*Computers, Materials & Continua*
DOI: 10.32604/cmc.2022.028632
***Article***

Tech Science Press

# High-Movement Human Segmentation in Video Using Adaptive N-Frames Ensemble

**Yong-Woon Kim[1], Yung-Cheol Byun[2,*], Dong Seog Han[3], Dalia Dominic[1] and Sibu Cyriac[1]**

[1]Centre for Digital Innovation, CHRIST (Deemed to be University), Bangalore, 560029, India
[2]Department of Computer Engineering, Jeju National University, Jeju, 63243, Korea
[3]School of Electronics Engineering, Kyungpook National University, Daegu, 41566, Korea
*Corresponding Author: Yung-Cheol Byun. Email: ycb@jejunu.ac.kr
Received: 14 February 2022; Accepted: 08 May 2022

**Abstract:** A wide range of camera apps and online video conferencing services support the feature of changing the background in real-time for aesthetic, privacy, and security reasons. Numerous studies show that the Deep-Learning (DL) is a suitable option for human segmentation, and the ensemble of multiple DL-based segmentation models can improve the segmentation result. However, these approaches are not as effective when directly applied to the image segmentation in a video. This paper proposes an Adaptive N-Frames Ensemble (AFE) approach for high-movement human segmentation in a video using an ensemble of multiple DL models. In contrast to an ensemble, which executes multiple DL models simultaneously for every single video frame, the proposed AFE approach executes only a single DL model upon a current video frame. It combines the segmentation outputs of previous frames for the final segmentation output when the frame difference is less than a particular threshold. Our method employs the idea of the N-Frames Ensemble (NFE) method, which uses the ensemble of the image segmentation of a current video frame and previous video frames. However, NFE is not suitable for the segmentation of fast-moving objects in a video nor a video with low frame rates. The proposed AFE approach addresses the limitations of the NFE method. Our experiment uses three human segmentation models, namely Fully Convolutional Network (FCN), DeepLabv3, and Mediapipe. We evaluated our approach using 1711 videos of the TikTok50f dataset with a single-person view. The TikTok50f dataset is a reconstructed version of the publicly available TikTok dataset by cropping, resizing and dividing it into videos having 50 frames each. This paper compares the proposed AFE with single models and the Two-Models Ensemble, as well as the NFE models. The experiment results show that the proposed AFE is suitable for low-movement as well as high-movement human segmentation in a video.

**Keywords:** High movement; human segmentation; artificial intelligence; deep learning; ensemble; video instance segmentation

## 1 Introduction

Video segmentation techniques have become more attractive to a variety of real-time applications. Their objective is to segment the target objects in a set of given video frames. Several online video conferencing services support the feature of changing the background in real-time for various reasons. These solutions are increasingly used by many organizations as many people work from home because of the COVID-19 pandemic. Real-time human segmentation is a key technology for enabling these kinds of applications. There is a wide range of applications using human segmentation. Camera apps are one of the most popular apps on mobile phones, and many users decorate their selfies using the human segmentation function. Content-based image retrieval system requires human segmentation when it stores human portrait images and retrieves a specific human from the stored human image database. The human segmentation enables the system to save storage by removing the background from human portrait images. It also can be used in image editing and video editing program. Users can extract human areas from a photo or a video automatically using the human segmentation function. Augmented Reality (AR), Virtual Reality (VR) and Metaverse are the applications where human segmentation can play an important role in displaying a human body in virtual space.

Human segmentation is a branch of image segmentation. There are various techniques for segmenting images that researchers have studied and developed. Many traditional image segmentation methods, including thresholding, watersheds, clustering with contours and edges, region growing, Markov random fields and graph cuts, were used [1,2]. For high-quality image segmentation, homogeneity and uniformity of the segmented area are important. However, this is not an easy task [3]. Deep-Learning (DL) based Segmentation Model (DSM) has opened up new possibilities for image segmentation since they have achieved significant improvements in both speed and precision over traditional methods [4–8]. The DL-based models use semantic labels to predict segmentation regions for every image pixel [9]. The results of several studies show that an ensemble of multiple models can help to enhance image segmentation precision. The key to the ensemble approach is combining different models to produce an effective model [10]. Collective decisions made by the ensemble help to generate less errors and make the prediction more accurate [11,12]. According to many studies, the blending of multiple segmentation models shows better performance than single segmentation models [13–15].

Recently, the image segmentation was extended to the Video Instance Segmentation (VIS) domain. The VIS technique classifies objects into predefined classes, tracks objects within a video, segments the classified objects, and classifies with localization throughout a video. There are a number of studies on image segmentation using a single DSM for a static image. These works, however, focus on single images. As a result, they are not well suited for VIS since video may contain characteristics that include sudden noises or interrelationships between consecutive video frames. Multiple studies have shown that the ensemble of multiple DSMs provides enhanced segmentation results compared to single models. Despite this, the ensemble method is not satisfying when it comes to VIS since segmentation speed is slower than a single model. To overcome the speed issue of the ensemble approach, the method called N-Frames Ensemble (NFE) uses the ensemble method to combine the segmented outputs of the present DSM with the segmented outputs from previous DSMs [16]. This approach has the advantage that a single DSM is required to produce a segmented output for a specific video frame. As a result, the NFE is as fast as a single DSM. However, NFE is not suitable for segmenting fast-moving objects. The high movement of objects in a video causes a high difference between adjacent frames, and it reduces the segmentation accuracy of the NFE since NFE fuses segmented results of adjacent video frames. To tackle these issues, this work proposes a novel ensemble approach called Adaptive N-Frames Ensemble (AFE) method by combining multiple segmentation models based on a decision

made by frame difference threshold. Fig. 1 shows the overview of the proposed AFE approach. The proposed approach has the following characteristics:

- It has a processing speed of a single DSM.
- It can improve segmentation performance using the ensemble of multiple DSMs.
- It can handle sudden video noises.
- It can be used for segmenting slow-moving objects as well as fast-moving objects in a video.
- It is suitable for high Frame Per Second (FPS) as well as low FPS videos

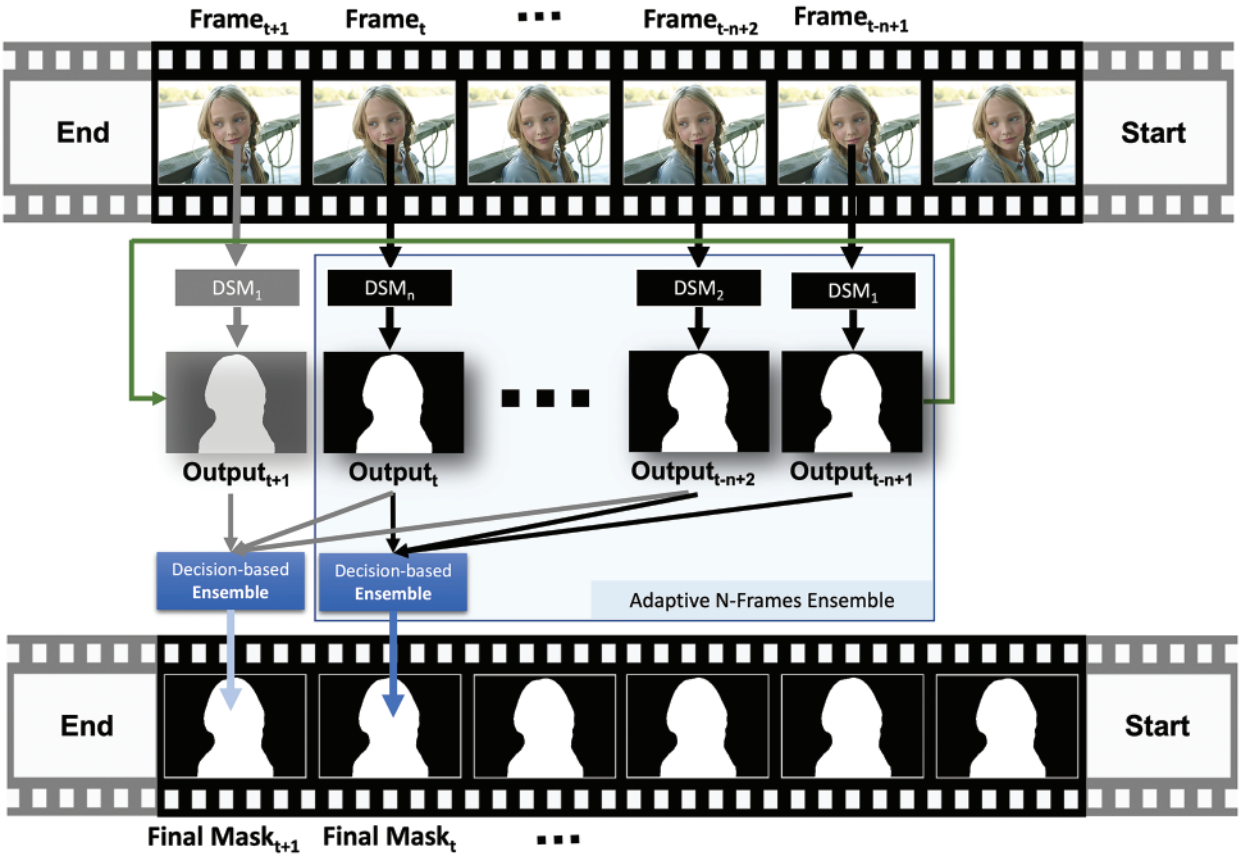


**Figure 1:** Overview of AFE approach

This work has the following major contributions:

- A novel AFE approach is proposed for high-movement human segmentation in a video. According to the experimental results, our approach efficiently deals with the issue of high-movement human segmentation in videos.
- This work suggests a novel approach to overcome the limitations of NFE and proves it with an empirical method. According to the experiment results, the proposed AFE shows better segmentation output than NFE.
- This work compares the proposed AFE with three image segmentation DL models and the ensemble of two models, as well as the NFE method by measuring Intersection over Union (IoU), variance error and bias error.

- This work reconstructs a new TikTok50f dataset to measure the effect of frame difference. This dataset cropped and resized the original TikTok dataset, and divided it into videos having 50 frames each.

Section 2 describes the related works of DSMs and human video segmentation. Section 3 presents the details of our proposed approach for high-movement human segmentation in a video. Section 4 explains the experimental result and its analysis. Section 5 provides a discussion of the results. Section 6 is the conclusion of this work.

## 2 Related Work

DL-based models offer outstanding potential in object classification and semantic segmentation areas. Long et al. [17] presented a general approach of adapting a Convolutional Neural Network for image classification into a "Fully Convolutional Network" (FCN), performing well in semantic segmentation. Singh et al. [18] and Jegou et al. [19] introduced a Dense convolutional Network (DenseNet) that consists of multiple densely connected blocks, which has proven to be effective for object classification and semantic segmentation. Chen et al. [20] proposed a DeepLab system that achieves semantic image segmentation using upsampled filters and atrous convolution, and a thoroughly connected Conditional Random Field (CRF) to get a better localization. Later, Chen et al. [21] suggested the DeepLabv3+ architecture, which uses the encoder-decoder structure in which, Deeplabv3 encodes the rich contextual information, and an effective decoder is chosen. MobileNets is a lightweight and efficient segmentation algorithm for limited performance devices like a mobile phone. Sandler et al. [22] proposed MobileNetV2 which is an enhanced edition of MobileNets. The model used an inverted residual block to bind the thin bottleneck layers to reduce the number of computations, and it shows enhanced performance across multiple applications. Howard et al. [23] presented MobileNetV3 based on hardware-aware Network Architecture Search (NAS) and NetAdapt that demonstrate state-of-the-art performance in semantic segmentation on mobile devices. MobileNetV2 and MobileNetV3 are highly regarded for their performance and efficiency, and these networks are adopted in many other models as a backbone. Zhang et al. [24] proposed PortraitNet for a real-time segmentation specifically designed for mobile phones. It consists of a decoder and an encoder. MobileNetV2 was used as the backbone for PortraitNet's encoder, and U-shape architecture was used for the decoder. PortraitNet uses two auxiliary losses to enhance segmentation precision. Mehta et al. [25] proposed ESPNet, a convolutional neural network using the Efficient Spatial Pyramid (ESP) structure. It can segment high-resolution images without consuming many computational resources. ESPNetv2 [26] is an extended version of ESPNet. It is a lightweight and power-efficient network for semantic segmentation, which makes it suitable for edge devices. Park et al. [27] proposed SINet, a highly lightweight portrait segmentation architecture, which can perform with 100.6 FPS, and 95.29% IoU on mobile phones. In comparison with traditional image segmentation methods, these DSMs show a significant performance improvement in image segmentation [28–33]. DSMs have proven to be useful for human segmentation purposes, according to some studies. These works, however, focus mainly on segmenting human bodies in a static image. Collective decisions made by the ensemble help to generate less errors and make the prediction more accurate. According to many studies, the blending of multiple segmentation models shows better performance than single segmentation models. Therefore, the ensemble of these DSMs can be regarded as a reasonable choice for human segmentation.

Warfield et al. [34] presented a new approach that combines multiple segmentation models and validates the performance of object segmentation. Rohlfing et al. [35] presented an approach of

combining multiple segmentation models using shape-based averaging. Holliday et al. [36] proposed a model compression method for the problem of semantic segmentation that uses the ensemble models to generate a training dataset for a single model and produce a real-time performance. Marmanis et al. [37] proposed the ensemble method with the FCN model for semantic segmentation of very high-resolution images and produced an outstanding performance. Kim et al. [38,39] suggested an ensemble of heterogeneous DSMs for portrait segmentation, and the efficiency of single models and ensemble models was analyzed. The authors demonstrated that some ensembles of DSMs produced better results than single models even with low consumption of computational resources. The reported works show that the ensemble approach can improve the performance of object segmentation. However, these approaches lack the consideration of real-time object segmentation in a video. Therefore, the consideration of segmentation speed with satisfactory segmentation output is required.

Gruosso et al. [40] presented the possibility of automatically segmenting humans in a surveillance video system using a DSM. The experiment was conducted using SegNet [41] model. Zhang et al. [42] presented a real-time framework for human segmentation in a video. The proposed model is composed of a fully convolutional network and a tracking module with a level set algorithm. Fully convolutional networks create human segmentations for specific frames in videos and feed them to tracking modules to capture human segmentations for the remaining frames. Perazzi et al. [43] proposed a convnet-based guided instance segmentation model that operates frame-by-frame by using the result of the previous frame to track the object of the following frame. Using one or a few segmentation masks, this approach segments a specific object in a video. The model combines offline and online learning strategies to improve the quality of video segmentation. However, these approaches focus on the structure of neural-network to produce better performance than other DSMs. Therefore, the ensemble of multiple DSMs for object segmentation in a video is a research domain that needs to be developed more.

Multiple papers have discussed the use of optical flow for video segmentation. Wang et al. [44] proposed a novel Portrait Video Segmentation (PVS) method which combines two segmentation networks. The proposed method produced highly accurate temporal-coherent segmentation results. Ding et al. [45] presented a collective estimation of VIS and optical flow. PSPNet and FlowNetS were used as the baseline networks unless otherwise specified. Video segmentation has also been attempted using differences between successive image frames. Liu et al. [46] presented a novel video segmentation model considering both accuracy and temporal consistency with real-time performance. This approach inferences each frame using a compact network. The proposed methods are tested using the three different DSMs to verify the efficiency of the model. Kim et al. [16] introduced a novel ensemble method called N-Frames Ensemble (NFE) for the automatic segmentation of selfies in a video using an ensemble of multiple DSMs to produce a high-performance segmentation. It uses a single DSM to produce a segmentation for a single video frame. This method uses the ensemble method to combine the segmented outputs of the current DSM with the outputs from previous DSMs. The advantage of this approach is that only a single DSM is required to segment a current video frame. As a result, the NFE has the segmentation speed of a single DSM. In addition, NFE has the benefits of the ensemble method. The NFE method has some limitations. It is not satisfactory for the segmentation of fast-moving objects in a video. Since it combines segmentation outputs of multiple video frames, there is a chance that it produces a blurred effect when combining the segmentation results of different frames. As a result, if the movement of objects in a video is high, the final segmentation quality will be affected. NFE is not suitable for segmenting videos with low FPS. A video with a low FPS is likely to have a longer time gap between frames, and it may have a significant difference between frames than one with a higher FPS. Since the NFE merges the segmentation outputs of neighboring video frames, segmentation accuracy is reduced due to the high difference between adjacent video frames. This work

proposes a novel ensemble approach called AFE for human video segmentation to overcome the issues that occur in the NFE model when the movement of a human in a video is high.

## 3 Our Approach

This section explains the proposed approach in detail. In the following, we first present the segmentation models used for the experiment in Section 3.1. In Section 3.2, the dataset used for testing the models is explained. Section 3.3 explains the proposed ensemble approach and its advantages. Finally, Section 3.4 describes the measurements used for the performance evaluation.

### *3.1 Segmentation Models*

This paper presents three pre-trained DSMs to compose the proposed ensemble model. There are several DSMs available on open-source websites [47–53]. Using these models, segmentation experiments can be performed immediately without training or parameter tuning. For the experiment, this work used three segmentation models, namely FCN, DeepLabv3 (DL3), and Mediapipe (MP). FCN and DL3 are publicly available [52,53] and are pre-trained models using the PASCAL VOC dataset. They use ResNet-101 as the backbone and require the PyTorch library in the Python environment. MP is publicly available [47] and uses MobilenetV3 as the backbone. It requires a Mediapipe library in the Python environment.

### *3.2 Datasets*

This work uses the TikTok dataset [54]. The TikTok dataset consists of 340 videos that capture a single person performing different types of dances. The duration is $10 \sim 15$ s. The resolution of videos is $1080 \times 604$ pixels. The frame rate for each video is 30 frames per second. For this experiment, we modified the TikTok dataset by cropping and resizing the images as well as making 50 frames per video. The resultant dataset contains 1711 videos, each containing 50 frames with $480 \times 360$ resolution. Finally, the dataset comprises around 85,550 video frames and ground-truth masks. We call it as TikTok50f dataset. The new TikTok50f dataset is divided into ten video groups based on the average difference of adjacent two frames of ground truth masks. Every 0.25% frame difference is considered as one group. The average frame difference of video group 1 is 0.25%, the average frame difference of video group 2 is 0.5%, and so on. Finally, the average frame difference of video group 10 is 2.5%.

### *3.3 Ensemble Approach*

In machine learning, an ensemble model is a combination of several single models to improve the overall predictions. For a single model segmentation approach, one model is used to generate the segmented outputs for an input image. Whereas, for the ensemble approach, several models are used to generate the segmented outputs to obtain an optimal result. In general, the ensemble approach involves combining heterogeneous models trained on the same dataset or homogeneous models trained on different datasets. Our work adopts the idea of the ensemble method to produce accurate human segmentation in a video. Among various ensemble methods, we follow the stacking ensemble approach in which multiple heterogeneous deep-learning models were trained using different datasets. The intermediate segmentation outputs of individual deep-learning models were combined using the soft-voting ensemble aggregator. The portrait segmentation using the ensemble approach for a still image was introduced in our previous publications [38,39]. These works show that the ensemble approach can improve the performance of object segmentation. However, it lacks the consideration

of real-time object segmentation in a video. Therefore, the consideration of segmentation speed with satisfactory segmentation output is required. The portrait segmentation using the ensemble of multiple deep-learning models was expanded to a video domain in another publication [16]. However, this approach merges the segmentation outputs of neighboring video frames. Therefore, segmentation accuracy is reduced due to the high difference between adjacent video frames. Finally, this work is an improved version of our previous works for portrait segmentation in a video. This paper uses pre-trained heterogeneous models for ensembles. FCN, DL3 and MP Models are used for ensemble in video segmentation to produce the segmented output. In Two-Models Ensemble (TME), all single models segment every single video frame, and then these segmented outputs are combined using the ensemble method to produce the optimal result.

The NFE [16] uses a single DSM to segment an object from a single video frame. This approach uses the ensemble method to combine the result of the current DSM with the result of previous DSMs. NFE produces almost the same result as TME if the movement of a target object is small enough. However, when the movement of the target object is high, the result is not as good as a single model or TME. This work presents a novel approach called AFE to overcome the limitations of NFE. The idea of AFE includes two conditions: the first condition is that, if the difference of neighboring video frames is less than a threshold value, AFE uses the ensemble method to combine segmented output results of the previous frame and current frame. The second condition is that, if the difference of neighboring video frames exceeds the threshold, the segmented output of the current frame is used without ensemble. Decision based on the difference of ground truth of adjacent video frames is named AFEGDT (AFE using Ground-truth Difference Threshold), and decision based on the difference of segmentation masks of sequent video frames is named AFESDT (AFE using Segmentation Difference Threshold). In AFE, only a single DSM is required to segment the current video frame at a given time. AFE combines the segmented results of adjacent video frames. Merging segmented results of neighboring video frames with high differences will affect the segmented output quality. In such a case, AFE uses segmented output result of the current video frame without ensemble. It merges the segmented results of neighboring video frames only if the frame difference is less than a threshold value. Fig. 2 presents the flow chart of human segmentation in video using the AFE approach. The AFE method uses a single DSM to produce a segmentation output of a single video frame. It combines the segmentation output of a current video frame with the segmentation of previous video frames only if the frame difference is less than a threshold. In detail, a video has a $Frame_t$ for a current video frame at a given time $t$, a $Frame_{t-1}$ for the previous video frame of a time $t-1$ and so on. When there are $n$ number of DSMs for the ensemble, the $Frame_{t-n+1}$ is fed to a $DSM_1$ to generate a segmented output, $Output_{t-n+1}$, the $Frame_{t-n+2}$ is to $DSM_2$ and so on. Finally, the $Frame_t$ is fed into a $DSM_n$ to generate a segmented output, $Output_t$. The proposed AFE method combines these outputs using the ensemble to generate the $Final\ Mask_t$ only when the frame difference $D$ is less than the threshold value $Th$. Alternatively, when $D$ exceeds the threshold $Th$, the segmented output of the current video frame is used for $Final\ Mask_t$ without ensemble.

The advantages of AFE are as follows

- AFE has a segmentation speed of a single DSM since it uses a single DSM for a current video frame. The segmentation output of the current DSM is combined with segmentation outputs of previous DSMs only if the frame difference is less than a particular threshold value.
- AFE has the benefits of the ensemble model because it uses the ensemble method to combine the output of multiple DSMs to produce optimal results when the frame difference is lesser than a threshold value.

- AFE is partially robust to sudden noises in a video. The mixing of sequent video frames can decrease the sudden noise of a specific video frame.

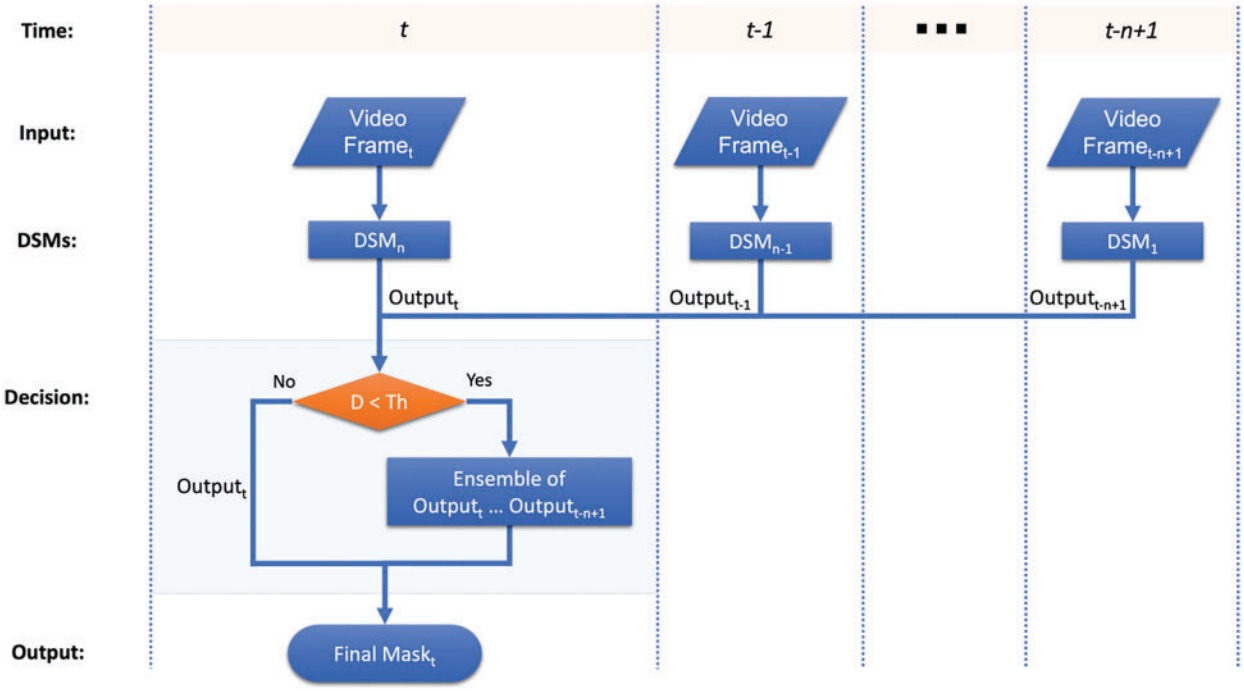


**Figure 2:** Flow chart of AFE approach

The advantage of AFE over NFE is as follows

- NFE is not suitable for the segmentation of fast-moving objects in a video. Fast-moving objects cause a high difference in neighboring video frames. The high difference between adjacent video frames will reduce the segmentation accuracy of the NFE ensemble since the NFE combines the segmentation results of neighboring video frames. However, AFE decides the ensemble of segmentation results based on the frame difference of adjacent video frames, and it is suitable for segmenting both slow-moving objects and fast-moving objects in a video.
- NFE is not appropriate for the video with low FPS because low FPS has a longer time slot between frames, and it is more likely to have a higher frame difference than a high FPS video. The high discord of neighboring video frames can reduce the precision of the NFE model. But AFE combines segmentation results of adjacent video frames only if the frame difference is lower than a threshold. As a result, the overall segmentation results are better than NFE in a low FPS video.

The two most popular ensemble methods are averaging and voting. There are several kinds of voting methods [55–57]. In this work, a simple soft voting method is used for the ensemble because it is simple and efficient. Here, individual classifiers are treated equally, and their outputs are averaged.

### *3.4 Performance Measurement*

A metric called Intersection over Union (IoU) is used to calculate segmentation accuracy.

The equation of mIoU is defined in (1) below

$$mIoU = mean\left(\frac{\mathrm{A} \cap \mathrm{B}}{A \cup B}\right) \tag{1}$$

where A and B are the predicted segmentation area and ground-truth area. The intersection area is denoted by A ∩ B, and the union area is denoted by A ∪ B. The IoU standard deviation measures the prediction variance error. False Negative Rate (FNR) and False Discovery Rate (FDR) are calculated to measure segmentation error. Areas that are smaller than the ground truth are measured by FNR. Areas that are larger than the ground truth are measured by FDR. The equations of FNR and FDR are defined in (2) and (3) below.

$$FNR = \frac{FN}{FN + TP} \tag{2}$$

$$FDR = \frac{FN}{FP + TP} \tag{3}$$

where FP, FN and TP mean False Positive, False Negative and True Positive each. In this paper, the Bias Error (BE) is measured with Eq. (4) and the Balanced Bias Error (BBE) with Eq. (5) as below

$$BE = FDR + FNR \tag{4}$$

$$BBE = |FDR - FNR| \tag{5}$$

## 4 Experiment Result

We conducted the experiment using three single models FCN, DL3 and MP. Evaluations are performed on the TikTok50f dataset. This section presents the experimental result and its analysis of single models, TME, NFE, and the proposed AFE (AFEGDT and AFESDT). The mIoU, IoU standard deviation, FNR, FDR, BE, and BBE are used to evaluate the performance of all models. Python, OpenCV, PyTorch and Mediapipe were used on an Ubuntu machine with GTX-1080 GPU and 16 GB RAM.

### *4.1 Experiments of Single Models and Ensemble Models*

Tab. 1 shows the experimental results of three single models FCN, DL3, and MP. The first column denotes ensemble methods, the second column is the combination of DSMs, and the remains are measurement metrics. In the results of no ensemble, MP shows 95.16% of mIoU and it is the highest mIoU value and shows 4.96% of BE. The value is the lowest BE among the experimented single models. FCN produces the lowest IoU standard deviation. A low IoU standard deviation indicates a low variance error, whereas a low BE indicates a low bias error. A low BBE means that FCN is well-balanced in false-positive and false-negative regions. Among all combinations of TME, FCN+MP produces the highest mIoU and the lowest false prediction rate and shows better mIoU than FCN or MP single model. But the IoU standard deviation and BBE are higher than the FCN. DL3+MP produces a better mIoU value than DL3 or MP single model, but IoU standard deviation is higher than DL3. FCN + DL3 also shows a better mIoU value than FCN or DL3 single model, BBE is higher than FCN and DL3, and IoU standard deviation is higher than FCN. TME results in all the combinations show higher mIoU than single models. In TME, the highest mIoU value is 95.64%, which is higher than the highest mIoU value of single models, 95.16%. Among all combinations of NFE, FCN+MP produces the highest mIoU and the lowest false prediction rate among all combinations. It shows lower mIoU than FCN or MP single model, and shows lower mIoU than the TME of FCN+MP. DL3+MP

shows lower mIoU than DL3 or MP single model, and shows lower mIoU than the TME of DL3+MP. FCN+DL3 shows lower mIoU than FCN or DL3 single model, and lower mIoU than the TME of FCN+DL3. NFE result is not good as a single model or TME in all three combinations. The highest mIoU value of the NFE is 94.04%, which is less than the highest mIoU value of single models, 95.16%. Among all combinations of AFEGDT in which the decisions are based on the difference of ground truth masks of adjacent video frames, FCN+MP produces the highest mIoU and the lowest false prediction rate among the three combinations. It shows better mIoU than FCN+MP combination involved in NFE and improved mIoU than FCN or MP single model. IoU standard deviation and BBE are lower than FCN+MP in NFE. DL3+MP produces better mIoU than the DL3+MP in the NFE model, and improved mIoU than DL3 or MP single model. IoU standard deviation and BBE are lower than DL3+MP in NFE. FCN+DL3 produces better mIoU than FCN+DL3 in NFE. It also produces improved mIoU than DL3 single model. IoU standard deviation and BBE are lower than FCN+DL3 in NFE. In the table, two of them show better mIoU than single models among three possible combinations. The highest mIoU value of the AFEGDT is 95.2%, which is higher than the highest mIoU value of NFE, 94.04%. Among all combinations of AFESDT in which the difference of segmented outputs of adjacent video frames is used for making decisions instead of the difference of ground truth outputs of adjacent video frames, FCN+MP produces the highest mIoU and the lowest false prediction rate among the three combinations. It shows a better mIoU than the FCN+MP combination involved in the NFE model and shows improved mIoU than FCN single model. IoU standard deviation and BBE are lower than FCN+MP in NFE. DL3+MP produces better mIoU than the DL3+MP in the NFE model and produces improved mIoU than DL3. IoU standard deviation and BBE are lower than DL3+MP in NFE. FCN+DL3 produces better mIoU than the FCN+DL3 in the NFE model and improved mIoU than DL3 single model. IoU standard deviation and BBE are lower than FCN+DL3 in NFE. In the table, all three possible combinations of AFESDT show better mIoU than single models and NFE.

**Table 1:** The experiment result of single models and ensemble models (%)

| | Models | mIoU | IoU Std | FNR | FDR | BE | BBE |
|---|---|---|---|---|---|---|---|
| No ensemble | FCN | 95.13 | 1.97 | 2.32 | 2.66 | 4.98 | 0.34 |
| | DL3 | 95.05 | 2.02 | 2.14 | 2.92 | 5.06 | 0.78 |
| | MP | 95.16 | 2.25 | 2.19 | 2.77 | 4.96 | 0.57 |
| TME | FCN+DL3 | 95.17 | 1.99 | 2.07 | 2.86 | 4.93 | 0.79 |
| | FCN+MP | 95.64 | 2.08 | 2.47 | 1.97 | 4.45 | 0.50 |
| | DL3+MP | 95.58 | 2.12 | 2.40 | 2.12 | 4.52 | 0.28 |
| NFE | FCN+DL3 | 93.68 | 2.53 | 3.23 | 3.28 | 6.52 | 0.05 |
| | FCN+MP | 94.04 | 2.60 | 3.65 | 2.49 | 6.14 | 1.15 |
| | DL3+MP | 94.01 | 2.62 | 3.55 | 2.63 | 6.18 | 0.93 |
| AFEGDT | FCN+DL3 | 95.09 | 2.04 | 2.22 | 2.80 | 5.02 | 0.58 |
| | FCN+MP | 95.20 | 2.37 | 2.35 | 2.56 | 4.91 | 0.22 |
| | DL3+MP | 95.16 | 2.40 | 2.25 | 2.70 | 4.95 | 0.44 |

(Continued)

**Table 1:** Continued

| | Models | mIoU | IoU Std | FNR | FDR | BE | BBE |
|---|---|---|---|---|---|---|---|
| AFESDT | FCN+DL3 | 95.09 | 2.04 | 2.23 | 2.79 | 5.02 | 0.56 |
| | FCN+MP | 95.14 | 2.43 | 2.26 | 2.71 | 4.97 | 0.45 |
| | DL3+MP | 95.10 | 2.46 | 2.16 | 2.84 | 5.01 | 0.68 |

### *4.2 Comparison Based on Frame Difference*

This work divided the TikTok50f dataset into ten video groups based on the average frame difference of adjacent two frames of ground truth images. Every 0.25% frame difference is considered as one group. There are ten video groups ranging from 0.25% to 2.5% frame differences. The results obtained from each group in all three combinations are given below. Tab. 2 shows the experimental result of the FCN+DL3 combinations for all video groups. The first column in the table shows all the ten groups. The second and third column shows the FCN and DL3 single model results. The fourth column is the FCN+DL3 results of TME. The fifth column is the FCN+DL3 results of NFE, and the sixth and seventh column shows the proposed AFEGDT and AFESDT model results. Group 1 produces the highest mIoU for single models, TME, NFE and AFE models, compared to other groups. In group 1, TME produces the highest mIoU of 96.29%, and the mIoU is almost similar to NFE, AFEGDT and AFESDT. In group 2, TME produces better mIoU than any of the single models, NFE, AFEGDT or AFESDT. NFE produces lower mIoU than FCN and DL3 single model from group 2. But, the proposed AFEGDT and AFESDT models produce better mIoU than NFE and similar mIoU as the single models. From group 3 to group 10, TME produces better mIoU than any of the single model or NFE or AFE models. Proposed AFE models produce higher mIoU than the NFE model and DL3 single model, but it shows lower mIoU than FCN single model. NFE and AFE show similar results in group 1, and it is as good as the mIoU of TME. From group 2 onwards, the proposed AFE is better than the NFE model. Tab. 3 shows the experimental result of the FCN+MP combinations for all groups. Group 1 produces the highest mIoU for single models as well as all ensemble models compared to other groups. In group 1, TME produces the highest mIoU of 96.59%, and the mIoU is almost similar to NFE or AFE. From group 2 to group 4, TME produces better mIoU than any of the single models or NFE or AFE. The proposed AFE produces better mIoU than any of the single model or NFE. NFE produces better mIoU than FCN or MP single model. From group 5 to group 10, TME produces better mIoU than any of the single models, NFE or AFE. The AFEGDT produces better mIoU than any of the single model or NFE. The proposed AFESDT also produces better mIoU than NFE but lower mIoU than one of the single models. NFE produces lower mIoU than FCN and MP single models. NFE and AFE show similar results in group 1and group 2. From group 3 onwards, the proposed AFE is better than the NFE model. Tab. 4 shows the experimental result of the DL3+MP combinations for all groups. Group 1 produces the highest mIoU for single models as well as ensemble models compared to other groups. In group 1, TME produces the highest mIoU of 96.56%, and the mIoU is almost similar to NFE or AFE. In group 2 and group 3, TME produces better mIoU than any of the single models or NFE or AFE. The proposed AFE produces better mIoU than any of the single model or NFE. NFE produces better mIoU than FCN or MP single model. From group 4 to group 10, TME produces better mIoU than any of the single models, NFE or AFE. The AFEGDT produces better mIoU than any of the single model or NFE. The AFESDT produces better mIoU than NFE but lower mIoU than one of the single models. NFE significantly produces lower mIoU than DL3 or

MP single model. NFE and AFE show similar results in group 1and group 2. From group 3 onwards, the proposed AFE is better than the NFE model.

**Table 2:** mIoU of FCN and DL3 combinations (%)

| Groups | FCN | DL3 | TME | NFE | AFEGDT | AFESDT |
|---|---|---|---|---|---|---|
| group 1 | 96.23 | 96.13 | 96.29 | 96.27 | 96.28 | 96.27 |
| group 2 | 95.25 | 95.17 | 95.33 | 95.15 | 95.25 | 95.22 |
| group 3 | 95.18 | 95.08 | 95.24 | 94.91 | 95.16 | 95.14 |
| group 4 | 95.32 | 95.26 | 95.41 | 94.90 | 95.31 | 95.30 |
| group 5 | 95.52 | 95.45 | 95.60 | 94.93 | 95.50 | 95.49 |
| group 6 | 95.43 | 95.38 | 95.51 | 94.70 | 95.42 | 95.41 |
| group 7 | 95.41 | 95.37 | 95.49 | 94.55 | 95.39 | 95.39 |
| group 8 | 95.33 | 95.26 | 95.39 | 94.30 | 95.29 | 95.29 |
| group 9 | 95.24 | 95.15 | 95.29 | 94.07 | 95.20 | 95.19 |
| group 10 | 95.19 | 95.11 | 95.24 | 93.87 | 95.15 | 95.15 |

**Table 3:** mIoU of FCN and MP combinations (%)

| Group | FCN | MP | TME | NFE | AFEGDT | AFESDT |
|---|---|---|---|---|---|---|
| group 1 | 96.23 | 95.90 | 96.59 | 96.56 | 96.56 | 96.56 |
| group 2 | 95.25 | 95.21 | 95.86 | 95.65 | 95.65 | 95.70 |
| group 3 | 95.18 | 95.37 | 95.93 | 95.55 | 95.61 | 95.59 |
| group 4 | 95.32 | 95.31 | 95.94 | 95.38 | 95.56 | 95.40 |
| group 5 | 95.52 | 95.39 | 96.01 | 95.28 | 95.62 | 95.47 |
| group 6 | 95.43 | 95.40 | 96.00 | 95.12 | 95.56 | 95.42 |
| group 7 | 95.41 | 95.40 | 95.94 | 94.91 | 95.51 | 95.41 |
| group 8 | 95.33 | 95.34 | 95.86 | 94.69 | 95.42 | 95.33 |
| group 9 | 95.24 | 95.29 | 95.79 | 94.47 | 95.34 | 95.26 |
| group 10 | 95.19 | 95.23 | 95.72 | 94.24 | 95.28 | 95.21 |

**Table 4:** mIoU of DL3 and MP combinations (%)

| Group | DL3 | MP | TME | NFE | AFEGDT | AFESDT |
|---|---|---|---|---|---|---|
| group 1 | 96.13 | 95.90 | 96.56 | 96.53 | 96.53 | 96.53 |
| group 2 | 95.17 | 95.21 | 95.84 | 95.64 | 95.64 | 95.69 |
| group 3 | 95.08 | 95.37 | 95.84 | 95.47 | 95.54 | 95.54 |
| group 4 | 95.26 | 95.31 | 95.86 | 95.30 | 95.50 | 95.34 |

(Continued)

**Table 4:** Continued

| Group | DL3 | MP | TME | NFE | AFEGDT | AFESDT |
|---|---|---|---|---|---|---|
| group 5 | 95.45 | 95.39 | 95.92 | 95.21 | 95.57 | 95.43 |
| group 6 | 95.38 | 95.40 | 95.93 | 95.07 | 95.52 | 95.39 |
| group 7 | 95.37 | 95.40 | 95.88 | 94.88 | 95.48 | 95.39 |
| group 8 | 95.26 | 95.34 | 95.80 | 94.65 | 95.38 | 95.30 |
| group 9 | 95.15 | 95.29 | 95.72 | 94.43 | 95.29 | 95.22 |
| group 10 | 95.11 | 95.23 | 95.65 | 94.20 | 95.23 | 95.17 |

### *4.3 Result Analysis*

Tab. 5 shows the average values of mIoU, IoU standard deviation, FNR, FDR, BE and BBE from single models, TME, NFE, AFEGDT and AFESDT. The average of all single models shows 95.11% of mIoU and 5.00% of BE. TME shows a higher mIoU value, lower IoU standard deviation, and lower BBE value than the single models. NFE shows lower mIoU than both single and TME. The mIoU value of the AFEGDT is 95.15%, higher than NFE, which is 93.91%. The difference in mIoU value between NFE and AFEGDT is 1.24%. AFEGDT produces the lowest BBE value than any other model. The mIoU value of the AFESDT is 95.11%. It is lower than the average value of the AFEGDT and similar to a single model. The proposed AFE models significantly improve mIoU, IoU standard deviation, BE, and BBE than the NFE model. AFEGDT shows better mIoU and BBE than single models, while AFESDT shows similar mIoU and BBE to single models.

**Table 5:** Average of all models (%)

| | mIoU | IoU Std | FNR | FDR | BE | BBE |
|---|---|---|---|---|---|---|
| Single | 95.11 | 2.08 | 2.22 | 2.78 | 5.00 | 0.57 |
| TME | 95.47 | 2.06 | 2.31 | 2.32 | 4.63 | 0.52 |
| NFE | 93.91 | 2.58 | 3.48 | 2.80 | 6.28 | 0.71 |
| AFEGDT | 95.15 | 2.27 | 2.27 | 2.69 | 4.96 | 0.41 |
| AFESDT | 95.11 | 2.31 | 2.22 | 2.78 | 5.00 | 0.56 |

Tab. 6 shows the average values of mIoU from single models, TME, NFE, AFEGDT and AFESDT models in all the ten video groups, which are divided based on frame difference. The bracket value of the first column is the frame difference rate of each group. The last row indicates the average values of all videos. TME shows a higher mIoU value than the single models in all the groups. In NFE, mIoU is decreasing from group 1 to group 10, while the frame difference is increasing from group 1 to group 10. From group 1 to group 3, NFE shows higher mIoU than single models, but from group 4 to group 10, it shows a lower value than single models. The AFEGDT produces higher mIoU than both single models and NFE models in all the groups. The AFESDT shows higher mIoU than NFE but lower than AFEGDT. From group 1 to group 9, the AFESDT performs with higher mIoU than single models. From the table, TME and the proposed AFE models have better mIoU than single models and NFE models in most of the groups. NFE, AFEGDT and AFESDT produce almost the same result as TME in group 1, where the difference between adjacent video frames is 0.25%. But the mIoU of the NFE model drops rapidly as the frame difference increases, and it becomes less than the single models

from group 4 onwards. On the other hand, as the frame difference increases, AFE models decrease mIoU and converge to single models. The result shows demonstrably that AFE models maintain a stable mIoU even as the frame difference increases.

**Table 6:** Average mIoU comparison of all groups (%)

| Group | Single | TME | NFE | AFEGDT | AFESDT |
|---|---|---|---|---|---|
| group 1 | 96.08 | 96.48 | 96.45 | 96.46 | 96.45 |
| group 2 | 95.21 | 95.68 | 95.48 | 95.51 | 95.54 |
| group 3 | 95.21 | 95.67 | 95.31 | 95.43 | 95.43 |
| group 4 | 95.30 | 95.74 | 95.20 | 95.46 | 95.35 |
| group 5 | 95.46 | 95.84 | 95.14 | 95.56 | 95.46 |
| group 6 | 95.41 | 95.81 | 94.96 | 95.50 | 95.41 |
| group 7 | 95.39 | 95.77 | 94.78 | 95.46 | 95.39 |
| group 8 | 95.31 | 95.68 | 94.55 | 95.37 | 95.31 |
| group 9 | 95.22 | 95.60 | 94.32 | 95.27 | 95.23 |
| group 10 | 95.18 | 95.54 | 94.10 | 95.22 | 95.18 |
| All videos | 95.11 | 95.47 | 93.91 | 95.15 | 95.11 |

Fig. 3 compares the mIoU ratio of single models with all ensemble models at various frame differences, represented as in Tab. 6. The dotted line at 100% represents the average mIoU of single models, and the other curves represent TME, NFE, AFEGDT, and AFESDT. The TME has the highest mIoU ratio among all the models, and it is always more accurate than single models. Moreover, it is not heavily dependent on frame differences. The NFE is more accurate than single models and is almost as accurate as of the TME when the frame difference is 0.25%. However, as the frame difference increases, the mIoU decreases significantly, and it becomes lower than single models when the frame difference is more than 0.75%. The AFEGDT and AFESDT show higher mIoU than single models and are almost as accurate as the TME and NFE when the frame difference is 0.25%. If the frame difference increases, the mIoU decreases and converges to single models. However, it always shows better mIoU than NFE. AFEGDT is slightly better than AFESDT. All ensemble models show comparable mIoU and are better than single models until 0.75%. NFE shows a high dependency on the frame difference. This is the disadvantage of NFE, and the proposed AFE overcomes this limitation of NFE. Even though the mIoU of AFE decreases when the frame difference increases, it exhibits a stable pattern and approaches to the mIoU of single models. As a result, NFE does not work well for videos with high frame differences because the NFE has a high dependency on the frame difference. But AFE maintains a stable mIoU even as the frame difference increases.

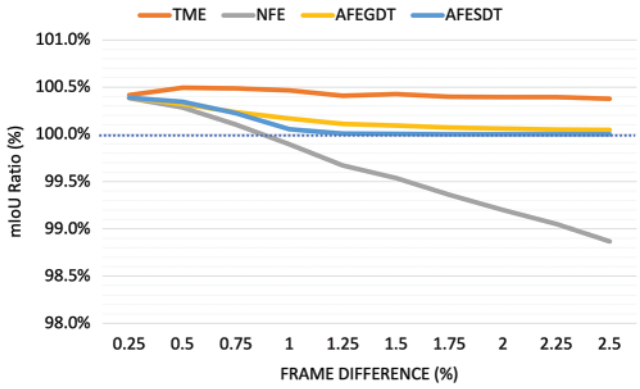


**Figure 3:** Comparison of all ensemble models based on mIoU ratio to single models

### *4.4 Examples of Video Segmentation*

Fig. 4 shows the segmentation results of single models, TME, NFE and the proposed AFE model in which the frame difference of adjacent video frames is less. Column (a) shows the original video frames; (b) is ground truth; (c) and (d) are the segmentation results using single models; (e) is the result using the TME; and (f) is NFE and (g) is our AFE. The proposed AFE shows better segmentation results than single models. The DSM-1 of the first image shows a huge region of false prediction marked with a red circle.

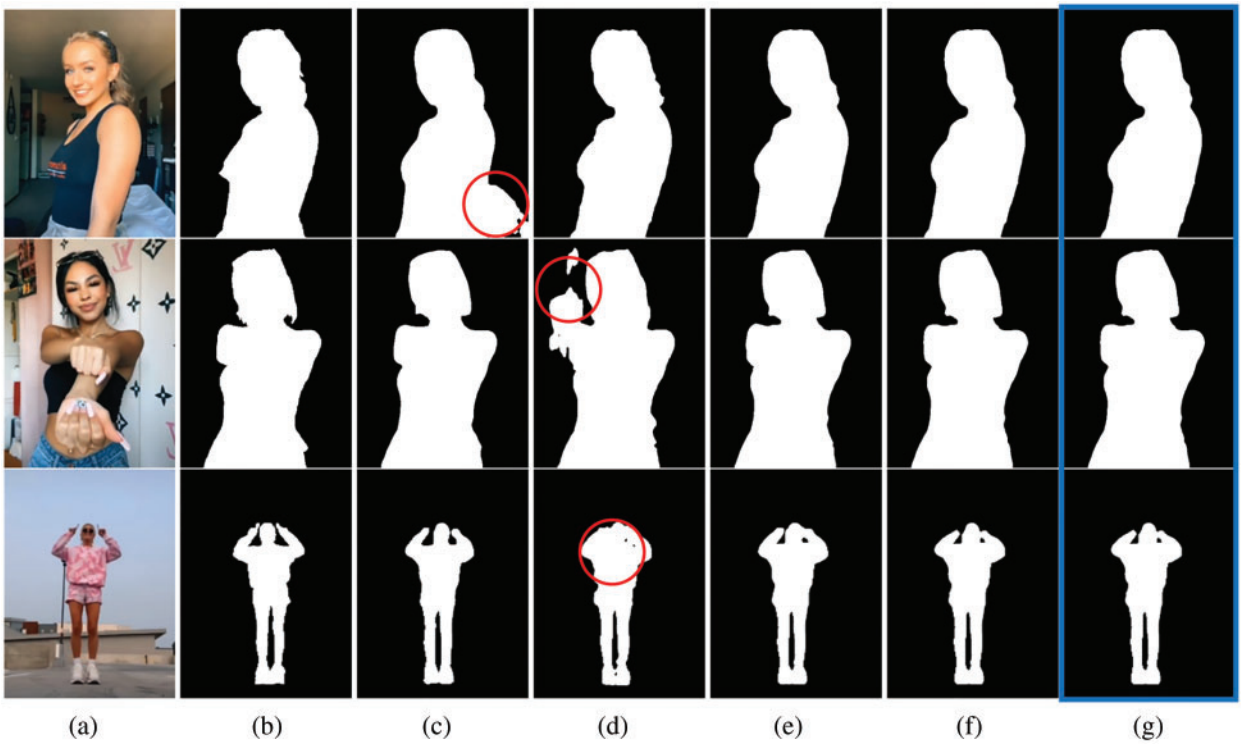


**Figure 4:** Image segmentation example of various videos with low frame difference; (a) original video frame (b) ground truth (c) DSM-1 (d) DSM-2 (e) TME (f) NFE (g) AFE

DSM-2 of the second and the third images also show regions of false prediction. However, TME, NFE and AFE are showing almost similar results and are better than single models. Fig. 5 shows the segmentation result of single models, TME, NFE and the proposed AFE models. In which the frame difference of adjacent video frames is high. Column (a) is the original video frames; (b) is ground truth; (c) and (d) are the segmentation results using single models; (e) is the result using the TME; and (f) is NFE and (g) is our AFE. TME is showing the best result, and it is better than single models. NFE is showing poor results than single models. AFE is not as good as TME. It is almost similar to single models but better than NFE.

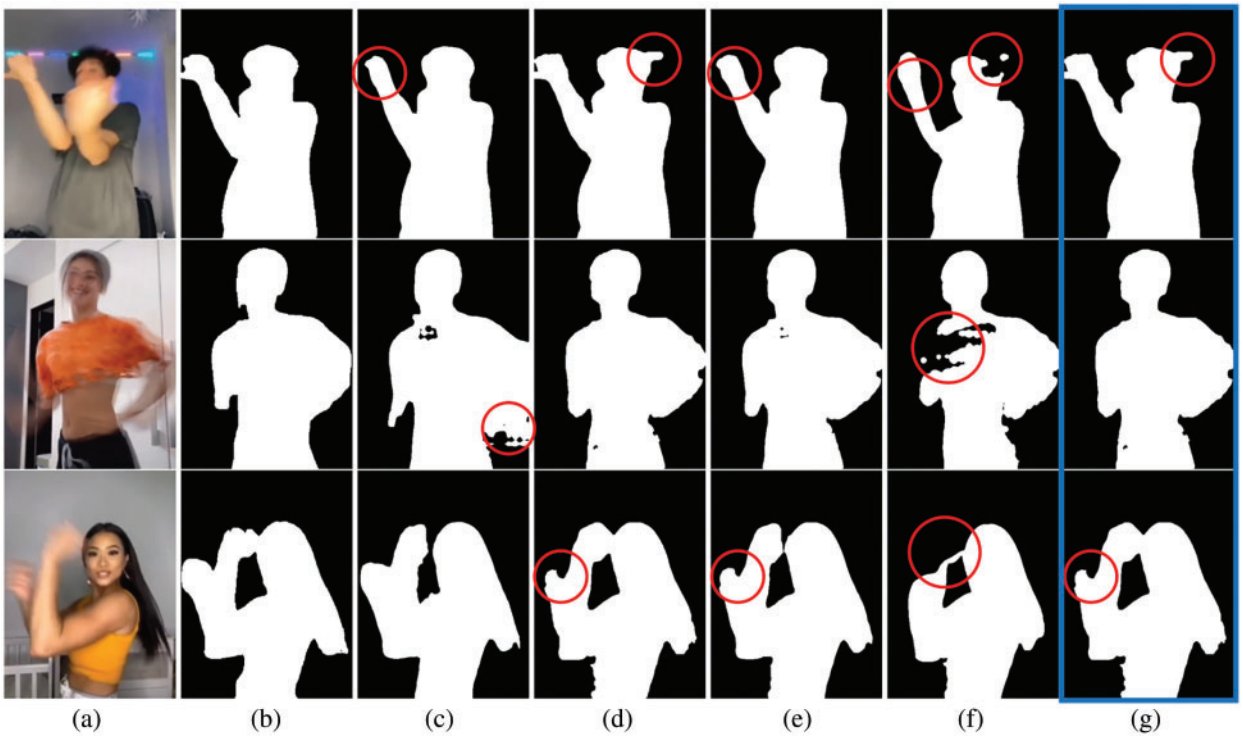


**Figure 5:** Image segmentation example of various videos with high frame difference; (a) original video frame (b) ground truth (c) DSM-1 (d) DSM-2 (e) TME (f) NFE (g) AFE

## 5 Discussion

The experiment results show that the proposed AFE approaches show higher mIoU of human segmentation in a video than single models and the existing NFE model. NFE shows a significant reduction of mIoU when the frame difference is increasing. If the average frame difference is small enough, then NFE is as accurate as the TME and shows higher mIoU than single models. But if the frame difference becomes larger, then mIoU becomes lower than single models. The AFEGDT shows better results than the AFESDT, but both approaches show a similar pattern. When the frame difference is small, the mIoU is higher than single models and almost as good as the TME and the NFE. However, with the increasing frame difference, mIoU approaches that of single models. The overall results present that NFE, the proposed AFE and TME have the highest mIoU when the frame difference is as low as 0.25%. But as NFE highly depends on frame difference, when the frame difference is higher than 0.75%, the mIoU decreases significantly and becomes lower than single

models. In AFE, even though the mIoU decreases when the frame difference increases, it remains stable and approaches the mIoU of single models. It can be clearly seen in group 1 and group 10 results. When the average frame difference is 0.25%, the mIoU value for NFE is 96.45%, and for AFE, it is 96.46%. Both the models show a similar result. But, when the average frame difference is 2.5%, the mIoU values for single models, NFE and AFESDT are 95.18%, 94.10%, and 95.18% each. In this case, NFE shows much less mIoU than single models, but AFESDT shows similar mIoU to single models.

## 6 Conclusion

In this work, we proposed an AFE approach for human video segmentation that is as efficient as a single DSM. TME, NFE, AFEGDT and AFESDT were experimented with and compared with the single DSMs. In this study, we used a simple soft voting method for the ensemble of multiple DSMs. Video frames of 1711 videos from the TikTok50f dataset that have a single-person view were used as a test dataset. This work divided the TikTok50f dataset into ten video groups based on the average frame difference of adjacent two frames of ground truth masks to explore the relationship between frame differences and the mIoU of ensemble models. We analyzed the performance of single models and experimented with ensemble approaches by measuring the mIoU value for precision, the IoU standard deviation for variance error, and the BE and BBE for bias error. The experiment result shows that the proposed AFE and NFE are as accurate as the TME and show higher mIoU than single models when the average frame difference is as less as 0.25%. But the mIoU of the NFE model drops rapidly as the frame difference increases more than 0.75% and becomes less than the single models. But in AFE, even though the mIoU decreases when the frame difference increases, it remains stable and approaches the mIoU of single models. The mIoU value for NFE is 96.45% and for AFEGDT is 96.46% when the average frame difference is 0.25%. The mIoU value for NFE is 94.10%, and for AFEGDT is 95.22% when the average frame difference is 2.5%. The results clearly demonstrate that increasing frame difference significantly reduces the mIoU of NFE, and this is the limitation of NFE. However, the proposed AFE approach addresses this limitation. Therefore, AFE is suitable for low-movement as well as high-movement human segmentation in a video as it makes an adaptive ensemble based on the frame difference of sequent video frames. This work experimented with a single person video dataset, and this is the limitation of this work. However, our approach is applicable to the segmentation of multiple persons in a video also.

**Acknowledgement:** Authors are thankful to the Centre for Digital Innovation of CHRIST (Deemed to be University), India and WIZnet India Pvt. Ltd., for the support and providing resources to carry out this research work.

**Funding Statement:** This research was financially supported by the Ministry of Small and Medium-sized Enterprises (SMEs) and Startups (MSS), Korea, under the "Regional Specialized Industry Development Program (R&D, S3091627)" supervised by the Korea Institute for Advancement of Technology (KIAT).

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] Y. Guo, Y. Liu, T. Georgiou and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 87–93, 2018.

[2] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz *et al.,* "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1, 2021.
[3] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 1, pp. 100–132, 1985.
[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: Practical Innovations, Open Solutions*, vol. 8, pp. 136361–136373, 2020.
[5] X. Liu, Z. Deng and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1089–1106, 2019.
[6] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez *et al.,* "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, no. 9, pp. 41–65, 2018.
[7] M. Gruosso, N. Capece and U. Erra, "Human segmentation in surveillance video with deep learning," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 1175–1199, 2021.
[8] 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.
[9] B. Zhao, J. Feng, X. Wu and S. Yan, "A survey on deep learning based fine-grained object classification and semantic segmentation," *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017.
[10] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1–2, pp. 1–39, 2010.
[11] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*. Cambridge (EE. UU.): MIT Press, 2016.
[12] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.
[13] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.
[14] Y. Zuo and T. Drummond, "Fast residual forests: Rapid ensemble learning for semantic segmentation," *Proceedings of the 1st Annual Conference on Robot Learning*, vol. 78, no. 13–15 November, pp. 27–36, 2017.
[15] Y. Koren, *The BellKor solution to the Netflixgrand prize*, pp. 1–10, 2009.
[16] Y. -W. Kim, Y. -C. Byun, A. V. N. Krishna and B. Krishnan, "Selfie segmentation in video using N -frames ensemble," *IEEE Access*, vol. 9, pp. 163348–163362, 2021.
[17] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, 2015.
[18] D. Singh, V. Kumar and M. Kaur, "Densely connected convolutional networks-based COVID-19 screening model," *Applied Intelligence*, vol. 51, no. 5, pp. 1–8, 2021.
[19] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Honolulu, HI, USA, 2017.
[20] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
[21] 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*. Springer International Publishing, Cham, Switzerland, pp. 833–851, 2018.
[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, UT, USA, 2018.
[23] A. Howard, M. Sandler, G. Chu, L. -C. Chen, B. Chen *et al.,* "Searching for MobileNetV3," in *2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, IEEE, Seoul, Korea (South), 2019.
[24] S.-H. Zhang, X. Dong, H. Li, R. Li and Y.-L. Yang, "PortraitNet: Real-time portrait segmentation network for mobile device," *Computers & Graphics*, vol. 80, no. 12, pp. 104–113, 2019.

[25] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilatedconvolutions for semantic segmentation," in *Computer Vision–ECCV 2018*. Springer International Publishing, Cham, Switzerland, pp. 561–580, 2018.
[26] S. Mehta, M. Rastegari, L. Shapiro and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, CA, USA, 2019.
[27] H. Park, L. L. Sjosund, Y. Yoo, N. Monet, J. Bang *et al.,* "SINet: Extreme lightweight portrait segmentation networks with spatial squeeze modules and information blocking decoder," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Snowmass Village, CO, USA, 2020.
[28] J. Yang, B. Price, S. Cohen, H. Lee and M. H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, 2016.
[29] X. Liu, Z. Deng and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1089–1106, 2019.
[30] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez *et al.,* "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, no. 9, pp. 41–65, 2018.
[31] M. Gruosso, N. Capece and U. Erra, "Human segmentation in surveillance video with deep learning," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 1175–1199, 2021.
[32] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He and X. Chen, "TBE-Net: A three-branch embedding network with part-awareability and feature complementary learning for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems: A Publication of the IEEE Intelligent Transportation Systems Council*, pp. 1–13, 2021.
[33] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 92, no. 6, pp. 1–16, 2021.
[34] 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 Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.
[35] T. Rohlfing, C. R. Maurer and R. C. Jr, "Shape-based averaging for combination of multiple segmentations," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 8, no. Pt 2, pp. 838–845, 2005.
[36] A. Holliday, M. Barekatain, J. Laurmaa, C. Kandaswamy and H. Prendinger, "Speedup of deep learning ensembles for semantic segmentation using a model compression technique," *Computer Vision and Image Understanding: CVIU*, vol. 164, no. 1, pp. 16–26, 2017.
[37] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu *et al.,* "Semantic segmentation of aerial images with an ensemble of CNNs," *ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. III-3, pp. 473–480, 2016.
[38] Y. W. Kim, J. Innila Rose and A. V. N. Krishna, "Accuracy enhancement of portrait segmentation by ensembling deep learning models," in *2020 Fifth Int. Conf. on Research in Computational Intelligence and Communication Networks (ICRCICN)*, IEEE, Bangalore, India, 2020.
[39] Y. -W. Kim, Y. -C. Byun and A. V. N. Krishna, "Portrait segmentation using ensemble of heterogeneous deep-learning models," *Entropy (Basel)*, vol. 23, no. 2, pp. 197, 2021.
[40] M. Gruosso, N. Capece and U. Erra, "Human segmentation in surveillance video with deep learning," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 1175–1199, 2021.
[41] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
[42] T. Zhang, C. Lang and J. Xing, "Realtime human segmentation in video," in *MultiMedia Modeling*, Springer International Publishing, Cham, Switzerland, pp. 206–217, 2019.

[43] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele and A. SorkineHornung, "Learning video object segmentation from static images," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, USA, 2017.
[44] Y. Wang, W. Zhang, L. Wang, F. Yang and H. Lu, "Temporal consistent portrait video segmentation," *Pattern Recognition*, vol. 120, no. 108143, pp. 108143, 2021.
[45] M. Ding, Z. Wang, B. Zhou, J. Shi, Z. Lu *et al.,* "Every frame counts: Joint learning of video segmentation and optical flow," *Proceedings of the . AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 10713–10720, 2020.
[46] Y. Liu, C. Shen, C. Yu and J. Wang, "Efficient semantic video segmentationwith per-frame inference," in *Computer Vision–ECCV 2020*. Springer International Publishing, Cham, Switzerland, pp. 352–368, 2020.
[47] Selfie Segmentation, "Github.io," (Accessed: 29-De-2021). *Available:* https://google.github.io/mediapipe/solutions/selfie_segmentation.html.
[48] PortraitNet,"Github.com," (Accessed: 29-Dec-2021). *Available:* https://github.com/dong-x16/PortraitNet.
[49] TensorflowLite-UNet,"Github.com,"(Accessed:29-Dec-2021).*Available:*https://github.com/PINTO0309/TensorflowLite-UNet.
[50] Portrait-Segmentation,"Github.com,"(Accessed:29-Dec-2021).*Available:*https://github.com/anilsathyan7/Portrait-Segmentation.
[51] SelfieSeg, "Github.com," (Accessed: 29-Dec-2021). *Available:* https://github.com/Innovation4x/SelfieSeg.
[52] Fully-Convolutional Network model with ResNet-50 and ResNet-101 backbones, "Pytorch.org," (Accessed: 29-Dec-2021). *Available:* https://pytorch.org/hub/pytorch_vision_fcn_resnet101/.
[53] DeepLabV3 models with ResNet-50, ResNet-101 and Mobile Net-V3 backbones, "Pytorch.org,"(Accessed:29-Dec-2021).*Available:*https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/.
[54] TikTok Dataset "Kaggle.com,"(Accessed:29-Dec-2021). *Available:* https://www.kaggle.com/yasaminjafarian/tiktokdataset.
[55] Z.-H. Zhou, "Ensemble methods: Foundations and algorithms," in *Caithness*, UK: Whittles Publishing, 2012.
[56] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
[57] P. Smyth and D. Wolpert, "Linearly combining density estimators via stacking," *Machine Learning*, vol. 36, no. 1/2, pp. 59–83, 1999.