Selfie Segmentation in Video Using N-Frames Ensemble

Yong-Woon Kim¹, Yung-Cheol Byun², Addapalli V. N. Krishna³, K. Balachandran³

¹Centre for Digital Innovation, CHRIST (Deemed to be University), Bangalore, Karnataka, 560029, India
²Department of Computer Engineering, Jeju National University, Jeju 63243, South Korea
³Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bangalore, Karnataka, 560029, India

Corresponding author: Yong-Woon Kim (Jonathan.kim@christuniversity.in); Yung-Cheol Byun (ycb@jejunu.ac.kr).

Following are the results of a study on the "Leaders in INdustry -university Cooperation +" Project, supported by the Ministry of Education and National Research Foundation of Korea. And, this work was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0016977, The Establishment Project of Industry-University Fusion District)

ABSTRACT Many camera apps and online video conference solutions support instant selfie segmentation or virtual background function for entertainment, aesthetic, privacy, and security reasons. A good number of studies show that Deep-Learning based segmentation model (DSM) is a reasonable choice for selfie segmentation, and the ensemble of multiple DSMs can improve the precision of the segmentation result. However, it is not fit well when we apply these approaches directly to the image segmentation in a video. This paper proposes an N-Frames (NF) ensemble approach for a selfie segmentation in a video using an ensemble of multiple DSMs to achieve a high-performance automatic segmentation. The proposed NF ensemble approach executes only one segmentation model upon a current video frame and combines segmentation results of previous frames to produce the final result. For the experiment, we use four state-of-the-art image segmentation models and 81 videos dataset with a single-person view from publicly available websites. This paper calculates Intersection over Union (IoU), IoU standard deviation, false prediction rate, Memory Efficiency Rate and Computing power Efficiency Rate to measure the performance of segmentation models. The average IoU value of the Two-Frames NF ensemble was 95.1253%, and the Three-Frames NF ensemble was 95.1734%, whereas the average IoU value of single models was 92.9653%. The result shows that the proposed approach improves the accuracy of selfie segmentation by more than 2% on average. The result of cost efficiency measurement shows that the proposed method consumes less computing power like single models. To sum up the overall work, the proposed approach is as fast as single selfie segmentation models. At the same time, it produces optimized and improved segmentation results like the ensemble of multiple segmentation models at once.

INDEX TERMS Deep learning, ensemble, image segmentation, multi-frames, neural network, selfie, soft voting, video.

I. INTRODUCTION Increased use of mobile phones caused the popularity of “selfies” (self-portrait photographs) these days. Popular camera apps support automatic background change of selfie photos. Many online video conferencing solutions support real-time background change functions for aesthetic, privacy, and security reasons. Organizations are using these solutions more and more since many people are working from home because of the COVID-19. Selfie segmentation is a crucial technology to enable such kinds of functions. Selfie segmentation is a subset of image segmentation. Many image segmentation methods have already been studied and developed. Many of them use traditional image segmentation methods, such as thresholding, watershed, region growing, clustering using contour and edge, graph cut, and Markov random fields [1], [2]. A segmented area should be homogeneous and uniform for good image segmentation, but it remains a challenging task [3]. Recently, Deep-Learning (DL) based Segmentation Model (DSM) has opened a new era of image segmentation [4]. DSMs have made remarkable improvements in speed and accuracy for
image segmentation compared to traditional methods [5]-[9]. These models predict segmentation regions using semantic labels for every pixel of the image [10].

In general, the ensemble of multiple methods can improve the performance of image segmentation. The key of the ensemble approach is to combine various models to create a more effective model [11]. The collective decision produced by the ensemble can reduce generalization errors of prediction and improve the performance accuracy [12], [13]. Many studies show that the combination of multiple segmentations can produce better predictive performance than the individual segmentation model [14]-[16].

Video Instance Segmentation (VIS) is a task that can classify objects into the pre-defined classes, trace objects throughout a video, segment classified objects, and classify objects with localization within a video. The extension of instance image segmentation to the video domain is a natural step as the use of video applications is increasing. Selfie segmentation in a video is a subset of VIS, and it can recognize and extract the human body, normally the upper body, from continuous video frames [17]-[18].

A. MOTIVATION
The use of selfie segmentation is increasing these days. However, high performance and real-time selfie segmentation in the video remains a challenging problem even with recent developments on image segmentation and matting using DSMs [5], [18]-[23]. There are several studies on selfie segmentation using single DSMs for single images. However, these studies focus on a single image. They are not nicely fit for image segmentation in a video considering the characteristics of a video, such as sudden noises or the continuity of adjacent video frames. Several studies on the ensemble method of multiple DSMs show better image segmentation accuracy compared to single DSMs. However, the ensemble approach is not satisfactory when we directly apply it to selfie segmentation in a video because the speed of segmentation is slower than a single DSM.

B. OUR METHOD
To address these issues, this paper proposes a novel approach for the selfie segmentation in a video using an N-Frames ensemble of multiple DSMs to achieve a high-performance automatic segmentation. The proposed method is not limited to selfie segmentation but can also be applied to segment any objects in a video. The characteristics of the proposed ensemble approach are as below:

- It is as fast as a single DSM.
- It can generate optimized results using an ensemble of multiple DSMs.
- It is robust to the sudden noises of a video.
- It is suitable for image segmentation of slowly moving objects such as people in a video.

C. CONTRIBUTIONS
The followings are major contributions of this study for selfie segmentation in a video.

- A novel N-Frames ensemble method for selfie segmentation in a video is proposed. The experiment result shows that the proposed method is as efficient as a single DSM, and it resolves the aforementioned challenges of the selfie segmentation in a video satisfactorily. To the best of our knowledge, no previous works have used the same approach for selfie segmentation in a video.

- This paper compares four state-of-the-art DSMs, the ensemble of multiple DSMs and the proposed N-Frames ensemble to measure the performance of selfie segmentation in a video. The DSMs experimented in this work were developed recently and show state-of-the-art performance in image segmentation. According to our best knowledge, no previous works use the ensemble of these DSMs for selfie segmentation in a video.

- A quantitative experiment with approximately 40,000 test images from 81 videos is conducted to evaluate the segmentation results. The size of the sample is an important factor to ensure the precision of the experiment. The number of test datasets used in this work is good enough to assure the accuracy of the experiment. According to our related works, it is rare to use such a massive number of test datasets for the experiment of image segmentation.

D. STRUCTURE
The following Section II explains the related works of DSM and human segmentation in a video. Section III introduces our ensemble approach for selfie segmentation in a video. Section VI presents experiment results and analysis of results. Section V presents the overall discussion and inspirations. Section VI is the conclusion part, covering the summary of experiments and the contribution of this paper.

II. RELATED WORKS
This section explores the detail of related works for image segmentation in a video.

A. IMAGE SEGMENTATION USING DEEP-LEARNING
Long et al. [24] proposed the fully convolutional network, FCN has shown good performance in semantic segmentation. Adam Paszke et al. [25] proposed an ENet (efficient neural network) model. It has an optimized deep neural network architecture to perform a real-time semantic segmentation on GPU hardware. Huang et al. [26] and Simon Jegou et al. [27] proposed a dense convolutional network (DenseNet) with several densely connected blocks, which has shown excellent results on image classification tasks. Deeplab [28] architecture uses atrous convolutions with upsampled filters for dense feature extraction and uses CRF.
to get a better localization, especially along the edge of objects. Chen et al. [29] proposed a network architecture called DeepLabv3. It adopts a dilated (atrous) convolution for the downsampling layer and upsampled filters for a dense feature map extraction and for long-range context capturing. MobileNets [2] is an efficient and lightweight segmentation model for mobiles and embedded systems. Mark Sandler et al. [30] introduced MobileNetV2, an improved version of MobileNets. It is based on an inverted residual structure that connects the thin bottleneck layers, and it improves the state-of-the-art performance on multiple tasks. Andrew Howard et al. [31] proposed MobileNetV3-Large and MobileNetV3-Small models using hardware-aware Network Architecture Search and NetAdapt algorithm. It was developed to achieve the best semantic segmentation on mobile devices. Due to the performance and efficiency of MobileNetV2 and MobileNetV3, these models are used as a backbone with other networks such as DeepLabv3, U-Net, LR-ASPP, etc., for semantic segmentation. Lingyu Zhu et al. [32] proposed an end-to-end portrait segmentation architecture with unique cross-granularity categorical attention and boundary enhancement mechanisms in a unified framework. Song-Hai Zhang et al. [33] proposed a real-time portrait segmentation model called PortraitNet for mobile devices. The model includes two modules, the encoder module and the decoder module. PortraitNet utilized MobileNetV2 as a backbone in the encoder module and U-shape architecture as a decoder. Sachin Mehta [34] introduced a fast and efficient ESPNet based on an efficient spatial pyramid (ESP) architecture. It can efficiently perform the semantic segmentation of high-resolution images using limited resources in terms of computation, memory, and power. ESPNetv2 [35] is a lightweight architecture for semantic segmentation that can be easily deployed on edge devices. Hyojin Park et al. [36] introduced an extremely lightweight portrait segmentation model, SINet, which used an information blocking decoder to measure the confidence score. It blocks the flow of the decoder utilizing the information.

Compared to traditional image segmentation methods, these DSMs achieved remarkable performance improvement in image segmentation [5]-[9]. Some studies showed that DSMs could be utilized for a selfie segmentation purpose.

B. IMAGE SEGMENTATION USING ENSEMBLE

Warfield et al. introduced an algorithm to combine multiple segmentations and validate image segmentation performance [37]. Rohlfing et al. introduced a shape-based averaging method to combine multiple segmentations and compared it to other ensemble methods [38]. Andrew Holliday et al. applied a compressed model technique for DL ensemble to the problem of semantic segmentation and achieved real-time speed [39]. D. Marmanis et al. applied the ensemble of multiple DL models using Fully Convolution Network (FCN) and achieved excellent segmentation results [40]. Y.-W. Kim et al. proposed an ensemble of multiple heterogeneous DSMs for portrait segmentation and analyzed the efficiency of the ensemble approach. The authors showed that some combinations of DSM could perform higher accuracy than single models while using low memory and computing power [41], [42].

In general, an ensemble model can generate optimized results using the combination of multiple machine-learning models or deep-learning models. Several studies applied the ensemble method using multiple DSMs to the image segmentation domain and achieved better performance compared to single DSMs. These works showed that the ensemble of various DSMs can be a reasonable choice for image segmentation applications.

C. IMAGE SEGMENTATION IN VIDEO

J. Li et al. [43] proposed a novel approach to segment an object in a video using a proposal-driven framework. The authors adopted the ResNet model for the proposal and the PSPNet model for object segmentation. Yifan Liu et al. [44] developed a real-time video segmentation method that considers accuracy and temporal consistency. The proposed method conducts per-frame inference using compact networks. The authors included the PSPNet18, the MobileNetV2 and a lightweight HRNet to verify that the proposed methods can improve the segmentation accuracy and the temporal consistency without extra computation and post-processing during inference. Mingyu Ding et al. [45] proposed a novel framework for joint estimation of semantic video segmentation and optical flow. The authors used the original PSPNet and the modified FlowNetS as the baseline network unless otherwise specified. Shan Lin et al. [46] proposed a Multi-frame Feature Aggregation (MFFA) module to improve instrument segmentation. It uses temporal and spatial relationships between frame pixels for feature aggregation. The authors applied the proposed approach to the real-time instrument segmentation using DeepLabV3+ with ResNet50 and MobileNet as backbone feature extractors. Federico et al. combined offline and online learning approaches. The method segments a particular object instance in a video by providing one or a few segmentation masks [47]. Several articles studied video portrait segmentation, such as “Automatic Real-time Background Cut for Portrait Videos” proposed by Xiaoyong Shen et al. [48]. They have segmented the person from the background on a video using a global background attenuation model. Monica Gruosso et al. [8] showed the possibility of automatic human recognition and segmentation in a surveillance video system. The authors used SegNet [49] encoder-decoder Convolution Neural Network (NCNN) model for the experiment. Tairan Zhang et al. [50] proposed a real-time single-person segmentation framework in a video. The framework combined a CNN model and a tracking system using a level set algorithm. The CNN model obtained a human segmentation result from a specific frame.
in a video and passed it to the tracking system to capture the human segmentation of the rest frames.

The extension of instance image segmentation to the video domain is a natural step as the requirement of image segmentation in a video is increasing. There are several studies on image segmentation in a video. These studies showed that DSMs could be used for image segmentation in a video. In these studies, deep-learning models for image segmentation were optimized by considering the characteristics of a video. However, the ensemble of multiple DSMs for the selfie segmentation in a video is an area that needs further study.

It is evident that DSMs are a reasonable choice for the selfie segmentation in a video, and the ensemble of multiple DSMs can improve the precision of the segmentation result. However, it is not fit well when we apply these approaches directly to the image segmentation in a video. For the same reason, many researchers have developed various DSMs having optimized architectures for the image segmentation in a video. Considering the advantage of the ensemble model, it is an appropriate attempt to utilize the ensemble of multiple DSMs for image segmentation in a video. This paper proposes a novel ensemble called “N-Frames ensemble” for the selfie segmentation in a video to address issues that occurred when we apply single DSMs or the ensemble of multiple DSMs to the image segmentation in a video.

III. APPROACH

This section describes the proposed approach and the experiment environment in detail.

A. ENSEMBLE APPROACH

A single model segmentation uses only one model to produce the segmented outputs for the input images. The ensemble is a machine learning technique that incorporates several single models for an optimized prediction result. In general, the ensemble method takes two ways; one way is to combine heterogeneous models which are trained on the same dataset or to combine homogenous models which are trained on different datasets. The diversity of models is generally believed to be one of the critical performance factors in an ensemble [51]. In this paper, we have used pre-trained heterogeneous models for the ensemble. Fig. 1 shows the conceptual diagram for selfie segmentation in video using the N-Models (NM) ensemble approach. All segmentation models generate output results for every single frame in a video, and the output results are combined using the ensemble method. In more detail, at a given time t, a video has Frame, and the Frame is fed to segmentation models Model 1 to Model n to generate segmented output Mask 1 to Mask n. Finally, the masks are combined to make optimized output Mask, as shown in Fig. 1.

This paper proposes a novel ensemble approach for selfie segmentation in the video. We call it an N-Frame (NF) ensemble. Fig. 2 shows the conceptual diagram for selfie segmentation in video using the NF ensemble approach. The NF ensemble method uses a single segmentation model to generate an output result for a single frame in a video. It combines the output results of previous frames with the current output result using the ensemble method. In detail, at a given time t, a video has Frame, and the Frame is fed to a segmentation model n to generate a segmented n-th output mask. The NF ensemble method combines the output mask with other output masks generated from previous segmentation model n-1, ..., model 1 to generate the final Mask. It rotates each segmentation model using a round-robin way. For example, if there are total n models for ensemble and model n was used for Frame, at a given time t. The NF ensemble uses model 1 for the next Frame. It combines the segmented output mask of model 1 with previous output masks generated from segmentation model n, ..., model 2 to create final output Mask.
The benefit of this approach is that only one segmentation model needs to make a segmentation mask for a current frame in a video at a given time \( t \). The advantages of the NF ensemble are as below:

- NF ensemble can proceed as fast as a single segmentation model. It requires only one segmentation model at a given time \( t \). It combines the segmented output of the current segmentation model with the segmented results of previous segmentation models. The speed of segmentation is the mean speed of all segmentation models involved in the ensemble.

- NF ensemble can enjoy the advantage of the ensemble approach. It combines segmented results of multiple segmentation models to generate optimized output results for every video frame.

- NF ensemble is robust to sudden noises of a video. It merges segmented results of multiple frames, and the effect of fusing multiple frames reduces the sudden noises of a video.

- NF ensemble is suitable for the segmentation of the human body or slowly moving object in a video. For example, the movement of humans is almost frozen when we observe it at a millisecond level. Considering a video of 30 FPS, a time slice of each frame will be approximately 33 milliseconds, and the difference of human movement among each frame will be almost ignorable.

The disadvantages of the NF ensemble are as below:

- NF ensemble requires as much memory space as that of a total number of segmentation models used in an ensemble.

- NF ensemble is not suitable for the segmentation of a video that has a fast-moving object. As it merges segmented results of multiple frames in a video, combining numerous segmented results always has the chance to produce a blurring effect. Therefore, if the difference of segmented objects in video frames is high, it will affect the quality of the final segmentation result.

- NF ensemble is not suitable for the segmentation of a low FPS video. Frames of a low FPS video have a longer time gap, and it has a chance to have a higher difference of each frame than a high FPS video.

In contrast to ordinary learning approaches using a single model, an ensemble approach combines the results of the first-level learners and generates final results from the second-level learner. Among many ensemble methods, averaging and voting are the most commonly used. There are different voting methods such as Majority voting, Plurality voting, Weighted voting, Simple soft voting, Weighted soft voting, etc. [51]-[53]. In this paper, a simple soft voting method is used for the ensemble. It treats the individual classifiers equally and averages the outputs of the individual. Equation (1) is for the Simple soft voting:

\[
H^1(x) = \frac{1}{T} \sum_{t=1}^{T} h^1_{r(i,t)}(x) \tag{1}
\]

where \(|r| < |T|\), \(|r(i, t)| < |T|\), \( T > 1 \) and \( t \geq T \) at given time \( t \). In a set of \( T \) individual classifiers \( \{h_1, ..., h_T\} \), the classifier \( h_i \) generates a \( l \)-dimensional vector \((h^1(x), ..., h^i(x))\) for the instance of \( x \), where \( h \) indicates the classifiers and \( h^i(x) \) \( \in \{0, 1\} \).

In general, a round-robin method can be presented by the following (2):

\[
r = t - n \left[ \frac{t}{n} \right] + 1 \tag{2}
\]

Equation (3) denotes a round-robin method for a different time frame \( t \).

\[
r(i, t) = t + i - n \left[ \frac{t + i}{n} \right] + 1 \tag{3}
\]

**B. PERFORMANCE MEASUREMENT**

The Intersection over Union (IoU) is a metric used to measure the accuracy of image segmentation. An IoU is defined as in (4) below:

\[
IoU = \frac{A \cap B}{A \cup B} \tag{4}
\]

where \( A \) is an image segmentation result, \( B \) is a ground truth image, \( A \cap B \) is the intersection of \( A \) and \( B \), and \( A \cup B \) is the union of \( A \) and \( B \). IoU standard deviation is used to measure the prediction variance.

False Negative Rate (FNR) and False Discovery Rate (FDR) are used for validating the accuracy measurement. FNR is used to measure lesser regions than the ground truth. FDR is used to measure larger regions than the ground truth. FNR and FDR are defined as in (5) and (6) below:

\[
FNR = \frac{FN}{FN + TP} \tag{5}
\]

\[
FDR = \frac{FP}{FP + TP} \tag{6}
\]

where \( FN \) is False Negative, \( TP \) is True Positive, and \( FP \) is False Positive. The bias error means the amount of difference between the ground truth and the prediction. In general, the ensemble of multiple models can reduce bias. In this paper, FDR+FNR is used to measure the bias error. [FDR-FNR] is used to measure the variance of prediction. In this paper, Memory Efficiency Ratio (MER) and Computing Efficiency Ratio (CER) are used to measure the cost efficiency rate of segmentation models. MER indicates the memory efficiency of the model. Since an efficiency ratio can be denoted as costs over gain, MER measures the required memory size to gain accuracy, and it is calculated as in (7):

\[
MER = \frac{M}{IoU} \tag{7}
\]
where \( M \) is the number of parameters and \( \text{IoU} \) is the accuracy of a given model. CER indicates the computing power efficiency of a model. CER measures required computing power to gain accuracy. CER is calculated using (8) as below:

\[
CER = \frac{C}{\text{IoU}} \quad (8)
\]

where \( C \) is Floating-Point Operations (FLOP), and \( \text{IoU} \) is the accuracy of a given model.

### C. EXPERIMENT ENVIRONMENT

This work uses four selfie DSMs for the ensemble. There are several pre-trained DSMs for the selfie segmentation on open-source websites [54]-[59]. These DSMs can be used for selfie segmentation experiments immediately without any training or parameter tuning process. This paper chooses four state-of-the-art pre-trained selfie DSMs which are publicly available [59]. These DSMs were verified by measuring the accuracy and comparing with other papers [36], [60]. We have used the EG1800+CDI test dataset [42] for the verification, where EG1800 [61] is a dataset used in reference papers. Among the four DSMs, MNV2 uses MobileNetV2 as an encoder and an upsampling block with a transpose convolution as a decoder. Its input/output resolution is 128x128. MNV3 uses MobileNetV3 with a depth multiplier 0.5 as an encoder and an upsampling block with a transpose convolution as a decoder. It has 224x224 input/output resolution. PN is a pre-trained model of PortraitNet [33]. It has 224x224 input/output resolution. SN is a pre-trained model of SINet [36]. It has 320x320 input/output resolution. Table I shows the backbone of each model and the accuracy comparison using the verification dataset EG1800+CDI.

| Model | Backbone | \( \text{IoU} \) (%) of ref. | \( \text{IoU} \) (%) of our test |
|-------|----------|-----------------------------|-----------------------------|
| MNV3  | MobileNetV3 | 94.19                       | 94.76                       |
| MNV2  | MobileNetV2 | *95.76                      | 96.21                       |
| PN    | MobileNetV2 | 95.99                       | 95.7                        |
| SN    | Custom     | 95.29                       | 95.43                       |

(* The value was calculated based on the accuracy ratio to SINet [36].)

We experimented with these four pre-trained selfie DSMs to segment the body of a human from video dataset. The tested video dataset was collected from publicly available websites. The dataset is a collection of captured frames from 81 videos having a single person view. The resolution of videos is 480x360 pixels, the duration is 10 seconds ~ 20 seconds, and the frame rate of videos is in the range of 25 to 30 frames per second (FPS). Finally, the test dataset consists of around 40,000 frame images generated from the above mentioned 81 videos. Single models, NM ensemble and NF ensemble models are evaluated using the test dataset. We experimented with all ensembles of four single models. The simple soft voting is the combining method for the ensemble of single models. The experiment was conducted using Python with Keras, PyTorch, OpenCV and TensorFlow libraries on Ubuntu operating system and GeForce GTX 1080 GPU hardware.

### IV. EXPERIMENT

This paper uses four DL-based selfie segmentation models, namely MNV2, MNV3, PN, and SN, to compose proposed ensemble models. Four single models and proposed ensemble models are experimented with to evaluate the accuracy, variance, and bias errors. \( \text{IoU} \) is used for accuracy measurement, \( \text{IoU} \) standard deviation for variance error measurement, and \( \text{FNR+ FDR} \) for bias error measurement. \( |\text{FNR} - \text{FDR}| \) means the absolute value of \( \text{FNR-FDR} \), which measures the difference between \( \text{FNR} \) and \( \text{FDR} \). MER and CER are also calculated to measure the efficiency of single models and proposed ensemble models.

#### A. SINGLE MODEL RESULT

Table II shows the experimental results of four single models MNV3, MNV2, PN, and SN. In the table, MNV2 produces the highest \( \text{IoU} \) value, the lowest \( \text{IoU} \) standard deviation, the lowest \( \text{FNR+ FDR} \), and the lowest \( |\text{FNR} - \text{FDR}| \). It indicates that MNV2 is the most accurate in selfie segmentation among four single models. The lowest \( \text{IoU} \) standard deviation indicates that the variance error is lowest, and the lowest \( \text{FNR+ FDR} \) indicates that the bias error is also the lowest. The lowest \( |\text{FNR} - \text{FDR}| \) means MNV2 is well balanced in false positive regions and false negative regions. MNV3 shows the highest \( |\text{FNR} - \text{FDR}| \) value, and MNV3 tends to predict false positive regions more than false negative regions. SN is the second most well-performing model compared to others.

| Model | IoU (%) of ref. | IoU (%) of our test |
|-------|----------------|---------------------|
| MNV3  | 92.3115        | 2.5137              |
| MNV2  | 95.1166        | 1.7565              |
| PN    | 90.2161        | 2.3212              |
| SN    | 94.2172        | 1.8655              |

#### B. TWO-MODELS AND THREE-MODELS NM ENSEMBLE RESULT

In this paper, the NM ensemble was introduced in Section III-A. To evaluate the NM ensemble, we experimented with the Two-Models\((N=2)\) NM ensemble and Three-Models\((N=3)\) NM ensemble model. Table III shows the result of the Two-Models NM ensemble. In the table, MNV2+SN produces the highest \( \text{IoU} \) value indicating the most accurate in segmentation among Two-Models NM ensembles and shows better accuracy than MNV2 or SN single models. MNV2+PN produces an improved \( \text{IoU} \) value than MNV2 or PN single model, but \( \text{IoU} \) standard deviation and \( |\text{FNR} - \text{FDR}| \) are higher than MNV2, which indicates the variance error and the balance of error is not better than MNV2. MNV3+MNV2 shows improved results in \( \text{IoU} \),
FNR+FDR, and $|FNR-FDR|$ than MNV3 and MNV2. MNV3+PN shows much improvement in IoU, IoU standard deviation, FNR+FDR, and $|FNR-FDR|$ than MNV3 and PN. The range of improved IoU is $1.822\% - 3.9174\%$, and the $|FNR-FDR|$ value of MNV3+PN shows $3.4143\%$ improvement than MNV3 single model. MNV3+SN also shows much improvement in IoU and $|FNR-FDR|$ than MNV3 and SN, but it shows higher IoU standard deviation than SN, which indicates the combined model has a higher variance error than SN single model. PN+SN shows a better IoU value and IoU standard deviation than PN. But its IoU value is lower, and IoU standard deviation is higher than SN single model.

Table IV shows the experiment result of the Three-Models NM ensemble. Three combinations out of four (75%) show improved IoU value, IoU standard deviation, and FNR+FDR than the best single model MNV2. All Three-Models NM ensembles show improved results than any single model that participated in the ensemble. The range of IoU values is between 94.0351% and 96.5587%, and the difference of maximum and minimum is 2.5236%. It indicates that the variance of IoU values of the Three-Models NM ensemble is less than the Two-Models NM ensemble.

The same trend can be observed in IoU standard deviation. The range of IoU standard deviation of the Two-Models NM ensemble is between 1.5435% and 2.2678%, and the difference between the two values is 0.7245%. The range of IoU standard deviation of the Three-Models NM ensemble is between 1.5445% and 1.7601%, and the difference between the two values is 0.2216%. It indicates that the variance of the Three-Models NM ensemble is less than the Two-Models NM ensemble.

To verify the idea of NF ensemble, the difference of ground truth segmentation masks among adjacent frames in the test dataset is calculated. The below (9) calculates the difference of n number of consecutive frames in a video.

$$D(n) = \bigcup_{i=0}^{n-1} G_{t-i} - \bigcap_{i=0}^{n-1} G_{t-i}$$

where $G_i$ is a ground truth segmentation mask of a Frame at a given time $t$, $n > 1$ and $t \geq n$. This equation can be expressed using bit operation such as $D(n) = (G_i \text{ OR } G_{t-1} \text{ OR } ... \text{ OR } G_{t-n+1})$ XOR $(G_i \text{ AND } G_{t-1} \text{ AND } ... \text{ AND } G_{t-n+1})$. Table VI shows the difference of consecutive 2-frames, 3-frames and 4-frames in the test dataset. The result shows that the difference of consecutive 2-frames is less than 0.35%, consecutive 3-frames is less than 0.65%, and consecutive 4-frames is less than 1%.

| TABLE IV. THREE-MODELS NM ENSEMBLE (1-FRAME) (%) |
|-----------------|--------|------------|--------------|--------|
| MNV3+MNV2+PN(1F) | 95.3514 | 1.7661 | 4.7705 | 0.7186 |
| MNV3+MNV2+SN(1F) | 96.0695 | 1.6852 | 4.0005 | 1.2286 |
| MNV3+PN+SN(1F)  | 94.5861 | 1.7298 | 5.5494 | 1.3640 |
| MNV2+PN+SN(1F)  | 95.4597 | 1.5445 | 4.6341 | 0.0747 |

C. TWO-FRAMES AND THREE-FRAMES NF ENSEMBLE RESULT

In this paper, we proposed a novel ensemble approach called an N-Frame (NF) ensemble in Section III.A. The conceptual diagram of selfie segmentation in a video using NF ensemble is shown in Fig. 2. The idea of NF ensemble is:

- If the difference of segmented outputs of adjacent video frames is small enough, then the ensemble of segmented outputs of adjacent video frames using NF ensemble will produce almost the same result as the ensemble of segmented outputs of a single frame using NM ensemble.

To evaluate the proposed NF ensemble, Two-Frames(2F) and Three-Frames(3F) NF ensemble were experimented. In Table VII, the Two-Frames NF ensemble improves IoU, IoU standard deviation, FNR+FDR, and $|FNR-FDR|$ compared to single models. The performance of the Two-Frames NF ensemble is almost the same as the Two-Models NM ensemble. MNV2 + SN performs the highest IoU value indicating the most accurate segmentation among Two-Frames NF ensemble models and shows better accuracy than any other single model. IoU value is $96.4729\%$, which is slightly less than $96.5587\%$ of MNV2+SN(1F) NM ensemble. MNV3 + PN shows much improvement in IoU, IoU std, FNR+FDR and $|FNR-FDR|$ than MNV3 and PN. MNV3 + SN also shows much improvement.
improvement in IoU value than any other single model. PN + SN has better accuracy than MNV3 and PN single model and better IoU standard deviation than MNV3.

Table VIII shows Three-Frames NF ensemble models. All NF ensemble models show an improvement in IoU value than MNV3, PN, and SN models. Among four possible ensemble combinations, three of them show better IoU accuracy than any single model. The highest IoU value of the Three-Frames NF ensemble is 95.8342%, which is less than the highest IoU value of the Two-Frames NF ensemble, 96.4729%. The difference between these two IoU values is 0.6389%, and it is almost the same as the difference value of 3-frames in Table VI.

| TABLE VIII. THREE-FRAMES NF ENSEMBLE (%) | IoU | IoU Std | FNR+FDR | |FNR-FDR|
|------------------------------------------|-----|---------|---------|---------|
| MNV3+MNV2+PN(3F)                        | 95.1681 | 1.8852 | 4.9650 | 0.7186  |
| MNV3+MNV2+SN(3F)                        | 95.8342 | 1.8147 | 4.2465 | 1.2328  |
| MNV3+PN+SN(3F)                          | 94.4035 | 1.8741 | 5.7444 | 1.3692  |
| MNV2+PN+SN(3F)                          | 95.2878 | 1.6990 | 4.8154 | 0.0686  |

D. AVERAGE RESULT OF ALL MODELS

Table IX shows the average values of IoU, IoU standard deviation, FNR+FDR and [FNR-FDR] from single, Two-Models (1F) NM ensemble, Two-Frames (2F) NF ensemble, Three-Models (1F) NM ensemble, and Three-Frames (3F) NF ensemble models. Two-Models (1F) NM ensemble and Two-Frames (2F) NF ensemble models significantly improve IoU, IoU std, FNR+FDR, and [FNR-FDR] than single models. The difference of IoU value between Two-Models (1F) NM ensemble and Two-Frames (2F) NF ensemble models is 0.0615%, and it is much less than that of 2-frames in Table VI. Three-Models (1F) NM ensemble and Three-Frames (3F) NF ensemble models also show significant improvement in IoU, IoU std, FNR+FDR, and [FNR-FDR] than single models. The difference of IoU value between Three-Models (1F) NM ensemble and Three-Frames (3F) NF ensemble models is 0.1933%, and it is much less than the difference of 3-frames in Table VI.

| TABLE IX. AVERAGE OF ALL MODELS (%) | IoU | IoU Std | FNR+FDR | |FNR-FDR|
|-------------------------------------|-----|---------|---------|---------|
| Single Models                       | 92.9653 | 2.1149 | 7.2717 | 1.6290  |
| Two-Models (1F)                     | 95.1868 | 1.8978 | 4.9354 | 1.4894  |
| Two-Frames (2F)                     | 95.1253 | 1.9385 | 5.0001 | 1.4980  |
| Three-Models (1F)                   | 95.3667 | 1.6814 | 4.7386 | 0.8465  |
| Three-Frames (3F)                   | 95.1734 | 1.8182 | 4.9428 | 0.8473  |

Fig. 3 shows the average value of IoU and IoU standard deviation. Proposed ensemble models produce a higher IoU value than a single model. Two-Models (1F) NM ensemble and Two-Frames (2F) NF ensemble models show lower IoU standard deviation than a single model. Three-Models (1F) NM ensemble model shows the highest IoU and the lowest IoU standard deviation in the chart. Three-Frames (3F) NF ensemble model shows lower IoU standard deviation than the Two-Models (1F) NM ensemble and Two-Frames (2F) NF ensemble models. It indicates that the ensemble of more models can produce a lower IoU standard deviation.

| TABLE X. COMPARISON OF MER AND CER FOR SINGLE MODELS | Params (M) | FLOPs (G) | IoU (%) | MER | CER |
|-----------------------------------------------------|------------|-----------|---------|-----|-----|
| SN                                                   | 0.087      | 0.15      | 94.2172 | 0.092 | 0.159 |
| PN                                                   | 2.115      | 0.209     | 90.2161 | 2.344 | 0.232 |
| MNV3                                                 | 1.192      | 2.3724    | 92.3115 | 1.291 | 2.570 |
| MNV2                                                 | 3.625      | 7.227     | 95.1166 | 3.811 | 7.598 |

| TABLE XI. COMPARISON OF MER AND CER FOR ENSEMBLE MODELS | IoU (%) | MER | CER |
|--------------------------------------------------------|---------|-----|-----|
| PN+SN (1F)                                             | 94.0351 | 2.342 | 0.382 |
| *MNV3+SN (1F)                                          | 95.7374 | 1.336 | 2.635 |
| MNV3+PN (1F)                                           | 94.1335 | 3.513 | 2.742 |
| MNV2+SN (1F)                                           | 96.5587 | 3.844 | 7.640 |
| MNV2+PN (1F)                                           | 95.1321 | 6.034 | 7.816 |
| MNV3+MNV2 (1F)                                         | 95.5236 | 5.043 | 10.049 |
| PN+SN (2F)                                             | 93.9890 | 2.343 | 0.191 |
| *MNV3+SN (2F)                                          | 95.6587 | 1.337 | 1.318 |
| MNV3+PN (2F)                                           | 94.0867 | 3.515 | 1.372 |
| *MNV2+SN (2F)                                          | 96.4729 | 3.848 | 3.823 |
| MNV2+PN (2F)                                           | 95.0812 | 6.037 | 3.910 |
| *MNV3+MNV2 (2F)                                        | 95.4631 | 5.046 | 5.028 |
| MNV3+PN+SN (1F)                                        | 94.586 | 3.588 | 2.888 |
| MNV2+PN+SN (1F)                                        | 95.460 | 6.104 | 7.947 |
| MNV3+MNV2+SN (1F)                                      | 96.069 | 5.105 | 10.148 |
| MNV3+MNV2+PN (1F)                                      | 95.351 | 7.270 | 10.287 |
| MNV3+PN+SN (3F)                                        | 94.404 | 3.595 | 0.964 |
| *MNV2+PN+SN (3F)                                       | 95.288 | 6.115 | 2.654 |
| *MNV3+MNV2+SN (3F)                                     | 95.834 | 5.117 | 3.391 |
| *MNV3+MNV2+PN (3F)                                     | 95.168 | 7.284 | 3.435 |

(*) indicates more cost-efficient ensemble model than MNV2 single model.
E. EFFICIENCY COMPARISON
Table X shows the MER and the CER efficiency metrics of single models, Table XI shows the efficiency metrics of ensemble models. Among Two-Models (1F) NM ensemble models, MNV3+SN(1F) shows higher IoU with lower MER and CER value than the best single model MNV2. It indicates that MNV3+SN(1F) NM ensemble can produce better accuracy consuming lesser memory and lesser computing power than MNV2 single model. Among Two-Frames (2F) NF ensemble models, MNV3+SN (2F), MNV2+SN (2F), and MNV3+MNV2 (2F) show higher IoU value and lower CER than MNV2. It indicates these three NF ensemble models can produce better accuracy consuming lesser computing power than MNV2 single model. MNV3+SN (2F) NF ensemble model even requires lesser memory and computing power to perform higher accuracy than MNV2. Among Three-Frames (3F) NF ensemble models, MNV2+SN+PN (3F), MNV3+MNV2+SN (3F), and MNV3+MNV2+PN (3F) show higher IoU value and lower CER than MNV2. It indicates that these Three-Frames NF ensemble models can perform higher accuracy consuming lesser computing power than MNV2 single model. The overall result shows that the proposed NF ensemble approach is cost-efficient which can produce higher IoU accuracy consuming lesser computing power than a single model.

Fig. 4 compares the single model MNV2 and ensemble models which have higher IoU and lower CER than MNV2. In the chart, MNV2+SN (2F) NF ensemble model shows the highest IoU value compared to any other model. MNV3+SN (2F) NF ensemble model requires the lowest computing power while performing better IoU accuracy than MNV2.

TABLE XII. AVERAGE OF MER AND CER

| Model                        | IoU (%) | MER   | CER   |
|------------------------------|---------|-------|-------|
| Single model                 | 92.965  | 1.885 | 2.640 |
| Two-Models (1F)              | 95.187  | 3.685 | 5.211 |
| Two-Frames (2F)              | 95.125  | 3.688 | 2.607 |
| Three-Models (1F)            | 95.367  | 5.517 | 7.817 |
| Three-Frames (3F)            | 95.173  | 5.528 | 2.611 |

Fig. 5 compares IoU and FLOPs of single models in table X and representative models in table XI. There are four groups in the chart. Group#1 is the upper right part of the chart, and it shows high accuracy and high use of computing power. MNV3, MNV3+MNV2+SN (1F), and MNV2+SN (1F) models belong to this group. Group#2 is the lower right part of the chart, and it shows high accuracy and low use of computing power. SN, MNV2+SN (2F), MNV3+MNV2+SN (3F), MNV2+PN+SN (3F), and MNV3+SN (2F) models belong to this group. Group#3 is the lower left part of the chart, and it shows a low accuracy and low use of computing power. PN and MNV3 models belong to this group. In general, a model that produces high accuracy using low computing power is desirable, and the models of Group#2 belong to this category. The result shows that proposed NF ensemble models can perform high accuracy using low computing power than other single and NM ensemble models.
Table XII and Fig. 6 show the average of MER and CER of single models, NM ensemble, and NF ensemble models. All ensemble models show higher IoU values than a single model. Two-Frames (2F) NF ensemble and Three-Frames (3F) NF ensemble show lower CER than a single model. It indicates that proposed NF ensemble models can perform better accuracy in a cost-efficient way than single models. However, it is also observed that the requirement of memory for the ensemble is increased as the number of models for the ensemble is increased.

**F. EXAMPLES OF SELFIE SEGMENTATION**

Fig. 7 and 8 show the selfie segmentation result of single models and the proposed NF ensemble model. Column (a) is the original frames of a video; (b) and (c) are the segmentation results using single models; (d) is the result using the proposed NF ensemble; and (e) is ground truth. Fig. 7 shows the segmentation result of various videos. The proposed NF ensemble shows better segmentation result than single models. Fig. 8 shows the segmentation result of continuous frames. The proposed NF ensemble shows stable segmentation result while single models show irregular segmentation results between adjacent frames.
V. DISCUSSION

The experiment results show that proposed ensemble approaches improve accuracy, variance, and bias errors of selfie segmentation in a video than single models. In Two-Models (1F) NM ensemble, four combinations out of six (approximately 66%) show higher accuracy and lower bias errors than the best single model MNV2. It indicates that more than half of Two-Models (1F) cases ensemble model can produce improved segmentation accuracy than single models. For the Three-Models (1F) NM ensemble, three combinations out of four (75%) show higher accuracy, lower variance, and lower bias errors than the best single model MNV2. Comparing the Two-Models (1F) NM ensemble and Three-Models (1F) NM ensemble, MNV2+SN (1F) NM ensemble shows the highest IoU value. However, more percentage of the Three-Models (1F) NM ensemble shows better accuracy than single models. It indicates that Three-Models (1F) NM ensemble has higher reliability for accuracy improvement than Two-Models (1F) NM ensemble. The average accuracy in table IX shows that the Three-Models (1F) NM ensemble gives us the highest accuracy, lowest variance, and lowest bias errors than all experimented ensemble models.
In table VI, the average difference of segmented output masks of adjacent four frames is less than 1%. It is an important observation that the difference of segmented output masks of adjacent video frames is almost ignorable considering the expected performance improvement of segmentation using NF ensemble. Due to this characteristic, proposed NF ensemble models show almost equal accuracy to NM ensemble models. In Two-Frames (2F) NF ensemble, three combinations out of six (50%) showed higher accuracy and lower bias errors than the best single model MNV2.

Overall performance of Two-Frames (2F) NF ensemble shows similar to Two-Models (1F) NM ensemble. The difference of average IoU between the Two-Frames (2F) NF ensemble and Two-Models (1F) NM ensemble is 0.0515%. For the Three-Frames (3F) NF ensemble, three out of four combinations (75%) show improvement in accuracy, variance, and bias errors than the best single model MNV2. The average IoU value of the Three-Frames (3F) NF ensemble is 0.1933% less than Three-Models (1F) NM ensemble. However, it is 2.2081% higher than that of single models. It indicates that the Three-Frames (3F) NF ensemble can perform as accurately as the Three-Models (1F) NM ensemble.

MER and CER are helpful methods to measure the cost efficiency of single and ensemble models. In general, the use of memory and computing power is increased when multiple models are combined to generate ensemble results. However, the result in table XI shows that it is possible to combine multiple models to produce better accuracy than a single model yet having better memory efficiency (MER) and computing power efficiency (CER). Especially, proposed NF ensemble models show almost the same CER value to a single model. It indicates that the proposed NF ensemble model can produce as high accuracy as the NM ensemble model; simultaneously, it uses less computing power as a single model. It is the most significant contribution of this paper.

VI. CONCLUSION

In this paper, we proposed NF ensemble approach that produce better accuracy, lower variance, and lower bias errors than single DL-based selfie segmentation models. Two-Models (1F) and Three-Models (1F) NM ensemble, and Two-Frames (2F) and Three-Frames (3F) NF ensemble models were experimented and compared with the single models. A simple soft voting method was used to combine multiple DL-based selfie segmentation models. Captured video frames of 81 videos collected from TED Talk, Talk Show, News, etc., having a single-person view, were used as a test dataset. Intersection over Union (IoU) for accuracy, IoU standard deviation for variance error, and false prediction rate for bias error were used to measure the performance of single models and proposed ensemble models. Memory Efficiency Rate (MER) and Computing power Efficiency Rate (CER) were used to analyze the cost efficiency of single models and proposed ensemble models. The experiment result shows that the NM ensemble and NF ensemble could produce better accuracy than single models, and the Three-Models combination showed higher accuracy than the Two-Models combination.

The ensemble of multiple DL-based selfie segmentation models improved the performance of segmentation, and it required higher memory and computing power than single models. However, the analysis of efficiency rate showed that some combinations of single models could perform better accuracy than single models yet having better memory efficiency (MER) and computing power efficiency (CER). Proposed NF ensemble models showed almost the same CER value as single models. It indicates that the proposed NF ensemble approach can produce as high accuracy as NM ensemble models, at the same time it uses as less computing power as single models.

REFERENCES

[1] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, “A review of semantic segmentation using deep neural networks,” Int. J. Multimed. Inf. Retr., vol. 7, no. 2, pp. 87–93, 2018, doi: 10.1007/s10462-017-0141-z.

[2] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” 2017, [Online]. Available: http://arxiv.org/abs/1704.04861.

[3] R. M. Haralick and L. G. Shapiro, “Image segmentation techniques,” Comput. Vision, Graph. Image Process., vol. 29, no. 1, pp. 100–132, 1985, doi: 10.1016/S0734-189X(85)90153-7.

[4] S. Minace, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” pp. 1–23, 2020, [Online]. Available: http://arxiv.org/abs/2001.05566.

[5] J. Yang, B. Price, S. Cohen, H. Lee, and M. H. Yang, “Object contour detection with a fully convolutional encoder-decoder network,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 193–202, 2016, doi: 10.1109/CVPR.2016.28.

[6] Liu, X., Deng, Z. & Yang, Y. Recent progress in semantic image segmentation. Artif Intell Rev 52, 1089–1116 (2019). https://doi.org/10.1007/s10462-018-9641-3.

[7] García-García, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martínez, V.; Martínez-González, P.; García-Rodríguez, J. A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput. 2018, 70, 41–65, doi:10.1016/j.asoc.2018.05.018.

[8] Gruosso, M., Capece, N. & Erra, U. Human segmentation in surveillance video with deep learning. Multimed Tools Appl 80, 1175–1199 (2021). https://doi.org/10.1007/s11042-020-09425-0.

[9] A. García-García, S. Orts-Escolano, S. Oprea, V. Villena-Martínez, and J. Garcia-Rodríguez, “A review on deep learning techniques applied to semantic segmentation,” 2017, arXiv:1704.06857. [Online]. Available: http://arxiv.org/abs/1704.06857.

[10] B. Zhao, J. Feng, X. Wu, and S. Yan, “A survey on deep learning-based fine-grained object classification and semantic segmentation,” Int. J. Autom. Comput., vol. 14, no. 2, pp. 119–135, 2017, doi: 10.1007/s11633-017-1053-3.

[11] L. Rokach, “Ensemble-based classifiers,” Artif. Intell. Rev., vol. 33, no. 1–2, pp. 1–39, 2010, doi:10.1007/s10462-009-9124-7.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning.

[13] B. Breiman, “Stacked regressions,” Mach. Learn., vol. 24, no. 1, pp. 49–64, 1996, doi: 10.1023/A:1010684112532.

[14] L. K. Hansen and P. Salamon, “Neural Network Ensembles,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 10, pp. 993–1001, 1990, doi: 10.1109/34.105871.

[15] Y. Zhu and T. Drummond, “Fast Residual Forests: Rapid Ensemble Learning for Semantic Segmentation,” Proc. 1st Annu. Conf. Robot Learn. PMLR 78, no. CoRL, pp. 27–36, 2017, [Online]. Available: http://proceedings.mlr.press/v78/zhu17a.html.
Yong-Woon Kim was born in Seoul, South Korea in 1971. He received the M.S. degree in computer science from Yonsei University, South Korea in 1997. He is an Associate Professor and pursuing Ph.D. degree in computer science at CHRIST (Deemed to be University), Bangalore, India.

Before joining CHRIST (Deemed to be University), he worked with Samsung Electronics Co. Ltd., LG Electronics Co. Ltd., and many global IT companies as a principal embedded software engineer for ten years and as an entrepreneur in India from 2007. He is currently serving as a coordinator of Centre for Digital Innovation at CHRIST (Deemed to be University). He completed research collaborations with the Electronics and Telecommunications Research Institute (ETRI) and other global R&D organizations. His research interests include computer vision, artificial intelligence, and the internet of things.

Yung-Cheol Byun studied at the University of Florida as a Visiting Professor from 2012 to 2014. He directs the Machine Learning Laboratory, Department of Computer Science. Before joining Jeju National University, he worked as a Special Lecturer with Samsung Electronics Company Ltd., in 2000 and 2001. From 2001 to 2003, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI). He was promoted to join Jeju National University as an Assistant Professor in 2003. He is currently serving as a Director for the Information Science Technology Laboratory and other academic societies. He has been hosting an international conference, Computer, Network, Systems, and Industrial Engineering (CNSI), and serving as a program chair, a workshop chair, and a session chair for various international conferences and workshops.

Addapalli V. N. Krishna was born in 1965. He is currently a Professor in Computer Science and Engineering Department at CHRIST (Deemed to be University). He received his PhD in Computer Science and Engineering from Acharya Nagarjuna University, Andhra Pradesh in 2010. He also holds Bachelor of Engineering in Mechanical Engineering, Master of Engineering in Mechanical Engineering and Master of Technology in Computer Science. His area of research is Cryptography & Network Security, Mathematical Modelling and has published around fifty papers. He has also been associated with one MRP, one Monograph funded by CHRIST.

K. Balachandran is working as Professor, in Department of Computer Science and Engineering, Christ University, Bangalore has done his PhD in the Information and Communication Engineering in Anna University, Chennai, Tamil Nadu. He has done his post-graduation in molecular physics, Computer application and in Information Technology. His research includes developing an Adaptive weight-based ensemble model for the Lung cancer pre-diagnosis. His research interests include Artificial Intelligence, Data Science and Computer Networks.