Computational Recognition of Facial Expressions in Sculpture

Abbas Khan
Department of Archaeology
University of Cambridge, UK
Queen Mary University of London
akr54@cam.ac.uk; acw676@qmul.ac.uk

Liliana Janik
Department of Archaeology
University of Cambridge
Cambridge, United Kingdom
lj102@cam.ac.uk

Hatice Gunes
Department of Computer Science and Technology
University of Cambridge
Cambridge, United Kingdom
hatice.gunes@cl.cam.ac.uk

***Abstract*—The art of understating emotions across different cultures is a subjective experience; however, what cross-cuts the cultural expression is our human neuro-physiological ability to communicate our feelings via facial expressions. Combined with AI know-how, such capacity provides us with unique opportunities to trace the artists' intentions to communicate particular emotions to the viewer, from those in the deep past c. 25,000 years ago to the contemporary sculpture. Here, we present a computational approach to analyzing facial expressions depicted in artwork of numerous regions, specifically sculptures. We collected a large dataset of sculptures' faces from online collections of various museums and used the existing methods to assign labels to each image. Each instance may have more than one label predicted by different methods, so we treated facial expression recognition as a multi-label classification problem. We designed deep learning-based frameworks using different backbones to categorize facial expressions in sculptures. We also implemented GradCAM to visualize the attributes in each image contributing the most to the predicted labels.**

***Index Terms*—computational art, facial expressions, deep learning, sculptures, buddhas, multi-label classification**

I. Introduction

Artificial Intelligence (AI) has had a substantial impact on many domains of research, and art is no exception [1]. AI technologies in the context of art can be categorized into two contexts. (i) Analysis of the art (ii) Generating novel artwork. The AI technologies which boosted applications related to "AI and art" include convolutional neural network (CNN) [2] to anatomize the art, and generative adversarial networks (GANs) [3] to develop novel artwork. The proposed study is focused on the former research theme to unravel some artistic mysteries. Creating and analyzing art is mainly considered only a human activity only; however, the current revolution of digitization and AI have strengthen their roots in the artwork. The digital repositories enable us to view and explore art collections worldwide. In recent years, because of COVID-19 restrictions, it has been impossible to visit each museum physically, and digital data practices in museums and art galleries have become more relevant. To avoid this sudden rupture of closing the museums for visitors, many museums around the world have published online versions of their collections [4]. It also made it possible for the computational methods to analyze these collections more precisely [5] and applications of deep learning-based methods which require a large number of images for training. A more comprehensive overview of 'how AI and Art are related?' is presented in many recent review articles including, [6], [7], [8], [9] and [10].

The project was funded by the Cambridge Humanities Research Grants Scheme, University of Cambridge. The project was housed at the McDonald Institute for Archaeological Research, University of Cambridge, United Kingdom.

Facial expression recognition (FER) has been a widely explored topic in the fields of pattern recognition and computer vision [11]. However, these techniques are rarely used to evaluate their applications in human-made faces to classify facial expressions expressed by artefacts. This research addresses this exciting topic of recognizing emotions in visual arts, particularly in the sculpture faces. The main contributions of this work can be stated as follows:

- We utilized deep learning-based methods to analyze facial expressions in sculpture faces.
- We collected images from different museums worldwide and introduced a publicly available automatically labelled dataset for the research community.
- The existing FER methods, pre-trained on human facial expressions, are used to generate soft and multi-label classes for the collected dataset.
- We designed deep convolutional neural network (DCNN) based models, using different strategies to classify the sculpture's expressions into multiple classes.
- We also generated coarse localization maps to visualize the important image regions that contributed to the predictions.

II. Related Work

The use of machine learning-based methods in the arts enables us to look into the past to investigate, interpret, preserve or manage the ancient cultures and their values. One of the recent works, introduced by Cowen et al. [12] used a traditional machine learning-based method known as Principal Component Analysis (PCA) to investigate universal facial expressions from the art of the ancient Americas. In a specific social context, the authors classified five facial expressions from 63 sculptures (pain, strain, anger, elation, and sadness). A more recent study "unpublished" [13] on the facial expressions of Terracotta Warriors is proposed by Tian et al. GANs [3] based

approach is used to generate synthetic images, and the ResNet-18 [14] backbone network classifies the facial expressions into one of the seven emotions categories. Some AI technologies study the visual content of artworks; for example, Alameda-Pineda et al. [15] predicted the emotional content triggered by the abstract paintings of MART dataset [16], labelled as either the painting evokes positive emotional response or negative. Milani and Fraternali [17] introduced paintings dataset for iconography classification, and a ResNet-50 [14] based backbone classifies the artwork based on iconographic elements depicted from the painting. Other more related studies which address recognition of emotion from images include [18], [19], and [20]. A huge body of AI work is devoted to generating synthetic artwork through statistical modelling within the computational creativity literature. Wombo, an AI startup based in Canada, introduced a mobile app called Dream [21] to convert the text prompt or a pre-defined art style into AI-powered painting. VQGAN-CLIP [22] is another neural network architecture which uses CLIP [23] and VQGAN [24] to generate higher visual quality novel images. It prompts the user to enter text, and then VQGAN generates the images while CLIP assesses the quality of images corresponding to the inputted text. Elgammal et al. "unpublished" [25] introduced AICAN, which is built over GAN. The authors argue that the generation of images by simulating over a given distribution will look like existing art. To generate novel images, they propose to maximize the deviation from existing styles while staying within the art distribution. A detailed overview of recently created artworks using AI are discussed by [26], [27], [28], and [29]. FER algorithms for human faces are designed for security, surveillance, social media platforms, and recommendation systems. These FER systems analyze the contraction and relaxation of facial muscles known as Action units (AUs), using which we can describe different facial expressions. Many recent approaches [30] [31] are using deep learning-based models to learn these facial features. To further enrich the facial expressions analysis, valence-arousal models [32] are proposed, which use the two-dimensional affective space. Valence indicates positive or negative emotion (from pleasant to unpleasant), while arousal refers to its intensity. Face landmarks detection is another popular task in FER, for which all the key points of a human face are tracked and detected, and then using these points as feature vectors, facial expressions are predicted [33]. The overall task of FER is divided into three subproblems: (i) Finding the face, (ii) Extracting the facial features, and (iii) Classifying these features into facial-expression-interpretative. A more detailed overview of these FER systems is presented in [34] [35], and [36].

III. Methodology

In the proposed research, we have used a multi-label classification approach to address the facial expression recognition in sculpture's faces. As the deep learning-based models require millions of training images, and to the best of our knowledge, there is currently no publicly available dataset

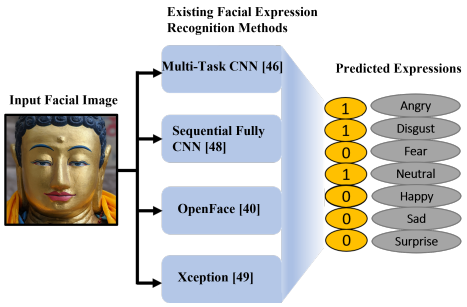

Fig. 1. Generating labels for multi-label facial expressions classification.

for sculpture's expressions classification. So, we relied on the existing deep learning methods designed to classify human facial expressions for labelling. We used several existing methods (discussed in section VI) to classify the expressions from sculpture faces and then used their predictions as soft ground truth to train our models. While predicting the facial expression for each image, we found that a single image can be classified into more than one expression category by different methods, as shown in figure 1. So instead of training our model on a single label, we used multi labels predicted by the four existing methods.

We implemented different state-of-the-art backbones networks VGG16 [37], MobileNet(v1 and v2) [38], and EfficientNet(B0,B1,B4 and B6) [39] with different training strategies to extract the features from input images and used fully connected layers as classifier. The larger networks such as VGG16 could not preform well as compared to smaller networks and usually over-fit the data in few epochs. So we used smaller size network such as Mobile-Nets and EfficientNets series. For the task at hand, we found that MobileNet-v1 architecture produces the best result(to be discussed in section VIII). A general overview of the proposed method is shown in figure 2. The proceeding subsections include the details of the architectures implemented in this study.

A. VGG16 Architecture

VGG16 [37] is the winner of the Image-Net Large Scale Visual Recognition Challenge; it downsamples the input images with a factor of 32 through a series of five convolutional blocks. Each block applied 3 x 3 convolutions with Batch-Normalization and ReLU activation and 2 x 2 max pooling on the incoming features. It consistently follows this arrangement of 3 x 3 successive convolutions and 2 x 2 max-pooling throughout the architecture.

B. MobileNet Architecture

MobileNets [38] are lightweight deep neural networks with streamlined structures and use depth-wise separable convolutions, comprising two layers: i-depth-wise convolution layer and ii-the point-wise convolution layer. The first one filters the input, and the second combines these features to generate features. These convolution layers are computationally less expensive than the standard convolution operations. Stride

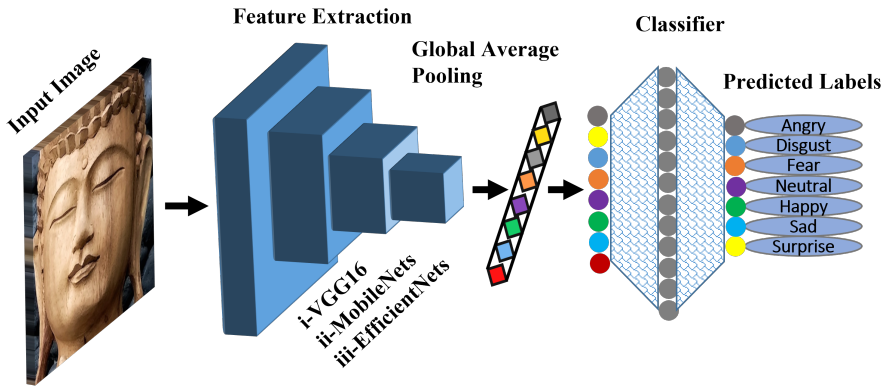

Fig. 2. The proposed architecture for multi-label classification of facial expressions in sculptures.

convolutions are used to handle downsampling, and a Batch-Normalization and ReLU activation follow each convolution layer.

C. EfficientNet Architecture

The EfficientNet [39] architecture is implemented with a novel model scaling method that uses a simple and effective compound coefficient to scale up CNNs. Unlike conventional scaling approaches in CNNs, the authors introduced scaling coefficients to uniformly scale all dimensions of depth, width, and resolution. Using 'neural architecture search', a family of models called EfficientNets [from B0 to B7] is developed by choosing different compound coefficients.

D. Proposed Architecture

The proposed architecture is shown in figure 2, where backbone network extracts features from the input image, followed by global average pooling of the bottleneck. A multi-layer perceptron is used as a classifier with Batch-Normalization and ReLU activation in the first layer and sigmoid activation in the last output layer.

IV. Data Annotation

The collected dataset is labelled using existing state-of-the-art facial expression recognition algorithms. We used four different methods for this purpose, as shown in figure 1. We observed that most of the expressions in sculpture's faces overlap, and according to the experts, it is difficult to assign a single absolute expression category to each face. Hence, we followed the multi-label classification approach and relied on these four labelling methods. The following subsections include the details of these methods and some intermediate results obtained from them.

A. OpenFace 2.0: Toolkit

The OpenFace 2.0 (extended version of OpenFace toolkit) [40] accomplishes the tasks of facial action unit recognition, facial landmark detection, face detection, head pose estimation, and eye-gaze estimation. As a first step, it detects the face in the given input image using Multi-task Cascaded Convolutional Network [41], followed by facial landmark detection and tracking through Convolutional Experts Constrained Local Model [42]. Eye gaze estimation is performed using a Constrained Local Neural Field (CLNF) landmark detector [43] [44] to find the iris, eyelids, and pupil in the cropped faces, finally it predicts the facial expressions through facial action unit (AU) intensity and presence. The AU recognition method is inspired by Baltrusaitis et al. [45] which applies linear kernel Support Vector Machines for AU detection. OpenFace predicts 18 AUs based on the presence (if AU is visible in the face) and 17 AU based on the intensity of AU (minimum to maximum on a 5 point scale). We used both types of AUs provided by OpenFace and then combined them to predict the facial expression labels for the proposed dataset. In some cases, where OpenFace could not detect the faces in the inputted images, we applied different scaling strategies to present the face at different scales. Figure 3 explains the overall working procedure adopted to process the data using the OpenFace toolkit. A comma-separated values (CSV) file is generated for each image in the dataset, which stores all the relevant information predicted by OpenFcae and the facial expression labels inferred from the AUs.

B. Multi-task Cascaded Convolutional Network

The Multi-task Cascaded Convolutional Network (MTCNN) is a three-stage network that detects faces and landmark locations [41]. In the first stage, a shallow CNN (P-Net) produces bounding box regression vectors and candidate windows, followed by non-maximum suppression (NMS) to remove the highly overlapped windows. In the second stage, another CNN called (R-Net) further refines these predictions by rejecting highly false positive candidates and calibrating the bounding boxes. The third stage CNN namely (O-Net), further regresses the bounding boxes and predicts a single bounding box for each face along with five facial landmarks positions. In the proposed research, we used MTCNN [41] to detect the faces as shown in figure 4.

C. Multi-task CNN and CNN-RNN Networks

A teacher-student based approach is proposed by [46] to jointly perform facial action unit detection, expression clas-

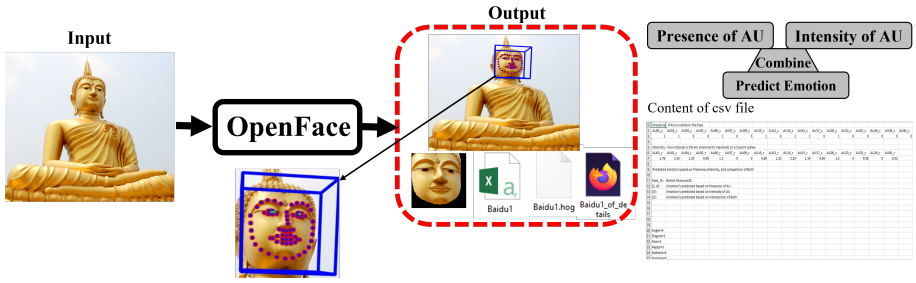

Fig. 3. Overview of processing the dataset through OpenFace toolkit, labelling strategy, and details of each file.

sification, and valence-arousal estimation. A teacher model is trained for each task with the corresponding ground truth. The predictions of the teacher model (soft ground truths) and the actual ground truths are further used to train the student models. The authors argue that the student models outperform the teacher model, and five student models are ensembled to boost the performance. Two different architectures, CNN and CNN-RNN based on ResNet50 [14], multilayer perceptron, and Gated recurrent units, are used. In the proposed study, we used Multi-task CNN architecture of five ensembled student models to assign labels to the dataset. Similarly to the previous experiment, we saved all the information in CSV files, including facial action units, emotion category, and valence-arousal values.

D. Real-time CNN for Facial Expression and Gender Classification

Arriaga et al. "unpublished [47] used two CNN based architectures, 'sequential fully CNN' and Xception [48] in their work to design real-time vision systems which simultaneously detect the faces and classify their gender and facial expressions. For the first model, they removed the fully connected layers and used a global average pooling layer followed by softmax to predict the emotions and gender. XceptionNet inspires the second method, they incorporated depth-wise separable convolutions in architecture to reduce the number of parameters, and OpenCV face detection module [49] is used as a pre-processing step to detect the faces for both models.

E. Proposed Labelling Strategy

An image is passed through each of the four methods mentioned previously, as shown in figure 1, to assign the facial expression labels (Angry, Disgust, Fear, Happy, Neutral, Sad or Surprise). Each instance may have multiple labels allotted by different methods, so our final label is called 'multi-class label', referring to more than one facial expression. The overall procedure for obtaining these labels is explained in figure 4. As a first step, we used face detectors to find a face in the image, followed by deep learning-based algorithms which predict the facial expressions label for the detected face. For the proposed study, we found that the OpenCV face detection module often fails to detect the faces in sculpture images, so we also used MTCNN [41]. In some cases, MTCNN cannot find the face if it is too tiny or blurred. We applied different strategies zooming in the centre, rotation, or cropping to make a face visible in a given input image for detection. However, if still face is not detected, we used VoTT [50], a free and open-source image annotation and labelling tool introduced by Microsoft. In this way, we increased the number of cropped faces for the proposed study.

V. Implementation Details

All experiments described in this paper are performed using a PC equipped with a cluster of Nvidia P100 GPUs. Keras framework with a Tensorflow backend is used to develop and evaluate the proposed methodology. Pre-trained weights trained on the ImageNet dataset are loaded for each backbone network. All the networks are trained using Adam optimization [51] with $\beta 1$ = 0.9 and $\beta 2$= 0. 99. The learning rate was set to 1e-4, and binary cross-entropy, represented by equation 1, was used as a loss function to train the models;

$$Loss = -(y \log(p) + (1 - y) \log(1 - p)) \quad (1)$$

The early stopping regularization technique is used to stop the training if the loss is not reducing for consecutive five training epochs.

VI. Evaluation Metrics

Accuracy is the most commonly used method to measure the performance of deep learning-based classification models. However, the proposed dataset is imbalanced, so accuracy cannot be a suitable metric for this problem. We used other rigorous metrics, i.e. precision, recall, and F1-Score, to evaluate its performance. Moreover, we also computed micro, macro, and weighted averages for each evaluation metric to consider Intra and inter-class variations. Micro average refers to when all samples equally contribute to the final averaged metric, macro average ensures the equal contribution from each class (seven facial expression classes), and weighted-average weights contributions from each class based upon its size.

VII. Dataset Details

Thanks to the large-scale digitization efforts, the number of art collections available online have increased. The dataset used in the proposed study is collected from different museum websites and search engines. Table I summarizes the details

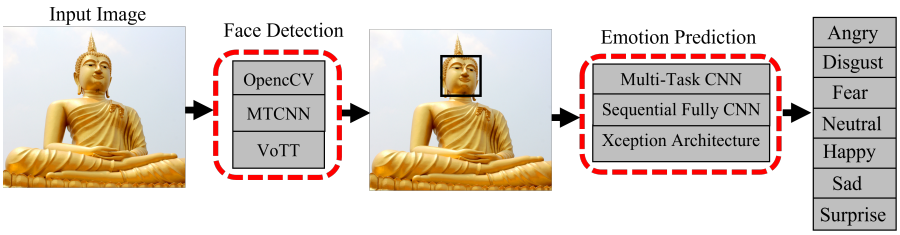

Fig. 4. Proposed strategies of face detection and facial expression classification.

of each dataset and the number of images contained in each museum's collections. While collecting the images for the proposed dataset, we considered the following points.

- It should contain a facial expression or a face.
- It should contain images with a frontal face, although some side views are also included.
- Should be a result of keyword search using; (Statues, Masks, Busts, Relief, Figure, Spirit, God, Body, King, Queen, Portraits, Model, Poetry, Guardian, Buddha, Guru, Durga, Maya, and Casts).

The following section lists all the sources/search engines/online collections, their brief history, and details of images collected. **Archaeological Haniwa faces** are clay figures that were made for ritual use in the history of Japan. They were built layer by layer using the Wazumi technique. In the proposed study, we have used 51 Haniwa faces. **Baidu Search Engine**, a dominant search engine in China, is used to collect the images of Asian Buddhas. Overall, we collected 440 images, and after data filtering (removing images with distorted faces), we left with 346 images.

British Museum provides 100 images of its online collections, used initially to detect faces. These images contain the upper part of the sculpture body, known as the bust. **The Buddha Dataset** consists of 194 images, and after deleting the images with no faces, we had 151 images. All these Buddhas are created in Japan and are based in different museums worldwide. **The Fitzwilliam Museum** of the University of Cambridge encompasses collections of modern western Europe and antiquities. Using the keywords mentioned previously, we successfully collected 184 sculpture images from their website's objects and artworks collection. **The Museum of Archaeology and Anthropology** known as

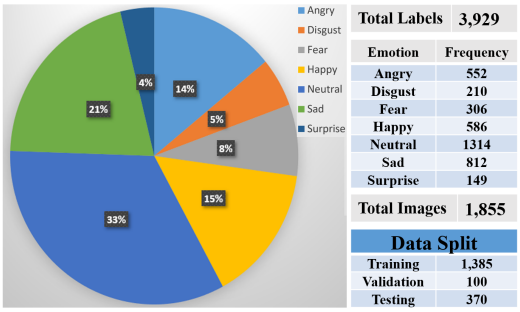

Fig. 5. Dataset distribution: Each segment of pie chart represents percentage of facial expressions for a particular class.

MAA Museum has a collection from different parts of the world. It was founded in 1884, and its initial collections came from Polynesia and the Cambridge Antiquarian Society. Later on, artwork from South Pacific regions, collections from the eighteenth century, and James Cook's three expeditions were added to its collections. **The National Museum of Asian Art** is home to 45,000 objects originating from China, Japan, Korea, and Southeast Asia. We were able to get 155 images from their online collection of objects belonging to different regions. **The Palace Museum** of Beijing contains 1.8 million pieces of art. Most of these belong to the ancestry of Ming and Qing the imperial collection. From various categories of its collections, we collected 61 images of sculptures. **The Rubin Museum** preserves the arts from different cultures, including Central Asia, the Himalayas, the Indian subcontinent, and Eurasia, 74 images of sculpture faces were collected from its online collections.

The Sainsbury Centre for Visual Arts is home to the art collection of Robert and Lisa Sainsbury, donated to the Uni-

TABLE I
DESCRIPTION OF COLLECTED DATASET.

No.	Source	# of images	No.	Source	# of images
1	Archaeological Haniwa faces	51	9	Palace Museum	61
2	Baidu Search Engine	346	10	Rubin Museum	74
3	British Museum	100	11	Sainsbury Centre	67
4	Buddha Dataset	151	12	Tokyo National Museum	14
5	Fitzwilliam Museum	184	13	Xiangtangshan Caves	35
6	Maa Museum	99	14	Jomon Figurines	125
7	National Museum of Asian Art	155	15	Ancient Yenisei Masks	29
8	Metropolitan Museum of Art	306		**Total No. of Images**	**1,855**

versity of East Anglia in 1973. It contains several thousands of artworks and 67 images of Buddha's faces are collected from their web page. **The Tokyo National Museum** is the oldest Japanese museum and preserves a wide range of artworks from Japan and Asia. **The Xiangtangshan Caves** are located in Hebei province, China, and were part of the Northern Qi dynasty (550-577). The content from the caves is categorized into three categories:i-Architecture, ii- Carved images, and iii-Scripture. From the Xiangtangshan Caves project, we collected 35 images from their sculpture collections.

The Jomon Figurines belong to Japan's Neolithic period, and its name is derived from the "cord markings".The Jōmon period lasted between ca.10,500 to 300 B.C. and was divided into six different phases. We collected 125 Jomon Figurines from various online collections and books. **The Ancient Yenisei Masks** are the death masks of the Tashtyk people belonging to the Yenisei Kyrgyz (Old Turkic), from the 3rd century BCE to the 13th century C.E. According to the Archaeologists, the suture was stitched after his death or to remove the brain during an elaborate burial rite. These masks are very close to real faces; we could collect 29 of these Ancient Yenisei masks for the proposed study by exploring different books and websites. **The Metropolitan Museum of Art**, also known as "the Met", is the largest museum of artwork in the West and contains a collection of over two million paintings. It was founded in 1870 and became the home for artwork from ancient Egypt, classical antiquity, sculptures from Europe, and American and modern art. We collected 306 images from its open-access collections available online.

Figure 5 visualizes the distribution of different facial expression categorizes from the dataset,as well as the data split for training, validation, testing, and frequency of each class for the multi-label classification approach.

VIII. Results and Discussion

This section sheds light on the visual and quantitative results of different methods implemented in the proposed study and the proposed multi-label classification approach. Table II provides classification results of three different architectures, MobileNet-v1, VGG16, and EfficientNet-B6, which performed the best from the EfficientNet family for this problem. Bold represents the best result, and an underlined-bold number shows performance tie. Overall, we can see that MobileNet-v1 performance is superior to the other two architectures. During training, we found that the networks with a much higher parameter, such as VGG16 (138 million), over-fits the data and results in reduced performance on the test set. In comparison, the CNN models, such as MobileNet-v1 with only 4.2 million parameters, can generalize well. We also designed custom CNN's backbones with fewer parameters than MobileNet-v1; however, it was unable to learn the useful features for expressions recognition and underwent under-fitting. Hence, our final architecture comes with MobileNet-v1 (determined empirically) as a backbone.

Some more visual results are shown in figure 6, where 6(a) contains the outputs obtained from the OpenFace, by drawing all the facial landmarks, detected faces, head pose and eye-gaze estimation. OpenFace can detect multiple faces in the image, assigns landmarks and bounding boxes, and predicts AUs for each face individually. Figure 6(b) depicts the results from Real-time CNN for facial expression and gender classification. As a pre-processing step, firstly, it draws a bounding box across the face in the image and then predicts facial expression followed by gender classification for the detected face.

Figure 6(c) represents the results of our proposed multi-label classification architecture. For a given input face, our method can generate multiple facial expression labels by looking into different regions such as noise, eyes, lips, and mouth. The propped strategy is able to analyze these facial region simultaneously and incorporate their corresponding AUs while predicting the labels for each face. To reveal the attention patterns for a particular facial expression category, we visualized the activation of pre-trained models using Gradient-Weighted Class Activation Mapping (Grad-CAM) [52].

TABLE II
Comparison of the results obtained from different methods.

	MobileNet-v1 Results			VGG16 Results			EfficientNet-B6 Results		
Class	Precision	Recall	F1-Score	Precision	Recall	f1-Score	Precision	Recall	f1-Score
Angry	**0.50**	0.42	**0.42**	0.37	0.30	0.33	0.43	**0.45**	0.40
Disgust	**0.36**	**0.19**	**0.24**	0.33	0.08	0.12	0.29	0.12	0.16
Fear	**0.40**	0.34	**0.37**	0.39	**0.37**	**0.37**	0.37	0.22	0.24
Happy	0.53	**0.52**	**0.53**	**0.56**	0.42	0.44	0.51	0.42	0.42
Neutral	**0.81**	0.81	0.80	0.76	0.87	0.81	0.75	**0.90**	**0.82**
Sad	**0.64**	**0.48**	**0.54**	0.56	**0.48**	0.51	0.55	0.42	0.47
Surprise	**0.39**	**0.36**	**0.37**	0.36	0.29	0.31	0.34	0.17	0.22
Weighted Avg.	**0.61**	**0.55**	**0.56**	0.59	0.53	0.55	0.46	0.39	0.39
Macro Avg.	0.52	0.45	0.47	**0.56**	**0.53**	**0.53**	0.55	**0.53**	0.52
Micro Avg.	**0.61**	**0.55**	**0.58**	0.48	0.40	0.41	0.58	0.53	0.55

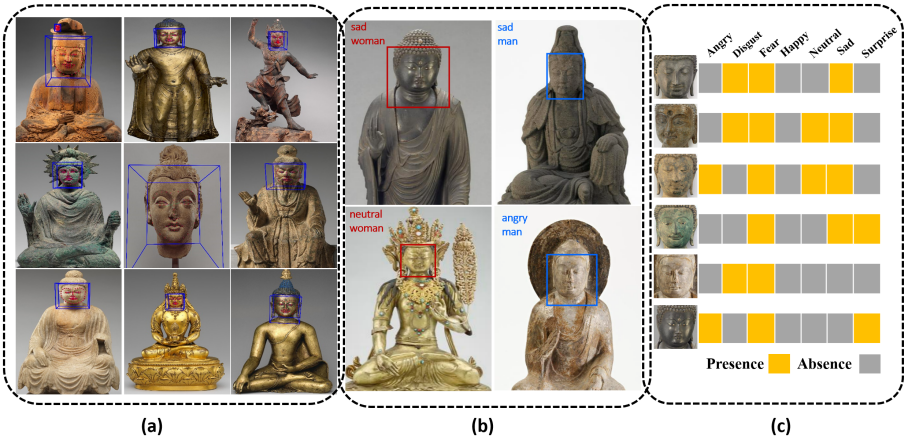

Fig. 6. Visual results of proposed Study: **(a)** OpenFace facial landmarks, face detection and eye-gaze estimation **(b)** Face detection, gender and facial expression prediction by Real-time CNNs **(c)** Multi-label classification results of the proposed methodology.

IX. Grad-CAM Visualization

Explainable AI can determine how an AI system makes the decision. In the proposed study, we used the Grad-CAM [52] to visualize the image regions analyzed by the deep learning models. This methods uses the gradient of any ground truth flowing into the last convolution layer to generate a coarse localization map weighing the image regions that the model considered to predict the target score. Grad-CAM Visualizations shown in figure 7 demonstrate that the network can capture the most relevant information from face regions, including noise, eyes, lips, and mouth. These attention maps advocate the previous research [53], [54] that the mouth, nose and eye areas are considered to be the most distinctive for recognizing facial expressions.

X. Conclusion

We used state-of-the-art computational approaches to recognize facial expressions in sculptures. A dataset of sculptures' bodies and faces is also presented to accelerate research in the proposed area, that will allow us to access the emotional potency of sculpture which in turn, for the first time, provides us a tool to look for artist indication in non-verbally communicating with the viewer that goes beyond art's historical interpretation of past and contemporary sculpture. The dataset is labelled using four different machine learning-based facial expression recognition methods, and deep learning based CNN architectures are proposed to classify the facial expressions. Most facial expression categories are ambivalent, so we used a multi-label classification approach to categorize them into more than one class. A coarse localization map is also generated using Grad-CAM to highlight the essential regions of the face images analyzed by the network to make predictions.

References

[1] J.-W. Hong and N. M. Curran, "Artif. intell., artists, and art: attitudes toward artwork produced by humans vs. artif. intell." *ACM Trans. on Multimed. Computing, Commun., and Applications (TOMM)*, vol. 15, no. 2s, pp. 1–16, 2019.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. in neural information Process. Syst.*, vol. 25, 2012.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Adv. in neural information Process. Syst.*, vol. 27, 2014.

[4] L. Noehrer, A. Gilmore, C. Jay, and Y. Yehudi, "The impact of -19 on digital data practices in museums and art galleries in the uk and the us," *Humanities and Social Sciences Commun.*, vol. 8, no. 1, pp. 1–10, 2021.

[5] S. Lang and B. Ommer, "Reflecting on how artworks are processed and analyzed by comput. vis.: Supplementary material," pp. 0–0, 2018.

[6] D. Grba, "Deep else: A critical framework for ai art," *Digital*, vol. 2, no. 1, pp. 1–32, 2022.

[7] D. Fessenko, "Can artif. intell. (re) define creativity?"

[8] G. Castellano and G. Vessio, "Understanding art with ai: Our research experience," 2021.

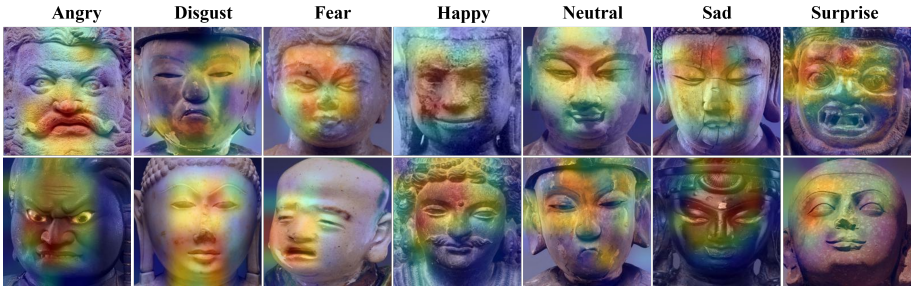

Fig. 7. Grad-CAM visual explanations for facial expression recognition.

[9] L.-H. Lee, Z. Lin, R. Hu, Z. Gong, A. Kumar, T. Li, S. Li, and P. Hui, "When creators meet the metaverse: A survey on computational arts," *arXiv preprint arXiv:2111.13486*, 2021.

[10] E. v. d. Peijl, A. Najjar, Y. Mualla, T. J. Bourscheid, Y. Spinola-Elias, D. Karpati, and S. Nouzri, "Toward xai & human synergies to explain the history of art: The smart photobooth project," Springer, pp. 208–222, 2021.

[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. on affective computing*, 2020.

[12] A. S. Cowen and D. Keltner, "Universal facial expressions uncovered in art of the ancient americas: A computational approach," *Science Adv.*, vol. 6, no. 34, p. eabb1005, 2020.

[13] W. Tian, Y. Xie, T. Ma, and H. Zhang, "Uncover common facial expressions in terracotta warriors: A deep learning approach," *arXiv preprint arXiv:2105.04826*, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[15] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe, "Recognizing emotions from abstract paintings using non-linear matrix completion," pp. 5240–5248, 2016.

[16] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe, "In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings," pp. 349–358, 2012.

[17] F. Milani and P. Fraternali, "A dataset and a convolutional model for iconography classification in paintings," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 4, pp. 1–18, 2021.

[18] H.-R. Kim, Y.-S. Kim, S. J. Kim, and I.-K. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE Trans. on Multimed.*, vol. 20, no. 11, pp. 2980–2992, 2018.

[19] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Process. letters*, vol. 51, no. 3, pp. 2043–2061, 2020.

[20] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with cnns," *IEEE Trans. on Multimed.*, vol. 22, no. 2, pp. 515–523, 2019.

[21] ""Wombo AI:," https://app.wombo.art/, accessed: 2022-05-20.

[22] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "Vqgan-clip: Open domain image generation and editing with natural language guidance," *arXiv preprint arXiv:2204.08583*, 2022.

[23] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," pp. 2085–2094, 2021.

[24] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," pp. 12 873–12 883, 2021.

[25] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, "Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms," *arXiv preprint arXiv:1706.07068*, 2017.

[26] E. Cetinic and J. She, "Understanding and creating art with ai: Review and outlook," *ACM Trans. on Multimed. Computing, Commun., and Applications (TOMM)*, vol. 18, no. 2, pp. 1–22, 2022.

[27] M. Cheng, "The creativity of artif. intell. in art," MDPI, p. 110, 2022.

[28] A. Hertzmann, "Can computers create art?" Multidisciplinary Digital Publishing Institute, p. 18, 2018.

[29] B. Agüera y Arcas, "Art in the age of machine intelligence," Multidisciplinary Digital Publishing Institute, p. 18, 2017.

[30] N. Churamani and H. Gunes, "Clifer: Continual learning with imagination for facial expression recognition," IEEE, pp. 322–328, 2020.

[31] J. Premaladha, M. Surendra Reddy, T. Hemanth Kumar Reddy, Y. Sri Sai Charan, and V. Nirmala, "Recognition of facial expression using haar cascade classifier and deep learning," in *Inventive Communication and Computational Technologies*. Springer, 2022, pp. 335–351.

[32] T. T. A. Höfling, A. Gerdes, U. Föhl, and G. W. Alpers, "Read my face: automatic facial coding versus psychophysiological indicators of emotional valence and arousal," *Frontiers in psychology*, vol. 11, p. 1388, 2020.

[33] M. Munasinghe, "Facial expression recognition using facial landmarks and random forest classifier," IEEE, pp. 423–427, 2018.

[34] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020.

[35] S. Seema, B. Sowmya, P. Chandrika, D. Kumutha, and N. Krishna, "Efficient facial expression recognition using deep learning techniques," in *Deep Learning Applications for Cyber-Physical Systems*. IGI Global, 2022, pp. 99–118.

[36] S. Bhattacharya, "A survey on: Facial expression recognition using various deep learning techniques," in *Advanced Computational Paradigms and Hybrid Intelligent Computing*. Springer, 2022, pp. 619–631.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[39] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," PMLR, pp. 6105–6114, 2019.

[40] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," IEEE, pp. 59–66, 2018.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal Process. letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[42] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for 3d facial landmark detection," pp. 2519–2528, 2017.

[43] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," pp. 354–361, 2013.

[44] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," pp. 3756–3764, 2015.

[45] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," IEEE, pp. 1–6, 2015.

[46] D. Deng, Z. Chen, and B. E. Shi, "Multitask emotion recognition with incomplete labels," IEEE, pp. 592–599, 2020.

[47] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017.

[48] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," pp. 1251–1258, 2017.

[49] "OpenCV Face Detection using haar feature-based cascade classifiers," https://docs.opencv.org/3.4.1/d7/d8b/tutorial_py_face_detection.html, accessed: 2022-05-20.

[50] "VoTT:Free and open source image annotation tool developed by microsoft:," https://github.com/microsoft/VoTT, accessed: 2022-05-20.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," pp. 618–626, 2017.

[53] M. L. Smith, G. W. Cottrell, F. Gosselin, and P. G. Schyns, "Transmitting and decoding facial expressions," *Psychological science*, vol. 16, no. 3, pp. 184–189, 2005.

[54] L. L. Kontsevich and C. W. Tyler, "What makes mona lisa smile?" *Vision research*, vol. 44, no. 13, pp. 1493–1498, 2004.