Yoga Posture Image Classification Using Big Transfer(BiT)

K.M Nafiur Rahman Fuad
Dept. of Computer Science and Engineering
Bangladesh University of Business and Technology
Dhaka,Bangladesh
nafiurrahman373@gmail.com

Uland Rozario
Dept. of Computer Science and Engineering
Bangladesh University of Business and Technology
Dhaka,Bangladesh
rajiebbb@gmail.com

***Abstract*—Yoga, known for its holistic approach to well-being, has gained widespread popularity, offering physical and mental benefits. Recent integration of technology into yoga practice has opened innovative avenues for enhanced learning experiences. One such advancement is the use of deep learning models for yoga posture classification, facilitating better alignment and comprehension for practitioners and instructors. This paper explores the application of the cutting-edge Big Transfer model to classify yoga posture images. BiT, recognized for its exceptional performance in visual recognition tasks, leverages pre-trained representations to achieve high accuracy, even with limited data. Our study involves an extensive dataset encompassing diverse yoga postures and variations. We fine-tune the BiT model on this dataset and rigorously evaluate its performance. The results underscore BiT's effectiveness in accurately identifying yoga postures, making it a valuable tool for automating posture assessment and providing real-time feedback during yoga sessions. Furthermore, we discuss the potential transformative applications of this technology in remote yoga instruction and telehealth. Automatic classification of yoga postures enhances personalized guidance and broadens access to yoga instruction. In conclusion, this paper presents a novel approach to yoga posture image classification using BiT, with the potential to revolutionize how practitioners learn and instructors teach yoga in the digital age. This research bridges the gap between technology and tradition, enhancing the practice of this ancient art.**

***Index Terms*—Image Classification, Deep Learning, Big Transfer(BiT) Model, Computer Vision, Posture Assessment**

I. Introduction

Yoga is one of the well-known and most practiced spiritual, mental, and physical exercises in India [1]. Even though the idea of yoga was founded in India [2] but by the time is passing by, a lot of people all over the world now practice yoga. A study by Zhang et al. found that the popularity of yoga among people in the United States is increasing day by day [3]. In 2002 approximately 10,386,456 (5.1%) people were practicing yoga, it nearly tripled to 32,761,194 people (13.7%). It is mainly because of the health benefits of yoga. K Upadhyay et al. [4] discussed in their study the link between yoga and diabetes, highlighting how complex a condition it is and how it causes metabolic abnormalities related to insulin and glucose control. Due to its support of a healthy diet and way of life, as well as its capacity to balance the endocrine system, stimulate abdominal organs, and relieve stress, it is suggested that yoga is useful in both preventing and treating diabetes.

In recent years, the field of computer vision has witnessed remarkable advancements, revolutionizing the way we interact with and understand visual data. One prominent application of computer vision is image classification, a fundamental task with broad implications across diverse domains, including healthcare, autonomous systems, and content recommendation. Yoga, as an ancient practice that promotes physical and mental well-being, has gained widespread recognition for its positive impact on human health and relaxation [5]. In this context, the automated classification of yoga poses from images plays a pivotal role I in enabling the development of innovative applications and tools to support yoga practitioners and instructors.

The task of yoga pose classification is inherently challenging due to the variations in lighting, pose variations, and background clutter that can be present in images. Traditional image classification techniques often struggle to achieve high accuracy in such scenarios. However, recent advancements in deep learning and transfer learning have shown tremendous promise in improving the accuracy and efficiency of image classification tasks [6]. Transfer learning, in particular, leverages pre-trained models on large-scale datasets and fine-tunes them for specific tasks, enabling the extraction of high-level features from images that can be valuable in addressing the intricacies of yoga pose classification.

This paper introduces a novel approach to yoga pose classification using transfer learning, specifically employing the Big Transfer (BiT) model, a state-of-the-art architecture known for its outstanding performance on various computer vision tasks. Our methodology leverages the knowledge encoded within the BiT model to recognize and categorize yoga poses accurately. The proposed system addresses the challenges posed by the diverse yoga posture dataset, making it a valuable tool for yoga practitioners, instructors, and researchers.

II. Related Work

In Transfer Learning, Image classification is one of the primary issues. Transfer learning has been extensively conducted for numerous computer vision tasks [7]. Transfer learning has been extensively applied in computer vision, particularly for image classification tasks. Researchers have explored using pretrained convolutional neural networks (CNNs) like VGG, ResNet, and Inception for various image recognition tasks II, including object detection, semantic segmentation, and facial recognition.It has also been applied to medical imaging for tasks [8] such as disease diagnosis and lesion detection II [9]. Models pretrained on large image datasets can be fine-tuned for medical image analysis with limited labeled medical data.Transfer learning gives the opportunity to,

- **Faster Training:** Transfer learning often requires less training time and fewer labeled examples compared to training a model from scratch, making it more efficient.
- **Better Generalization:** Pre-trained models have learned valuable representations from large and diverse datasets, which can lead to improved generalization on the target task, especially when the target task has limited data.
- **Reduced Overfitting:** The knowledge transferred from the source task can help in preventing overfitting on the target task, as the model starts with a more robust initialization.

Ashraf et al. [10] studied on yoga pose classification. They implemented YoNet, a deep learning-based network created for yoga pose classification, was presented by the researchers. A thorough performance comparison between YoNet and top picture classification models, such as ResNet, InceptionNet, InceptionResNet, and Xception, was carried out. YoNet used a cutting-edge technique by separating the spatial and depth information from input photos and applying them to the classification procedure. YoNet identified five different yoga positions with an accuracy rate of 94.91% and a precision score of 95.61%, demonstrating the remarkable outcomes of the experiment. YoNet fared notably better than the other models, demonstrating the potency of their suggested architecture for classifying yoga poses. Palanimeera et al. [11] used machine learning techniques to categorize yoga positions. To identify and categorize yoga poses, particularly the sun salutation set of postures, they employed a pose estimation approach in conjunction with four machine learning models (KNN, SVM, naive Bayes, and logistic regression). One male participant provided information about his height, weight, and age. Based on the angles obtained from the skeletal joint of the pose estimation algorithm, the pose detection techniques were applied to identify the yoga postures. It was assessed how accurate the machine learning approaches' classification findings were. To build training and test datasets for the machine learning models, the researchers stored the data. A formula involving the coordinates of two points was used to determine the joint angles. In order to improve posture detection, Liaqat et al. [12] in their study proposed a novel hybrid approach that combines deep learning classifiers, such as 1D-CNN, 2D-CNN, LSTM, and bidirectional LSTM, with machine learning classifiers, such as SVM, logistic regression, decision tree, Naive Bayes, random forest, Linear discrete analysis, and Quadratic discrete analysis. By combining the best features of deep learning and machine learning, this hybrid approach achieves an impressive accuracy rate of more than 98% on a well-known benchmark dataset. Along with choosing the best techniques for posture recognition, the researchers also introduced a revolutionary CNN and LSTM architecture for automated posture detection. Their hybrid methodology, which combined the strengths of machine learning and deep learning, beat separate DL and ML techniques, and the experimental outcomes indisputably confirmed the effectiveness of this cutting-edge hybrid strategy. Sharma et al. [13] developed a self-assistance computer vision-based yoga posture identification system that allows for real-time correction and recognition of yoga poses. In order to facilitate this effort, they have assembled the YOGI dataset, which consists of 10 yoga poses and 5 mudras (hand gestures) for recognition; each stance and mudra has between 400 and 500 photos. They used joint angles as characteristics for several machine learning and deep learning models, using a skeleton-based technique to extract features from the body for yoga positions and the hand for mudra poses. After much work, they were able to use XGBoost with RandomSearch CV as the best model, which produced an amazing accuracy of 99.2%. They also created a piece of software that functions as a teacher, providing spoken assistance and pose correction in real-time. This was a novel addition to the field as it allowed for the simultaneous detection and recognition of yoga mudras and postures. A variety of computer vision techniques, including color segmentation, contour detection, infrared segmentation, and hand tracking using glove sensors and artificial vision techniques, were further utilised in their study. In an effort to close the technological gap by offering yoga practitioners support via their smartphones, they also investigated how Microsoft Kinect sensors and depth cameras might improve hand tracking and detection precision. Jose et al. [14] created a system that makes use of cutting-edge deep learning methods, such as transfer learning and convolutional neural networks (CNN), to make it easier to identify yoga poses from still photos or moving videos. After carefully training the model on photos of ten different yoga poses, its predicted accuracy test produced encouraging findings, with an accuracy rate of 85%. This innovative technique demonstrates the potential of contemporary technology in this field and is a first step towards developing an automated tool for the analysis of yoga photos and movies. Ten classes of yoga poses—bridge, children, downward dog, mountain, plank, sitting forward bend, tree, triangle pose, warrior1, and warrior2—with approximately 700 uniformly spaced photos each were included in the dataset to aid with the research. The researchers used the Softmax function in the last dense layer for network output normalization after using dropout layers for regularization and the ReLU activation function on all dense layers (apart from the output layer). The network was compiled using the ADAM optimizer and classification cross-entropy as the loss function. Notably, in terms of accuracy, their suggested transfer learning architecture fared better than other machine learning models and conventional CNN architectures. Ogundokun et al. [15] presented a thorough three-phase approach designed to improve neural network models' ability to recognize human posture. This novel method effectively addressed issues such as overfitting and poor performance by combining CNN transfer learning, image data augmentation, and hyperparameter optimization (HPO). They used classic CNN and MLP as baseline classifiers for comparison, but they also leveraged the capability of transfer learning algorithms like AlexNet and VGG16 in conjunction with HPO to detect human posture. The incorporation of picture data augmentation methodologies was crucial in augmenting the training dataset, reducing overfitting, and improving the classification performance as a whole. By using a random-based search approach, the four models—AlexNet, VGG16, CNN, and MLP—had their hyperparameters carefully optimized. The suggested models achieved accuracy rates of 91.2% for AlexNet, 90.2% for VGG16, 87.5% for CNN, and 89.9% for MLP, which were pretty good results. Notably, this work represented a major leap in the study of human posture detection since it was the first time that hyperparameter optimization was used to the MPII human pose dataset.

III. Methodology

This research study conducts a comprehensive investigation of the efficacy and practicality of different transfer learning architectures in the field of picture categorization. This study's main goal is to thoroughly benchmark and evaluate these architectures in order to highlight their potential for picture classification. The next subsections will offer a comprehensive examination of the fundamental mechanisms and complex structures that make up various transfer learning approaches. The scientific strategy used to categorize photos of yoga poses is very remarkable. Yoga posture classification presents a distinct set of difficulties because of the variety of positions, body types, and settings. We use the robust Big Transfer (BiT) Medium model, which provides a strong solution for this difficult task, to tackle these problems. Our goal is to provide a high-level overview of our approach, emphasizing the crucial actions and factors to take into account in order to produce reliable and correct yoga posture image categorization. In this research study, we aim to test transfer learning architectures and develop a framework to improve the classification of yoga poses, with potential applications in wellness evaluation, health and fitness monitoring, and other domains. By utilizing the capabilities of the BiT Medium model, we hope to further the development of cutting-edge techniques for picture classification with a particular emphasis on yoga postures. This will open the door for the creation of more effective and efficient posture identification systems.

A. Data Collection and Preparation

The basis of this study is the compilation of a broad and heterogeneous dataset that includes a variety of photos of yoga postures, including various asanas, positions, and expressive modifications. The careful curation of this dataset provides a solid foundation for the other stages of our research. Preprocessing the data is a necessary step in order to enable efficient use of the dataset. Among the many tasks that are included in this preliminary phase are image resizing to a standard format, cropping for best framing, and pixel normalization to guarantee consistency in pixel values and dimensions throughout the dataset. By carrying out these crucial preprocessing procedures, we produce a consistent and cohesive dataset, laying the groundwork for reliable and insightful analysis in the stages of our study that follow. Stratifying the dataset into three separate subsets—the training set, validation set, and test set—is a crucial step in our research process. In order to ensure that the produced models are thoroughly evaluated for their performance and generalizability, partitioning is essential for facilitating effective model evaluation. To promote an exact and methodical approach to image classification, we use a laborious labeling procedure. This involves carefully annotating each image with the appropriate category for a yoga stance. Furthermore, the dataset goes through a thorough cleaning process as part of our dedication to data quality and trustworthiness. By reducing noise and irregularities in the dataset, this cleansing procedure not only guarantees data integrity but also improves model evaluation.

B. Big Transfer (BiT) Medium Model

The state-of-the-art BiT model,based on ResNet50-v2 architecture, known for its exceptional performance, is chosen as the foundation for yoga posture classification.ResNet-50v2 is a deep convolutional neural network architecture. It is an improved version of the original ResNet-50.ResNet-50v2 has 7x7 convolutional layer with 64 filters.The final layer is a fully connected layer with 1,000 neurons.It uses a softmax activation function to produce class probabilities.The input size of Bit-M model on our provided datasets are 256x256.Batch normalization is applied after convolution.ReLU (Rectified Linear Unit) activation function follows batch normalization.ResNet-50v2 is composed of multiple residual blocks. In total, there are 16 residual blocks in the network, organized into four groups.

Transfer Learning: BiT-M is utilized as a pre-trained model,pre-trained on ImageNet-21k,leveraging its pre-existing knowledge across visual domains.We fine-tuned BiT-M to get better results.By using computer vision,we modified our dataset images to increase their sharpness and gamma so that BiT-M can acheive great results.

C. VGG16

The VGG16 model, also known as the Visual Geometry Group 16 architecture, is a well-known and widely used convolutional neural network design for a variety of computer vision tasks, most notably image categorization. VGG16 is praised for being straightforward and efficient, making it a useful tool in the deep learning field. The VGG16 design, which has 16 layers altogether, is distinguished by its methodical approach to feature extraction and classification.

TABLE I
RELATED WORKS SUMMARY

Author	Year	Method	Dataset	Result	Limitation
Ashraf et al. [10]	2023	Yonet	Yoga-82	94.91% accuracy with 95.61% precision	Limited dataset,Comparison with limited models & Lack of detailed analysis
Palanimeera et al. [11]	2021	KNN, SVM, Naïve bayes, Logistic Regression	Webcam images	The KNN provided 99.02% of accuracy	Limited comparison with other models
Liaqat et al. [12]	2021	KNN, SVM, Naïve bayes, Logistic Regression,1-D CNN,2-D CNN,LSTM	Webcam images	Hybrid approach provided 98% of accuracy	Lack of dataset details,Empirical Parameter Setting,Unaddressed Computational Resources,Limited Posture Detection
Sharma et al. [13]	2022	XGBoost with RandomSearch CV	Yogi	XGBoost with RandomSearch CV provided 99.2% of accuracy	Lack of real-time testing,comparison with other models,limited generalizability,limited dataset.
Jose et al. [14]	2021	CNN & Transfer learning	Smartphone images	Transfer learning provided 85% & 3D-CNN provided 91.5% of accuracy	limitations in terms of dataset size, model architecture, and prediction accuracy
Hussain et al. [16]	2019	Inception-v3	CIFAR-10 & Caltech Faces	CIFAR-10 provided 70.1% & Caltech Faces provided 65.7% of accuracy	Narrow Dataset Evaluation,Lack of Parameter Exploration,Limited Model Evaluation,Incomplete Comparison
Ogundokun et al. [15]	2022	AlexNet and VGG16	MPII	AlexNet provided 91.2% & VGG16 provided 90.2% of accuracy	Overfitting and Poor Performance,High Data Requirements,Ambiguous Hyperparameter Criteria,Lack of Model Architecture Details
Ogundokun et al. [17]	2022	DeneSVM	Human Posture Recognition dataset	DeneSVM provided test precision of 94.72%, validation accuracy of 93.79%, and training accuracy of 97.06%	Lacks computational capacity,Demands extra computation power

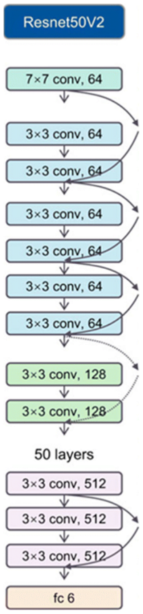

Fig. 1. Resnet50-v2 architecture

13 convolutional layers and 3 fully connected layers make up the 16 layers total. The fully connected layers are essential for making the final classification determinations, while the convolutional layers are crucial for identifying and abstracting characteristics from the input images. The continuous application of 3x3 convolutional filters across the whole network makes VGG16 stand out. This decision enables VGG16 to catch finer and more nuanced characteristics within images as opposed to the 7x7 filters utilized in prior architectures like AlexNet. When working with large and varied datasets, this fine-grained feature extraction is very helpful since it enables the model to recognize minute patterns and structures. Additionally to filter size selection, VGG16 makes use of max-pooling layers with 2x2 pooling windows. Max-pooling is a subsampling method that aids in reducing the spatial dimensions of the feature maps created by convolutional layers, increasing the computational efficiency of the network while preserving critical information. The final three layers of the VGG16 model are fully connected layers, where the extracted features are transformed and combined to make the ultimate classification decision. These layers contribute to the model's ability to distinguish and categorize objects within images, a fundamental aspect of image classification tasks.

Pretrained versions of the VGG16 model, which was initially trained on the ImageNet dataset, were heavily used in our paper. In the context of deep learning for computer vision, this tactical approach offers a number of noteworthy benefits. The ImageNet dataset is a sizable collection of pictures representing a wide range of item categories, making it an invaluable resource for image identification applications. The abundance of features and patterns the VGG16 model discovered during its initial training can be used by pretraining it on ImageNet. These properties cover a wide range of visual concepts, giving the model the ability to recognize objects, textures, forms, and more. By utilizing this vast knowledge base, we are able to significantly reduce the amount of time as well as computational resources that would otherwise be needed to train a model from scratch on our particular dataset.

A crucial step in our strategy is fine-tuning, which entails adjusting the pretrained VGG16 model to the particulars of our target task. Whether our goal is scene categorization, object recognition, or another image-related task, fine-tuning makes sure the model adjusts its learnt representations to match the specifics of our data. Through this procedure, the model's efficiency, precision, and capacity to create exact classifications based on the particular traits and patterns pertinent to our application are improved.

As a result, VGG16 is a strong and trustworthy convolutional neural network design that excels at image classification tasks. In the field of computer vision, the consistent application of 3x3 convolutional filters along with fully connected layers has cemented its position. We unlock the full power of VGG16 by pretraining on ImageNet and then optimizing on our dataset, enabling it to shine in a wide range of real-world applications. The model creation process is streamlined by this method, which also equips deep learning researchers and practitioners to take on a variety of image identification issues with efficiency and confidence.

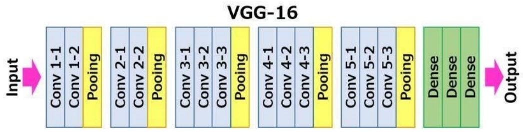

Fig. 2. VGG16 architecture

D. *MobileNetV3*

An outstanding member of the MobileNetV3 family, MobileNetV3-Large, stands out for its capacity to offer a compact yet effective solution for various mobile and on-device vision applications. MobileNetV3-Large provides a compelling framework that excels in achieving the ideal balance between model size, processing speed, and accuracy as the demand for real-time picture identification, object detection, and scene analysis continues to rise.

The Hard Swish activation function is one of MobileNetV3-Large's distinguishing characteristics. Hard Swish, in contrast to conventional ReLU, has been specifically designed to increase computing efficiency while maintaining the non-linearity necessary for capturing complex data patterns. This decision speeds model execution on hardware with limited resources while also enhancing the model's comprehension of complicated visual data.

We used the 28-layer MobileNetV3 Large model in our collaborative research. For tasks like object recognition, semantic segmentation, and image classification, this architecture's depth enables it to extract and encode increasingly complex and abstract characteristics from input images. Additionally, batch normalization and the Hard Swish activation function are included in this design after the convolutional layers. Batch normalization speeds up convergence and improves model performance by lowering internal covariate shift, which stabilizes and speeds up the training process. By using Hard Swish, the network's capacity to identify subtle correlations within the data is strengthened. Additionally, as part of our strategy, the MobileNetV3-Large is pre-trained on the principles of MobileNetV2. This transfer learning technique makes use of the knowledge and features that MobileNetV2 learned while being trained on a larger dataset. This improves the MobileNetV3-Large's performance and makes it possible for it to produce reliable findings even in situations with a dearth of labeled data.

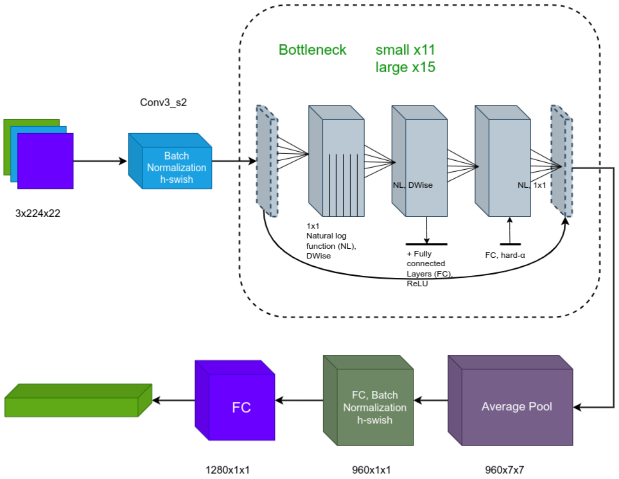
Bottleneck
small x11
large x15
Conv3_s2
Batch Normalization h-swish
3x224x22
1x1
Natural log function (NL), DWise
NL, DWise
+ Fully connected Layers (FC), ReLU
NL, 1x1
FC, hard-α
FC
FC, Batch Normalization h-swish
Average Pool
1280x1x1
960x1x1
960x7x7

Fig. 3. Resnet50-v2 architecture

The capacity of MobileNetV3 models to input photos of varied sizes is a remarkable feature. We used 224x224 pixels as the input size for this study. This decision might have been influenced by the particular specifications of our application or the available computational power. MobileNetV3 models, however, continue to be adaptable and can handle a variety of input dimensions, making them useful for a variety of image-processing jobs. The MobileNetV3-Large neural network design is strong and effective, and it excels in mobile and embedded vision applications. With 28 layers of depth, the clever integration of the Hard Swish activation function, and pre-training on MobileNetV2, this architecture is ideally suited for jobs requiring a trade-off between model size, processing speed, and accuracy. Its capacity to adjust to various input image sizes further strengthens its suitability for a wide range of use cases, from edge computing applications to smartphone-based image identification.

E. *Xception*

The convolutional neural network design known as Xception, which stands for ”Extreme Inception,” has received a great deal of attention and praise in the deep learning community. With a total of 71 layers, Xception stands out for its incredible depth. The network's ability to extract complex and hierarchical properties from input photos is demonstrated by its depth, which makes it especially well-suited for jobs requiring a deep comprehension of visual data. The Inception family of neural networks, which were innovators in the development of multi-branch, or ”inception,” modules, served as a model for the architecture of Xception. These modules can record and analyze features at various spatial scales. With the help of several depthwise separable convolutions, Xception brings this idea to its logical conclusion. This method drastically lowers the number of variables and calculations needed, improving efficiency without compromising performance. By doing this, Xception increases the network's capacity for representation while keeping the computational load under control.

The enormous ImageNet dataset used for its pre-training is one of the main reasons Xception is successful. ImageNet is a huge database of annotated photos that includes a wide range of item types. By using the knowledge and characteristics discovered from this substantial dataset during pre-training on ImageNet, Xception can lay a strong basis for subsequent fine-tuning or application-specific training. The network's capacity to classify photos into 1000 different object categories as a result is astounding. Xception is a flexible and effective tool for picture recognition, object detection, and many computer vision applications because of this feature. In conclusion, Xception is a powerful convolutional neural network for image-related tasks thanks to its outstanding depth, creative architectural design influenced by Inception, and pre-training on ImageNet. Many machine learning practitioners and researchers turn to it because of its capacity to interpret visual data effectively while keeping high performance. Xception is a valuable asset in the field of deep learning due to its adaptability and utility across a range of computer vision applications.

For the purposes of our particular solution, we set Xception to accept input photos with a 256x256 pixel size. The specifications of our application or the available computational resources may have had an impact on this decision. Modern deep learning architectures are known for their adaptability in input size, which enables Xception to be used efficiently for a variety of image-related tasks, from in-depth object detection to fine-grained image analysis. Rectified Linear Unit (ReLU) activation function was chosen for our implementation. Due to its computational effectiveness and capacity to add non-linearity into the model, ReLU is an activation function that is frequently used in deep neural networks. It aids in the network's ability to recognize intricate patterns and characteristics in the data, which enhances classification accuracy and speeds up learning. Xception is a great choice for applications requiring a high level of feature extraction and abstraction due to its exceptional depth. For difficult computer vision tasks like picture segmentation, object detection, and scene comprehension, this depth makes it possible to build a hierarchy of progressively complicated features. The multi-branch architecture of the network's foundation in Inception further enhances its capacity to record and display a variety of nuanced visual data.

In conclusion, Xception is a powerful convolutional neural network that excels in picture classification tasks by utilizing its depth, drawing inspiration from Inception, and pre-training on ImageNet. It is a versatile and effective tool for a variety of computer vision applications, including but not limited to object recognition, scene analysis, and image categorization. This is due to its adaptability to diverse input sizes and the use of ReLU as the activation function.

F. Model Fine-Tuning

As a specialized feature extractor, the BiT model works with photos of yoga poses to extract important and instructive information. A unique categorization layer is carefully incorporated into the model as part of our pursuit of accuracy. This specific classification layer is carefully crafted to align with the many categories of yoga poses that are included in our dataset. A thorough fine-tuning procedure is used to improve the model's performance and guarantee that it complies with the yoga posture recognition requirement. In this stage, sharpness and gamma correction on the training data are optimized by the use of computer vision techniques. Gradient descent principles guide the fine-tuning procedure, which updates model weights deliberately to reduce classification loss. In the context of yoga posture image categorization, this meticulous fine-tuning process gives the model the ability to adapt and specialize, which in turn allows it to produce precise and accurate predictions.

G. Performance Evaluation

By assessing the model's prediction power on the validation set, close observation of its performance is kept throughout the training phase. In order to make sure the model not only learns from the training data but also generalizes to new, unknown data efficiently, this continuous assessment is an essential quality control technique. Early halting techniques may be used as a protection against excessive training in order to prevent overfitting and encourage robust model training. After the training phase is complete, the completed model is thoroughly assessed on the designated test set. This assessment stage offers a reliable gauge of the model's performance in the real world and acts as the final straw for determining whether or not it can be generalized. Important assessment parameters include recall, accuracy, precision, and the F1-score. Taken as a whole, these provide a comprehensive picture of how well the model classifies photographs of yoga poses. These thorough evaluation methods guarantee that the model performs well not just in terms of accuracy but also in terms of precision and recall, which are essential for accurate and well-informed classification results.

H. Model Deployment and Applications

The use of the trained BiT model for real-time yoga posture classification is the research endeavor's apex. The model serves as a useful tool in this practical application, providing fast feedback to yoga practitioners, enabling precise posture recognition and supporting the quest of proper alignment and form. Our concept has practical applications not only in the field of research but also in the ever-changing domains of telemedicine and remote yoga instruction. This part of the study explores the concrete applications of our research, highlighting how our paradigm might transform yoga teaching and remote health care. Our methodology improves the remote yoga experience by offering individualized assistance based on real-time posture analysis, which promotes better accessibility and quality in yoga practice. This creative method not only empowers practitioners but also makes yoga accessible to a wider audience, democratizing its health advantages and advancing telehealth and remote learning in the field of wellness and fitness.

I. Baseline Architecture

One powerful deep neural network architecture designed to handle a variety of challenging computer vision applications is the Big Transfer (BiT) model. This architectural wonder is distinguished by a sophisticated fusion of several layers, each of which contributes differently to the network's operation. Although we can't cover all of the details of BiT's intricate architecture, including its intricate equations, in this talk, we can give you a general idea of the essential elements that are frequently present in convolutional neural networks (CNNs), like BiT. We can also provide equations for some of the important operations, which will help to clarify how these layers operate internally.

An overview of the layers and their equations typically found in a CNN like BiT:
Input Layer:: The input layer represents the raw image data, typically with dimensions (height, width, channels), where channels represent the color channels (e.g., red, green, blue). There's no specific equation for the input layer, as it just passes the input data to the subsequent layers.
Convolutional Layers:: Convolutional layers apply convolution operations to the input data to extract features. These layers have learnable filters or kernels. The convolution operation is defined as follows for a 2D convolution:

$$(I * K)(x, y) = \sum_{i=1}^{h} \sum_{j=1}^{w} I(x+i, y+j) \cdot K(i, j) \tag{1}$$

Where:

- I is the input feature map.
- K is the kernel or filter.
- h and w are the height and width of the kernel.

Activation Functions:: After convolution, an activation function is applied element-wise to introduce non-linearity. Common activation functions include ReLU (Rectified Linear Unit) and variants like Leaky ReLU and Parametric ReLU.Equation for ReLU:

$$f(x) = max(0, x) \tag{2}$$

Pooling Layers:: Pooling layers downsample the feature maps to reduce spatial dimensions and computational complexity. Max pooling is a commonly used operation:

$$MaxPool(x, y) = max(I(x+i, y+j)) \tag{3}$$

Normalization layers like Batch Normalization are used to improve training stability and convergence. Batch Normalization equation (simplified):

$$\text{BN}(x) = \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} \cdot \gamma + \beta \tag{4}$$

where:

- x is the input.
- μ is the mean.
- σ is the standard deviation.
- γ and β are learnable parameters
- $\in$ is a small constant for numerical stability.

Fully Connected Layers:: Fully connected layers are used for high-level feature aggregation and decision making. The equation for a fully connected layer operation is typically a matrix multiplication: $Y = XW + B$ where:

- X is the input.
- W is the weight matrix.
- B is bias vector.

Softmax Layer (for Classification):: In classification tasks, a softmax layer is applied to produce class probabilities. Softmax equation for class i:

$$P(y = i \mid X) = \frac{e^{X_i}}{\sum_{j=1}^{C} e^{X_j}} \tag{5}$$

where:

- X_i is the i-th element of the input.
- C is the number of classes.

These are the fundamental components and equations found in a typical CNN architecture, including the BiT model. The actual architecture of BiT can be quite deep and complex, with multiple blocks and variations to improve performance on various computer vision tasks. The specific architecture and equations for BiT would depend on the particular variant (e.g., BiT-S, BiT-M, BiT-L) we are using.

IV. Evaluation

We take a guided tour through the important aspects of our research in this part. We first define the condition of the dataset and describe the particular assessment criteria used in our investigation to give a thorough knowledge of our methodology. This fundamental stage guarantees that our data is ready and that we have the resources available to precisely assess the effectiveness of our models. The empirical setting is next covered, providing insight into the finer points of our experimental design. Here, we provide a clear and transparent view of the techniques and instruments utilized to carry out our research by explaining the decisions taken regarding model architecture, training settings, and data preprocessing. The part concludes with a careful assessment and a thorough examination of the findings. This stage examines our models' performance and determines how well they match our goals and metrics. We hope to get important insights and conclusions from this thorough examination and analysis that advance the main objectives of our study.

A. Dataset

1) Dataset 1: The Yoga Posture Dataset from Kaggle, which has a set number of classes and a predetermined amount of photos for every yoga stance, serves as the foundation for the suggested design. We used a dataset of 2,756 photos from 47 different classes, including positions such as Adho Mukha Svanasana, Ashta Chandrasana, and Setu Bandha Sarvangasana, to evaluate the accuracy and efficiency of the BiT model. These photos constitute the architecture's input data, and in order to maximize the efficiency of the model, the dimensions of each image are preprocessed to 256 by 256 pixels. We offer a preview of a few example photos from the dataset below.

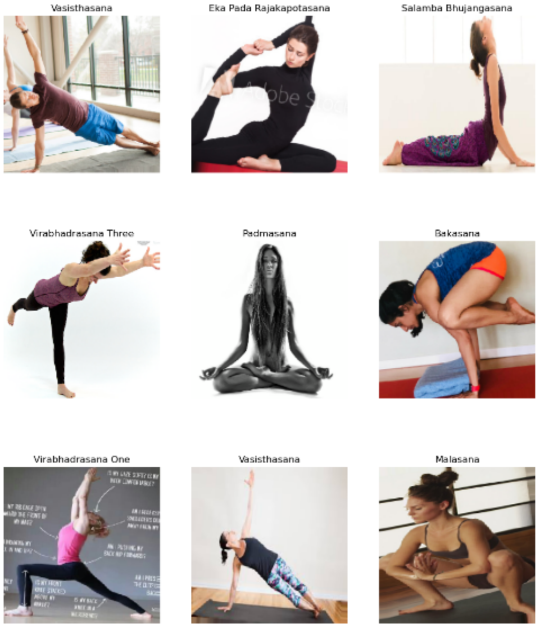

Fig. 4. The sample images of various yoga poses dataset 1

2) Dataset 2: The suggested architecture is also influenced by another dataset that is devoted to Yoga Poses and has a predetermined set of five different classes, each with a certain number of photos. We used a dataset of 954 photos from these five classes, including stances like Downdog, Goddess, and Plank, to evaluate the effectiveness and precision of the BiT-M model. The model uses these photographs as input, and in order to maximize the model's performance, each image is preformatted to a resolution of 256x256 pixels before processing. For reference, we've included some visual samples from this dataset below.

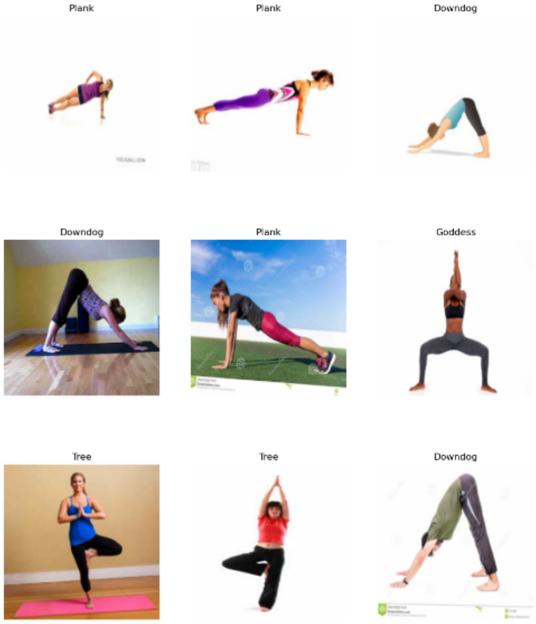

Fig. 5. The sample images of various yoga poses dataset 2

B. Evaluation metric

We must compare our model to other models that are currently in use in order to perform an exhaustive and comprehensive evaluation of our model. We can see the distinct advantages and benefits that the BiT model offers with clarity and discernment thanks to this comparative assessment, which acts as an essential benchmark. We may make informed decisions about our model's performance, suitability for the given task, and efficacy by comparing it to other well-established models. This thorough comparative analysis not only confirms the excellence of our model but also provides insightful information about its unique advantages and potential areas for development, thereby advancing the discipline.

TABLE II
DATASET 1

Model	Training accuracy	Validation accuracy
VGG16	66.30	59.38
BiT-M	80.00	69.14
MobileNetV3	81.06	79.52
Xception	66.68	54.25

TABLE III
DATASET 2

Model	Training accuracy	Validation accuracy
VGG16	90.51	89.46
BiT-M	**98.83**	**95.69**
MobileNetV3	84.83	87.82
Xception	96.60	89.95

One important factor to take into account with these models is the quantity of training epochs. The Big Transfer Model (BiT) was effectively trained for just 10 epochs, compared to

100 epochs for the other models. The BiT model's inherent efficiency, which allows it to make the best use of its training time, is the reason for this difference in training duration. It is important to emphasize that selecting a learning rate has a big impact on training with fewer epochs, which is why it is essential to getting the desired outcomes faster. It's crucial to remember that, due to the short training period, the results of the other models would probably be even worse than those shown in the table above if they had likewise been trained for only 10 epochs. Furthermore, BiT-M struggled to sustain validation accuracy due to Dataset 1's inherent class distribution imbalance, but MobileNetV3 fared better under these circumstances.

However, while working with Dataset 2, when an equal amount of photos are dispersed across each class, the true strengths of the BiT model become evident. Here, the BiT model proves its usefulness and surpasses other models, exhibiting its flexibility and efficiency, especially on datasets with well-balanced class distributions.

Result & Discussion

We commence a thorough assessment of the Big Transfer model's performance outcomes in this part. Through an in-depth examination of its many aspects, our analysis clarifies its advantages, disadvantages, and general efficacy within the framework of our investigation. Our goal in doing this review is to present a comprehensive analysis that highlights the model's accomplishments while also pointing out possible areas for development and optimization.

We will examine loss curve analysis in the next part, which is a useful tool for a more thorough assessment of the Big Transfer (BiT) model's performance. These loss curves give us vital information on the training and convergence of the model, enabling us to evaluate its efficacy and pinpoint any areas in need of improvement. We can better understand how the model learns and adapts during training by closely examining these loss curves, which ultimately leads to a more thorough evaluation of the model's overall performance.

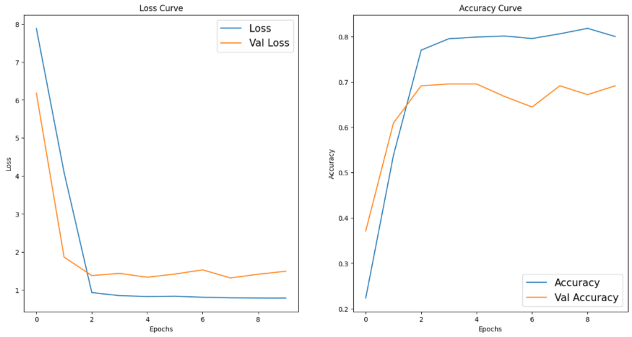

Fig. 6. loss curve & accuracy curve of BiT on dataset 1

We have evaluated the Big Transfer (BiT) model and found truly state-of-the-art results, with a special emphasis on learning curve analysis across various computer vision benchmarks.

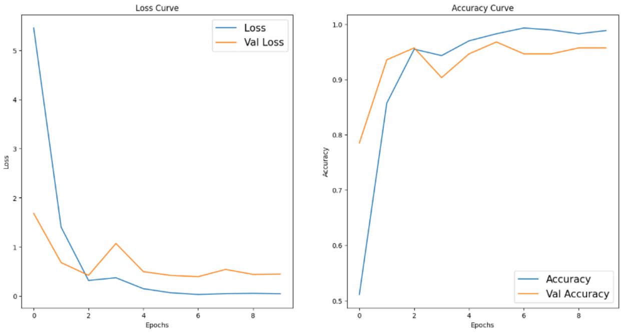

Fig. 7. loss curve & accuracy curve of BiT on dataset 2

Notably, the amount of time needed for training to achieve our particular research goals has been significantly decreased with the use of pretrained models, such BiT. BiT's pre-established learning rate setting, which maximizes the model's adaptability and quick convergence during training, is responsible for its exceptional efficiency. It's also important to note that BiT-M, which is built on the ResNet-50V2 architecture and has a significant depth of 50 layers, performed remarkably well in our tests. Even with only 10 training epochs, we were able to obtain excellent results with carefully adjusted learning curves. Surprisingly, the model may provide results as good in as few as 5 epochs because to its high efficiency. This is an extremely useful tool, especially when time or computational resources are limited.

To put it simply, the BiT model performs admirably "out of the box," requiring little fine-tuning to yield excellent outcomes. For academics and practitioners working with limited time or resources, its quick adaptation and low training epoch performance offer a useful benefit.

Comparing BiT-M results with other models:

In contrast to the Bit-M model, other models like VGG16, MobileNetV3, and Xception did not produce results in this short amount of time.They therefore consume more computing power than the Bit-M model.

VGG16

After 100 epochs of operation, the VGG16 model revealed its final results, which were not very encouraging.With 16 layers, it is believed that this model won't achieve its potential.

Xception

The Xception model was run over 40 epochs.Although it required fewer epochs to accomplish its task, the BiT-M model outperformed it in the datasets we gave.The computationally effective depthwise separable convolutions form the foundation of the Xception architecture. As a result, fewer parameters and computations are required while yet achieving good performance. Depthwise and pointwise convolutions are the two distinct layers that make up a depthwise separable convolution. Typically, this architecture lessens overfitting and expedites training.However, when we were using Xception to train our models, we noticed an increase in overfitting issues, which is why Xception's performance on the datasets we gave is not very encouraging.

MobileNetV3

MobileNetV3 is a family of convolutional neural network (CNN) architectures designed for efficient deep learning on mobile and embedded devices.We used MobileNetV3 Large and it showed promising result in dataset 1 but falls behind BiT in dataset 2.It is 28 layers deep model and it was pretrained over MobileNetV2 model.We ran this model for 100 epoch then we got results given in above table.We obversed increased training time compared to BiT-M model which is quite inefficient.

So,after comparing between each model we used in our study,we found BiT-M model much more efficient and better at giving us promising results.

FUTURE WORK

There are many opportunities to improve the Big Transfer Model even more so that it can identify a wider range of image collections more accurately. The following are the potential paths for further project work:

Fine-Tuning and Model Variations:

In our model, the BiT model was adjusted using a wide range of yoga posture datasets.In order to improve model performance specifically for yoga posture categorization, future study could examine the effects of fine-tuning hyperparameters and architecture modifications, such as various deep learning models.

Expand the Dataset:

Optimizing Dataset Efficiency by Trimming.The performance of the model is increasingly resilient as our dataset becomes more varied and large. To account for variations in lighting, backdrops, and practitioner apparel, also think about gathering data from several sources or implementing data augmentation techniques.

Multi-Modality Integration:

Including more multi-modal data in our research.We will attempt to combine picture data with sensor data (such as IMU sensors on the practitioner's body) to help us better comprehend yoga poses and improve the performance of the model.

Real-Time Feedback and Interaction:

Investigate how the model can be incorporated into apps or systems for yoga instruction to give practitioners feedback in real-time during their yoga sessions.This can entail creating a mobile or online application that makes use of the model to help users adopt the right posture.

Explainability and Transparency:

Look at ways to make the model's conclusions easier to understand and more open to scrutiny. Explainability is essential in AI models, particularly in the context of yoga, where alignment and safety are top priorities.

User Studies and Feedback:

Perform user studies and get feedback from yoga students and instructors who make use of the technology.This could enhance the user experience and reveal areas where the technology might benefit from development.

Remote Yoga Instruction and Telehealth:

Evaluate the viability and efficiency of applying our paradigm to telehealth and remote yoga instruction.How can our methodology accommodate the special requirements of distant learners and patients while also enabling personalized guidance?

Robustness and Generalization:

Evaluate the model's performance under difficult circumstances, including a range of lighting, practitioner attire, and yoga studio settings. For the model's applicability in the actual world, robustness and generalization abilities must be guaranteed.

Ethical Considerations:

Address privacy and ethical issues linked to the classification of yoga postures using deep learning models. This covers issues including permission, data privacy, and potential biases in the model's forecasts.

Comparison with Traditional Methods:

Evaluate how well deep learning models classify yoga postures in comparison to conventional approaches. Gaining knowledge of the advantages and disadvantages of both strategies can be extremely beneficial.

Long-Term Health Monitoring:

Investigate the possibilities for long-term health monitoring by employing these models to track practitioners' advancement and spot areas of growth or worry over time.

V. Conclusion

This research offers a creative solution that leads the way for yoga posture classification in a fast-changing world where old traditions and cutting-edge technology collide.We set out to overcome the numerous difficulties involved in accurately categorizing yoga postures by strategically implementing transfer learning and, in particular, by utilizing the impressive capabilities of the cutting-edge Big Transfer (BiT) model. Our goal is to offer a priceless resource to the yoga community at large, including practitioners, teachers, and academics.

We found that Big Transfer, our model, performed exceptionally well on the datasets presented, paving the way for future developments in posture categorization. Our research revealed that the optimized Big Transfer Medium model achieves 80%

training accuracy and 69.14% validation accuracy on the first dataset while requiring much less processing power and training faster. It also produces satisfactory results with fewer resources. On the second dataset, it subsequently obtained 95.69% validation accuracy and 98.83% training accuracy. For better outcomes, we increased the dataset and adjusted the BiT-M model. To improve the quality of the dataset, redundant photos were eliminated. Alongside the original dataset, we also used Computer Vision algorithms to improve the photos' gamma and sharpness. We then saved the improved images in another folder that we called the "modified dataset." The aforementioned outcomes were influenced by these improvements. This evolution provides enough space for the creation of transfer learning models, and in the near future, we can expect the birth of more effective models such as Big Transfer. There are plenty of chances for in-depth research in the field of posture categorization using transfer learning because there aren't many academics working in this area. As transfer learning continues to progress, we can expect ever more accurate findings for posture categorization.

As our research develops, it provides a look into a day where technology and yoga coexist together, increasing the health of yoga practitioners around the world. The way forward is to seamlessly combine these two realms, where technology enhances yoga's advantages and yoga, in turn, enhances the technological environment. Our research is an important step in realizing this synergistic vision, which has the potential to transform how we interact with this traditional practice in the modern era.

Acknowledgment

The authors thank Advanced Machine Intelligence Research (AMIR) Lab for resource sharing and their precious support.

References

[1] B. Z. Patel, "Yoga as a present and future in india," *International Journal of Health, Physical Education & Computer Science in Sports*, vol. 37, no. 1, p. 72, 2020.

[2] D. G. White, "Yoga, brief history of an idea," *Yoga in practice*, vol. 5, no. 1, pp. 1–23, 2012.

[3] Y. Zhang, R. Lauche, H. Cramer, N. Munk, and J. A. Dennis, "Increasing trend of yoga practice among us adults from 2002 to 2017," *The journal of alternative and complementary medicine*, vol. 27, no. 9, pp. 778–785, 2021.

[4] A. K. Upadhyay, A. Balkrishna, and R. T. Upadhyay, "Effect of pranayama [voluntary regulated yoga breathing] and yogasana [yoga postures] in diabetes mellitus (dm): A scientific review," *Journal of Complementary and Integrative Medicine*, vol. 5, no. 1, 2008.

[5] Y.-H. Byeon, J.-Y. Lee, D.-H. Kim, and K.-C. Kwak, "Posture recognition using ensemble deep models under various home environments," *Applied Sciences*, vol. 10, no. 4, 2020.

[6] Y. Ke, C. ZENG, X. LU, and Y. CUI, "Recognition technology of human body movement behavior in fitness exercise based on transfer learning," in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 1002–1006, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[8] H. E. Kim, A. Cosa-Linan, N. Santhanam, *et al.*, "Transfer learning for medical image classification: a literature review," *BMC Med Imaging 22*, vol. 69, 2022.

[9] M. Fraiwan, Z. Audat, L. Fraiwan, and T. Manasreh, "Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images," *PLOS ONE*, vol. 17, pp. 1–21, 05 2022.

[10] F. B. Ashraf, M. U. Islam, M. R. Kabir, and J. Uddin, "Yonet: A neural network for yoga pose classification," *SN Computer Science*, vol. 4, no. 2, p. 198, 2023.

[11] J. Palanimeera and K. Ponmozhi, "Classification of yoga pose using machine learning techniques," *Materials Today: Proceedings*, vol. 37, pp. 2930–2933, 2021.

[12] S. Liaqat, K. Dashtipour, K. Arshad, K. Assaleh, and N. Ramzan, "A hybrid posture detection framework: Integrating machine learning and deep neural networks," *IEEE Sensors Journal*, vol. 21, no. 7, pp. 9515–9522, 2021.

[13] A. Sharma, Y. Shah, Y. Agrawal, and P. Jain, "Real-time recognition of yoga poses using computer vision for smart health care," *arXiv preprint arXiv:2201.07594*, 2022.

[14] J. Jose and S. Shailesh, "Yoga asana identification: a deep learning approach," in *IOP Conference Series: Materials Science and Engineering*, vol. 1110, p. 012002, IOP Publishing, 2021.

[15] R. O. Ogundokun, R. Maskeliūnas, and R. Damaševičius, "Human posture detection using image augmentation and hyperparameter-optimized transfer learning algorithms," *Applied Sciences*, vol. 12, no. 19, 2022.

[16] M. Hussain, J. J. Bird, and D. R. Faria, "A study on cnn transfer learning for image classification," in *Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK*, pp. 191–202, Springer, 2019.

[17] R. O. Ogundokun, R. Maskeliūnas, S. Misra, and R. Damasevicius, "A novel deep transfer learning approach based on depth-wise separable cnn for human posture detection," *Information*, vol. 13, no. 11, 2022.