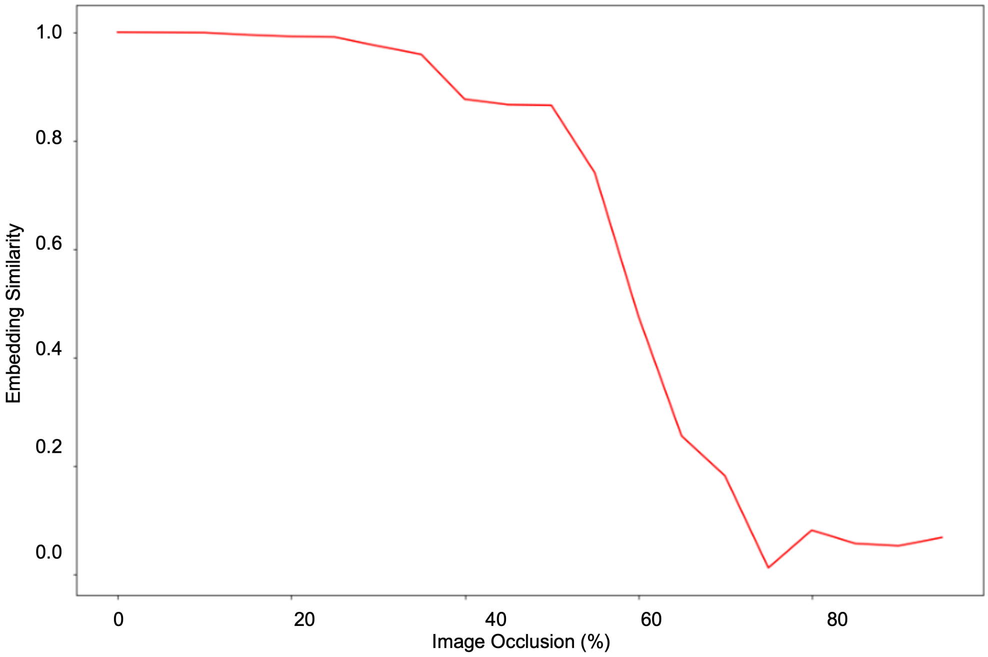

Figure 8. Test of the limit of occlusion robustness and symmetry capture capability. The embeddings of an original image and an occluded image are compared at various levels of horizontal occlusion. The y-axis represents the cosine similarity between the embeddings whereas the x-axis represents the percent of the image occluded

The additional use cases further highlight the versatility of the approach. Anti-spoofing, expression recognition, and deepfake detection examine more specialized and localized regions. The demonstrated capability on these tasks in addition to the global identity representation shows the hierarchical transformer's ability to pivot to more specialized facial representation applications without much alteration.

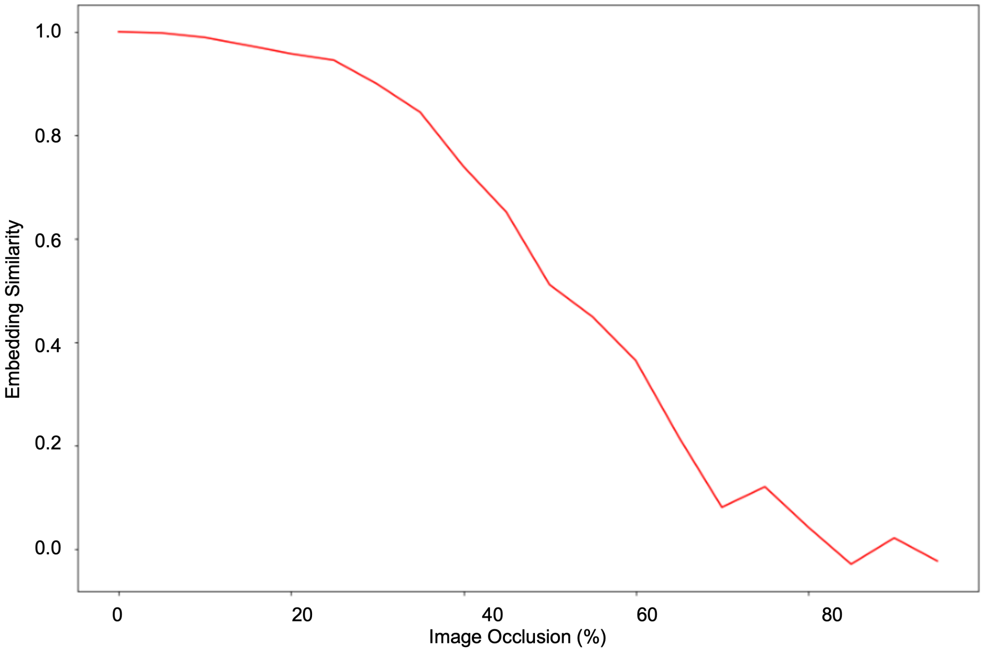

Figure 9. Testing gradual vertical occlusion robustness. Note the cosine similarity falls off far sooner than the horizontal occlusion experiment

5.2. Limitations and Future Work

The third SiW protocol for the FAS task showed comparatively poor results. This protocol involves testing on unseen attack types (print vs. video). This domain generalization problem is a common and difficult problem that often requires specialized model augmentation to address it. Investigating how the hierarchical transformer can be augmented to deal with this domain generalization problem is a topic of further study.

The facial geometry modeling of this method is generated from a single image or video frame. This speeds up computation and makes the model usable with a larger variety of inputs. However, it loses the ability to capture time-related facial features such as motion. Various facial representation tasks involve motion that could be useful for classification, such as the movement of the mouth or eyes in expression recognition. To extend the functionality of this method to capture these details, our approach could be expanded to include temporal features in its decision-making. It may be worth exploring the merits of either a direct temporal expansion of the model or some augmented approach such as optical flow.

6. Conclusions

In this paper, we proposed a hierarchical architecture for capturing facial geometry features. This model's ability to model both fine-grained details and global relationships makes it versatile in addressing a wide range of facial representation tasks. We demonstrated its symmetry and robust modeling capability through a series of experiments. First, we compared embeddings of various circumstances (occlusions, pose variation, and differences in the background). The consistency of the embeddings generated demonstrated the robustness of our approach to disturbances. Next, we tested symmetry modeling with a sliding window experiment. The sharp contrast between occluding less than half the face and more than half illustrated the facial symmetry modeling capabilities. Finally, we further demonstrated the flexibility of the approach by applying it to various facial representation tasks. These tasks, anti-spoofing, facial expression recognition, and deepfake detection, showcase the different ways this facial geometry modeling can be applied to problems. The results on anti-spoofing and deepfake detection showed its ability on fine-grained details whereas facial expression recognition demonstrated its ability on broader facial representation.

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

Funding: This material is based on research sponsored by the Department of Homeland Security (DHS), United States Secret Service, and National Computer Forensics Institute (NCFI) via contract number 70US0920D70090004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The SiW dataset can be found and used with permission at http://cvlab.cse.msu.edu/siw-spoof-in-the-wild-database.html. The CK+ dataset can be found and used with permission at http://www.jeffcohn.net/Resources. The FF++ dataset can be found and used with permission at https://github.com/ondyari/FaceForensics. Paper code is located at https://secureaiautonomylab.github.io.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ACER	Average Classification Error Rate
APCER	Attack Presentation Classification Error Rate
BPCER	Bona fide Presentation Classification Error Rate
CNN	Convolutional Neural Network
DOAJ	Directory of Open Access Journals
FAS	Face Anti-Spoofing
FER	Facial Expression Recognition
MDPI	Multidisciplinary Digital Publishing Institute
MSA	Multi-headed Self-Attention
ViT	Visual Transformer

References

1. Zhao, W.; Chellappa, R.; Phillips, P.J.; Rosenfeld, A. Face recognition: A literature survey. *ACM Comput. Surv. CSUR* **2003**, *35*, 399–458. [CrossRef]
2. Kortli, Y.; Jridi, M.; Al Falou, A.; Atri, M. Face recognition systems: A survey. *Sensors* **2020**, *20*, 342. [CrossRef] [PubMed]
3. Galbally, J.; Marcel, S.; Fierrez, J. Biometric antispoofing methods: A survey in face recognition. *IEEE Access* **2014**, *2*, 1530–1552. [CrossRef]
4. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
5. Thevenot, J.; López, M.B.; Hadid, A. A survey on computer vision for assistive medical diagnosis from faces. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 1497–1511. [CrossRef]
6. Meng, Q.; Zhou, F.; Ren, H.; Feng, T.; Liu, G.; Lin, Y. Improving Federated Learning Face Recognition via Privacy-Agnostic Clusters. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
7. Liu, C.T.; Wang, C.Y.; Chien, S.Y.; Lai, S.H. FedFR: Joint optimization federated framework for generic and personalized face recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 22 February–1 March 2022; Volume 36, pp. 1656–1664.

8. Shome, D.; Kar, T. FedAffect: Few-shot federated learning for facial expression recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4168–4175.
9. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. Sphereface: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 212–220.
10. Hsu, H.K.; Yao, C.H.; Tsai, Y.H.; Hung, W.C.; Tseng, H.Y.; Singh, M.; Yang, M.H. Progressive domain adaptation for object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 2–5 March 2020; pp. 749–757.
11. Tian, J.; Hsu, Y.C.; Shen, Y.; Jin, H.; Kira, Z. Exploring Covariate and Concept Shift for Out-of-Distribution Detection. In Proceedings of the NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, Virtual Event, 13 December 2021.
12. Wang, Z.; Wang, Z.; Yu, Z.; Deng, W.; Li, J.; Gao, T.; Wang, Z. Domain Generalization via Shuffled Style Assembly for Face Anti-Spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 4123–4133.
13. Jia, Y.; Zhang, J.; Shan, S.; Chen, X. Single-side domain generalization for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8484–8493.
14. Yang, C.; Lim, S.N. One-shot domain adaptation for face generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5921–5930.
15. Cao, D.; Zhu, X.; Huang, X.; Guo, J.; Lei, Z. Domain balancing: Face recognition on long-tailed domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5671–5679.
16. Wang, G.; Han, H.; Shan, S.; Chen, X. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6678–6687.
17. Zhu, X.; Lei, Z.; Yan, J.; Yi, D.; Li, S.Z. High-fidelity pose and expression normalization for face recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Quebec City, QC, Canada, 27–30 September 2015; pp. 787–796.
18. Dou, P.; Wu, Y.; Shah, S.K.; Kakadiaris, I.A. Robust 3D face shape reconstruction from single images via two-fold coupled structure learning. In Proceedings of the British Machine Vision Conference, Vancouver, BC, Canada, 22–24 August 2014; pp. 1–13.
19. Kemelmacher-Shlizerman, I.; Basri, R. 3D face reconstruction from a single image using a single reference face shape. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 394–405. [CrossRef]
20. Murugappan, M.; Mutawa, A. Facial geometric feature extraction based emotional expression classification using machine learning algorithms. *PLoS ONE* **2021**, *16*, e0247131.
21. Blanz, V.; Vetter, T. A morphable model for the synthesis of 3D faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 8–13 August 1999; pp. 187–194.
22. Breuer, P.; Kim, K.I.; Kienzle, W.; Scholkopf, B.; Blanz, V. Automatic 3D face reconstruction from single images or video. In Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, San Diego, CA, USA, 12–15 October 2008; pp. 1–8.
23. Saito, S.; Wei, L.; Hu, L.; Nagano, K.; Li, H. Photorealistic facial texture inference using deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5144–5153.
24. Hassner, T.; Harel, S.; Paz, E.; Enbar, R. Effective face frontalization in unconstrained images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4295–4304.
25. Passalis, G.; Perakis, P.; Theoharis, T.; Kakadiaris, I.A. Using facial symmetry to handle pose variations in real-world 3D face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1938–1951. [CrossRef]
26. Singh, A.K.; Nandi, G.C. Face recognition using facial symmetry. In Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology, Coimbatore, India, 26–28 October 2012; pp. 550–554.
27. Galterio, M.G.; Shavit, S.A.; Hayajneh, T. A review of facial biometrics security for smart devices. *Computers* **2018**, *7*, 37. [CrossRef]
28. Meng, Q.; Zhao, S.; Huang, Z.; Zhou, F. Magface: A universal representation for face recognition and quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14225–14234.
29. Wang, C.Y.; Lu, Y.D.; Yang, S.T.; Lai, S.H. PatchNet: A Simple Face Anti-Spoofing Framework via Fine-Grained Patch Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 20281–20290.
30. Brahnam, S.; Chuang, C.F.; Sexton, R.S.; Shih, F.Y. Machine assessment of neonatal facial expressions of acute pain. *Decis. Support Syst.* **2007**, *43*, 1242–1254. [CrossRef]
31. Das, A.; Mock, J.; Huang, Y.; Golob, E.; Najafirad, P. Interpretable self-supervised facial micro-expression learning to predict cognitive state and neurological disorders. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 818–826.
32. Xue, F.; Wang, Q.; Guo, G. Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 3601–3610.

33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 10012–10022.
34. Haar, A. *Zur Theorie der Orthogonalen Funktionensysteme*; Georg-August-Universitat: Gottingen, Germany, 1909.
35. Lee, B.Y.; Tang, Y.S. Application of the discrete wavelet transform to the monitoring of tool failure in end milling using the spindle motor current. *Int. J. Adv. Manuf. Technol.* **1999**, *15*, 238–243. [CrossRef]
36. Papageorgiou, C.P.; Oren, M.; Poggio, T. A general framework for object detection. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 7 January 1998; pp. 555–562.
37. Lienhart, R.; Maydt, J. An extended set of haar-like features for rapid object detection. In Proceedings of the International Conference on Image Processing, Bordeaux, France, 16–19 October 2002; Volume 1, p. I.
38. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
39. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
40. Mita, T.; Kaneko, T.; Hori, O. Joint haar-like features for face detection. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Nice, France, 13–16 October 2005; Volume 2, pp. 1619–1626.
41. Pham, M.T.; Cham, T.J. Fast training and selection of haar features using statistics in boosting-based face detection. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 4–21 October 2007; pp. 1–7.
42. Hu, G.; Yang, Y.; Yi, D.; Kittler, J.; Christmas, W.; Li, S.Z.; Hospedales, T. When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 142–150.
43. He, R.; Wu, X.; Sun, Z.; Tan, T. Wasserstein CNN: Learning invariant features for NIR-VIS face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1761–1773. [CrossRef]
44. Sharma, S.; Shanmugasundaram, K.; Ramasamy, S.K. FAREC—CNN based efficient face recognition technique using Dlib. In Proceedings of the 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), Ramanathapuram, India, 25–27 May 2016; pp. 192–195.
45. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015.
46. 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 Jane 2016; pp. 770–778.
47. Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houlsby, N.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
48. Wang, W.; Cao, Y.; Zhang, J.; Tao, D. Fp-detr: Detection transformer advanced by fully pre-training. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
49. Song, H.; Sun, D.; Chun, S.; Jampani, V.; Han, D.; Heo, B.; Kim, W.; Yang, M.H. ViDT: An Efficient and Effective Fully Transformer-based Object Detector. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
50. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
51. Tang, S.; Zhang, J.; Zhu, S.; Tan, P. Quadtree Attention for Vision Transformers. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
52. Chen, R.; Panda, R.; Fan, Q. RegionViT: Regional-to-Local Attention for Vision Transformers. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
53. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
54. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
55. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
56. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Wei, X.; Xia, H.; Shen, C. Conditional positional encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.
57. Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; et al. Rethinking attention with performers. *arXiv* **2020**, arXiv:2009.14794.
58. Wang, Y.C.; Wang, C.Y.; Lai, S.H. Disentangled Representation with Dual-stage Feature Learning for Face Anti-spoofing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 1955–1964.

59. George, A.; Marcel, S. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, 4–7 August 2021; pp. 1–8.
60. Wang, Z.; Wang, Q.; Deng, W.; Guo, G. Learning multi-granularity temporal characteristics for face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 1254–1269. [CrossRef]
61. Zhang, W.; Ji, X.; Chen, K.; Ding, Y.; Fan, C. Learning a facial expression embedding disentangled from identity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 6759–6768.
62. Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; Wang, H. Feature decomposition and reconstruction learning for effective facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 7660–7669.
63. Hong, J.; Lee, C.; Jung, H. Late Fusion-Based Video Transformer for Facial Micro-Expression Recognition. *Appl. Sci.* **2022**, *12*, 1169. [CrossRef]
64. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 2185–2194.
65. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 83–92.
66. Li, Y.; Lyu, S. Exposing deepfake videos by detecting face warping artifacts. *arXiv* **2018**, arXiv:1811.00656.
67. Bappy, J.H.; Simons, C.; Nataraj, L.; Manjunath, B.; Roy-Chowdhury, A.K. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Trans. Image Process.* **2019**, *28*, 3286–3300. [CrossRef]
68. Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; Guo, B. Protecting Celebrities from DeepFake with Identity Consistency Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 9468–9478.
69. Mazaheri, G.; Roy-Chowdhury, A.K. Detection and Localization of Facial Expression Manipulations. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 1035–1045.
70. Hosler, B.; Salvi, D.; Murray, A.; Antonacci, F.; Bestagini, P.; Tubaro, S.; Stamm, M.C. Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 1013–1022.
71. Ilyas, H.; Javed, A.; Malik, K.M. Avfakenet: A Unified End-to-End Dense Swin Transformer Deep Learning Model for Audio-Visual Deepfakes Detection. Available online: https://www.scopus.com/record/display.uri?eid=2-s2.0-85138317182&origin=inward&txGid=925378ef2e24c5aebd9db8ca01390b3c (accessed on 1 December 2022).
72. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
73. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, Vancouver Convention Center, Vancouver, BC, Canada, 30 April–3 May 2018.
74. Wang, J.; Liu, Y.; Hu, Y.; Shi, H.; Mei, T. Facex-zoo: A pytorch toolbox for face recognition. In Proceedings of the 29th ACM International Conference on Multimedia, Ottawa, ON, Canada, 28 October–3 November 2021; pp. 3779–3782.
75. Zhang, S.; Wang, X.; Liu, A.; Zhao, C.; Wan, J.; Escalera, S.; Shi, H.; Wang, Z.; Li, S.Z. A Dataset and Benchmark for Large-Scale Multi-Modal Face Anti-Spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
76. Boulkenafet, Z.; Komulainen, J.; Li, L.; Feng, X.; Hadid, A. OULU-NPU: A Mobile Face Presentation Attack Database with Real-World Variations. In Proceedings of the 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 612–618. [CrossRef]
77. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 1–11.
78. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
79. Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. DeepFake detection based on discrepancies between faces and their context. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6111–6121. [CrossRef] [PubMed]
80. 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 425–442.
81. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020; pp. 5001–5010.
82. Zeng, J.; Shan, S.; Chen, X. Facial expression recognition with inconsistently annotated datasets. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 222–237.

83. Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; Wang, J. Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 18710–18719.
84. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2168–2177.
85. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; Yu, N. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 772–781.
86. Ding, H.; Zhou, S.K.; Chellappa, R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 118–126.
87. Li, X.; Lang, Y.; Chen, Y.; Mao, X.; He, Y.; Wang, S.; Xue, H.; Lu, Q. Sharp multiple instance learning for deepfake video detection. In Proceedings of the 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, 12–16 October 2020; pp. 1864–1872.
88. Ruan, D.; Yan, Y.; Chen, S.; Xue, J.H.; Wang, H. Deep disturbance-disentangled learning for facial expression recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, 12–16 October 2020; pp. 2833–2841.
89. Liu, A.; Tan, Z.; Wan, J.; Escalera, S.; Guo, G.; Li, S.Z. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2021; pp. 1179–1187.
90. Yang, X.; Luo, W.; Bao, L.; Gao, Y.; Gong, D.; Zheng, S.; Li, Z.; Liu, W. Face anti-spoofing: Model matters, so does data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3507–3516.
91. Wang, Z.; Yu, Z.; Zhao, C.; Zhu, X.; Qin, Y.; Zhou, Q.; Zhou, F.; Lei, Z. Deep spatial gradient and temporal depth learning for face anti-spoofing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5042–5051.

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.