Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.0000/ACCESS.20xx.DOI

Deepfake Detection: Analysing Model Generalisation Across Architectures, Datasets and Pre-Training Paradigms

SOHAIL AHMED KHAN[1] and DUC-TIEN DANG-NGUYEN[1] (Member, IEEE)
[1]MediaFutures, Department of Information Science and Media Studies, University of Bergen, Norway

Corresponding authors: Sohail Ahmed Khan and Duc-Tien Dang-Nguyen (e-mail: {sohail.khan, ductien.dangnguyen}@uib.no)

ABSTRACT

As deepfake technology gains traction, the need for reliable detection systems is crucial. Recent research has introduced various deep learning-based detection systems, yet they exhibit limitations in generalizing effectively across diverse data distributions that differ from the training data. Our study focuses on understanding the generalization challenges by exploring specific aspects such as deep learning model architecture, pre-training strategy and datasets. Through a comprehensive comparative analysis, we evaluate multiple supervised and self-supervised deep learning models for deepfake detection. Specifically, we evaluate eight supervised deep learning architectures and two transformer-based models pre-trained using self-supervised strategies (DINO, CLIP) on four different deepfake detection benchmarks (FakeAVCeleb, CelebDF-V2, DFDC and FaceForensics++). Our analysis includes intra-dataset and inter-dataset evaluations, examining the best performing models, generalisation capabilities and impact of augmentations. We also investigate the trade-off between model size, efficiency and performance. Our main goal is to provide insights into the effectiveness of different deep learning architectures (transformers, CNNs), training strategies (supervised, self-supervised) and deepfake detection benchmarks. Through our extensive analysis, we established that Transformer models outperform CNN models in deepfake detection. Also, we show that FaceForensics++ and DFDC datasets equip models with comparably better generalisation capabilities, as compared to FakeAVCeleb and CelebDF-V2 datasets. Our analysis also show that image augmentations can be helpful in achieving better performance, at least for the Transformer models.

INDEX TERMS deepfakes; image classification; convolutional neural networks; transformers; video processing.

I. INTRODUCTION

DEEPFAKES, or deepfake media, are digital media that have been generated or modified using deep learning algorithms [1]. They have gained notoriety in recent years due to their potential to manipulate and deceive by producing fraudulent and deceptive media content. While deepfakes can serve innocent or even entertaining purposes, they also harbor substantial dangers when harnessed for malicious intentions, like crafting convincing fraudulent media to sway public opinion, manipulate electoral outcomes, or incite violence [2], [3], [4]. Also, given the prevalence of powerful and budget-friendly computing resources along with the widespread accessibility of paid, as well as open-source software, the creation of deepfakes has become increasingly straightforward [5]. This accessibility extends to individuals with limited technical knowledge, facilitating the production of remarkably convincing deepfakes that closely resemble genuine content.

The research community has been actively proposing novel AI-based automated deepfake detection models, trying to address these issues posed by deepfake media [6], [7], [8], [9], [10], [11]. However, a significant issue associated with current deepfake detection models is their lack of generalisation capability [1], [7], [12]. This means that these detection systems work very well when dealing with deepfakes that come from the same data distribution as they were trained on. However, they struggle to perform well when exposed to deepfakes generated using different methods than the ones used for training.

Previous research efforts have introduced a multitude of

carefully designed deep learning models for deepfake detection, accompanied by novel techniques for training (e.g., augmentations, multi-modal training setup, diverse set of training features etc) and evaluating these models on well-known deepfake datasets. However, given the vast volume of research publications, it has become increasingly challenging to discern which kind of architectures yield optimal results and which datasets are most effective in facilitating robust model performance, thus enhancing generalisation to unseen data. In light of these considerations, we contend that a comprehensive analysis that unites a diverse range of deep learning architectures, trained and assessed across multiple prominent deepfake datasets in a unified manner, is imperative to gain a deeper understanding of this issue. We also think that such a comparative analysis has the potential to uncover valuable insights for identifying the most suitable architecture and dataset(s) to enhance the effectiveness of deepfake detection. Consequently, we believe that this analysis can contribute significantly to addressing the current challenge of model generalisation in the realm of deepfake detection.

In this study, we carry out a comprehensive comparative analysis of several widely recognised deep architectures for image and video recognition, aiming to assess their efficacy in detecting deepfakes. Our primary goal is to determine which among these models achieves superior performance on unseen, out-of-distribution data, i.e., exhibit impressive generalisation capability. The models selected for our study comprise of both Convolutional Neural Networks (CNNs) and Transformer models. The rationale behind incorporating transformer models is rooted in their recent notable achievements across a spectrum of computer vision tasks such as image classification [13], [14], [15], object detection [16], [17], image segmentation [18], video classification, multi-modal learning [19], [20], 3D analysis [21], [22] and beyond [23].

For our analysis we train all participating models on four deepfake detection datasets and evaluate them in both intra-dataset [1] and inter-dataset [2] configurations (see Figure 1). Additionally, we evaluate the difficulty level of each benchmark and investigate whether a more challenging benchmark leads to better generalisation performance on unseen data. To this end, we train participating models on all four datasets twice: first, without any image augmentations and then with various image augmentations to find out if they improve models' performance.

Since recently, transformer models trained using self-supervised methods have exhibited their capability to produce robust visual features [24], [25], [26]. Subsequently, models trained through self-supervised methods have been shown to achieve excellent performance on new tasks, often without the need for additional training or with minimal training [18], [24], [25], [27], [28]. Owing to this, we also analyse Vision Transformer (ViT) architecture pre-trained using two well-known self-supervised learning strategies: **DINO** [24] and **CLIP** [25]. We choose to ViT in this case since it is shown to achieve better performance as compared to CNN architecture, i.e., ResNet [24]. To study these models and find out how good the self-supervised features are, we use self-supervised ViT-Base models (DINO and CLIP) as feature extractors and train a classification head on top of them. It is important to note that we only train the classification head and freeze the weights of the feature extractors to avoid backpropagating gradients through them.

In summary, our study aims to provide insights into various aspects, including: (1) identifying the most effective models for detecting deepfakes among those being tested, (2) pinpointing the model with the highest ability to adapt to new and unseen data, (3) assessing the difficulty of different datasets for model training, (4) determining the dataset that best facilitates generalisation to unseen data, (5) evaluating the performance of self-supervised training strategies and (6) examining the impact of augmentations on enhancing model performance.

This next parts of this paper is organised as follows. In Section II we present a brief literature review on the topic of deepfake detection. Section III presents the proposed framework. In Section IV we present the results and discussion of our findings and finally Section V concludes this study by summarising our analysis and presents future research direction.

II. LITERATURE REVIEW

Since recently quite a large number of research studies focused on deepfake media detection have been proposed. Most studies employ CNN models trained on large amounts of data in order to detect deepfake media. The proposed studies also employ different strategies e.g., novel augmentation techniques [29], hybrid models [9], [30], biological features [31], multi-modal features [6], [9], temporal features along with spatial information [9], [10], [32], recurrent networks, transformer models [8], [9] etc to detect deepfake images/videos while trying to increase the models' generalisation capabilities. Below we present some well-known, as well as some of the recently proposed deepfake detection studies. We chose to review studies in this section that share similarities with ours, focusing on common aspects such as the selection of detection models and the datasets used to train and evaluate the proposed models.

A. CNN BASED DETECTION MODELS

In 2019, Rossler *et al.* released FaceForensics++, a dataset for deepfake detection [33]. The dataset, containing over 1.8 million manipulated images, was made publicly available. Using the dataset authors also conducted an extensive analysis of several data-driven forgery detection methods. The methods included traditional machine learning models (SVMs) trained on handcrafted features, as well as contemporary deep learning architectures including MesoNet [34] and XceptionNet [35]. By conducting a thorough analysis,

[1] models trained and evaluated on the same dataset
[2] models trained on one dataset and evaluated on another dataset

authors discovered that a vanilla deep CNN model, XceptionNet [35], outperforms other participating models significantly in the context of detection task on compressed low quality data. Authors additionally demonstrated, via experiments and surveys, that the data-driven models outperform humans in detecting deepfakes. Nevertheless, the paper lacks a cross-dataset analysis of the models, which could have been beneficial in understanding the generalisation performance across diverse and unseen domains.

In [32] Sabir *et al.* proposed a deepfake detection system by focusing on the temporal information present in video streams to exploit temporal discrepancies across multiple frames. In order to analyse temporal data authors employed a recurrent convolutional architecture [36], [37] comprising of a CNN for feature extraction and a BiDirectional Recurrent Neural Network (BiDir RNN) to analyse temporal information present in videos. Specifically, authors studied two different CNN architectures, ResNet [38] and DenseNet [39] for feature extraction. The authors also employed a carefully crafted pre-processing regime to preprocess facial frames before inputting them into the models. The models were evaluated using the renowned FaceForensics++ deepfake detection benchmark [33], showing excellent results in an intra-dataset evaluation regime. Authors do not carry out a cross-dataset analysis in their study.

Ciftci *et al.* [31], shifted away from traditional image features and proposed to employ biological signals (i.e., photoplethysmography or PPG signals which detects subtle alterations in color and motion within RGB videos) to train their models. The proposed model was comprised of a CNN, as well as a SVM. The CNN and SVM models made their individual predictions on the provided features sets, which were then fused together in order to get a final classification score. The proposed deepfake detection scheme achieved promising results when tested using both intra-dataset as well as inter-dataset configurations on multiple different deepfake detection benchmarks including, CelebDF [40], FaceForensics [41] and FaceForensics++ [33] datasets.

In study [6], Zhu *et al.* introduced a deepfake detection framework that leveraged 3D face decomposition features for detecting deepfakes. The authors demonstrated that the fusion of 3D identity texture and direct light features notably enhanced the detection performance, simultaneously promoting the model's generalisation ability when assessed across different datasets. The training of the detection model involved both a cropped facial image and its corresponding 3D attributes. Authors employed XceptionNet [35] for feature extraction. The study also provides an extensive analysis of various feature fusion strategies. The proposed model was trained on the FaceForensics++ [33] benchmark and subsequently evaluated on three datasets: (1) FaceForensics++, (2) Google Deepfake Detection Dataset [42] and (3) DFDC [43] dataset. The reported evaluation statistics showed promising results across all three datasets, highlighting the model's robust generalisation capability in comparison to previously proposed deepfake detection methods.

B. TRANSFORMER BASED DETECTION MODELS

In study [9], Khan *et al.* introduced the utilisation of transformer architecture for the purpose of deepfake detection, presenting two novel models: (1) Image Transformer and (2) Video Transformer. Both models were trained using 3D face features [44] in addition to standard cropped face images. The integration of 3D face features aimed to swiftly obtain accurately aligned facial details, enhancing the learning process. The combination of these aligned features with conventional cropped face data contributed to the acquisition of pertinent facial details. To harness temporal information within videos, authors modified the standard Vision Transformer (ViT) [13] to accommodate multiple successive face frames. Notably, the proposed model exhibited incremental learning capabilities, accommodating new data without forgetting prior knowledge. The authors conducted comprehensive evaluations of their models across prominent deepfake detection benchmarks, including FaceForensics++ [33], DFDC [43] and Google DFD [42]. Their models showed impressive performance across all these datasets, underscoring their efficacy in deepfake detection.

Wang *et al.,* [8] introduced a Multi-modal Multi-scale TRansformer (M2TR) model, which processes patches of multiple sizes to identify local abnormalities in a given image at multiple different spatial levels. M2TR also utilises the frequency domain information along with RGB information using a sophisticated cross-modality information fusion block to detect forgery related artifacts in a better way. Through extensive experiments authors establish the effectiveness of M2TR and show their model outperforms SOTA Deepfake detection models by acceptable margins.

Coccomini *et al.,* in [30] propose a video deepfake detection model based on a hybrid transformer architecture. Authors used an EfficientNet-B0 as feature extractor. The extracted features were then used to train two different types of Vision Transformer models in their study, e.g., (1) Efficient ViT and (2) Convolutional Cross ViT. Through experimentation, authors established that the model comprising of EfficientNet-B0 feature extractor and Convolutional Cross ViT achieved the best performance scores as compared to other models that they tested.

Zhao *et al.,* [10] propose an Interpretable Spatial-Temporal Video Transformer (ISTVT) for deepfake detection was proposed. The proposed model incorporates a novel decomposed spatio-temporal self-attention as well as a self-subtract mechanism to learn forgery related spatial artifacts and temporal inconsistencies. ISTVT can be also visualise the discriminative regions for both spatial and temporal dimensions by using the relevance propagation algorithm [10]. Extensive experiments on large-scale datasets were conducted, showing a strong performance of ISTVT both in intra-dataset and inter-dataset deepfake detection establishing the effectiveness and robustness of proposed model.

Through this literature review it becomes apparent that the research community actively employs deep learning models along with other techniques to try develop robust and

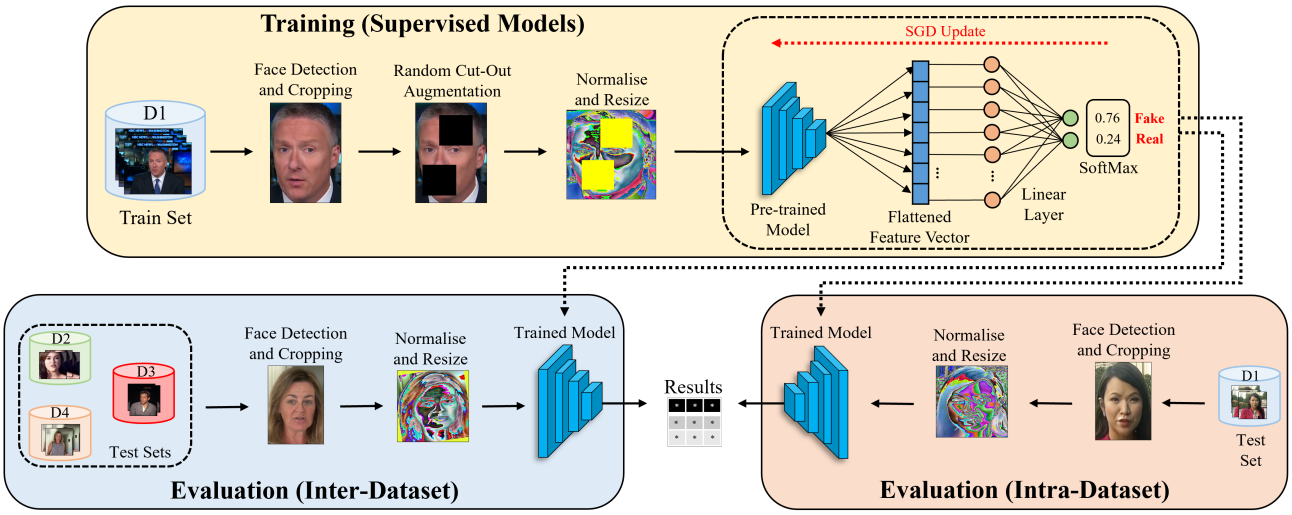

FIGURE 1. The proposed framework. *The process involves several steps, starting with the extraction and cropping of face frames from videos, followed by augmentation, normalisation and resizing. The pre-trained models are then used as feature extractors, with a new classification head (linear layer) added on top for supervised models. During training, the weights of both the feature extractor and the classification head are updated for supervised models, while only the newly added classification head is updated for self-supervised models. The models are evaluated through both intra-dataset and inter-dataset evaluations to test their performance and generalisation capabilities. For image models, the input data is a single cropped face image, while for video models, it is a tensor containing eight consecutive cropped face images from a given video.*

efficient deepfake detectors. However, while carefully reading the research studies it also becomes noticeable that the models perform poorly on unseen, out-of-distribution data. In addition to this, there is a lack of comparative studies which aim to identify which specific family of deep learning architectures is better for the task of deepfake detection. Furthermore, it's not easy to determine without thorough experimentation that which of the well-known datasets offer improved generalisation potential to the models, i.e., allow models to better handle new and unseen data.

To address this, we study some of the most frequently used architectures (EfficentNets, XceptionNet, Vision Transformers) in the literature of deepfake detection in this research. We also employ widely known datasets for experimentation and try to find out the datasets offering best generalisation capabilities to the models. We also analyse some of the understudied approaches for deepfake detection i.e., we train and evaluate the performance of self-supervised models on deepfake detection and compare their performance with that of the supervised models.

III. THE PROPOSED FRAMEWORK

The workflow followed in this study for training and evaluating the models is illustrated in Figure 1. On top we show the training pipeline where we start by extracting and cropping faces from videos. The cropped face frames are then augmented, normalised and resized before being fed to the model for training. We load pre-trained models as feature extractors, i.e., we remove the last layer from the loaded models and add a new classification head (linear layer) on top. For supervised models, during training we update weights of both feature extractor as well as the classification head.

For self-supervised models, our objective is to assess the quality of the representations they produce since they were initially trained through self-supervised training strategies. Consequently, for these models we only update weights of the newly added classification head while maintaining the frozen weights of the feature extractor backbone. This strategy enables us to directly compare the self-supervised feature representations with those obtained from supervised models. Since we deal DINO and CLIP as feature extractors, we follow the guidelines provided in their respective code repositories to extract features.

For DINO, we extract features from the last four encoder blocks, as this configuration yielded optimal results. On the other hand, for CLIP, we exclusively extract features from the last encoder block. We then feed these features into the classification head.

For intra-dataset evaluation we evaluate models on the same dataset (test set) it was trained on, e.g., model trained on dataset D1 is evaluated on the test set of D1. The primary objective of intra-dataset evaluation is to discern which model achieves the highest performance score as compared to other participating models on each of the dataset. Moreover, this evaluation will offer insights into which dataset presents the greatest learning challenge for the models and which dataset is comparatively easier to learn.

In the context of inter-dataset evaluation, we evaluate models that were trained on one dataset across the remaining three datasets. For instance, a model trained on dataset D1 is tested on D2, D3 and D4 datasets. The objective of inter-dataset evaluation is two-fold: first, to investigate the models' ability to generalise across datasets and second, to understand how effectively the training dataset empowers models to generalise well on unseen data.

The input data for training and evaluating image models

is a single face cropped image ($[3, 224, 224]$), whereas, input data for training and evaluating video models is a tensor containing 8 consecutive face cropped images ($[8, 3, 224, 224]$) from any given video.

A. DATASETS

In this study we train and evaluate several different deep learning models on four deepfake detection datasets/benchmarks: FakeAVCeleb [45], CelebDF-V2 [40], DFDC [43] and FaceForensics++ [33]. All of the four datasets comprise of real and fake videos, where fake videos are generated using different deepfake generation methods. In upcoming sections, we present a brief description of these datasets.

FaceForensics++ [33] is one of the most widely studied deepfake detection benchmarks. FaceForensics++ comprises of 1000 real video sequences (mostly from YouTube) of mostly frontal faces and without any occlusions. These real videos were then manipulated using four different face manipulation methods: (1) FaceSwap [46], (2) Deepfakes [47], (3) Face2Face [48] and (4) NeuralTextures [49] to have four subsets. Each subset comprises of 1000 videos each. In total, the dataset contains 5000 videos, i.e., 1000 real and 4000 fake videos. FaceForensics++ offers 3 different qualities of data, (1) Raw, (2) High-Quality and (3) Low-Quality. In our study, we experimented the high-quality videos.

FaceSwap and Deepfakes subset contains videos generated using what is called, face swapping. As the name suggests, face of the target person is replaced with the face of source person and results in transferring the identity of the source person onto the target. Face2Face and NeuralTextures subsets are generated by a different process called, face re-enactment. In contrast to face swapping, face re-enactment swaps the faces of source and target, however, keeps the original identity of the target face.

Deepfake Detection Challenge (DFDC) dataset [43] comprises of around 128k videos, out of which, around 104k are fake. Similar to the FaceForensics++, the DFDC also comprises of videos generated using more than one face manipulation algorithms. Five different methods were employed to generate fake videos, namely, (1) Deepfake Autoencoder [43], (2) MM/NN [50], (3) NTH [51], (4) FSGAN [52] and (5) StyleGAN [53]. In addition to these, a random selection of videos also underwent a simple sharpening post-processing operation which increases the videos' perceptual quality. Unlike FaceForensics++ dataset, the DFDC dataset also contains videos having undergone audio-swapping. However, in this study we do not use audio features to train and evaluate our models.

Since DFDC dataset is huge as compared to other participating datasets, we only use a subset of the full dataset to train and evaluate our models i.e., to keep the number of training, validation and test data nearly similar. For training we use roughly around 19500 (around 16500 fake and 3100 real) randomly selected videos from which we get 100k face cropped images (50k real and 50k fake). We use 20k images as validation set. For testing the models we use 4000 face frames randomly selected from 3500 (3200 fake and 300 real) videos.

CelebDF-V2 [40] contains 5639 fake and 590 real videos. The real videos are collected from YouTube and contain interview videos of 59 celebrities having diverse ethnic backgrounds, genders, age groups. CelebDF-V2 dataset comprises of fake videos generated using Encoder-Decoder models. Post-processing operations are also employed to circumvent color mismatch, temporal flickering and inaccurate face masks.

FakeAVCeleb [45] is the most recently proposed deepfake detection dataset. FakeAVCeleb dataset contains 19500 fake and 500 real videos. This dataset also contains audio modality and manipulates audio as well as video content to generated deepfake videos. For video manipulation, FaceSwap [54] and FSGAN [52] alogrithms are used. For audio manipulation, a real-time voice cloning tool called SV2TTS [55] and Wav2Lip [56] are used. The dataset is divided into 4 subsets, i.e., (1) FakeVideo/FakeAudio, (2) RealVideo/RealAudio, (3) FakeVideo/RealAudio and (4) RealVideo/FakeAudio.

In this study, we only employ 2 of the mentioned subsets to train our models, i.e., (1) FakeVideo/FakeAudio and (2) RealVideo/RealAudio.

TABLE 1. The amount of `real/fake` images used to train, validate and test our image models.

Train/Test Data						
Dataset	Train		Validation		Test	
	Real	Fake	Real	Fake	Real	Fake
FakeAVCeleb [45]	47,808	47,808	5,360	5,360	2,000	2,000
CelebDF-V2 [40]	50,000	50,000	10,000	10,000	1,000	1,000
DFDC [43]	50,000	50,000	10,000	10,000	2,000	2,000
FaceForensics++ [33]	50,000	50,000	10,000	10,000	2,000	2,000

B. DATASET PREPARATION

The data preparation process was notably time-consuming due to two main factors: firstly, the datasets being substantial in size and secondly, some selected datasets lacking clear dataset preparation guidelines. For instance, FakeAVCeleb does not provide predefined train/validation/test splits. Consequently, we had to manually develop a strategy to effectively partition the dataset into train, validation and test sets. Ensuring that a single identity didn't appear in multiple splits added another layer of complexity to this task.

Additionally, all the datasets exhibit an imbalance, with a significantly higher number of "fake" videos compared to "real" ones. To address this, we took steps to ensure that the resulting datasets of cropped face images are balanced by extracting faces from videos. Our efforts aimed to include at least one frame from every video selected for training and evaluation. You can refer to Table 1 for detailed information

regarding the number of face frames used for training, validation and model evaluation from each dataset.

The data provided in Table 1 clearly illustrates that the training and validation sets for FakeAVCeleb contain a relatively smaller number of frames. This discrepancy can be attributed to the dataset containing a limited count of real videos (only 500), while the number of fake videos is substantially larger (19500). As the video clips are of shorter duration, this translates to 47,808 frames being extracted from the chosen 300 real videos for the training set and 5360 frames from 100 videos for the validation set. Despite this slight variance in the size of the training and validation sets, we assume that it has a minimal impact on the models' performance. This assumption is supported by our observations from training and evaluating the models using even fewer frames (approximately 25k real and 25k fake frames), which resulted in no significant deviations in the final test scores.

In addition, the test set for CelebDF-V2 contains fewer frames for the same underlying reason – the test set of the dataset includes only 50 real and 50 fake videos. In response, we meticulously extracted a total of 2000 frames from this set of 100 test videos for the purpose of evaluation.

C. PREPROCESSING AND AUGMENTATIONS

We adopt two distinct approaches to train our models in this study. Initially, we train models without applying any image augmentations. Subsequently, we train the models using a range of randomly chosen image augmentations, such as horizontal flips, affine transformations and random cut-out augmentations. All the cropped face images are then normalised according to the same method used for pre-training models on the ImageNet [57]. For implementing the augmentations, we utilise the *imgaug*[3] library.

D. MODELS

We opt to explore six supervised image recognition models, equally divided into three CNNs and three transformer-based models. Furthermore, we assess two variations of transformer models trained via self-supervised methods, namely (1) DINO [24] and (2) CLIP [25]. In addition to the image classification models, our study encompasses the training and evaluation of two distinct video classification models: (1) ResNet-3D [58], a CNN model for video classification and (2) TimeSformer [59], a transformer model tailored for video classification.

We choose models based on their performance on the ImageNet benchmark [57], their parameter count and, in the case of certain models like Xception [35] and EfficientNet [43] their established performance in deepfake detection, as reported by some of the previous studies [33], [43].

1) Image Models

Deepfake detection is typically treated as an image classification problem. In this context, deep learning models are trained and evaluated on images independently, dealing with each image on its own. This differs from video-based deepfake detection, where models are trained and tested on consecutive video frames to capture temporal discrepancies between frames along with spatial cues within each frame.

Below, we provide a brief introduction to the image models employed in this study.

- **Xception** [35] is a convolutional neural network (CNN) architecture that builds upon the Google's Inception CNN architecture [60]. It distinguishes itself by using depth-wise separable convolutions in place of conventional Inception modules. Unlike standard convolutions applied across all N channels at once, depth-wise convolutions operate sequentially on individual image channels. This characteristic reduces Xception's trainable parameters compared to other prominent deep CNN models. Despite this reduction, Xception's performance remains on par with models having more parameters, as evidenced on the ImageNet benchmark [57]. Furthermore, its smaller parameter count enhances resistance to overfitting on unseen data and decreases computational load, making it an efficient choice. Figure 2A illustrates the concept of depth-wise convolution, the fundamental building block of Xception. Xception not only demonstrates excellence on the ImageNet benchmark but also boasts significant achievements in previous deepfake detection studies [6], [33], [43]. Based on its proven track record in this domain, we include Xception for analysis in this study.
- **Res2Net-101** [61] is a CNN architecture which is built upon the widely adopted ResNet architecture [38]. Res2Net introduces a new building block named the "Res2Net Block," which replaces the conventional bottleneck residual blocks utilised in ResNet models. By operating at a granular level, the Res2Net architecture captures multi-scale features and extends the receptive field range for every network layer. As a result, the network becomes more potent and efficient, leading to enhanced performance across diverse computer vision tasks, including image classification, segmentation and object detection [61]. The innovative Res2Net block can be seamlessly integrated into other leading-edge backbone CNN models, such as ResNet [38], DLA [62], BigLittleNet [63] and ResNeXt [64]. We visualise the Res2Net block in Figure 2B. In this study, we employ Res2Net-101 to explore whether multi-scale CNN features contribute to improved deepfake detection performance. Additionally, we investigate whether these enhancements extend to cross-dataset performance, gauging the model's generalisation capability.
- **EfficientNet-B7** [65] belongs to the EfficientNet family of CNN architectures. In their paper, the authors propose a scaling technique that uniformly adjusts depth, width and resolution using a compound coefficient. The central concept revolves around systematically scaling the model's architecture and parameters to achieve better efficiency. Unlike the conventional approach of arbitrarily scaling individual dimensions, the proposed strategy employs a consistent set of scaling coefficients across all dimensions. Consequently, the

[3]https://imgaug.readthedocs.io/en/latest/

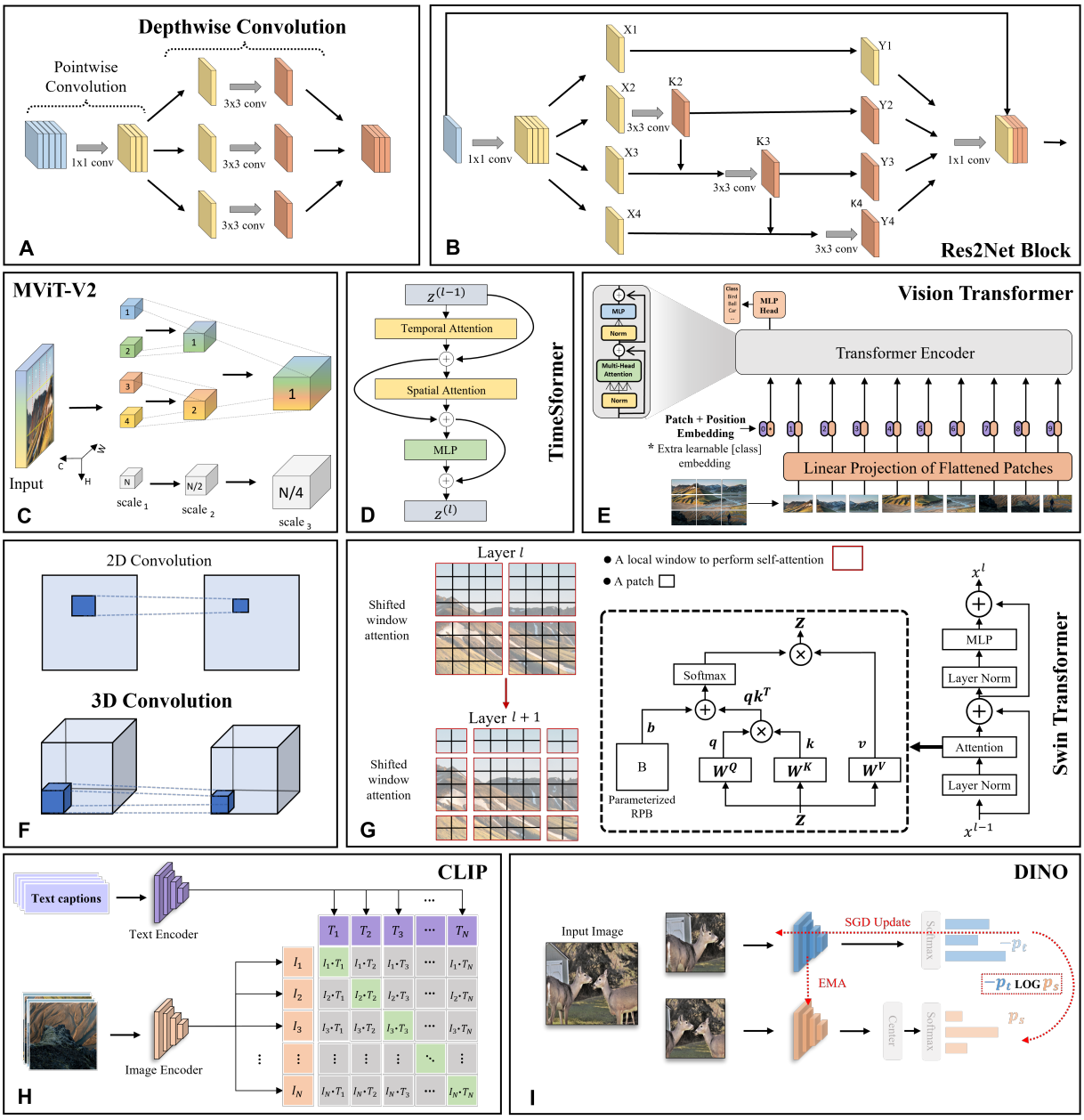

FIGURE 2. Visual representation of the models used for analysis in this study. Due to space limitations, only basic, key concepts for each model are illustrated instead of the whole model. For optimal understanding of the essential components of each model, we recommend viewing this figure in color and at a higher magnification.

architecture offers a family of seven models spanning various scales [65]. Impressively, EfficientNet achieves top-notch performance across several image classification benchmarks, while maintaining computational efficiency that surpasses other architectures like ResNet and Inception [65]. In a manner similar to Xception, a specific variant of EfficientNet, namely EfficientNet-B7, has also demonstrated remarkable prowess in deepfake detection tasks. Notably, the triumphant solution of the Google-sponsored Deepfake Detection Challenge (DFDC) was built upon the strengths of EfficientNet-B7 models [43]. Given this notable track record, our research aims to delve into the potential of this model in our study.

- **Vision Transformer (ViT-Base)** [66] belongs to the family of transformer models which were initially designed for natural language processing tasks. In the realm of computer vision, the Vision Transformer (ViT) emerged as a pioneering transformer-based architecture designed specifically for image classification tasks [13]. ViT harnesses the power of self-attention mechanisms to processes visual data. Its methodology employs a deceptively simple yet impactful strategy: the division of images into smaller patches, which are then fed into a transformer model as a unified entity. These patches are enriched with positional embeddings, enabling them to retain their spatial context within the original image.

A classification token is introduced at the outset of this input, which is subsequently processed by the transformer encoder—a mechanism reminiscent of the encoders in text-oriented transformer models. This approach empowers the model to better capture the context and relationships between different parts of the image. As a result, the network effectively captures contextual nuances and interrelationships across distinct segments of the image, achieving performance comparable to state-of-the-art CNN models on the ImageNet dataset, especially when trained on giant datasets like ImageNet-21k or JFT-300M. The ViT architecture is visually depicted in Figure 2E. In our analysis, we undertake the training and evaluation of the base version of ViT-Base model for the deepfake detection task and subsequently compare its performance against other models participating in the study.

- **Swin Transformer (Swin-Base)** [14] is a class of Vision Transformer models. Swin Transformer architecture comes with a hierarchical structure, utilising a shifted windows approach for computing image representations. The shifted windowing strategy enhances efficiency by confining self-attention computation to non-overlapping local windows, while still enabling cross-window connections. This hierarchical design offers flexibility for modeling at different scales and maintains linear computational complexity concerning image size. Swin Transformers achieve competitive performance, comparable to other state-of-the-art image classification models like EfficientNets [14], [65] and even outperform Vision Transformers and ResNets [13], [38]. Not only limited to image classification, Swin Transformers also excel in tasks such as image segmentation and object detection [14]. Figure 2G provides an illustration of the window generation and attention calculation process in Swin Transformers. Because of the excellent performance Swin Transformer achieve on ImageNet, we use it for the task of deepfake detection and try to study how it performs as compared to other participating models.
- **Multiscale Vision Transformer (MViT-V2-Base)** [15] is another class of ViT models. Unlike traditional ViTs, the MViTs have multiple stages that vary in both channel capacity and resolution. These stages create a hierarchical pyramid of features, where initial shallow layers focus on capturing low-level visual information with high spatial resolution, while deeper layers extract complex, high-dimensional features at a coarser spatial resolution. This approach allows the network to capture the context and relationships between different parts of the image in a better way, which results in improved performance on a broad range of computer vision tasks including image classification, image segmentation [15]. A broad overview of the architecture of MViT is shown in Figure 2C. Since MViTs are relatively new and achieve excellent performance on different vision tasks, we employ these in our study to analyse how well they perform on the task of deepfake detection.
- **DINO** [24] is a self-supervised training method, which is interpreted as self-***DI***stillation with ***NO*** labels. Authors train ViT using DINO and show interesting properties which emerge from the ViT model. Authors make the following observations in their study, i.e., (1) self-supervised ViT features (DINO) incorporate explicit visual information within an image, useful for computer vision tasks such as semantic segmentation, which does not come along as evidently with supervised ViTs and also not with CNNs; (2) self-supervised ViT features are also shown to achieve excellent performance when tested as k-NN classifiers, attaining 78.3% top-1 on ImageNet with a ViT-small architecture. For more details about this strategy, we would like to point readers towards the original paper [24]. The DINO training strategy is shown in Figure 2I. Inspired from these findings, we also employ ViT-Base [13] architecture trained using DINO [24]. In our study, we use the ViT-Base as feature extractor and add a classification head on top. We only train the added classification head on participating deepfake detection datasets, while freezing the weights of the ViT-Base feature extractor.
- **Contrastive Language-Image Pre-Training (CLIP)** [25] is a neural network that has been trained on a diverse set of (image, text) pairs in a self-supervised contrastive manner. It has the ability to infer the most suitable text excerpt for a given image using natural language, without explicit supervision for this task. It exhibits zero-shot capabilities similar to the ones exhibited by GPT-2/GPT-3 [67], [68]. In CLIP's original research paper, authors show that it achieves performance scores equivalent to the original ResNet50 [38] CNN model when evaluated on ImageNet [57] in a "zero-shot" fashion, i.e., even though CLIP does not use any of the 1.28 million labelled examples from the original dataset it achieves comparable performance as a ResNet50 model trained on ImageNet in a supervised manner. CLIP is illustrated in Figure 2H. For more details on CLIP, we refer readers to [25]. We employ a ViT-Base model trained using CLIP as a feature extractor for our study. Similar to DINO, we add a classification head on top of ViT-Base trained using CLIP. For our analysis, we only train the classification head and keep the CLIP ViT-Base features frozen i.e., we do not update its weights during training.

2) Video Models

We examined two distinct video classification models in this paper: (1) ResNet-3D [58], a CNN-based video classifier and (2) TimeSformer [59], a transformer-based video classification model. Our investigation encompasses assessing the performance of both these models in both intra-dataset and inter-dataset contexts across four renowned deepfake detection benchmarks. Our decision to include video-based models alongside image-based detection models stems from our curiosity about the potential impact of temporal information present in videos for the deepfake detection task.

- **ResNet-3D** [58] is based on the same principles as the original ResNet architecture [38], but they are specifically designed to work with 3D data, such as videos and volumetric medical images. These models use 3D convolutions, instead of 2D layers, for feature extraction. In addition to that, ResNet-3D models generally use a large number of layers,

which allows them to learn complex and abstract features in the data. ResNet-3D models have been utilised for a variety of computer vision tasks, including video classification, action recognition and medical image segmentation [58], [69]. For reference, we illustrate both 2D and 3D convolutions in Figure 2F. We choose to employ ResNet-3D model for our study because, (1) it is widely studied in regards of video recognition and (2) pre-trained models are easily available. We chose ResNet-3D model pre-trained on 8 frames per video to experiment in this study.

• **TimeSformer** [59] is a video recognition model based on the transformer architecture. TimeSformer utilises self-attention over space and time, instead of traditional convolutional layers, or the spatial attention as employed by ViT for image recognition. The TimeSformer model modifies the transformer architecture, generally used for image recognition, by directly learning the spatio-temporal features from a sequence of frame-level patches. This is accomplished by extending the self-attention mechanism from the image space to the 3D space-time volume. Similar to the Vision Transformer (ViT) model, the TimeSformer employs linear mapping and positional embeddings to interpret ordering of the resulting sequence of features. In TimeSformer paper [59], authors experimented with different self-attention techniques. Out of different techniques, the "divided attention" technique which calculates temporal and spatial attention separately within each block, was found to perform better than other self-attention calculation techniques and thus we choose to analyse the same architecture in this study. Divided space-time attention is illustrated in Figure 2D. We opt to evaluate TimeSformer on the task of deepfake detection and compare it with convolutional video classification network, ResNet-3D. We also chose 8 frame per video version of the TimeSformer model, same as the ResNet-3D model we described above.

E. EVALUATION METRICS

In order to analyse the performance of our models in a comprehensive way, we employ multiple widely used classification metrics, e.g., (1) LogLoss, (2) AUC and (3) Accuracy. Below we briefly introduce the chosen evaluation metrics.

1) LogLoss

LogLoss, also known as logarithmic loss or cross-entropy loss, is used to measure the classification performance of machine/deep learning models. LogLoss calculates the dissimilarity between the predicted probability score with the true label (0, 1 in case of binary classification). The LogLoss score is computed as the negative logarithm of the likelihood of the true labels given a set of predicted probabilities. The range of the LogLoss function is from 0 to infinity, with 0 representing the ideal outcome and higher values representing worse outcomes.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (1)$$

Where L is the LogLoss, N is the total number of samples in the dataset, y_i is the true label of the *i-th* sample, p_i is the predicted probability for the *i-th* sample.

It is worth noting that Logloss is a widely used evaluation metric in machine learning competitions such as Kaggle competitions, as it gives a general idea of how good the predictions of the model are. We use LogLoss as one of the evaluation metrics in this study as other previously proposed deepfake detection research studies often use it as their evaluation metric and thus would allow us to compare our results with them.

2) Area Under the Curve (AUC)

AUC is another widely known metric used to evaluate classification models. AUC basically refers to calculating the entire two-dimensional area under the Receiver Operating Curve (ROC). AUC gives hints about how well a model has made a certain prediction. Quite understandably, the higher the area falling under the ROC, i.e., AUC, the better the performance of the model at discriminating between "real" and "fake" samples in our case. Most of the recently proposed deepfake detection studies employ AUC as the evaluation metric to study the performance of their models.

Note that the ROC curve is created by varying the threshold used to make predictions from 0 to 1, so the AUC provides a summary of the model's performance across all possible thresholds.

3) Accuracy

Accuracy is another prominent classification metric. Accuracy score is basically the measure of correct predictions made by a model in relation to all the predictions made by the model. Accuracy does not indicate how well a model has made a certain classification, as was the case with LogLoss and AUC. Accuracy score can be obtained by dividing the number of correct predictions by total predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

Where TP is the number of true positives, TN refers to the number of true negatives, FP refers to the number of false positives and FN refers to the number of false negatives.

It is worth noting that accuracy is the proportion of correctly classified samples out of the total number of samples. It is a common evaluation metric used in binary classification tasks, however, it can be misleading in cases where the classes (real, fake) are imbalanced, or if the cost associated with the false positives and false negatives is different. In such cases, other evaluation metrics like F1 score, precision, recall, or AUC may provide a more accurate evaluation of

the classification model's performance. In our study however, since we have balanced number of samples both for **real** and **fake** classes, we can use accuracy as one of the evaluation metric.

TABLE 2. This table presents a detailed account of efficiency metrics of all the participating supervised models, including, the parameter count, inference times both on CPU and GPU and the number of floating point operations per second (FLOPs).

Model Efficiency				
Model	**Parameters**	**CPU**	**GPU**	**FLOPs**
Xception	21 million	49.28ms	7.65ms	4.6G
Res2Net-101	43 million	110.23ms	31.81ms	8.2G
EfficientNet-B7	64 million	148.77ms	37.37ms	5.4G
ViT	86 million	239.18ms	6.18ms	16.9G
Swin-Base	87 million	254.03ms	27.31ms	15.5G
MViT-V2-Base	51 million	238.65ms	43.79ms	10.2G
ResNet-3D	32 million	392.04ms	10.07ms	41.92G
TimeSformer	121 million	2498.54ms	36.56ms	196.1G

4) Efficiency Comparison

To gain a comprehensive understanding of models' performance in deepfake detection, we conduct an in-depth analysis using three distinct classification performance metrics outlined in earlier sections. Additionally, we provide efficiency metrics (see Table 2) for each model to offer insights into the trade-off between a model's effectiveness in detecting deepfakes and its efficiency in real-world deployment. This analysis highlights the financial implications of deploying detection models on cloud services, emphasising the trade-off between efficiency and detection performance. For example, while models like Xception or ViT demonstrate high efficiency (on GPU), the forthcoming sections show that slower, more heavy models often outperform faster, lighter models in deepfake detection. For visual depiction of these efficiency scores, please see Figure 12 and 13 in the Appendix.

We employ fvcore[4] library to compute GFLOPs for our models. While various libraries exist for GFLOPs measurement, it's crucial to acknowledge that results may exhibit slight variations.

To determine CPU and GPU inference times, we execute inference on 300 random images and then calculate the average time spent on each image in milliseconds. For GPU, a warm-up phase precedes inference, involving the processing of 10 images to ensure optimal GPU performance before the actual inference on 300 images commences. Our machine is equipped with an RTX 3080 GPU, Ryzen 5800X CPU and 32GB RAM.

[4]https://github.com/facebookresearch/fvcore

F. IMPLEMENTATION DETAILS

We use *PyTorch*[5] framework to facilitate the training and testing of our models. In our training approach, we employ a batch size of 16 for image models and 4 for video models. The learning rate remains constant at 3×10^{-3} for both image and video models. Our chosen loss function is CrossEntropyLoss and we utilise Stochastic Gradient Descent (SGD) as the optimiser for model training. Our models undergo training for a span of 5 epochs, with final selection of the model having lowest validation loss for subsequent testing and evaluation purposes. For the evaluation stage, we use *Scikit-Learn* library [70]. We use *Scikit-Learn* to calculate and report LogLoss, AUC, Accuracy scores, as well as ROC and DET[6] (Detection Error Tradeoff) curves [70].

To facilitate our model implementations and leverage pre-trained weights, we heavily rely on the *PyTorch Image Models*[7] repository by Ross Wightman. Additionally, we adapt certain code snippets from [24] to train linear classification heads on top of self-supervised feature extractors like DINO and CLIP. We augment images for training using the *imgaug*[8] library.

IV. RESULTS

We conducted extensive experimentation and evaluation on six image recognition models and two video classification models, which we specifically trained for deepfake detection. These evaluations are conducted across four different datasets, as outlined in Section III.The analysis includes evaluating all models under both intra-dataset conditions (trained and evaluated on the same dataset) and inter-dataset conditions (trained on one dataset and evaluated on other datasets, excluding the training dataset). Subsequent sections present the performance outcomes of all participating models within both intra-dataset and inter-dataset contexts.

In addition to the supervised models, our investigation includes two vision transformer (ViT-Base) models that have been pre-trained using the self-supervised techniques DINO [24] and CLIP [25], as previously outlined in Section III. We then compare these two self-supervised models against a supervised Vision Transformer (ViT) [13]. It's important to note that all three models - DINO, CLIP and the supervised ViT - are all ViT-Base models. By training a classification head on top of these three models, our goal is to discern whether self-supervised features offer superior representations in comparison to supervised features.

A. FAKEAVCELEB

FakeAVCeleb [45] is a newly released deepfake detection dataset containing four different categories of videos as given in section III-A earlier. Since we focus only on visual deepfakes in this study, we do not use the audio data (real

[5]https://pytorch.org/
[6]https://scikit-learn.org/stable/auto_examples/model_selection/plot_det.html
[7]https://github.com/huggingface/pytorch-image-models
[8]https://imgaug.readthedocs.io/en/latest/

TABLE 3. Intra-dataset performance comparison of image models. The table below presents scores achieved by image models when trained and evaluated on FakeAVCeleb [45] dataset. Best results are highlighted in yellow.

FakeAVCeleb						
Model	With Augs			No Augs		
	LogLoss	AUC	ACC	LogLoss	AUC	ACC
Xception	0.0047	**100.00%**	99.93%	0.0040	**100.00%**	99.85%
Res2Net-101	**0.0008**	**100.00%**	99.98%	0.0037	**100.00%**	99.93%
EfficientNet-B7	0.0132	**100.00%**	99.63%	0.0047	**100.00%**	99.83%
ViT	0.2073	99.29%	94.60%	0.3768	98.78%	92.43%
Swin-Base	0.0033	**100.00%**	99.88%	0.0058	**100.00%**	99.83%
MViT-V2-Base	**0.0008**	**100.00%**	**100.00%**	**0.0023**	**100.00%**	99.95%
ResNet-3D	0.0041	**100.00%**	**100.00%**	0.0066	**100.00%**	**100.00%**
TimeSformer	0.0796	99.96%	97.50%	0.1238	99.94%	97.00%

and fake) for training and evaluating the models. Thus out of the four subsets of FakeAVCeleb dataset, we only use two for our experiments i.e., (1) FakeVideo/FakeAudio, (2) RealVideo/RealAudio.

We present scores of intra-dataset evaluation in Table 3 showing that all models perform pretty well in distinguishing between fake and real faces. From Table 3, we can see that all of the participating models achieved almost 99% AUC and very low LogLoss score when tested in an intra-dataset configuration. The numbers in 3 suggest that FakeAVCeleb dataset is relatively easy and thus the models can accurately distinguish between real and fake samples.

Table 11 in Appendix reports results achieved by all the models when trained on FakeAVCeleb and evaluated on the remaining three datasets. When we look at the numbers in Table 11, it is apparent that almost all of the models perform poorly on all the other datasets. We can see that in terms of accuracy scores, the models are making random guesses. LogLoss and AUC scores are also not remarkably good in inter-dataset evaluation.

For self-supervised models, the intra-dataset evaluation scores are not as high as those achieved by the supervised models, however, they are still not bad. This is understandable as these models aren't trained in an end-to-end manner, rather only the classification heads are trained on frozen features, as previously mentioned. On this dataset, DINO outperforms the other two models, i.e., CLIP and supervised ViT, with a significant margin as indicated in Table 8.

In an inter-dataset evaluation setting, self-supervised models provide intriguing insights. Notably, DINO, trained on the FakeAVCeleb dataset and evaluated on CelebDF-V2 and FaceForensics++ datasets, demonstrates comparable results to supervised image models. It's worth highlighting that DINO achieves this performance while only training the classification head, in contrast to supervised models that undergo full training. Also, the results suggest that training more complex models on easier datasets do not yield good performance scores when tested on out-of-distribution data (overfitting).

From the results given in Tables 3, 8, 10, 11 and 12 we can infer that FakeAVCeleb dataset is not challenging enough for the models to learn and is fairly easy to distinguish between fake and real samples for both supervised and self-supervised models. In addition to that, this dataset does not enhance the models' ability to learn robust distinguishing features between real and fake faces, or in other words, it lacks at integrating the generalisation capability into the models, as is apparent from Tables 10, 11 and 12 in Appendix.

B. CELEBDF-V2

Table 4 presents the performance of supervised models when trained and evaluated on CelebDF-V2 [40] dataset. Same as it was the case with FakeAVCeleb dataset, almost all of the participating models achieve excellent scores i.e., more than 97% accuracy and more than 99% AUC score, while having a very small LogLoss. We can thus infer that the models quite comfortably learnt to discriminate between real/fake samples of the CelebDF-V2 dataset, similar to FakeAVCeleb dataset.

To gauge the extent to which this dataset aids models in acquiring robust features for enhanced generalisation, we carry out extensive inter-dataset evaluation involving all participating models trained on CelebDF-V2. The outcomes of this evaluation are presented in Table 13 in the Appendix. Surprisingly similar to the observations from models trained on the FakeAVCeleb dataset and assessed on other datasets, the models trained on CelebDF-V2 and subjected to inter-dataset evaluation also display suboptimal performance. This outcome could possibly be attributed to CelebDF-V2 not being particularly challenging for the models to differentiate, as they almost flawlessly categorise every real/fake sample. Nonetheless, this dominance in classification also renders the models less adept at handling unseen data, as evidenced by the performance metrics detailed in Table 13 in the Appendix.

The evidence of CelebDF-V2 being less challenging to learn is further substantiated by the outcomes obtained from the self-supervised models, as illustrated in Table 8. The numbers clearly demonstrate that even when training merely a classification head on feature extractors that remain frozen,

TABLE 4. Intra-dataset comparison of image models. The table below presents scores achieved by image models when trained and evaluated on CelebDF-V2 [40] dataset.

CelebDF						
Model	With Augs			No Augs		
	LogLoss	AUC	ACC	LogLoss	AUC	ACC
Xception	0.0712	99.73%	97.00%	0.0367	99.95%	98.55%
Res2Net-101	**0.0237**	**100.00%**	98.95%	0.0185	99.99%	99.45%
EfficientNet-B7	0.0433	99.95%	98.40%	0.0340	99.98%	98.75%
ViT	0.0336	99.96%	98.60%	0.0350	99.95%	98.60%
Swin-Base	0.0340	99.94%	98.80%	0.0202	99.97%	99.40%
MViT-V2-Base	0.0075	**100.00%**	**99.70%**	**0.0096**	**100.00%**	**99.70%**
ResNet-3D	0.0748	99.68%	97.00%	0.1525	98.68%	95.00%
TimeSformer	0.0309	**100.00%**	98.00%	0.0220	99.96%	99.00%

models still manage to achieve commendable results. For inter-dataset evaluation, self-supervised models trained on CelebDF-V2 and tested on the other datasets yield outcomes akin to those of supervised models, but in some cases, e.g., for DFDC self-supervised models show a considerable performance drop. For additional details, kindly consult Tables 13 and 14 in the Appendix.

C. FACEFORENSICS++

The performance metrics for all supervised models when trained and evaluated on the FaceForensics++ [33] dataset are presented in Table 5. These results are noticeably less favorable compared to those achieved with the previous datasets, FakeAVCeleb and CelebDF-V2. Few models managed to exceed 95% accuracy and LogLoss scores are also less impressive in comparison. The metrics imply that this dataset presents a relatively intricate challenge for the models to differentiate between real and fake samples. The self-supervised models also encounter difficulties in achieving good scores on the FaceForensics++ dataset, as evident from the numbers in Table 8. This reaffirms the notion that accurately distinguishing between fake and real faces in the FaceForensics++ dataset poses a formidable task. This prompts us to question whether a more demanding dataset corresponds to enhanced generalisation capabilities.

Consequently, we move forward with evaluating all supervised models trained on the FaceForensics++ dataset using an inter-dataset evaluation framework. The insights from this evaluation are outlined in Table 15 within the Appendix. The models exhibit satisfactory performance even when confronted with data originating from previously unseen domains. Noteworthy is the improved ability of models trained on the FaceForensics++ dataset and assessed on diverse datasets to generalise effectively. This contrasts with models trained on the FakeAVCeleb and CelebDF-V2 datasets, which tend to exhibit comparatively poor generalisation capabilities. To illustrate, the assessment of MViT trained on FaceForensics++ and evaluated on the FakeAVCeleb dataset yields an accuracy exceeding 80% and an AUC score exceeding 90%. Furthermore, not only on the FakeAVCeleb

TABLE 5. Intra-dataset comparison of image models. The table below presents scores achieved by image models when trained and evaluated on FaceForensics++ [33] dataset.

FaceForensics++						
Model	With Augs			No Augs		
	LogLoss	AUC	ACC	LogLoss	AUC	ACC
Xception	0.2342	96.96%	91.05%	0.2957	95.85%	89.03%
Res2Net-101	0.2165	97.87%	93.48%	0.3213	97.30%	91.85%
EfficientNet-B7	0.3111	96.92%	90.33%	0.3737	94.02%	86.95%
ViT	0.2445	97.27%	92.18%	0.3571	94.04%	85.15%
Swin-Base	**0.1573**	**98.58%**	**94.90%**	0.2191	97.60%	92.18%
MViT-V2-Base	0.1828	98.34%	94.10%	**0.1918**	**97.63%**	**93.00%**
ResNet-3D	0.3224	96.42%	90.36%	0.3085	96.19%	91.07%
TimeSformer	0.2807	97.10%	90.00%	0.2451	96.76%	90.71%

TABLE 6. Intra-dataset comparison of image models. The table below presents scores achieved by image models when trained and evaluated on DFDC [43] dataset.

DFDC						
Model	**With Augs**			**No Augs**		
	LogLoss	**AUC**	**ACC**	**LogLoss**	**AUC**	**ACC**
Xception	0.5613	88.75%	77.63%	0.5120	91.68%	80.65%
Res2Net-101	0.5570	90.64%	79.98%	0.5691	91.78%	83.45%
EfficientNet-B7	0.5542	89.97%	79.30%	**0.4263**	**93.30%**	**84.15%**
ViT	**0.4696**	**91.89%**	81.08%	0.5709	89.44%	78.35%
Swin-Base	0.5602	90.89%	82.60%	0.6650	87.77%	79.05%
MViT-V2-Base	0.6079	88.41%	78.90%	0.5491	90.65%	82.40%
ResNet-3D	0.5865	85.64%	75.75%	0.6739	84.69%	73.50%
TimeSformer	0.4870	91.18%	**83.25%**	0.6176	92.30%	81.75%

dataset, we can also see encouraging performance from all models trained on this dataset and evaluated on others. The results in Tables 15 and 16 in Appendix support the statement that more challenging datasets mean better generalisation capability. But we have to further re-enforce this statement after evaluating the models trained using DFDC [43] dataset in the upcoming section.

D. DFDC

DFDC is one of the biggest and widely adopted deepfake detection benchmarks. We present intra-dataset evaluation scores of our models trained and evaluated on DFDC in Table 6. Res2Net-101 turned out to be the best model in this evaluation, managing to achieve more than 84% accuracy score, 93% AUC score on the DFDC dataset. Self-supervised models also achieve relatively low scores when trained and evaluated on DFDC, as apparent from Table 8. This establishes that DFDC is comparably more challenging dataset out of all the four datasets in this study.

In Table 17 inside the Appendix section we present inter-dataset evaluation scores achieved by the supervised models trained on DFDC dataset. It is evident from the numbers that the models trained using DFDC dataset still achieve acceptable performance on unseen data, as compared to the scores achieved by the models which were trained on FakeAVCeleb and CelebDF-V2. Also, by looking at the results now, we can affirm the statement that models trained using more challenging datasets seem to achieve better results. This finding is evident from Tables 10, 15, 16, 17 and 18.

E. DISCUSSION

1) Supervised Models

In Figure 3, we illustrate a comparison of all participating supervised models based on their attained accuracy scores in an intra-dataset evaluation context. The visualisation clearly indicates that there exists minimal performance difference among the models. Across the majority of cases, the models achieve accuracy levels ranging from approximately 92% to 94%.

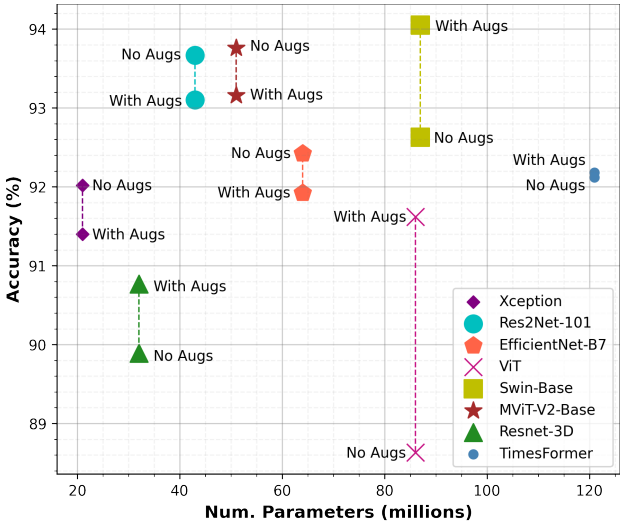

FIGURE 3. Performance (accuracy) comparison of participating models on all datasets. The reported scores result in an intra-dataset evaluation. Results in this figure are obtained by evaluating each model separately on each dataset and averaging the resulting scores. In addition to this, the figure presents the performance of each model trained with and without the augmentations, along with their parameter count.

Notably, the figure underscores that image augmentations do not always yield significant performance gains. For instance, XceptionNet, Res2Net-101, MViT-V2-Base and EfficientNet-B7 display superior performance when trained without image augmentations, as compared to their counterparts trained with augmentations. Nonetheless, the divergence in accuracy scores between models trained with and without image augmentations is generally modest, except in the case of ViT. Specifically, the ViT trained with image augmentations achieves an accuracy of 91.62%, whereas the ViT trained without augmentations records an accuracy of 88.63%. In addition to this, Figure 3 highlights that transformer models consistently perform better when trained

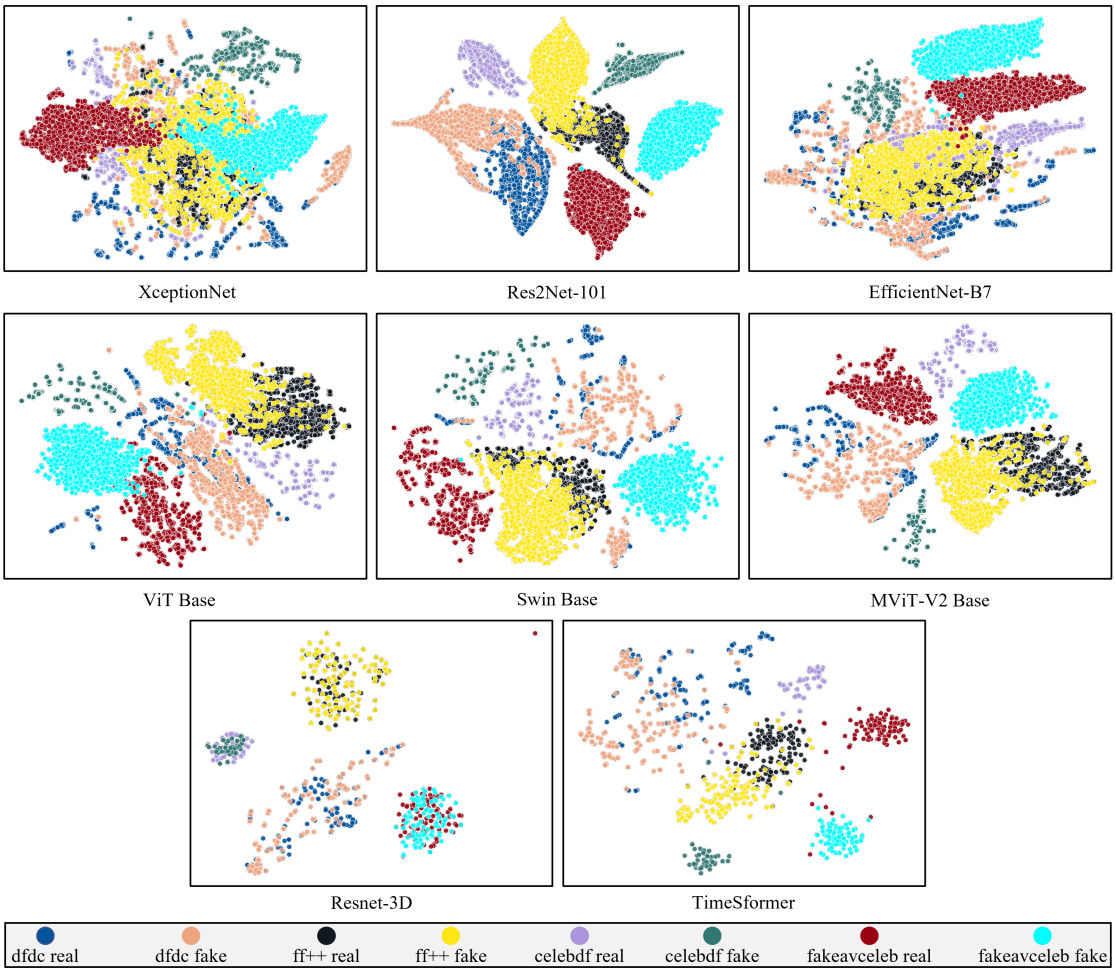

FIGURE 4. TSNE visualisations of the participating detection models. We chose the best performing models on all datasets (with/without image augmentations).

using augmentations. Additionally, video models also exhibit better performance when trained using image augmentations. An important insight is that the best-performing model, Swin-Base, attains its peak accuracy when trained with image augmentations, further advocating for the incorporation of augmentations in training protocols.

Furthermore, it's worth noting that the transformer models (Swin-Base and MViT-V2-Base, TimeSformer) demonstrate superior performance compared to their CNN counterparts. Interestingly, the Res2Net-101 model also achieves remarkable numbers in the intra-dataset evaluation context, despite having roughly half the number of parameters (43 million parameters) compared to the top-performing Swin-Base model (87 million parameters). Figure 3 and Table 7 collectively indicate a valuable observation: models equipped with multi-scale feature processing capabilities, such as Res2Net, MViT-V2 and Swin Transformer, exhibit the best performance among all the models.

Moving towards inter-dataset analysis, we present the outcomes attained by the supervised models when assessed in an inter-dataset context through Figure 7 in Appendix section. The figure showcases that the models exhibit noticeably reduced performance levels in inter-dataset evaluation compared to intra-dataset evaluation. This discrepancy is reasonable since detection models tend to experience performance degradation when confronted with data originating from unseen distributions. However, the Figure 7 in Appendix

TABLE 7. This table compares the performance of all the participating (supervised) models. We present scores after averaging the scores (LogLoss, AUC, Accuracy) achieved by each model when evaluated in an intra-dataset setting.

Performance Comparison of Supervised Models on All Datasets						
Model	With Augs			No Augs		
	LogLoss	AUC	ACC	LogLoss	AUC	ACC
Xception	0.2179	96.36%	91.40%	0.2121	96.87%	92.02%
Res2Net-101	**0.1395**	97.12%	93.10%	0.2282	**97.27%**	93.67%
EfficientNet-B7	0.2305	96.71%	91.92%	0.2097	96.83%	92.42%
ViT	0.2388	97.10%	91.62%	0.3350	95.55%	88.63%
Swin-Base	0.1494	**97.35%**	**94.05%**	0.2275	96.34%	92.63%
MViT-V2-Base	0.1998	96.68%	93.16%	0.1882	97.07%	**93.76%**
Resnet-3D	0.2470	95.44%	90.77%	**0.1620**	94.89%	89.89%
TimeSformer	0.2196	97.06%	92.18%	0.2521	97.24%	92.12%

TABLE 8. This table compares the performance of all the participating (self-supervised) models when evaluated in an intra-dataset setting. The statistics of this table are illustrated in Figure 10 in Appendix.

Performance Comparison of Self-Supervised Models on Individual Datasets							
Model	With Augs			No Augs			Dataset
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Supervised	0.4105	90.19%	82.50%	0.3727	91.77%	85.50%	FakeAVCeleb
Dino	**0.1444**	**99.00%**	**95.33%**	**0.0801**	**99.64%**	**97.25%**	
CLIP	0.4369	89.88%	81.20%	0.3715	93.37%	84.55%	
Supervised	**0.2941**	95.52%	88.05%	**0.2237**	**97.18%**	**91.80%**	CelebDF-V2
Dino	0.3655	**97.31%**	**90.90%**	0.3930	97.10%	88.90%	
CLIP	0.3750	91.43%	82.80%	0.3399	94.73%	85.40%	
Supervised	0.5182	83.11%	74.95%	**0.4971**	85.47%	77.43%	FaceForensics++
Dino	1.1758	**88.67%**	**80.60%**	1.1186	**89.48%**	**81.85%**	
CLIP	**0.5019**	82.75%	74.15%	0.5093	85.16%	75.80%	
Supervised	0.5836	79.19%	68.65%	0.5829	80.93%	72.63%	DFDC
Dino	2.2839	**80.72%**	**72.38%**	1.5812	83.03%	74.15%	
CLIP	**0.5601**	79.08%	71.75%	**0.5196**	**83.12%**	**75.00%**	

reports a useful finding: across all datasets, as compared to CNN models the transformers consistently emerge as the top-performing models. We refer readers to Table 10 in the Appendix section to examine the inter-dataset scores achieved by models on each of the dataset.

We also present the TSNE [71] plots of all the supervised models in Figure 4, to visually represent how the models separate real faces from the fake ones. Also, it gives us an idea about how the models group together faces coming from same datasets near to each other as compared to the faces coming from a different dataset. The TSNE plots also help us visualise which datasets are more challenging than the others. For example, if we look at the TSNE plots in Figure 4, we can see that the models tend to separate the easier datasets (FakeAVCeleb and CelebDF-V2) in a better way, as compared to how they separate the more challenging datasets (FaceForensics++ and DFDC).

Another notable observation is that image models tend to perform the separation task more effectively compared to video models. This is expected, considering our earlier mention that video models typically require larger amounts of training data (we trained both image and video models on the same dataset in this study). As part of our future research, we aim to explore video models on larger datasets to further validate this hypothesis. Despite this, the t-SNE visualizations reveal an interesting insight: while the video model ResNet-3D may struggle to distinguish between real and fake faces within the same dataset, it excels at effectively separating data from different datasets.

In addition to that, for a better diagnosis of the models we also visualise the predictions using Gradient-weighted Class Activation Mapping (Grad-CAM)[9] [72]. Figure 6 in appendix section presents Grad-CAMs of the supervised image models on all datasets. It is interesting to observe that all models, to varying degrees, concentrate on different facial regions when making predictions.

Furthermore, we provide the ROC, DET curves for the participating models assessed in an intra-dataset context, as illustrated in Figures 8 and 9 in the Appendix respectively. The corresponding AUC scores reinforce the notion that FakeAVCeleb and CelebDF-V2 datasets present less complexity to the models in comparison to FaceForensics++ and DFDC datasets. This underscores the idea that training the models on more challenging datasets, rather than easier ones, enhances their generalisation capabilities for deepfake detection.

The scores (LogLoss, AUC, ACC) reported in Tables 7 and 9 for each model are calculated by averaging the individual scores achieved by that specific model on each dataset. For example, s1, s2, s3, s4 are scores that a model achieved on datasets d1, d2, d3 and d4.

2) Self-Supervised Models

In Figure 5 we show a similar comparison involving self-supervised models. It is clear that DINO outperforms the other two models. A careful examination of the outcomes in Tables 8 and 9 enables us to deduce that self-supervised features, particularly DINO, yield superior feature representations in comparison to CLIP and supervised. To strengthen this finding further, we illustrate the ROC and DET curves in Figures 10 and 11 in the Appendix respectively.

[9]https://github.com/jacobgil/pytorch-grad-cam

TABLE 9. This table compares the performance of the self-supervised models. We present scores after averaging the scores (LogLoss, AUC, Accuracy) achieved by each model on the four datasets, when evaluated in an intra-dataset setting. In this table, ***Supervised*** refer to ViT-Base model pre-trained using supervised training scheme. ***DINO*** refers to ViT-Base model pre-trained using self-supervised scheme proposed in [24] and ***CLIP*** refers to ViT-Base pre-trained using self-supervised scheme prposed in [25]. All of these ViT-Base models are used as feature extractors, where we only train a classification head on top of each of the feature extractor and freeze the weights of feature extractors.

Performance Comparison of Self-Supervised Models on All Datasets						
Model	With Augs			No Augs		
	LogLoss	AUC	ACC	LogLoss	AUC	ACC
Supervised	**0.4516**	87.00%	78.54%	**0.4191**	88.84%	81.59%
Dino	0.9924	**91.43%**	**84.80%**	0.7932	**92.31%**	**85.54%**
CLIP	0.4685	85.78%	77.48%	0.4351	89.09%	80.19%

3) The Outcome

Answering the six questions that we posed at the beginning of this study in Section I:

- identifying the most effective models for detecting deepfakes among those being tested - **Ans: Models equipped with multi-scale feature representation capabilities, such as MViT-V2, Res2Net-101 and Swin Transformer (hierarchical representations)**.
- pinpointing the model with the highest ability to adapt to new and unseen data - **Ans: Upon examining the tables in the Appendix section, it becomes evident that MViT-V2 consistently achieves superior performance scores compared to other models in the majority of cases. Furthermore, these tables also highlight that Transformer models generally outperform CNN models in most scenarios**.
- assessing the difficulty of different datasets for model training - **Ans: DFDC and FaceForensics++ datasets pose greater challenges for the models to learn in comparison to CelebDF-V2 and FakeAVCeleb datasets**.
- determining the dataset that best facilitates generalisation to unseen data - **Ans: Table 10 in the Appendix confirms that the FaceForensics++ dataset promotes strong generalisation of models to unseen data, with the DFDC dataset ranking second in this regard**.
- evaluating the performance of self-supervised training strategies - **Ans: From Tables 8 and 9, it is evident that DINO [24] outperforms the other two competing strategies in intra-dataset evaluation across all datasets**.
- examining the impact of augmentations on enhancing model performance - **Ans: Within the scope of this study, the augmentations that we have employed have a minimal effect on models' performance i.e., in some cases, augmentations help models achieve better performance, while in other cases, they don't**.

V. CONCLUSIONS

We conducted a comprehensive study to assess the effectiveness of various image and video classification architectures for deepfake detection. Models were initially pre-trained using both supervised and self-supervised approaches and then evaluated on four prominent deepfake detection datasets. Our extensive experiments revealed that models adept at processing multi-scale features, such as Res2Net-101, MViT-V2 and Swin Transformer, consistently outperformed others in intra-dataset comparisons. Notably, MViT-V2-Base and Res2Net-101 achieved superior performance with approximately half the parameters of the Swin-Base transformer model. Regarding generalisation across datasets, transformer models consistently outperformed CNN models, with FaceForensics++ [33] and DFDC [43] enhancing generalisation capabilities.

Our investigation into models pre-trained using self-supervised strategies showed that the ViT-Base model, pre-trained using DINO [24], outperformed both supervised ViT-Base and self-supervised CLIP [25] ViT-Base models. Additionally, our findings indicate that the selected image augmentations lead to improved performance for Transformer models, while offering comparably less notable benefits for CNN models.

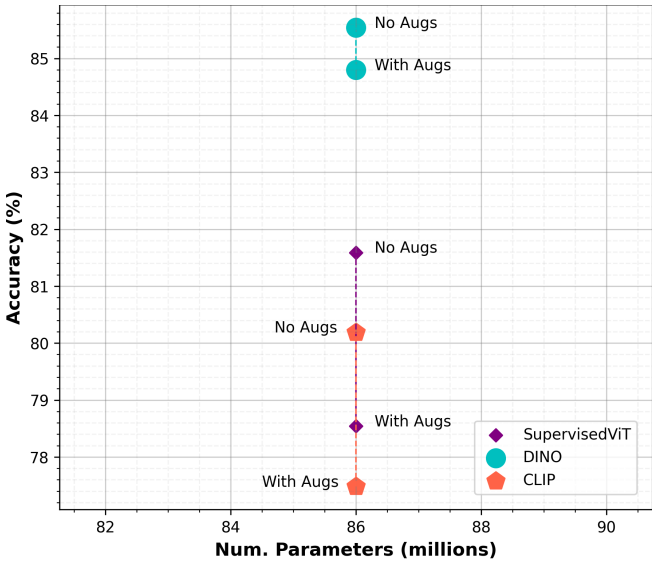

FIGURE 5. Performance (accuracy) comparison of two self-supervised ViT models and one supervised ViT. The reported scores result in an intra-dataset evaluation. Results in this figure are obtained by evaluating each model separately on each dataset and averaging the resulting scores. In addition to this, the figure presents the performance of each model trained with and without the augmentations. All of of the three models have the same amount of trainable parameters since they all are ViT-Base models and the only difference is the pre-training schemes used to train the models.

VI. ACKNOWLEDGMENTS

This research was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centres for Research-based Innovation scheme, project number 309339.

We acknowledge the use of ChatGPT [73] for checking and correcting the grammar of this paper. It's important to note that we did not use ChatGPT to generate totally new

text, rather we use it to check/correct our provided text.

REFERENCES

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[2] N. A. Diakopoulos and D. G. Johnson, "Anticipating and addressing the ethical implications of deepfakes in the context of elections," *New Media & Society*, vol. 23, pp. 2072 – 2098, 2020.

[3] A. De Ruiter, "The distinct wrong of deepfakes," *Philosophy & Technology*, vol. 34, no. 4, pp. 1311–1332, 2021.

[4] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[5] S. Thompson, "Making deepfakes gets cheaper and easier thanks to A.I." https://www.nytimes.com/2023/03/12/technology/deepfakes-cheapfakes-videos-ai.html, 2023, [Online; accessed 27-November-2023].

[6] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face forgery detection by 3D decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2929–2939.

[7] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 710–18 719.

[8] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2TR: Multi-modal multi-scale transformers for deepfake detection," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 615–623.

[9] S. A. Khan and H. Dai, "Video transformer for deepfake detection with incremental learning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1821–1828.

[10] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "Istvt: Interpretable spatial-temporal video transformer for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1335–1348, 2023.

[11] A. Khormali and J.-S. Yuan, "DFDT: an end-to-end deepfake detection framework using vision transformer," *Applied Sciences*, vol. 12, no. 6, p. 2953, 2022.

[12] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards generalizable deepfake detection with locality-aware autoencoder," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 325–334.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[15] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6824–6835.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[17] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 280–296.

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[19] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[21] N. Engel, V. Belagiannis, and K. Dietmayer, "Point transformer," *IEEE Access*, vol. 9, pp. 134 826–134 840, 2021.

[22] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1954–1963.

[23] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[24] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[26] M. Oquab, T. Darcet, T. Moutakanni, H. Q. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. B. Huang, S.-W. Li, I. Misra, M. G. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *ArXiv*, vol. abs/2304.07193, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258170077

[27] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[28] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[29] S. Das, S. Seferbekov, A. Datta, M. S. Islam, and M. R. Amin, "Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021, pp. 3776–3785.

[30] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *International Conference on Image Analysis and Processing*. Springer, 2022, pp. 219–229.

[31] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[32] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 80–87.

[33] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1–11.

[34] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *Proceedings of the IEEE international workshop on information forensics and security*. IEEE, 2018, pp. 1–7.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[36] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[37] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[38] 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*, 2016, pp. 770–778.

[39] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2261–2269.

[40] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3204–3213.

[41] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.

[42] N. Dufour and A. Gully, "Contributing data to deepfake detection research," https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html, 2019, [Online; accessed 08-March-2023].

[43] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[44] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *European Conference on Computer Vision*. Springer, 2020, pp. 152–168.

[45] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.

[46] "Faceswap github," https://github.com/MarekKowalski/FaceSwap/, [Online; accessed 02-December-2022].

[47] "Deepfakes github," https://github.com/deepfakes/faceswap, [Online; accessed 02-December-2022].

[48] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[49] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[50] D. Huang and F. D. la Torre, "Facial action transfer with personalized bilinear regression," in *European conference on computer vision*. Springer, 2012.

[51] E. Zakharov, A. Shysheya, E. Burkov, and V. S. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9458–9467.

[52] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7183–7192.

[53] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4396–4405.

[54] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, pp. 3697–3705.

[55] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, vol. 31, 2018.

[56] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492.

[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[58] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3154–3160.

[59] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the International Conference on Machine Learning*, vol. 2, no. 3, 2021, p. 4.

[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1–9.

[61] S. Gao, M.-M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 652–662, 2021.

[62] F. Yu, D. Wang, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2403–2412.

[63] C.-F. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. Feris, "Big-little net: An efficient multi-scale feature representation for visual and speech recognition," *arXiv preprint arXiv:1807.03848*, 2018.

[64] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5987–5995.

[65] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.

[67] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[68] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[69] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[71] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[72] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, 2016.

[73] OpenAI, "Introducing ChatGPT," https://openai.com/blog/chatgpt, 2021, [Online; accessed 08-August-2023].

SOHAIL AHMED KHAN is a PhD candidate at MediaFutures and University of Bergen. He holds a MSc in Cybersecurity and Artificial Intelligence from the University of Sheffield, UK. Prior to joining MediaFutures, Sohail worked as a research assistant at Mohamed bin Zayed University of AI, Abu Dhabi, UAE. Before that, he worked as a remote research assistant at CYENS Centre of Excellence, Nicosia, Cyprus. His research interests intersect deep learning, computer vision and multimedia forensics. Sohail is currently associated with the MediaFutures' Work Package 3, Media Content Analysis and Production.

DUC TIEN DANG NGUYEN is an associate professor of computer science at the Department of Information Science and Media Studies, University of Bergen. His main area of expertise is on multimedia forensics, lifelogging, multimedia retrieval and computer vision. He is the author and co-author of more than a hundred peer-reviewed and widely cited research papers. He is a PC member in a number of conferences in the fields of lifelogging, multimedia forensics and pattern recognition. He is co-organiser of over 50 special sessions, workshops and research challenges from ACM MM, ACM ICMR, NTCIR, ImageClef and MediaEval during the last 10 years. He is also the General Chair of MMM 2023, TPC Chair of MMM 2022 and ACM ICMR 2024.

APPENDIX

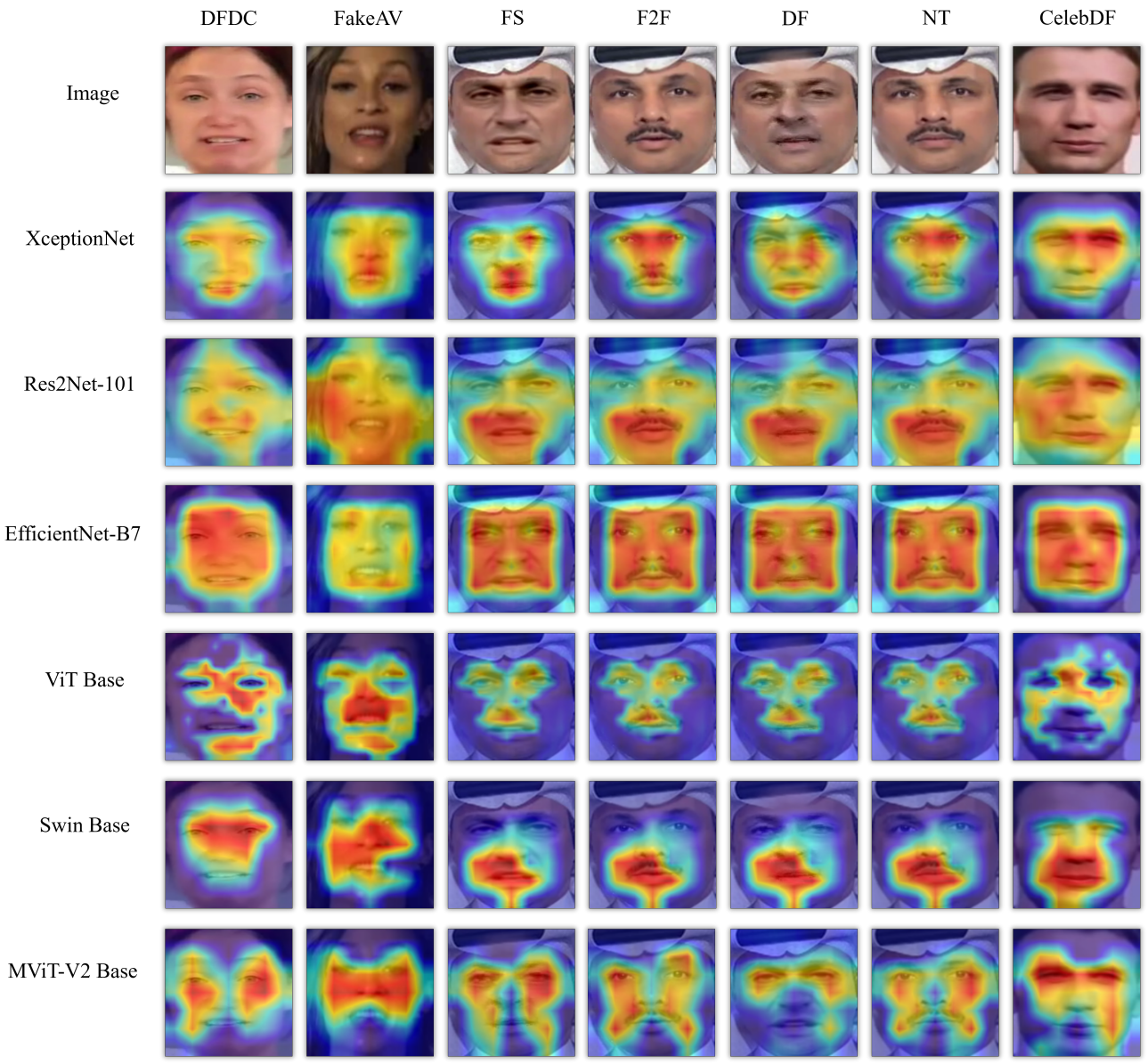

FIGURE 6. CBAM visualisations of the supervised image models.

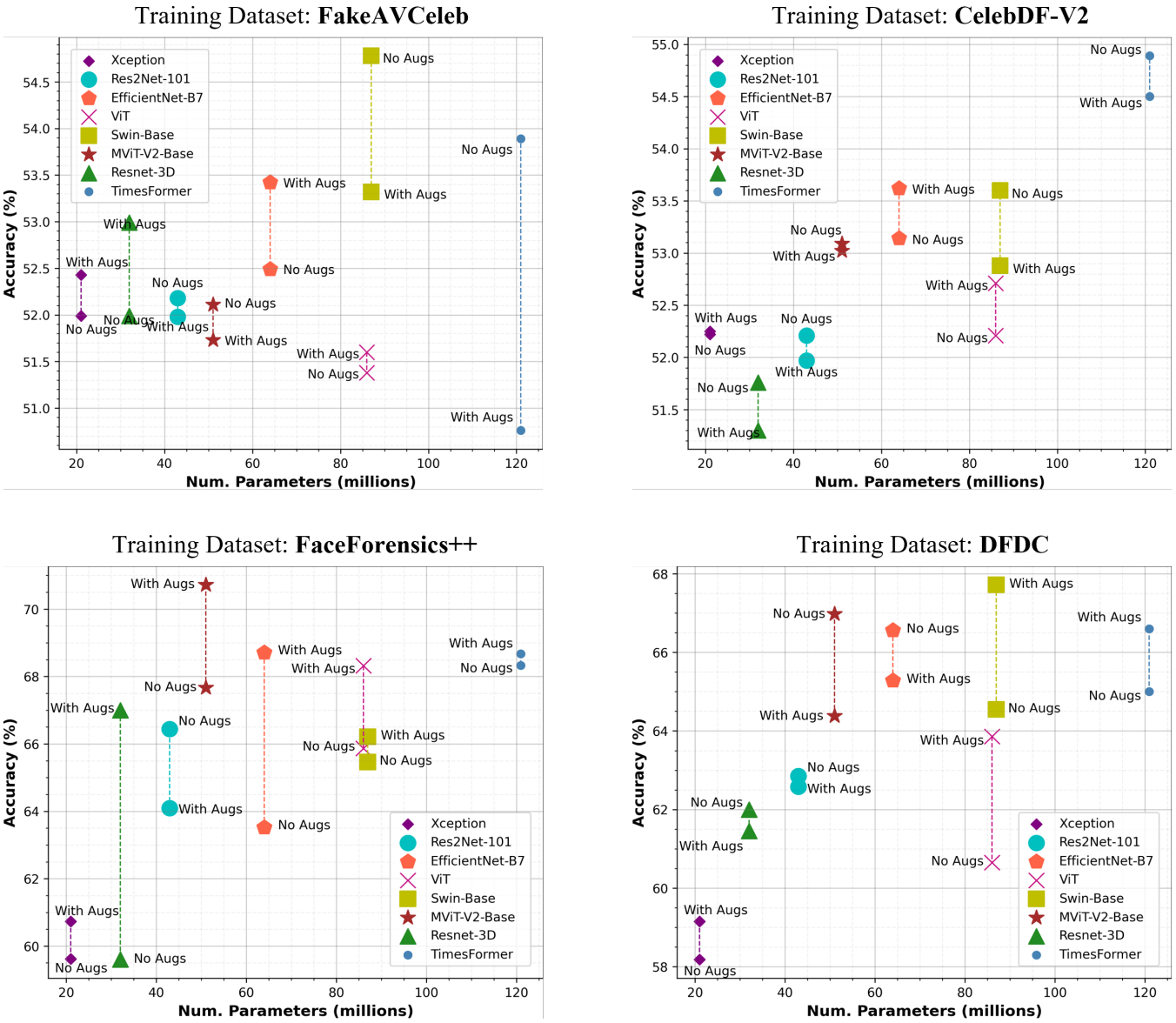

FIGURE 7. Performance (accuracy) comparison of participating models evaluated using inter-dataset scheme. Results in this figure are obtained by, (1) evaluating each model trained on one dataset on each of the remaining datasets and (2) averaging the achieved scores, i.e., add the 3 accuracy scores and divide by 3.

TABLE 10. This table compares the performance of all the participating (supervised) models evaluated in an inter-dataset setting. Results in this table are obtained by, (1) evaluating each model trained on one dataset on each of the remaining datasets and (2) averaging the achieved scores, i.e., add the 3 accuracy scores and divide by 3. Figure 7 illustrate the statistics of this table.

Inter-Dataset Evaluation							
Model	With Augs			No Augs			Training Dataset
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Xception	7.4484	57.69%	52.43%	6.8728	55.41%	51.99%	FakeAVCeleb
Res2Net-101	8.5574	58.77%	51.98%	7.0666	60.64%	52.18%	
EfficientNet-B7	8.5664	62.32%	**53.42%**	10.7718	60.05%	52.49%	
ViT	6.7348	61.01%	51.60%	9.1672	58.45%	51.38%	
Swin-Base	5.1077	**62.54%**	53.32%	4.3274	**64.88%**	**54.78%**	
MViT-V2-Base	4.7564	58.78%	51.73%	4.2891	59.38%	52.11%	
ResNet-3D	**4.4308**	57.61%	52.99%	**3.8206**	60.09%	51.99%	
TimeSformer	4.7334	61.55%	50.76%	4.7759	63.95%	53.89%	
Xception	**3.9439**	65.06%	52.25%	4.8776	66.40%	52.22%	CelebDF-V2
Res2Net-101	5.4266	65.90%	51.97%	5.6891	66.21%	52.21%	
EfficientNet-B7	5.9514	66.99%	53.62%	8.9668	67.13%	53.14%	
ViT	5.4921	68.52%	52.71%	8.9981	66.36%	52.21%	
Swin-Base	5.6007	70.06%	52.88%	4.8405	**70.56%**	53.60%	
MViT-V2-Base	4.8723	**70.71%**	53.02%	**4.6419**	67.20%	53.09%	
ResNet-3D	6.8365	61.57%	51.30%	5.0504	64.52%	51.76%	
TimeSformer	4.5629	69.04%	**54.50%**	4.9391	69.43%	**54.89%**	
Xception	1.0701	69.92%	60.73%	1.1262	67.78%	59.62%	FaceForensics++
Res2Net-101	1.0165	73.46%	64.09%	1.2360	73.61%	66.44%	
EfficientNet-B7	0.8792	79.51%	68.71%	1.0068	69.80%	63.52%	
ViT	**0.7899**	78.45%	68.32%	0.8301	73.24%	65.87%	
Swin-Base	0.8517	77.94%	66.21%	0.8482	**78.03%**	65.46%	
MViT-V2-Base	0.8407	**79.75%**	**70.72%**	**0.7292**	75.85%	67.67%	
ResNet-3D	1.0639	74.47%	67.00%	1.3331	66.61%	59.50%	
TimeSformer	1.0665	75.59%	68.67%	0.8492	77.03%	**68.33%**	
Xception	1.2959	63.62%	59.15%	1.6780	64.18%	58.18%	DFDC
Res2Net-101	2.0224	67.80%	62.58%	1.7396	69.50%	62.85%	
EfficientNet-B7	**1.0388**	71.32%	65.28%	1.2764	72.28%	66.56%	
ViT	1.2198	70.45%	63.86%	1.2498	64.71%	60.65%	
Swin-Base	1.2423	**73.49%**	**67.72%**	1.3802	69.49%	64.55%	
MViT-V2-Base	1.2329	72.37%	64.38%	1.2254	**72.68%**	**66.98%**	
ResNet-3D	1.1354	66.69%	61.45%	**1.1354**	66.27%	62.00%	
TimeSformer	1.1421	70.66%	66.60%	1.6584	71.77%	65.00%	

TABLE 11. Inter-dataset evaluation scores of models trained on FakeAVCeleb [45] dataset and evaluated on the remaining three datasets.

Training Dataset: FakeAVCeleb							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Xception	3.1366	50.52%	56.10%	3.9659	44.63%	54.25%	CelebDF-V2
Res2Net-101	3.9007	57.61%	54.60%	3.2127	56.73%	54.50%	
EfficientNet-B7	5.7925	59.31%	54.50%	12.3786	54.30%	51.65%	
ViT	4.7035	56.51%	51.60%	7.0900	52.18%	46.65%	
Swin-Base	3.5148	61.13%	**57.70%**	2.8360	**62.54%**	**61.45%**	
MViT-V2-Base	4.5526	**65.37%**	54.00%	3.6492	58.18%	55.05%	
ResNet-3D	**2.4752**	53.88%	51.00%	**1.7475**	57.36%	49.00%	
TimeSformer	3.9086	51.60%	48.00%	3.2374	58.52%	55.00%	
Xception	10.3539	62.90%	50.38%	8.9711	63.06%	50.33%	FaceForensics++
Res2Net-101	11.6456	59.23%	50.18%	9.7934	58.54%	50.53%	
EfficientNet-B7	10.5412	**63.80%**	**52.45%**	10.4825	63.13%	**52.05%**	
ViT	9.3036	61.70%	51.10%	12.3768	58.44%	50.93%	
Swin-Base	6.1675	62.97%	50.70%	5.5166	**64.87%**	51.23%	
MViT-V2-Base	**4.8833**	56.51%	50.65%	**4.7116**	63.40%	50.85%	
ResNet-3D	6.7343	51.30%	50.71%	5.9796	53.99%	50.71%	
TimeSformer	6.0010	62.61%	51.79%	6.1889	62.60%	51.43%	
Xception	8.8546	59.65%	50.80%	7.6813	58.53%	51.40%	DFDC
Res2Net-101	10.1260	59.48%	51.15%	8.1937	66.66%	51.50%	
EfficientNet-B7	9.3656	63.86%	53.30%	9.4543	62.71%	53.78%	
ViT	6.1972	64.81%	52.10%	8.0348	64.71%	56.58%	
Swin-Base	5.6410	63.51%	51.55%	4.6297	67.25%	51.65%	
MViT-V2-Base	4.8333	54.46%	50.55%	4.5065	56.55%	50.43%	
ResNet-3D	**4.0828**	67.65%	**57.25%**	**3.7347**	68.91%	**56.25%**	
TimeSformer	4.2907	**70.43%**	52.50%	4.9015	**70.74%**	55.25%	

TABLE 12. Inter-dataset evaluation scores of **self-supervised** models fine-tuned on FakeAVCeleb [45] dataset and evaluated on the remaining three datasets.

Training Dataset: FakeAVCeleb							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Supervised	**1.1051**	60.65%	58.50%	**1.2554**	58.82%	55.75%	CelebDF-V2
Dino	2.7798	**64.07%**	**60.65%**	2.7178	**63.39%**	**60.50%**	
CLIP	1.8561	50.22%	50.50%	3.2978	52.29%	50.15%	
Supervised	2.2969	**62.99%**	53.48%	2.2189	64.15%	55.08%	FaceForensics++
Dino	7.5589	61.59%	**57.73%**	8.4257	62.74%	52.05%	
CLIP	**1.1427**	58.52%	55.30%	**1.5146**	60.47%	**56.83%**	
Supervised	1.8896	**65.60%**	56.05%	**1.8893**	**68.37%**	**57.70%**	DFDC
Dino	8.2808	59.25%	54.10%	9.0520	62.46%	52.40%	
CLIP	**1.1260**	64.49%	**59.45%**	2.0150	65.06%	56.98%	

TABLE 13. Inter-dataset evaluation scores of models trained on CelebDF-V2 [40] dataset and evaluated on the remaining three datasets.

Training Dataset: CelebDF-V2							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Xception	**4.7313**	65.82%	51.68%	5.1136	67.77%	51.78%	**FakeAVCeleb**
Res2Net-101	5.7429	69.08%	52.30%	**4.3332**	**71.01%**	52.83%	
EfficientNet-B7	7.4940	63.86%	52.05%	9.6846	65.93%	51.45%	
ViT	5.0347	69.00%	**53.08%**	9.6735	61.89%	52.18%	
Swin-Base	6.0084	67.63%	52.20%	4.8922	68.28%	52.70%	
MViT-V2-Base	5.1980	**72.43%**	52.75%	5.3953	61.24%	51.55%	
ResNet-3D	6.0756	62.79%	50.50%	4.9703	61.58%	50.50%	
TimeSformer	4.8465	69.73%	53.00%	5.8829	68.77%	**54.00%**	
Xception	**4.2473**	63.26%	53.53%	5.6357	63.68%	53.58%	**FaceForensics++**
Res2Net-101	6.3947	64.79%	53.33%	6.9000	63.59%	52.90%	
EfficientNet-B7	6.3164	65.07%	54.80%	8.9065	66.31%	53.98%	
ViT	6.1010	68.14%	53.53%	9.9676	65.50%	53.50%	
Swin-Base	6.0278	**70.13%**	54.23%	5.5408	**68.45%**	54.30%	
MViT-V2-Base	5.2175	70.01%	53.15%	**4.6160**	67.88%	54.08%	
ResNet-3D	7.1877	60.00%	52.14%	5.6544	66.01%	54.29%	
TimeSformer	4.8228	68.84%	**57.50%**	5.0219	67.55%	**56.43%**	
Xception	**2.8532**	66.11%	51.55%	**3.8835**	67.74%	51.30%	**DFDC**
Res2Net-101	4.1424	63.83%	50.28%	5.8342	64.01%	50.90%	
EfficientNet-B7	4.0438	72.05%	**54.00%**	8.3092	69.17%	54.00%	
ViT	5.3405	68.41%	51.53%	7.3534	71.69%	50.95%	
Swin-Base	4.7659	**72.42%**	52.20%	4.0886	**74.95%**	53.80%	
MViT-V2-Base	4.2014	69.69%	53.15%	3.9144	72.48%	53.65%	
ResNet-3D	7.2461	61.91%	51.25%	4.5265	65.97%	50.50%	
TimeSformer	4.0195	68.56%	53.00%	3.9124	71.98%	**54.25%**	

TABLE 14. Inter-dataset evaluation scores of **self-supervised** models fine-tuned on CelebDF-V2 [40] dataset and evaluated on the remaining three datasets.

Training Dataset: CelebDF-V2							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Supervised	**1.8680**	65.64%	53.70%	**1.7462**	66.18%	**56.58%**	**FakeAVCeleb**
Dino	8.5606	**68.16%**	**57.35%**	10.6620	**66.48%**	53.10%	
CLIP	2.0280	61.23%	52.60%	2.2300	60.48%	53.28%	
Supervised	1.9176	63.57%	54.63%	2.0464	64.77%	55.63%	**FaceForensics++**
Dino	9.3992	**64.71%**	**56.60%**	9.2731	**66.18%**	56.25%	
CLIP	**1.5848**	60.23%	53.58%	**1.5040**	66.12%	**58.88%**	
Supervised	3.0170	53.42%	49.65%	3.3809	51.49%	50.20%	**DFDC**
Dino	13.8247	52.52%	50.60%	10.2818	54.76%	**52.60%**	
CLIP	**2.2078**	**59.35%**	**50.80%**	**2.6224**	**58.78%**	51.02%	

TABLE 15. Inter-dataset evaluation scores of models trained on FaceForensics++ [33] dataset and evaluated on the remaining three datasets.

Training Dataset: FaceForensics++							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Xception	0.8691	79.62%	65.88%	0.7795	76.14%	66.93%	FakeAVCeleb
Res2Net-101	0.7693	83.01%	71.28%	0.6527	85.48%	76.83%	
EfficientNet-B7	0.5782	89.59%	77.05%	0.7375	77.88%	70.08%	
ViT	0.6648	83.05%	70.65%	0.7419	76.40%	69.23%	
Swin-Base	0.5880	87.72%	72.95%	0.6373	89.10%	71.15%	
MViT-V2-Base	**0.3654**	**92.96%**	**84.65%**	**0.4047**	**90.25%**	**81.90%**	
ResNet-3D	0.7903	83.55%	68.00%	1.1338	73.34%	62.50%	
TimeSformer	0.9135	79.33%	75.00%	0.7900	76.65%	70.50%	
Xception	1.0426	65.92%	61.60%	1.2566	62.39%	58.65%	CelebDF-V2
Res2Net-101	1.0751	67.85%	62.40%	1.4218	65.46%	59.80%	
EfficientNet-B7	0.7759	78.46%	69.95%	1.0103	67.24%	61.25%	
ViT	**0.5915**	**82.44%**	**74.10%**	0.8504	75.11%	65.40%	
Swin-Base	0.7136	74.58%	67.05%	0.7879	70.94%	63.75%	
MViT-V2-Base	0.9791	76.66%	65.35%	0.7912	68.69%	62.70%	
ResNet-3D	1.1992	66.12%	65.00%	1.5866	59.44%	55.00%	
TimeSformer	1.1745	73.68%	63.00%	**0.7446**	**80.40%**	**71.00%**	
Xception	1.2988	64.22%	54.70%	1.3424	64.81%	53.28%	DFDC
Res2Net-101	1.2052	69.51%	58.60%	1.6336	69.89%	62.70%	
EfficientNet-B7	1.2835	70.49%	59.13%	1.2726	64.29%	59.23%	
ViT	1.1135	69.87%	60.20%	**0.8981**	68.20%	62.98%	
Swin-Base	1.2534	71.53%	58.63%	1.1194	**74.04%**	61.48%	
MViT-V2-Base	1.1775	69.63%	62.15%	0.9917	68.61%	58.40%	
ResNet-3D	1.2023	73.75%	**68.00%**	1.2788	67.04%	61.00%	
TimeSformer	**1.1116**	**73.77%**	**68.00%**	1.0129	**74.04%**	**63.50%**	

TABLE 16. Inter-dataset evaluation scores of **self-supervised** models fine-tuned on FaceForensics++ [33] dataset and evaluated on the remaining three datasets.

Training Dataset: FaceForensics++							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Supervised	**0.7400**	**69.21%**	**64.70%**	**0.7968**	**69.54%**	**64.35%**	FakeAVCeleb
Dino	3.4485	62.66%	60.32%	3.6146	64.73%	61.45%	
CLIP	0.8256	65.37%	59.63%	0.9682	66.54%	59.35%	
Supervised	**0.6256**	**74.47%**	**66.70%**	0.6993	71.45%	**66.00%**	CelebDF-V2
Dino	3.0070	68.39%	59.50%	3.4239	65.94%	60.10%	
CLIP	0.6627	68.77%	61.45%	**0.6678**	**73.97%**	65.45%	
Supervised	1.1463	61.54%	58.38%	1.1353	66.51%	61.90%	DFDC
Dino	7.5253	58.14%	54.35%	5.9902	61.83%	57.73%	
CLIP	**0.6647**	**71.49%**	**66.60%**	**0.8275**	**67.59%**	**62.85%**	

TABLE 17. Inter-dataset evaluation scores of models trained on DFDC [43] dataset and evaluated on the remaining three datasets.

Training Dataset: DFDC							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Xception	1.4046	58.38%	55.25%	1.8346	60.31%	53.63%	FakeAVCeleb
Res2Net-101	2.0891	59.23%	56.33%	1.6953	59.77%	55.43%	
EfficientNet-B7	**1.0800**	65.31%	61.63%	**1.0920**	**71.87%**	**65.40%**	
ViT	1.2515	59.31%	56.00%	1.1361	60.12%	57.43%	
Swin-Base	1.2053	**67.81%**	**62.90%**	1.2668	63.48%	60.25%	
MViT-V2-Base	1.2121	63.46%	60.05%	1.2139	65.75%	61.30%	
ResNet-3D	1.1114	63.19%	54.50%	1.2748	62.02%	56.50%	
TimeSformer	1.1582	65.34%	62.00%	1.6968	67.80%	61.00%	
Xception	1.1784	67.90%	61.25%	1.7465	64.95%	58.20%	CelebDF-V2
Res2Net-101	1.2293	83.01%	74.95%	1.1859	83.57%	72.35%	
EfficientNet-B7	0.8278	79.82%	70.15%	1.2972	74.27%	68.45%	
ViT	**0.7301**	85.62%	**76.45%**	0.9351	73.81%	67.25%	
Swin-Base	0.8411	84.36%	76.60%	1.1246	80.34%	73.20%	
MViT-V2-Base	0.8548	**87.83%**	71.55%	**0.7711**	**84.75%**	**76.75%**	
ResNet-3D	0.7638	79.60%	72.00%	0.7806	77.88%	72.00%	
TimeSformer	1.0558	76.48%	71.00%	1.3635	79.60%	74.00%	
Xception	1.3048	64.59%	60.95%	1.4530	67.29%	62.70%	FaceForensics++
Res2Net-101	2.7490	61.15%	56.48%	2.3375	65.15%	60.78%	
EfficientNet-B7	**1.2085**	68.82%	64.08%	1.4401	**70.71%**	**65.83%**	
ViT	1.6779	66.43%	59.13%	1.6781	60.19%	57.28%	
Swin-Base	1.6806	68.32%	63.65%	1.7493	64.66%	60.20%	
MViT-V2-Base	1.6317	65.82%	61.55%	1.6911	67.54%	62.88%	
ResNet-3D	1.5308	57.27%	57.86%	**1.3507**	58.91%	57.50%	
TimeSformer	1.2122	**70.17%**	**66.79%**	1.9148	67.90%	60.00%	

TABLE 18. Inter-dataset evaluation scores of **self-supervised** models fine-tuned on DFDC [43] dataset and evaluated on the remaining three datasets.

Training Dataset: DFDC							Evaluation Dataset
Model	With Augs			No Augs			
	LogLoss	AUC	ACC	LogLoss	AUC	ACC	
Supervised	1.2860	57.72%	52.25%	**0.9593**	60.67%	57.58%	FakeAVCeleb
Dino	3.5511	**60.45%**	**56.70%**	3.8135	**61.42%**	**59.20%**	
CLIP	**1.1978**	55.09%	52.33%	1.1680	55.04%	54.80%	
Supervised	0.8549	**72.57%**	**65.20%**	0.8149	69.37%	65.95%	CelebDF-V2
Dino	2.8856	69.28%	63.50%	3.4654	62.35%	57.50%	
CLIP	**0.7905**	66.77%	60.65%	**0.6538**	**76.81%**	**71.40%**	
Supervised	0.9295	64.05%	59.15%	**0.8214**	67.33%	62.10%	FaceForensics++
Dino	3.3117	63.87%	59.28%	2.9844	66.65%	61.95%	
CLIP	**0.7500**	**66.29%**	**61.90%**	0.8216	**68.50%**	**63.08%**	

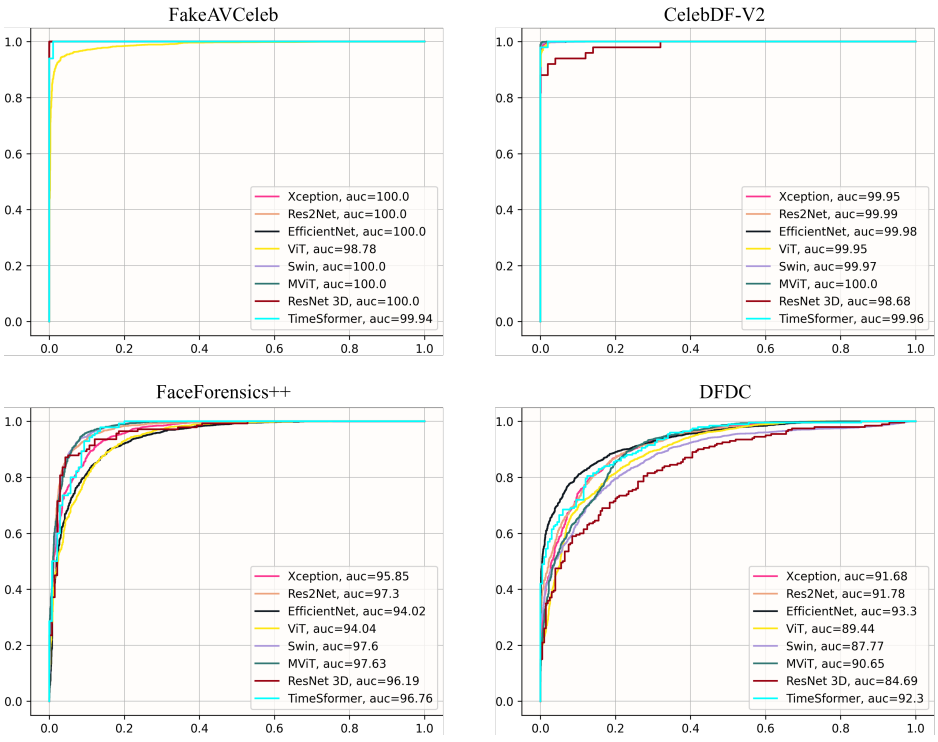

FIGURE 8. ROC curves of each of the model when evaluated on each of the 4 different participating datasets in an intra-dataset evaluation setting.

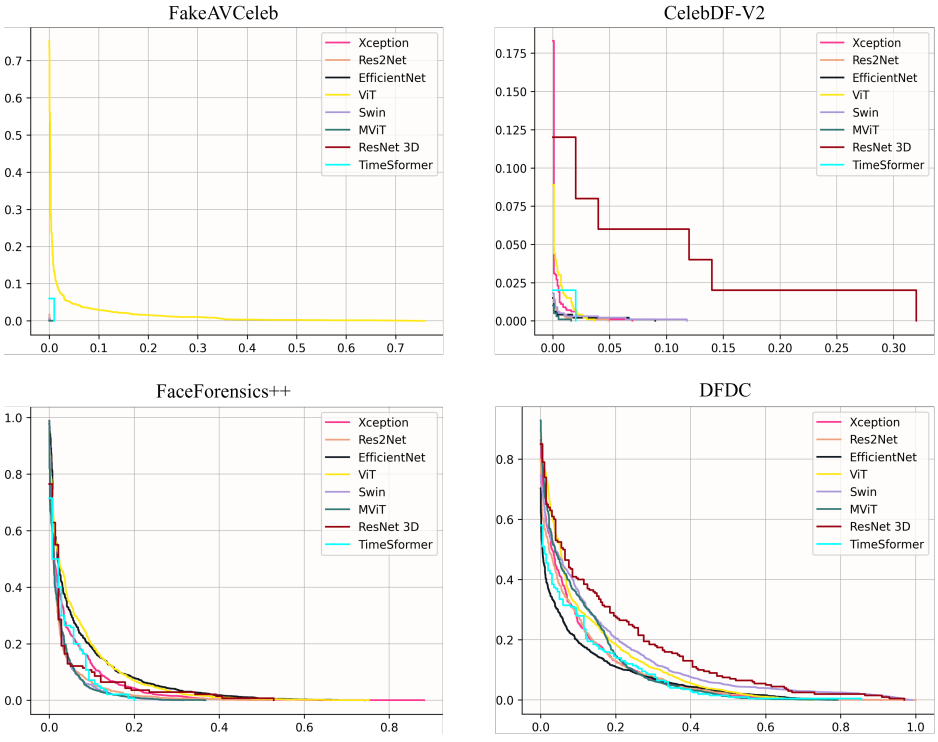

FIGURE 9. DET curves of each of the model when evaluated on each of the 4 different participating datasets in an intra-dataset evaluation setting.

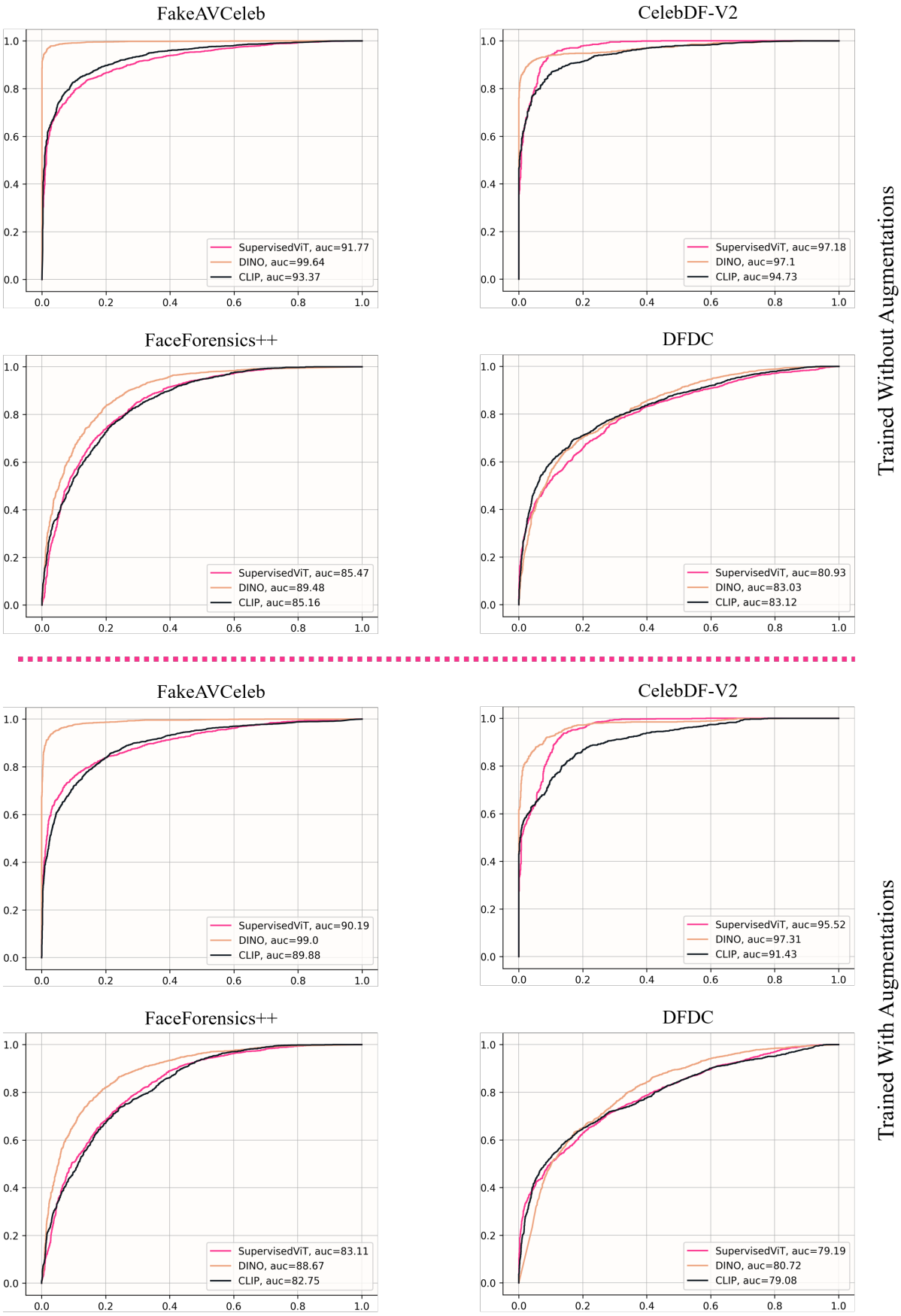

FIGURE 10. ROC curves of self-supervised models trained and evaluated on each dataset using the intra-dataset evaluation scheme.

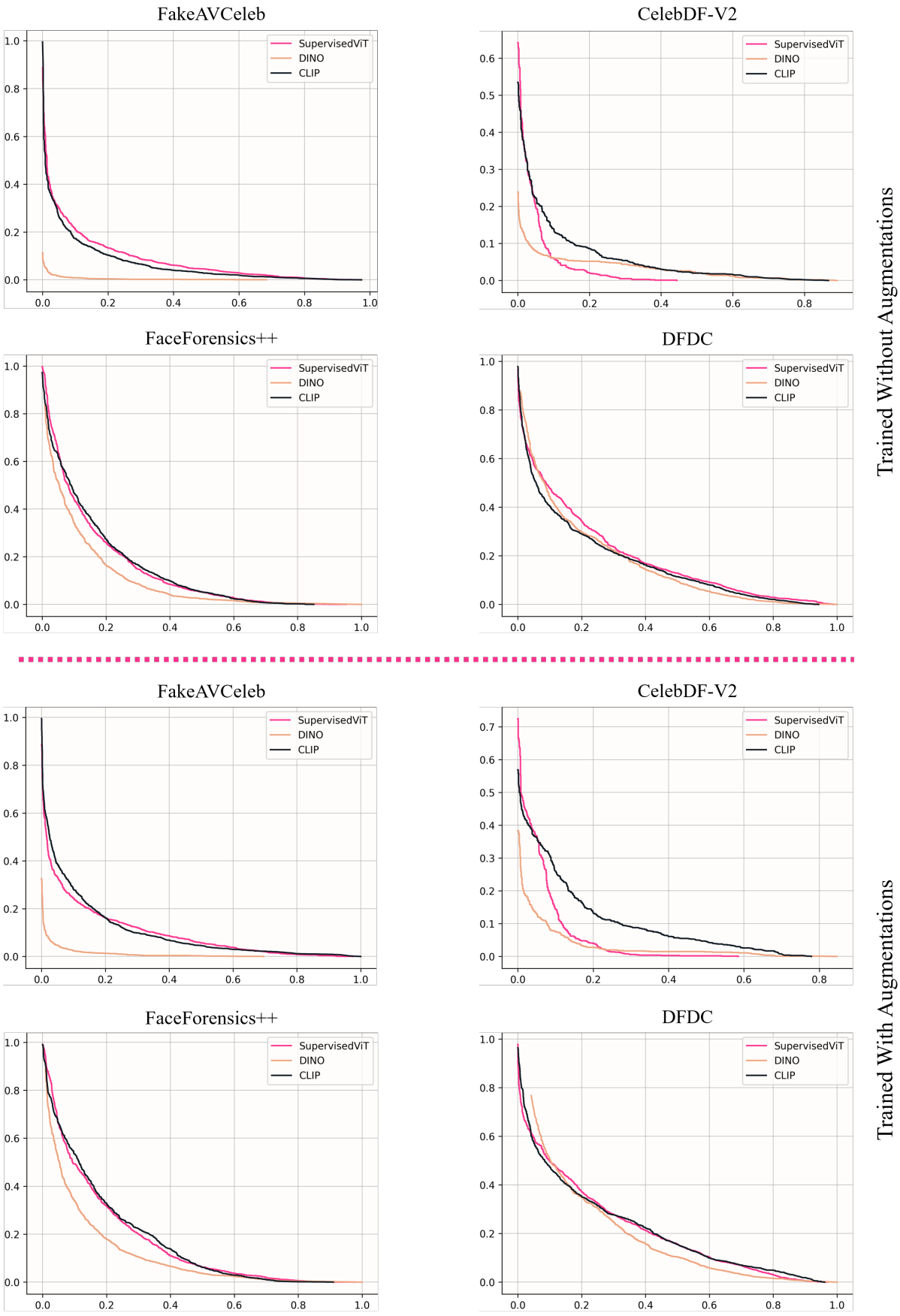

FIGURE 11. DET curves of self-supervised models trained and evaluated on each dataset using the intra-dataset evaluation scheme.

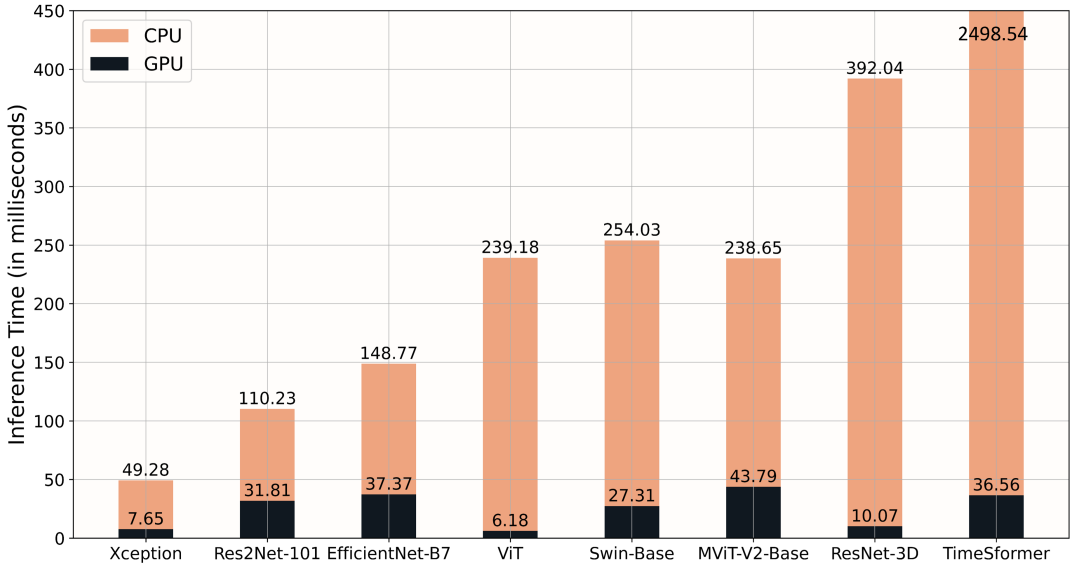

FIGURE 12. This bar chart highlights the efficiency of the supervised models in terms of inference time on both GPU and CPU devices. It reveals that CNN models outperform transformer models, taking nearly half the time for processing a single image frame on CPU. On GPU, the figure illustrates that all models achieve inference in less than 45 milliseconds at most. ViT and Xception models are the fastest among other models on GPU inference speeds, taking less than 10 milliseconds to process a single frame.

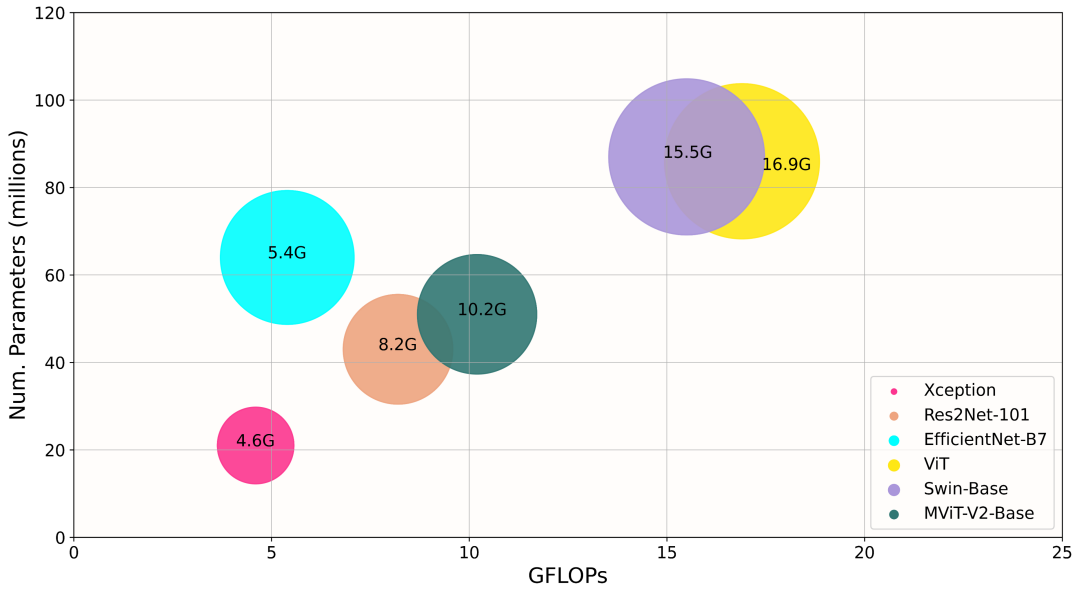

FIGURE 13. This figure illustrates the performance of supervised image models, showcasing both total parameters and the number of floating-point operations per second (GFLOPs). The results align with the preceding bar chart, emphasising the superior efficiency of CNN models, as compared to transformer models. It's important to note that video models, although not depicted here, exhibit a significantly higher number of floating-point operations per second, acting as outliers in the figure and slightly affecting its visual coherence. This disparity arises from the nature of video models processing more data at once, specifically 8 image frames, compared to image models that handle only one image at a time.