IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 30, 2021, accepted November 19, 2021, date of publication December 6, 2021,
date of current version December 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3133276

Selfie Segmentation in Video Using N-Frames Ensemble

YONG-WOON KIM[1], YUNG-CHEOL BYUN[2], ADDAPALLI V. N. KRISHNA[3], (Member, IEEE), AND BALACHANDRAN KRISHNAN[3], (Senior Member, IEEE)

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

Corresponding authors: Yong-Woon Kim (jonathan.kim@christuniversity.in) and Yung-Cheol Byun (ycb@jejunu.ac.kr)

This work was supported in part by the Ministry of Education and National Research Foundation of Korea through the "Leaders in Industry-University Cooperation +" Project, and in part by the South Korean Ministry of Trade, Industry and Energy (MOTIE) through the Korea Institute for Advancement of Technology (KIAT) Grant (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. Unlike the N-Models (NM) ensemble which executes multiple DSMs at once for every single video frame, the proposed NF ensemble executes only one DSM 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 to make an ensemble. We evaluated the proposed approach using 81 videos dataset with a single-person view collected from publicly available websites. To measure the performance of segmentation models, Intersection over Union (IoU), IoU standard deviation, false prediction rate, Memory Efficiency Rate and Computing power Efficiency Rate parameters were considered. The average IoU values of the Two-Models NM ensemble, Two-Frames NF ensemble, Three-Models NM ensemble and Three-Frames NF ensemble were 95.1868%, 95.1253%, 95.3667% and 95.1734% each, whereas the average IoU value of single models was 92.9653%. The result shows that the proposed NF ensemble 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.

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

I. INTRODUCTION

Self-portrait photographs (selfies) have become very popular among mobile phone users. 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. Some of the prominent image segmentation methods have been studied and developed by researchers. 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]. Deep-Learning (DL) based Segmentation Model (DSM) has

The associate editor coordinating the review of this manuscript and approving it for publication was Charalambos Poullis.

opened a new era of image segmentation. DSMs have made remarkable improvements in speed and accuracy for image segmentation compared to traditional methods [4]–[9]. These models predict segmentation regions using semantic labels for every pixel of the image [10].

In general, an 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, was necessitated due to the increased use of video applications such as online video conferencing solution. Selfie segmentation in a video is a subset of VIS. 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. CONTRIBUTIONS

To address these issues, we propose 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. The average Float Point Operations (FLOPs) of single DSMs and proposed model is equally 2.4896 Giga FLOPs.
- 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.

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.
- This paper compares **the proposed N-Frames ensemble with four state-of-the-art DSMs and the ensemble of multiple DSMs using various measurement metrics.** Intersection over Union and false prediction rate are used to measure accuracy, variance error and bias error. The efficiency rate of computing power and memory usage are measured to evaluate the efficiency of segmentation.
- We construct **a new video dataset to evaluate the performance of selfie segmentation in videos.**

C. 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 IV 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 WORK

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. Bolya *et al.* [25] proposed YOLAC++ framework based on their previous architecture YOLAC. It achieved real-time instance segmentation in one-stage process on GPU hardware. Singh *et al.* [26] and 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 is an efficient and lightweight segmentation model for mobiles and embedded systems. 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. 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. 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. 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. Mehta *et al.* [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 light-weight architecture for semantic segmentation that can be easily deployed on edge devices. 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. However, these works mainly aim to segment objects from a single image but does not perform well when it is applied to selfie segmentation in a video.

B. IMAGE SEGMENTATION USING ENSEMBLE

Warfield *et al.* introduced an algorithm to combine multiple segmentations and validated 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. However, these ensemble approaches require more computing power than single DSMs. For the selfie segmentation in a video, the segmentation speed is very important.

C. IMAGE SEGMENTATION IN VIDEO

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. 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. 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. 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 ''Temporal consistent portrait video segmentation'' proposed by Wang *et al.* [48]. They have achieved high accuracy temporal-coherent segmentation result using a soft correspondence network. 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. 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.

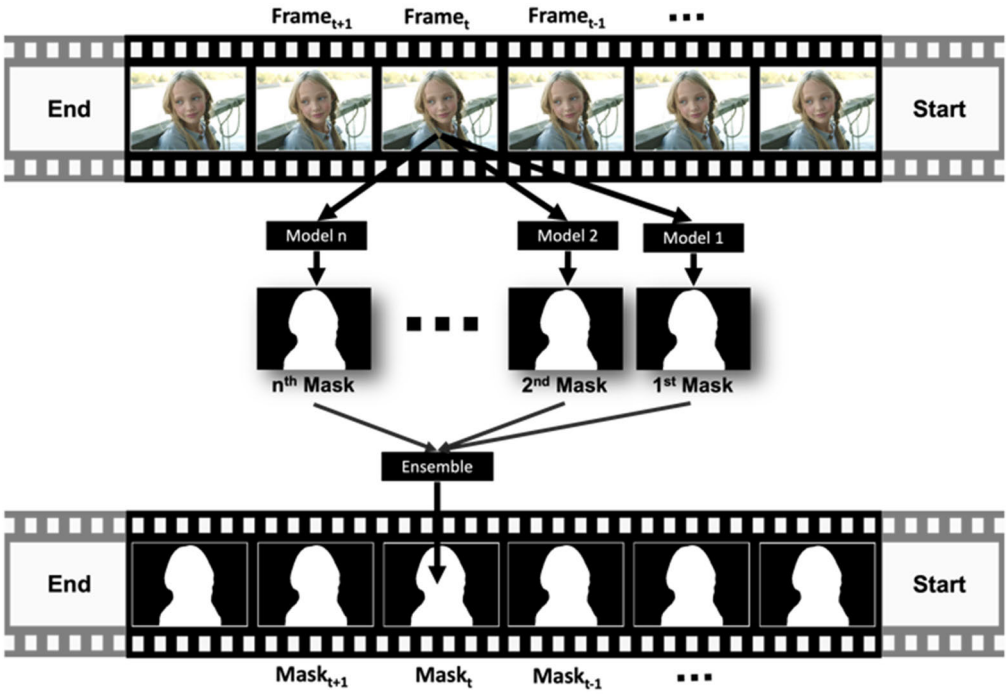

FIGURE 1. Conceptual diagram for selfie segmentation in video using NM ensemble approach.

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. THE PROPOSED APPROACH

This section describes the proposed approach 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 a Frame$_t$, and the Frame$_t$ 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$_t$, as shown in Fig. 1.

This paper proposes a novel ensemble approach for selfie segmentation in the video. We call it an N-Frames (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 a frame t, and the frame t is fed to a segmentation model n to generate a segmented nth 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 t. 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 the frame t at a given time t. The NF ensemble uses model 1 for the frame t + 1 at a given time t + 1. It combines the segmented output mask of model 1 with previous output masks generated from segmentation model n, . . . , model 2 to create the final output mask t + 1.

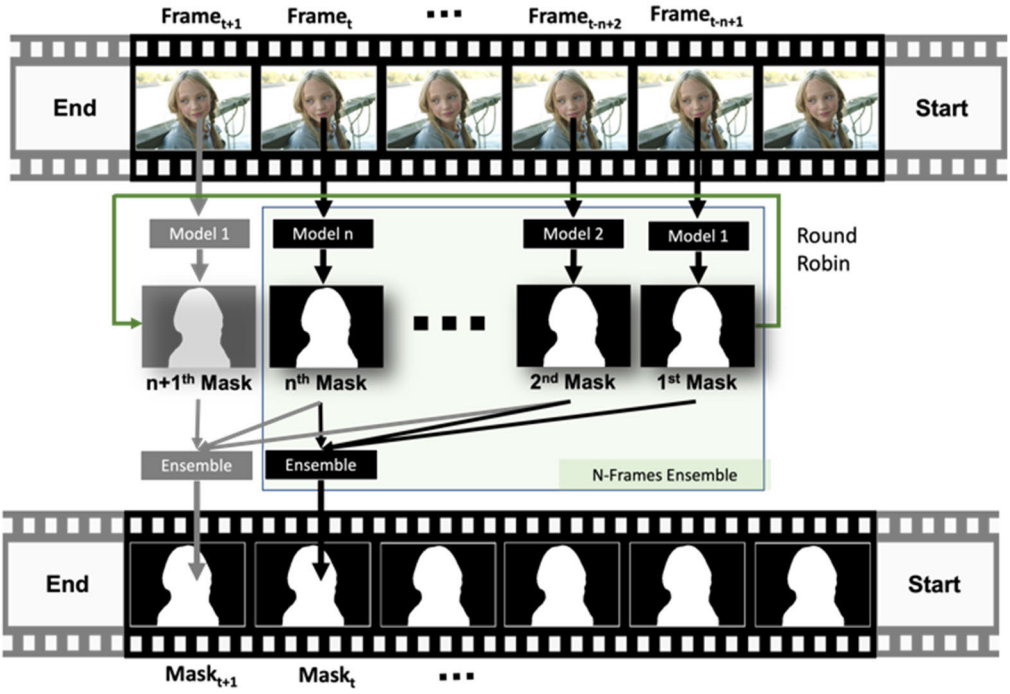

FIGURE 2. Conceptual diagram for selfie segmentation in video using NF ensemble approach.

The benefit of this approach is that only one segmentation model needs to make a segmentation mask for a current frame t 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. Section IV-F discusses the computing power efficiency.
- 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. Section IV-E and IV-F compare the proposed NF ensemble with NM ensemble.
- 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. In Section IV-G, the qualitative results are presented.
- 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. Section IV-G presents memory efficiency.
- 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. As the proposed NF ensemble merges the segmentation results of multiple adjacent video frames, the high difference of adjacent video frames will reduce the accuracy of segmentation result.

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 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 [51]:

$$H^j(x_t) = \frac{1}{T}\sum_{i=1}^{T} h^j_{r(t,i,T)}(x_{t-i+1}) \tag{1}$$

where |T| ≥ |r|, |T| ≥ |r(t, i, T)|, T > 1 and t ≥ T at given time t. In a set of **T** individual classifiers $\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T\}$, the classifier $\boldsymbol{h}_i$ generates a $\boldsymbol{l}$-dimensional vector $(\boldsymbol{h}^1(\boldsymbol{x}_t), \ldots, \boldsymbol{h}^l(\boldsymbol{x}_t))^T$ for the instance of $\boldsymbol{x}_t$ at given time t, where $\boldsymbol{h}$ indicates the classifiers and $\boldsymbol{h}^j(\boldsymbol{x}) \in [\boldsymbol{0}, \boldsymbol{1}]$.

A round-robin method can use modulo operation to get an index number. The modulo operation can be presented by the following equation (2) [54]:

$$r(a, n) = a - n\left\lfloor \frac{a}{n} \right\rfloor \tag{2}$$

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

$$r(t, i, T) = t - i + 1 - T\left\lfloor \frac{t-i}{T} \right\rfloor \tag{3}$$

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 [33]:

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

where A is an image segmentation result, B is a ground truth image, A ∩ B is the intersection of A and B, and A ∪ 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 [55], [56]:

$$FNR = \frac{FN}{FN + TP} \tag{5}$$

$$FDR = \frac{FP}{FP + TP} \tag{6}$$

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) [42].

$$MER = \frac{M}{IoU} \tag{7}$$

where M is the number of parameters and 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 [42]:

$$CER = \frac{C}{IoU} \tag{8}$$

where C is Floating-Point Operations (FLOPs), and IoU is the accuracy of a given model.

IV. EXPERIMENT

We used 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. IoU is used for accuracy measurement, IoU standard deviation for variance error measurement, and FNR + FDR for bias error measurement. |FNR-FDR| means the absolute value of FNR-FDR, which measures the difference between FNR and FDR. MER and CER are also calculated to measure the efficiency of single models and proposed ensemble models.

A. 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 [57]–[62]. 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 [62]. These DSMs were verified by measuring the accuracy and comparing with other papers [36], [63]. We have used the EG1800 + CDI test dataset [42] for the verification, where EG1800 [64] 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 128 × 128. MNV3 uses MobileNetV3 with a dept multiplier 0.5 as an encoder and an upsampling block with a transpose convolution as a decoder. It has 224 × 224 input/output resolution. PN is a pre-trained model of PortraitNet [33]. It has 224 × 224 input/output resolution. SN is a pre-trained model of SINet [36]. It has 320 × 320 input/output resolution. Table 1 shows the backbone of each model and the accuracy comparison using the verification dataset EG1800 + CDI.

We conducted the experiment with these four pre-trained selfie DSMs to segment the body of a human from video dataset. To the best of our knowledge, available video datasets are not dedicated to selfie segmentation purpose. For the experiment, we collected test video dataset from publicly available websites such as News, Talk Show, Interview, etc. The dataset is a collection of captured frames from 81 videos showing a single person upper body. The resolution of videos is 480 × 360 pixels, the duration is 10 seconds ∼20 seconds,

TABLE 1. Experimented DSMs.

Model	Backbone	IoU (%) of ref.	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].)

TABLE 2. Single models without ensemble (%).

	IoU	IoU Std.	FNR+FDR	|FNR-FDR|
MNV3	92.3115	2.5137	7.9439	4.2103
MNV2	95.1166	1.7565	4.9927	0.1648
PN	90.2161	2.3212	10.2217	0.5526
SN	94.2172	1.8685	5.9284	1.5882

TABLE 3. Two-models NM ensemble (1-frame) (%).

	IoU	IoU Std	FNR+FDR	|FNR-FDR|
MNV3+MNV2(1F)	95.5236	1.7644	4.5775	0.0610
MNV3+PN(1F)	94.1335	2.2678	6.0949	0.7960
MNV3+SN(1F)	95.7374	1.9854	4.3674	0.6074
MNV2+PN(1F)	95.1321	1.8627	4.9532	2.9590
MNV2+SN(1F)	96.5587	1.5433	3.4909	1.1258
PN+SN(1F)	94.0351	1.9633	6.1288	1.2653

TABLE 4. Three-models NM ensemble (1-frame) (%).

	IoU	IoU Std	FNR+FDR	|FNR-FDR|
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

and the frame rate of videos is in the range of 25 to 30 frames per second (FPS). The dataset was annotated using various tools and the annotated result was manually verified. Finally, the test dataset consists of around 40,000 frame images and ground truth masks. 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.

B. SINGLE MODEL RESULT

Table 2 shows the experimental results of four single models MNV3, MNV2, PN, and SN. In the table, MNV2 produces the highest IoU value, the lowest IoU standard deviation, the lowest FNR + FDR, and the lowest |FNR-FDR|. It indicates that MNV2 is the most accurate in selfie segmentation among four single models. The lowest IoU standard deviation indicates that the variance error is lowest, and the lowest FNR + FDR indicates that the bias error is also the lowest. The lowest |FNR-FDR| means MNV2 is well balanced in false positive regions and false negative regions. MNV3 shows the highest |FNR-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.

C. 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 3 shows the result of the Two-Models NM ensemble. In the table, MNV2 + SN produces the highest 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 IoU value than MNV2 or PN single model, but IoU standard deviation and |FNR-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 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 4 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 94.5861% and 96.0695%, and the difference of maximum and minimum is 1.4834%. In the Two-Models NM ensemble in Table 3, the range of IoU values are 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 1.5433% 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 1.7661%, 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

TABLE 5. Difference of maximum and minimum values (%).

	IoU	IoU Std	FNR+FDR	| FNR-FDR |
Single models	4.9005	0.7572	5.229	4.0455
Two-Models NM	2.5236	0.7245	2.6379	2.898
Three-Models NM	1.4834	0.2216	1.5489	1.2893

TABLE 6. Difference of n-frames (%).

2-frames	3-frames	4-frames
0.3423	0.6433	0.9234

NM ensemble. Table 5 shows the differences of maximum and minimum values of IoU, IoU standard deviation, FNR + FDR, and |FNR-FDR| values. Ensemble models show the lesser difference of maximum and minimum than that of single models, and Three-Models NM ensemble shows the lowest value in all aspects.

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

In this paper, we propose a novel ensemble approach called an N-Frames (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 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.

$$D(n) = \left| \bigcup_{i=0}^{n-1} G_{t-i} - \bigcap_{i=0}^{n-1} G_{t-i} \right| \tag{9}$$

where G_t is a ground truth segmentation mask of a Frame$_t$ at a given time t, n > 1 and t ≥ n. This equation can be expressed using bit operation such as $D(n) = (G_t$ **OR** G_{t-1} **OR** $\ldots$ $G_{t-n+1})$ **XOR** $(G_t$ **AND** G_{t-1} **AND** $\ldots$ $G_{t-n+1})$. Table 6 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%.

To evaluate the proposed NF ensemble, Two-Frames (2F) and Three-Frames (3F) NF ensemble were experimented. In Table 7, 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

TABLE 7. Two-frames NF ensemble (%).

	IoU	IoU Std	FNR+FDR	|FNR-FDR|
MNV3+MNV2(2F)	95.4631	1.7925	4.6408	0.0431
MNV3+PN(2F)	94.0867	2.2975	6.1461	0.8253
MNV3+SN(2F)	95.6587	2.0254	4.4499	0.5870
MNV2+PN(2F)	95.0812	1.9039	5.0066	2.9803
MNV2+SN(2F)	96.4729	1.5949	3.5799	1.1401
PN+SN(2F)	93.9890	2.0166	6.1770	1.2846

TABLE 8. 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

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 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 8 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 6.

E. RESULT ANALYSIS

Table 9 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 6. Three-Models (1F) NM ensemble and Three-Frames (3F) NF ensemble 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

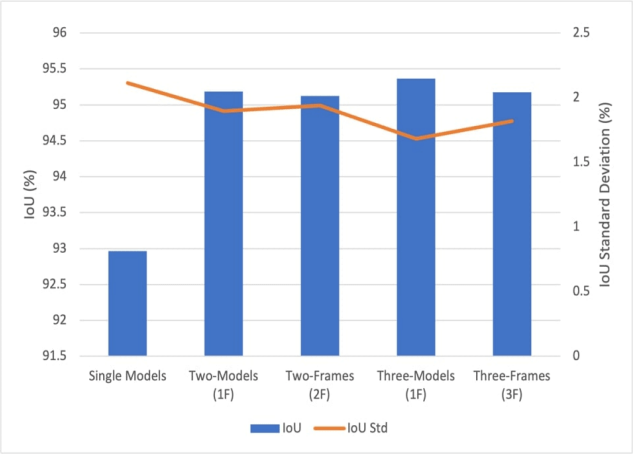

FIGURE 3. Average IoU and IoU standard deviation.

TABLE 9. 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

0.1933%, and it is much less than the difference of 3-frames in Table 6.

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.

F. EFFICIENCY COMPARISON

Table 10 shows the MER and the CER efficiency metrics of single models, Table 11 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

TABLE 10. 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 11. 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)

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.

Fig. 5 compares IoU and FLOPs of single models in table 10 and representative models in table 11. 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

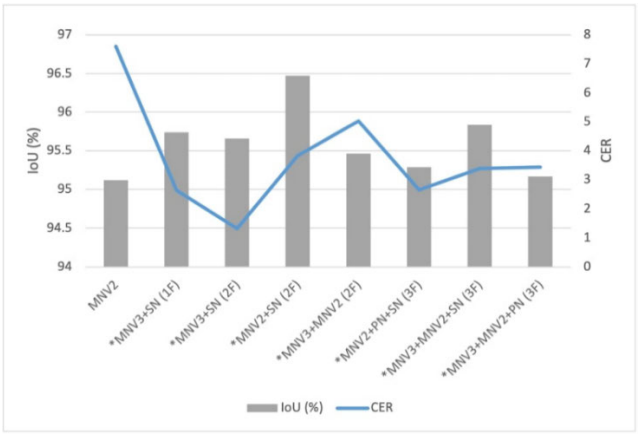

FIGURE 4. Comparison of MNV2 and ensemble models.

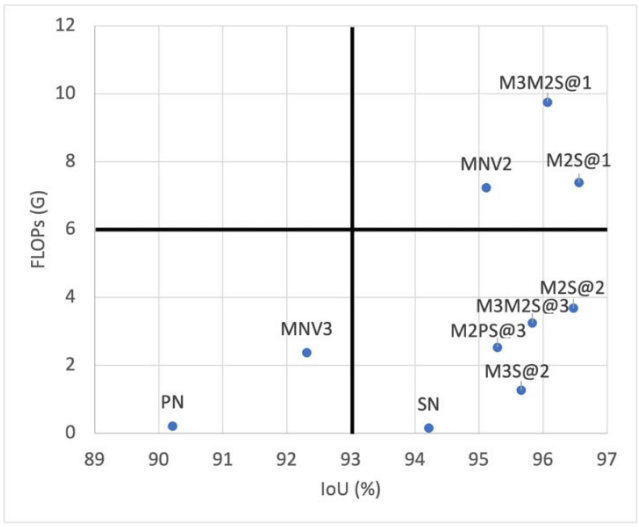

FIGURE 5. Comparison of IoU and FLOPs; M3M2S@1 = MNV3 + MNV2 + SN (1F); M2S@1 = MNV2 + SN (1F); M2S@2 = MNV2 + SN (2F); M3M2S@3 = MNV3 + MNV2 + SN (3F); M2PS@3 = MNV2 + PN + SN (3F); M3S@2 = MNV3 + SN (2F).

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.

Fig. 6 and Table 12 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.

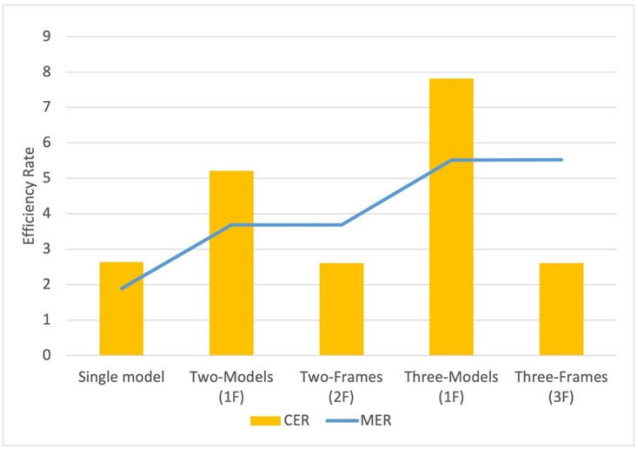

FIGURE 6. Average of CER and MER.

TABLE 12. 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

G. 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 Tow-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

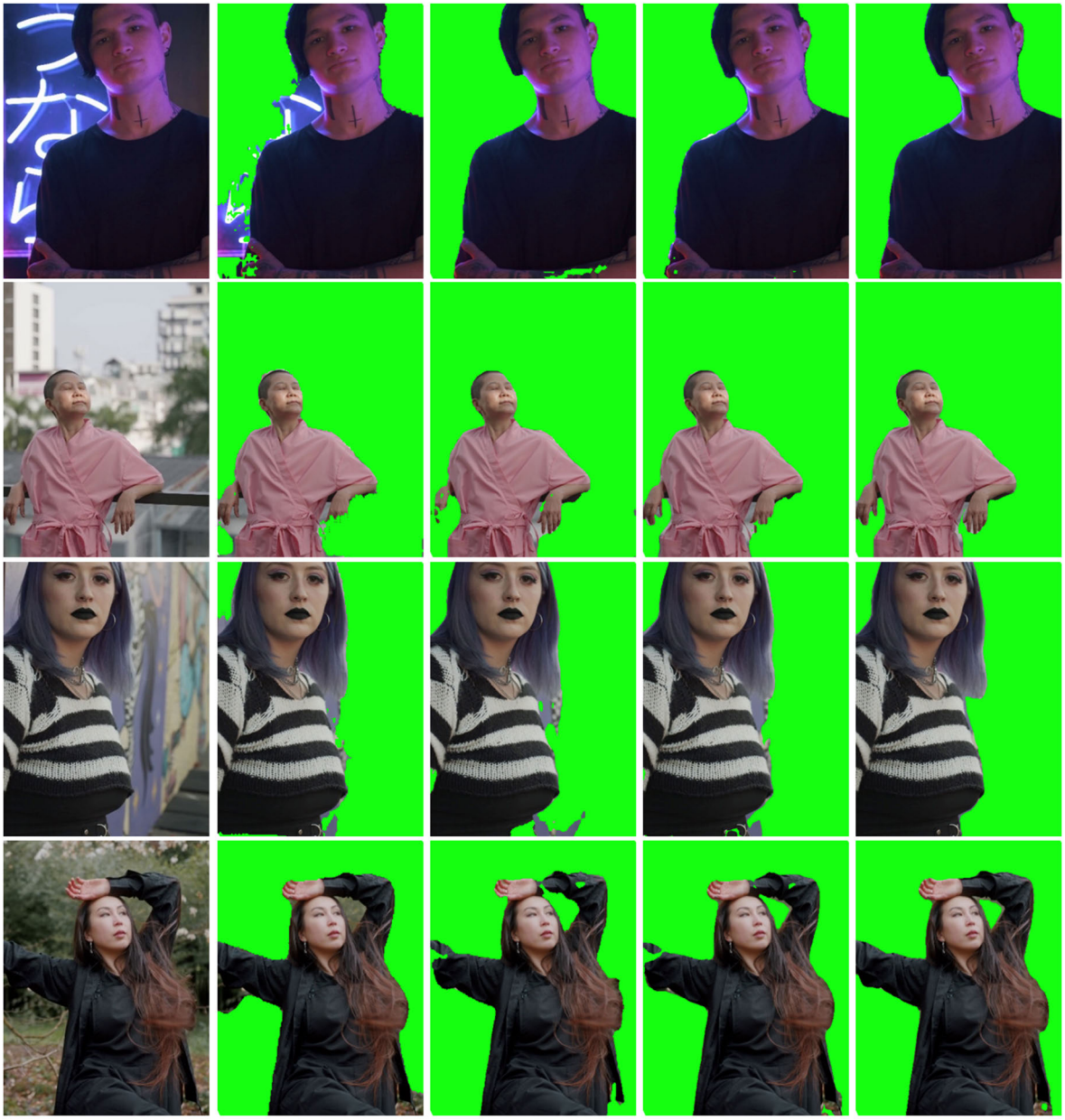

FIGURE 7. Selfie segmentation example of various videos; (a) original frame; (b) DSM1; (c) DSM2; (d) NF ensemble; and (e) ground truth.

Three-Models (1F) NM ensemble has higher reliability for accuracy improvement than Two-Models (1F) NM ensemble. The average accuracy in table 9 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 6, 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.

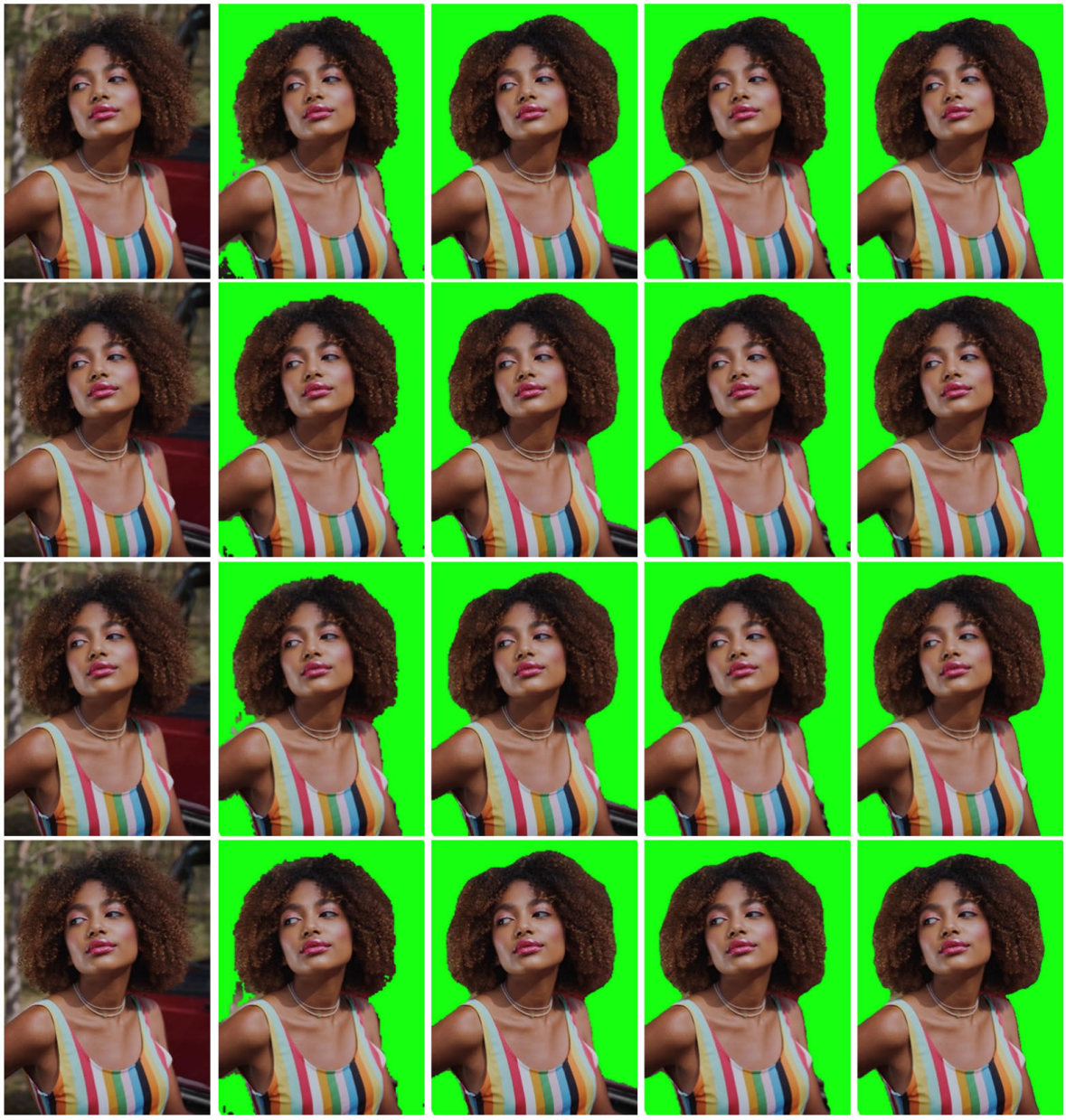

FIGURE 8. Selfie segmentation example of continuous frames; (a) original frame; (b) MNV2; (c) PN; (d) NF ensemble; and (e) ground truth.

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 11 shows that it is possible to combine multiple models to produce better accuracy than a single model yet having better memory efficiency and computing power efficiency. 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 Talk Show, News, Interview, 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.

The present work has several limitations. The proposed NF ensemble for selfie segmentation is limited to a single person segmentation in a video and it is not suitable for rapidly moving object. It is suitable for a video showing less movement of human body such like Talk Show, Interview, News, etc. The future work is to apply the proposed method to other objects in a video rather than a human body.

REFERENCES

[1] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimedia Inf. Retr.*, vol. 7, no. 2, pp. 87–93, Jun. 2018, doi: 10.1007/s13735-017-0141-z.

[2] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 17, 2021, doi: 10.1109/TPAMI.2021.3059968.

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

[4] I. Ahmed, M. Ahmad, F. A. Khan, and M. Asif, "Comparison of deep-learning-based segmentation models: Using top view person images," *IEEE Access*, vol. 8, pp. 136361–136373, 2020, doi: 10.1109/ACCESS.2020.3011406.

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

[6] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 1089–1106, Aug. 2019.

[7] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, Sep. 2018.

[8] M. Gruosso, N. Capece, and U. Erra, "Human segmentation in surveillance video with deep learning," *Multimedia Tools Appl.*, vol. 80, no. 1, pp. 1175–1199, Jan. 2021.

[9] J. Zhang, X. Zhao, Z. Chen, and Z. Lu, "A review of deep learning-based semantic segmentation for point cloud," *IEEE Access*, vol. 7, pp. 179118–179133, 2019, doi: 10.1109/ACCESS.2019.2958671.

[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. Automat. Comput.*, vol. 14, no. 2, pp. 119–135, Apr. 2017, doi: 10.1007/s11633-017-1053-3.

[11] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, nos. 1–2, pp. 1–39, Feb. 2010, doi: 10.1007/s10462-009-9124-7.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[13] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, Jul. 1996, doi: 10.1023/A:1018046112532.

[14] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct. 1990, doi: 10.1109/34.58871.

[15] Y. Zuo and T. Drummond, "Fast residual forests: Rapid ensemble learning for semantic segmentation," *Proc. 1st Annu. Conf. Robot Learn.*, vol. 78, 2017, pp. 27–36. [Online]. Available: http://proceedings.mlr.press/v78/zuo17a.html

[16] Y. Koren, "The BellKor solution to the Netflix grand prize," *Netflix Prize Documentation*, vol. 81, no. 2009, pp. 1–10, 2009.

[17] J. Luiten, P. Torr, and B. Leibe, "Video instance segmentation 2019: A winning approach for combined detection, segmentation, classification and tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 709–712, doi: 10.1109/ICCVW.2019.00088.

[18] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," in *Computer Vision—ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 92–107.

[19] X. Du, X. Wang, D. Li, J. Zhu, S. Tasci, C. Upright, S. Walsh, and L. Davis, "Boundary-sensitive network for portrait segmentation," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8, doi: 10.1109/FG.2019.8756516.

[20] B. Lyu, Y. Yang, S. Wen, T. Huang, and K. Li, "Neural architecture search for portrait parsing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 19, 2021, doi: 10.1109/TNNLS.2021.3104872.

[21] J. Miao, K. Sun, X. Liao, L. Leng, and J. Chu, "Human segmentation based on compressed deep convolutional neural network," *IEEE Access*, vol. 8, pp. 167585–167595, 2020.

[22] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6819–6829.

[23] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 93–102, 2016, doi: 10.1111/cgf.12814.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++: Better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 5, 2020, doi: 10.1109/TPAMI.2020.3014297.

[26] D. Singh, V. Kumar, and M. Kaur, "Densely connected convolutional networks-based COVID-19 screening model," *Appl. Intell.*, vol. 51, no. 5, pp. 3044–3051, 2021.

[27] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision—ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 833–851.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[31] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[32] L. Zhu, T. Wang, E. Aksu, and J.-K. Kamarainen, "Portrait instance segmentation for mobile devices," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1630–1635.

[33] S.-H. Zhang, X. Dong, H. Li, R. Li, and Y.-L. Yang, "PortraitNet: Real-time portrait segmentation network for mobile device," *Comput. Graph.*, vol. 80, pp. 104–113, May 2019.

[34] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Computer Vision—ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 561–580.

[35] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9190–9200.

[36] H. Park, L. L. Sjosund, Y. Yoo, N. Monet, J. Bang, and N. Kwak, "SINet: Extreme lightweight portrait segmentation networks with spatial squeeze modules and information blocking decoder," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2066–2074.

[37] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[38] T. Rohlfing and C. R. Maurer, Jr., "Shape-based averaging for combination of multiple segmentations," *Med. Image Comput. Comput.-Assist. Intervent.*, vol. 8, no. Pt 2, pp. 838–845, 2005.

[39] A. Holliday, M. Barekatain, J. Laurmaa, C. Kandaswamy, and H. Prendinger, "Speedup of deep learning ensembles for semantic segmentation using a model compression technique," *Comput. Vis. Image Understand.*, vol. 164, pp. 16–26, Nov. 2017, doi: 10.1016/j.cviu.2017.05.004.

[40] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNNs," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, Jun. 2016, doi: 10.5194/isprsannals-iii-3-473-2016.

[41] Y. W. Kim, J. Innila Rose, and A. V. N. Krishna, "Accuracy enhancement of portrait segmentation by ensembling deep learning models," in *Proc. 5th Int. Conf. Res. Comput. Intell. Commun. Netw. (ICRCICN)*, Nov. 2020, pp. 59–64, doi: 10.1109/ICRCICN50933.2020.9296196.

[42] Y.-W. Kim, Y.-C. Byun, and A. V. N. Krishna, "Portrait segmentation using ensemble of heterogeneous deep-learning models," *Entropy*, vol. 23, no. 2, p. 197, Feb. 2021.

[43] J. Li, S. He, H.-C. Wong, and S.-L. Lo, "Proposal-driven segmentation for videos," *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1098–1102, Aug. 2019, doi: 10.1109/LSP.2019.2921654.

[44] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *Computer Vision—ECCV 2020*. Cham, Switzerland: Springer, 2020, pp. 352–368.

[45] M. Ding, Z. Wang, B. Zhou, J. Shi, Z. Lu, and P. Luo, "Every frame counts: Joint learning of video segmentation and optical flow," in *Proc. Conf. AAAI Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10713–10720.

[46] S. Lin, F. Qin, H. Peng, R. A. Bly, K. S. Moe, and B. Hannaford, "Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6773–6780, Oct. 2021.

[47] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2663–2672.

[48] Y. Wang, W. Zhang, L. Wang, F. Yang, and H. Lu, "Temporal consistent portrait video segmentation," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108143.

[49] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.

[50] T. Zhang, C. Lang, and J. Xing, "Realtime human segmentation in video," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2019, pp. 206–217.

[51] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Caithness, U.K.: Whittles Publishing, 2012.

[52] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992, doi: 10.1016/S0893-6080(05)80023-1.

[53] P. Smyth and D. Wolpert, "Linearly combining density estimators via stacking," *Mach. Learn.*, vol. 36, no. 1, pp. 59–83, Jul. 1999, doi: 10.1023/a:1007511322260.

[54] D. E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, 2nd ed. London, U.K.: Addison-Wesley, 1973.

[55] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. B, Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[56] Wikipedia Contributors. (Oct. 18, 2021). *Precision and Recall, Wikipedia, the Free Encyclopedia*. Accessed: Dec. 1, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=1050491609

[57] *Selfie Segmentation*. Accessed: Oct. 8, 2021. [Online]. Available: https://google.github.io/mediapipe/solutions/selfie_segmentation.html

[58] *PortraitNet*. Accessed: Oct. 8, 2021. [Online]. Available: https://github.com/dong-x16/PortraitNet

[59] *Mobile_Phone_Human_Matting*. Accessed: Oct. 8, 2021. [Online]. Available: https://github.com/lizhengwei1992/mobile_phone_human_matting

[60] *TensorflowLite-UNet*. Accessed: Oct. 8, 2021. [Online]. Available: https://github.com/PINTO0309/TensorflowLite-UNet

[61] *Portrait-Segmentation*. Accessed: Oct. 8, 2021. [Online]. Available: https://github.com/anilsathyan7/Portrait-Segmentation

[62] *SelfieSeg*. Accessed: Oct. 8, 2021. [Online]. Available: https://github.com/Innovation4x/SelfieSeg

[63] B. Lyu, Y. Yang, S. Wen, T. Huang, and K. Li, "Neural architecture search for portrait parsing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 19, 2021, doi: 10.1109/TNNLS.2021.3104872.

[64] *EG1800.zip_Free*. Accessed: Oct. 21, 2021. [Online]. Available: https://pan.baidu.com/s/1myEB-dEmGz6ufniU3i1e6Uw

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 currently pursuing the Ph.D. degree in computer science with CHRIST University (Deemed to be University), Bengaluru, India.

He is an Associate Professor at CHRIST University. Before joining CHRIST University, he worked with Samsung Electronics Company Ltd., LG Electronics Company Ltd., and many global IT companies as a Principal Embedded Software Engineer for ten years and has been an entrepreneur in India, since 2007. He is currently serving as a Coordinator for the Centre for Digital Innovation, CHRIST University. He completed research collaborations with the Electronics and Telecommunications Research Institute (ETRI) and other global research and development 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, where he directs the Machine Learning Laboratory, Department of Computer Science. He worked as a Special Lecturer with Samsung Electronics Company Ltd., from 2000 to 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 the Director for the Information Science Technology Laboratory and other academic societies. He has been hosting an international conference, Computers, Networks, Systems and Industrial Engineering (CNSI), and serving as the program chair, the workshop chair, and the session chair for various international conferences and workshops.

ADDAPALLI V. N. KRISHNA (Member, IEEE) was born in 1965. He received the Bachelor of Engineering degree in mechanical engineering, the Master of Engineering degree in mechanical engineering, and the Master of Technology degree in computer science. He also received the Ph.D. degree in computer science and engineering from Acharya Nagarjuna University, Andhra Pradesh, in 2010.

He is currently a Professor with the Computer Science and Engineering Department, CHRIST University (Deemed to be University). He has also been associated with one MRP and one monograph funded by CHRIST University. He has published around 50 papers. His research interests include cryptography and networks security, and mathematical modeling.

BALACHANDRAN KRISHNAN (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Anna University, Chennai, Tamil Nadu, and the post-graduate degree in molecular physics, computer application, and information technology. He is working as a Professor with the Department of Computer Science and Engineering, CHRIST University, Bengaluru. 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.