Advanced Ultrasound

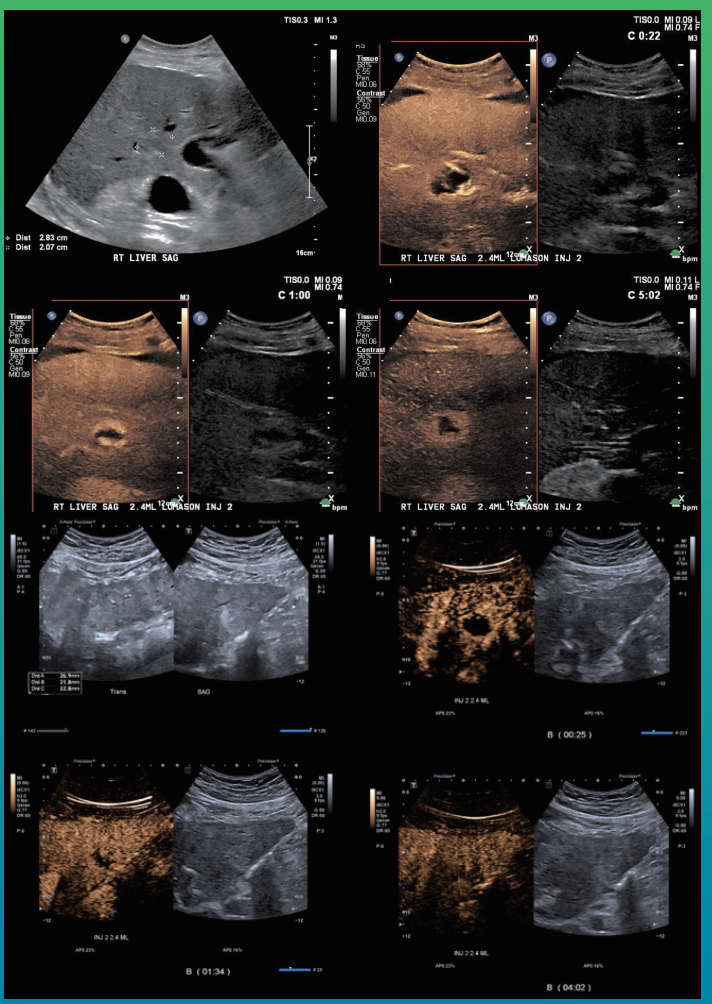
in Diagnosis and Therapy

December, 2023

Volume: 7

Issue: 4

Pages: 313–422



The official journal of the CAUSA
China-Amcrica Ultrasound Scholar Alliance

https://www.AUDT.org
ISSN: 2576-2508 (Print)
ISSN: 2576-2516 (Online)

CONTENTS

Case Reports

REVIEW ARTICLE

Semi-supervised Learning for Real-time Segmentation of Ultrasound Video Objects: A Review

Jin Guo, MD [a,1], Zhaojun Li, PhD [b,1], Yanping Lin, PhD [a,]*

[a] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China; [b] Department of Ultrasound, Shanghai General Hospital Jiading Branch, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Received March 30, 2023; revision received April 7, 2023; accepted April 22, 2023

Abstract: Real-time intelligent segmentation of ultrasound video object is a demanding task in the field of medical image processing and serves as an essential and critical step in image-guided clinical procedures. However, obtaining reliable and accurate medical image annotations often necessitates expert guidance, making the acquisition of large-scale annotated datasets challenging and costly. This presents obstacles for traditional supervised learning methods. Consequently, semi-supervised learning (SSL) has emerged as a promising solution, capable of utilizing unlabeled data to enhance model performance and has been widely adopted in medical image segmentation tasks. However, striking a balance between segmentation accuracy and inference speed remains a challenge for real-time segmentation. This paper provides a comprehensive review of research progress in real-time intelligent semi-supervised ultrasound video object segmentation (SUVOS) and offers insights into future developments in this area.

Key words: Ultrasound video segmentation; Semi-supervised learning; Real-time segmentation; Video object segmentation

Advanced Ultrasound in Diagnosis and Therapy 2023; 04: 333-347 DOI: 10.37015/AUDT.2023.230016

Ultrasound imaging is a widely used medical imaging technique that can dynamically display images of internal organs and tissues in real-time, playing a crucial role in image-guided interventions and clinical diagnosis [1]. However, the low signal-to-noise ratio and indistinct texture of ultrasound images make it challenging to achieve robust, high-precision, and rapid segmentation. Furthermore, acquiring well-labeled data is costly for medical image datasets, which require experts to provide reliable and accurate annotations [2]. To address the challenges of medical image segmentation tasks and reduce the burden of manual annotation, researchers have explored efficient utilization of medical image datasets and proposed semi-supervised learning (SSL) [3] solutions, such as generating pseudo-labels [4] and employing different data augmentation methods in consistency learning [5]. SSL is a practical approach for dealing with medical image data where data is scarce and labeling is expensive. It aims to use a small amount of labeled data while training the network with large-scale unlabeled data to achieve the same segmentation performance as supervised learning models. As medical images are more readily available than annotations, SSL methods are widely used in various medical image segmentation tasks. However, effectively exploiting unlabeled data and deriving useful information from it remains a challenge for SSL.

This paper focuses on a series of studies that have employed semi-supervised learning (SSL) methods to address real-time segmentation tasks in ultrasound videos, known as semi-supervised ultrasound video object segmentation (SUVOS). In recent years, SSL has

[1] *Jin Guo and Zhaojun Li have contributed equally to this study.*

[*] *Corresponding author: School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China*
e-mail: yanping_lin@sjtu.edu.cn

shown promising results in SUVOS scenarios. We first introduce the concepts and fundamental assumptions of SSL, its common approaches in segmentation tasks, and its applications in video object segmentation (VOS) tasks. Subsequently, we discuss recent research in SUVOS, comparing and summarizing the advantages and disadvantages of these methods. Finally, we explore the challenges and future research directions in this area, providing insights for researchers to develop more accurate and efficient real-time intelligent segmentation algorithms for SUVOS.

Preliminary investigation

Semi-supervised learning on image segmentation

Semi-supervised learning (SSL) is a machine learning method that combines supervised and unsupervised learning and is typically employed to enhance the accuracy of classification or regression tasks when labeled data is limited. In SSL, the training data consists of both labeled samples (with explicit output values or classifications) and unlabeled samples. Compared to supervised learning using only labeled data, SSL exploits the potential information of unlabeled data to improve learning performance. As medical image datasets are very expensive to annotate, SSL methods that can utilise small quantities of annotated data are widely used in medical image lesion segmentation and classification.

SSL methods are commonly based on three assumptions [6]: the cluster assumption, the manifold assumption and the smoothness assumption. The cluster assumption posits that samples from the same class should exhibit similar class labels within the feature space. The manifold hypothesis assumes that sample data within the same local neighbourhood have similar properties, and embedding high-dimensional data into a low-dimensional manifold should result in samples having similar labels when they are located within a small local neighbourhood of the low-dimensional manifold. Conversely, the smoothness assumption presumes that two samples in close proximity within a dense data region share similar class labels, meaning that when two samples are linked by edges in a dense data region, they likely share the same class label.

Semi-supervised segmentation with pseudo labels

The SSL model, which employs pseudo-labels, seeks to generate usable labels for unlabeled data to augment the dataset. This approach boasts the advantages of simplicity and ease of implementation, facilitating effective utilization of unlabeled data. Nonetheless, a drawback lies in the susceptibility to errors during the generation of pseudo-labels, leading to potential instability in their overall quality. Enhancing the confidence level of pseudo-labels remains a significant challenge for this method. Pseudo-labels-based methods can be categorized into self-training and co-training, contingent upon whether a single model is used for parameter updates or not.

Self-training [7], a simple semi-supervised method utilizing pseudo-labels, typically involves a network containing a single supervised segmentation model. Pseudo-labels are determined by the highest confidence prediction of this segmentation model. During training, data iteration occurs exclusively over this model. Bai et al. [8] refined segmentation results by implementing conditional random fields (CRF) and employed the optimized segmentation results as pseudo-labels for iterative processes. Chaitanya et al. [9] introduced local contrast loss to promote similarity in representations of pixels with identical labels (or pseudo-labels). He et al. [10] developed Distribution Alignment and Random Sampling (DARS) methods to acquire unbiased pseudo-labeling data that align with the true distribution. Some researchers discovered that incorporating data augmentation techniques into the self-training process yielded more advantageous model iterations. Yang et al. introduced ST++ [11], which prioritizes the most reliable pseudo-labels during iterative processes and mitigates overfitting of noisy pseudo-labels through data augmentation.

The self-training approach does not possess a mechanism for detecting inaccurate labels. As a result, co-training extends the self-training approach to enhance network performance by training multiple models concurrently.

Co-Training [12] is a disagreement-based approach that predominantly depends on the disparities in predictions between models and aims to minimize disagreement by assigning pseudo-labels across multiple views of unlabeled data. It posits that each training instance can be characterized by two complementary view feature sets. The method employs labeled data to train classifiers for each view, utilizes these classifiers to categorize unlabeled samples, selects the pseudo-labels with the highest confidence along with their original images to incorporate into the training set, and subsequently retrains the classifiers. The two primary objectives are to promote differences between views on the data and to estimate confidence in the predictions. Yao et al. [13] introduced a confidence-aware cross-pseudo-supervised algorithm to enhance the quality of pseudo-labels through a confidence-aware regularization, measuring the pseudo-variance between the original image and its Fourier-transform-enhanced counterpart. Lee D et al. [14] incorporated the concepts of low-density separation and entropy regularization to facilitate

the generation of pseudo-labels. Xia et al. [15] proposed an uncertainty-aware Mean Teacher [16] framework, which employs uncertainty estimation to select reliable pseudo-labels from the teacher model's predictions and provide them to the student model for learning.

Semi-supervised segmentation with consistency regularization

In accordance with the smoothness assumption, a model should produce consistent predictions for an image and its perturbed counterparts. Semi-supervised segmentation methods that leverage consistency regularization make use of unlabeled data by incorporating perturbations. These methods involve adding a regularization item to the loss function, which measures the discrepancy between the original image prediction and that of the perturbed image. Laine et al. [17] proposed two models: Pi Model and Temporal Ensembling. The Pi Model is forward-propagated twice during training using identical input data, with each propagation introducing a distinct random perturbation to the input data and calculating the consistency loss to ensure the model's predictions remain consistent on the unlabeled data after both perturbations. This method of forward propagation twice reduces computational efficiency, and the authors present an alternative model, Temporal Ensembling, which computes an exponential moving average (EMA) of the current model output predictions with the predictions obtained from past segmentation, with only single input to the model per epoch. In comparison to Temporal Ensembling, the Mean teacher [16] introduces a separate network of teachers, as opposed to relying solely on historical predictions. The teacher model updates the parameters in real-time by applying an EMA to the student network. This allows the teacher model to make better use of the dynamic information available to the model during the training process. In addition to exploring regularized input methods, some researchers have focused on incorporating perturbations. This approach typically involves adding perturbations directly to the input image using data augmentation techniques. Li et al. [18,19] transformed the input image and calculated the unsupervised loss by assessing the difference between network predictions under various transformations. Bortsova et al. [20] investigated the impact of the isotropy of elastic deformations on model performance improvement. CutOut [21] method is to randomly select an area in the original image and set its pixel value to zero or other fixed value, and the model is equivalent to adding some images with occlusion in training. New training samples are generated by merging these modifications. Following data augmentation with these two methods, a regularization item enforces consistency between the original image and the predicted results of the modified image. Olsson et al. [22] proposed a novel data augmentation mechanism, ClassMix, for semi-supervised semantic segmentation. Similar to CutMix, ClassMix generates new samples by blending pixel-level labels from two distinct images, also relying on mask blending.

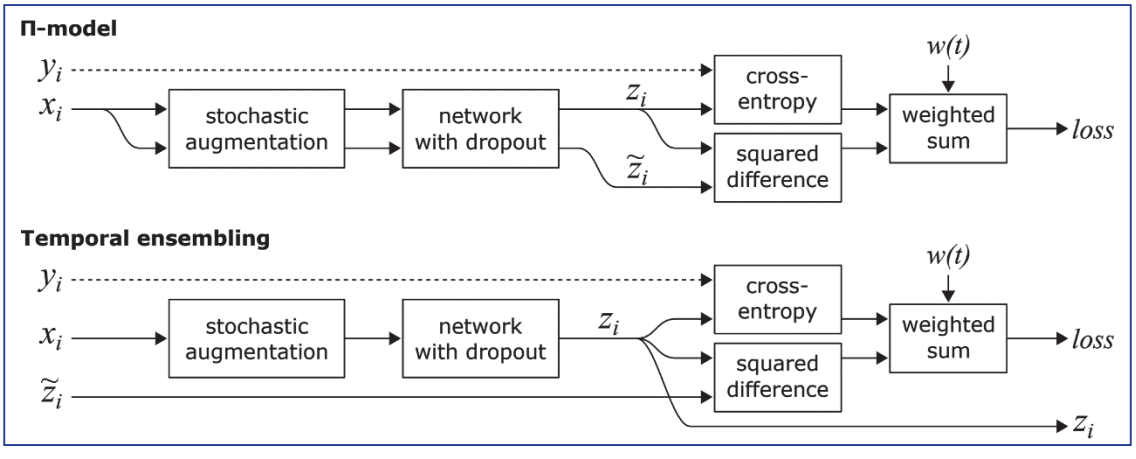


Figure 1 Pi Model and Temporal ensenmbling model structure [17].

Semi-supervised segmentation with comparative learning

The fundamental concept of contrast learning involves obtaining meaningful feature representations from unlabeled data by learning to differentiate between similar and dissimilar positive and negative sample pairs. Due to the absence of data annotations, similar samples are treated as augmentations of the same sample during training, while other data are considered distinct samples. Alonso et al. [25] applied contrast learning to

semi-supervised semantic segmentation tasks in order to learn similar class features between labeled and unlabeled datasets. A self-supervised learning structure was employed during the training process to learn feature representations using only positive sample pairs. Moreover, a memory bank was constructed to store high-confidence samples learned by the teacher model on the labeled dataset, with a contrast loss function ensuring that sample features were closely aligned with similar sample features stored in the memory bank. Lai et al. [26] primarily addressed the issue of semi-supervised semantic segmentation overfitting on a limited number of labeled samples. They proposed maintaining context-aware consistency across contexts and developed a directed contrast loss that achieves consistency on a pixel-to-pixel basis, requiring only lower quality features to be aligned with their corresponding objects.

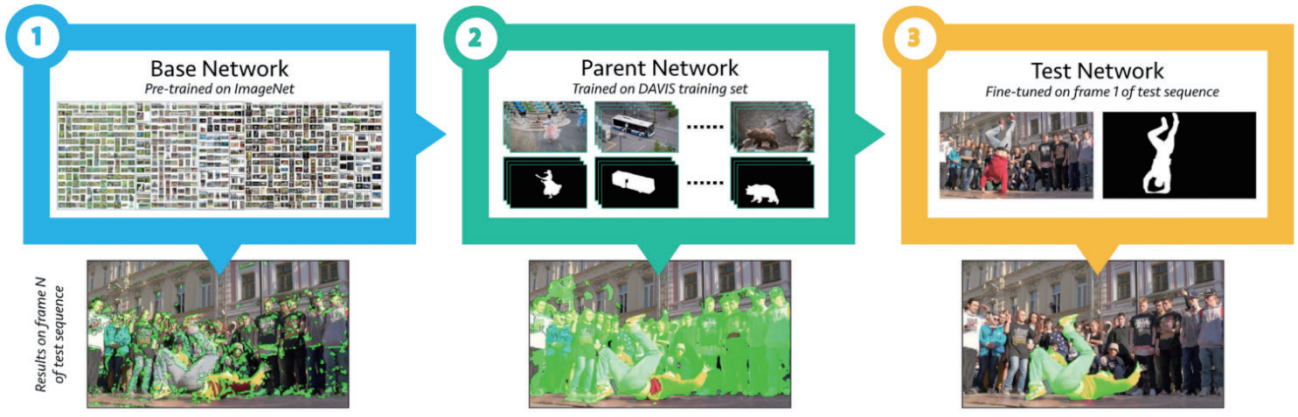


Figure 2 Overview of the OSVOS structure [27].

Semi-supervised Video Object Segmentation

Video object segmentation (VOS) involves the continuous localization and segmentation of user-defined objects of interest within video frames, achieved through the application of segmentation algorithms. Semi-supervised video object segmentation represents the predominant task within this category. Primarily, it tackles the one-shot learning problem, which entails providing pixel-level annotations for the initial frame of a video sequence and allowing the algorithm to automatically segment the specified object in subsequent frames. The prevailing algorithms encompass online fine-tuning-based methods, mask-based propagation techniques, and matching-based approaches.

Online fine-tuning-based methods

The concept of online fine-tuning methods was first introduced in OSVOS [27]. The core idea is to train a parent model on a large dataset so that the model can distinguish general features of foreground objects, fine-tune the model online at test time based on the image and its annotation in the first frame, and then segment the object frame by frame in subsequent frames. OnAVOS [28] attempts to address the limitations of the OSVOS method in adapting to drastic changes in object appearance by introducing an online adaptive mechanism. OSVOS-S [29] facilitates the segmentation of subsequent frames by extracting instance semantic information from the first frame through an instance segmentation network.

These methods do not consider the temporal association of video frames, processing each frame independently. Fine-tuning the parameters by training on the first frame results in high-quality segmentation in subsequent frames, demonstrating robustness to occlusion and object loss. However, the lack of available temporal information makes these networks less capable of handling changes in object appearance. Additionally, the requirement for online learning for each video during the testing phase to update model parameters poses challenges in meeting real-time requirements for segmentation speed. Thus, careful adjustment of the network structure and training strategy is necessary to achieve optimal performance in practical applications.

Propagation-based methods

The online learning approach faces challenges in achieving real-time segmentation; thus, some methods use mask propagation to accomplish semi-supervised VOS. These techniques primarily leverage the inter-frame temporal information of the video to predict the current frame mask by propagating the semantic information from the previous frame and its segmentation result to the current frame. One type of solution involves

utilizing optical flow to propagate the mask. For instance, MaskTrack [30] employs affine transformation and non-rigid deformation on the mask of the previous frame to obtain the current frame prediction. MaskRNN [31] predicts multiple object instance-level segmentation using estimated optical flow and the object's bounding box. PReMVOS [32] applies optical flow between consecutive image pairs to propagate the proposed mask to the next frame, computing the temporal consistency between the two proposed masks and enhancing segmentation accuracy by refining the network.

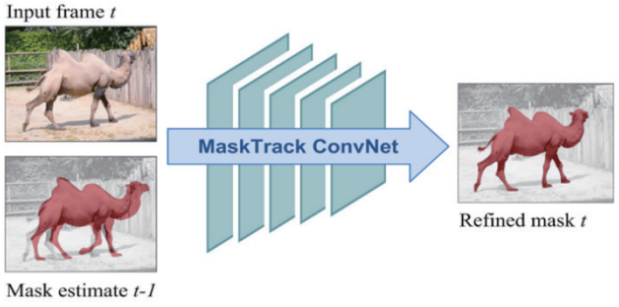


Figure 3 MaskTrack mask propagation process [30].

In addition to optical flow, a category of VOS methods is based on object tracking. FAVOS [33] developed a part-based tracking method that predicts object masks from multiple tracking frames within an object part. SAT [34] utilizes inter-frame consistency to treat each object as a tracelet, predicting the mask for each tracelet by cropping the search area around the object. SiamMask [35] considers the bounding box of a trace as an approximation of the mask, generating a segmentation mask by adding mask branches in SiamRPN [36]. This approach narrows the gap between target tracking and target segmentation, improving the speed of tracking and segmentation. Although these methods achieve favorable results in terms of segmentation accuracy and real-time performance, they are less robust to occlusions and drifts and may suffer from error accumulation in the segmentation of subsequent frames.

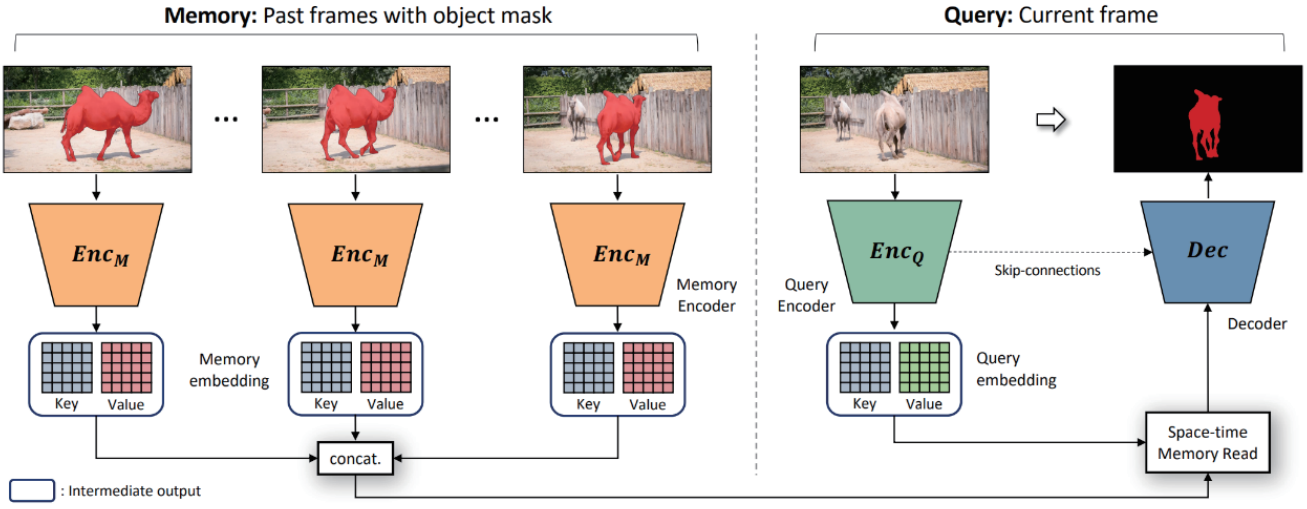


Figure 4 Overview of the STM framework [39].

Matching-based methods

In recent years, matching-based approaches have achieved exceptional results, garnering considerable attention as the most promising VOS solution. The fundamental concept involves capturing rich target information by learning pixel-level similarity or distance maps between query frames and historical frames in the embedding space. Most of these methods utilize only the first frame and adjacent frames or key frames sampled by parameters. PLM [37] conducts pixel-level matching of objects initialized in the first frame to subsequent images using twin networks. VideoMatch [38] devises a soft matching mechanism that computes the similarity between the current frame and the first frame at the feature level. Among the most well-known in recent years is STM [39], which proposes a space-time memory model that explicitly stores previously computed segmentation results in external memory. This memory mechanism facilitates the neural network in learning the temporal evolution of objects and exhibits remarkable performance, particularly when leveraging historical segmentation results. However, this approach suffers from space-time redundancy and performs unnecessary matching in untargeted regions, rendering the segmentation speed less than real-time. Several scholars have made improvements based on STM. Lu et al. [40]

perform in-memory updates by storing and extracting target information from external memory, and they fine-tune the segmentation model online to fit specific targets. STM-cycle [41] applies a cyclic approach to training and implementing segmentation networks to reduce incorrect segmentation in memory frames. Swiftnet [42] develops a more efficient memory mechanism to reduce the unnecessary accumulation of similar images through a Pixel-Adaptive Memory (PAM) module that adaptively selects whether to update memory frames. Matching-based approaches offer a great deal of flexibility in model design and can handle long-term correspondence. However, this approach relies on robust and generalized feature representation, which may limit its performance in some challenging scenarios.

Real-time Segmentation of Semi-supervised Ultrasound Video Objects (SUVOS)

Semi-supervised learning (SSL) has achieved impressive results in the field of natural image and video object segmentation. Ultrasound images are widely employed in clinical practice due to the excellent real-time performance, non-invasiveness, no radiation, and ease of operation. In recent years, some researchers have also applied SSL and semi-supervised object segmentation methods to ultrasound video segmentation tasks. In this section, we review and discuss representative methods in semi-supervised ultrasound object segmentation tasks, comparing the performance of various models in this domain.

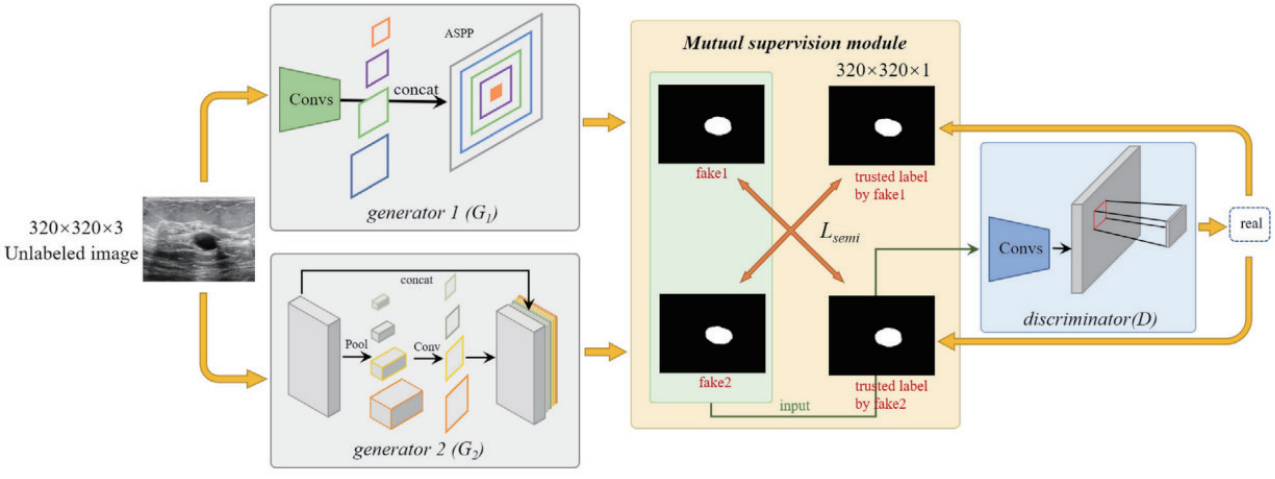


Figure 5 ASS-GAN semi-supervised flow using unlabeled images [48].

Pseudo-labels-based SUVOS

Semi-supervised segmentation algorithms utilizing pseudo-labels through iterative models are prevalent in the field of natural images, and some researchers have also applied these methods in ultrasound image segmentation scenarios.

Wang et al. [43] developed a Unet [44] based segmentation network to segment ultrasound thyroid nodules. Initially, the segmentation model was trained with labeled data; subsequently, unlabeled data was fed into the segmentation model to obtain predictions as pseudo-labels, which were used to retrain the network for updating parameters. Li et al. [45] designed a semi-supervised segmentation network based on the Temporal Ensembling (TE) method. After each training, the current model made predictions for the entire training set to generate pseudo-labels for each sample. These pseudo-labels were then averaged with the original labels to obtain the label average for the entire training set. Cao et al. [46] proposed an uncertainty TE segmentation model to reduce the negative impact of unreliable pseudo-labeled samples on the training process. This model employed an integration strategy to generate relatively stable pseudo-labels and incorporated an uncertainty mapping to further ensure the reliability of these pseudo-labels. Li et al. [47] generated pseudo-labels for unlabeled data by smoothing network predictions using a simple linear iterative clustering (SLIC) superpixel algorithm.

ASS-GAN [48] is an approach that incorporates two generators and a discriminator. In the first step, the generator and discriminator are trained with labeled data, endowing them with fundamental segmentation and discrimination capabilities. During the second step, unlabeled data is introduced to the two generators, which generate predictions. The discriminator evaluates its output predictions and employs high-confidence predictions as pseudo-labels for the other generator. These two generators form a mutual supervision

module, supplying each other with supervisory signals. Adversarial training is utilized to further constrain the distance between the pseudo-label and the true result, thereby facilitating the mutual supervision module's training. Huang et al. [49] integrated a fuzzy feature generator and a multi-scale fuzzy entropy module into the GAN-based network to discern the differences between the uncertainty map and the input.

In contrast to ASS-GAN, which employs generators as segmentation networks to produce pseudo-labels for unlabeled data, Iqbal et al. [50] and Pang et al. [51] utilized GAN networks to synthesize ultrasound breast tumor images to augment the dataset. Specifically, Pang et al. [51] employed a semi-supervised GAN architecture trained on unlabeled ultrasound breast data to generate synthetic breast tumor images. PDF-Unet [50] designed a self-encoder network, data expansion network (DEN), based on GAN and used the DEN to generate pseudo-images of ultrasound breast tumors to expand the dataset. The quality of the synthetic images was improved through adversarial training during the training process. In each iteration, the authors first fed unlabeled ultrasound data into the DEN, then employed a discriminator to differentiate between real and synthetic images, and updated the DEN parameters. A segmentation network was also trained using labeled data and used to generate probability maps for the pseudo-images. Finally, the generated images and probability maps were fused with the labeled data to train the primary segmentation network PDF-Unet.

However, insufficient supervision of the discriminator may impact the quality of the fake samples generated by the generator. Han et al. [52] effectively ensured the quality of predictions generated by the generator by designing a dual-attention fusion block to better extract representative features of the lesion region and background.

Consistency regularization SUVOS

Xu et al. [53] introduced a shadow consistency method for segmenting ultrasound prostate images based on consistency learning. This approach employs shadow enhancement and shadow removal mechanisms to encourage the segmentation network to extract features from shadow-free regions at both the image and feature levels. Chen et al. [54] developed a semi-supervised segmentation network based on the Mean Teacher method to assist robots in vascular puncture. While Mean Teacher has achieved some success in SSL tasks, its consistency loss is more sensitive to noise. Xie et al. [55] proposed a prior knowledge-guided consistent regularization method for semi-supervised breast cancer diagnosis in ultrasound images. This approach incorporates physicians' domain knowledge into sample perturbations, enabling the model to generate consistent predictions for unlabeled images and perturbed samples. Mendel et al. [56] transferred the Mean Teacher from consistent regularization to error correction for semi-supervised segmentation by adding a correction network. This network accepts the output of the segmentation network as input and produces a corrected segmentation result.

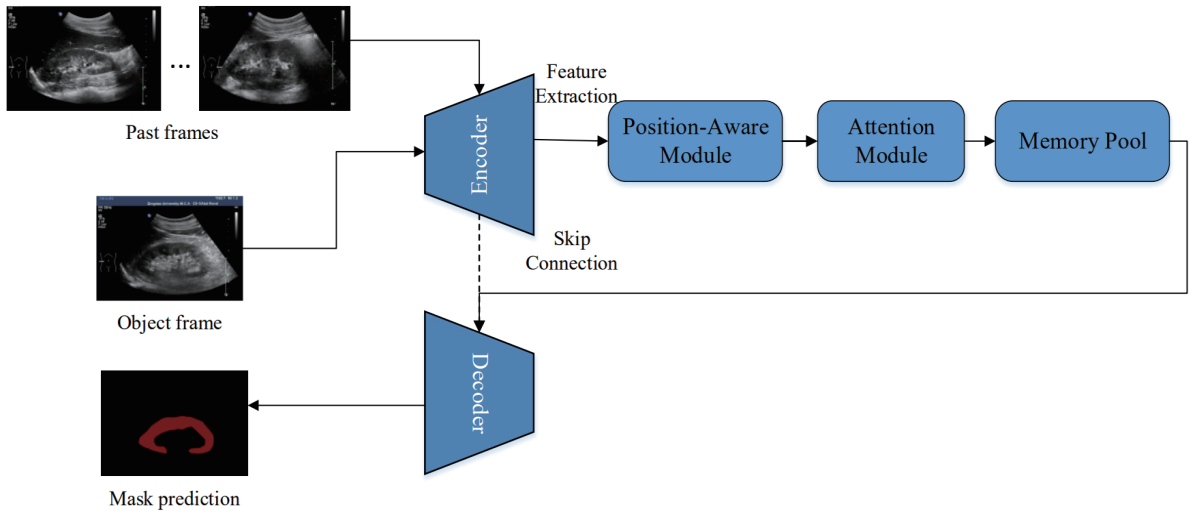


Figure 6 Ultrasonic renal parenchymal segmentation network based on the STM framework [57].

Based on video object segmentation SUVOS

The measurement of renal parenchyma area in the kidney is correlated with acute and chronic renal diseases, and the thickness of renal parenchyma serves as an indicator of chronic renal failure. This measurement can be obtained from ultrasound images. Wang et al. [57] proposed a deep learning method for kidney parenchyma segmentation in kidney ultrasound videos, capitalizing

on the temporal information present in video frames. Specifically, the authors designed a network based on the Space-Time Memory (STM) model and attention mechanism [58], in which video frames are input into an encoder, and an image mask is provided for the first frame. Subsequently, starting from the second frame, video frames are processed sequentially as query frames. The encoder enhances the model's robustness against changes in target appearance by extracting features from the input image and calculating the feature similarity between the query frame and the memory frame within a position-aware module. The model then determines whether each feature's mask information belongs to the foreground or background by incorporating an attention module. During the memory unit's reading process, the corresponding weights are initially computed by assessing the pixel similarity between the input and historical frames. Finally, a pre-trained decoder is employed for performing binary classification tasks for the foreground and background, while the decoder undergoes a limited number of iterations to fine-tune the parameters on the ultrasound data.

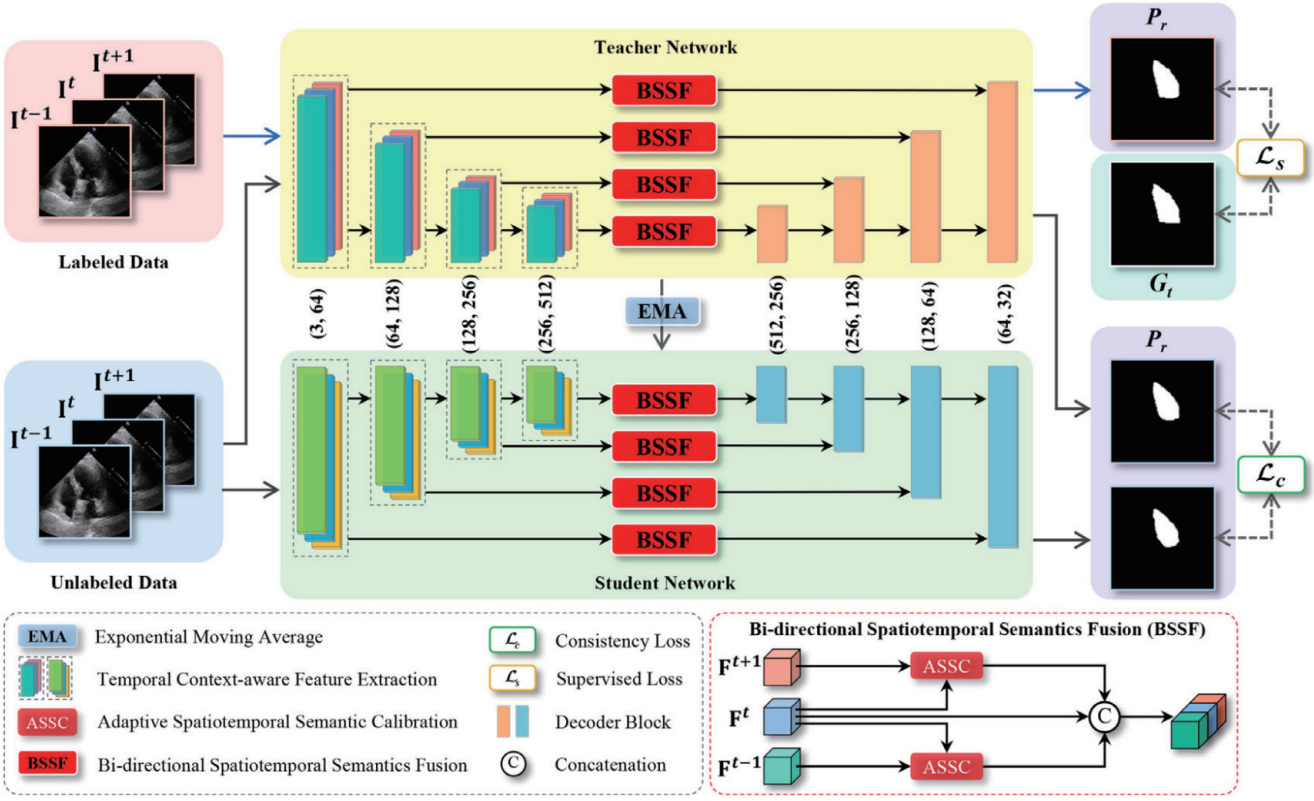


Figure 7 Semi-supervised echocardiography video segmentation framework based on Mean Teacher [60].

Echocardiography is extensively employed in diagnosing various cardiovascular diseases, ranging from heart failure to valvular heart disease, due to its ease of use, affordability, and absence of radiation [59]. Accurate diagnosis of cardiovascular disease based on echocardiography relies on the precise segmentation of several key features and the measurement of critical indicators of cardiac function. Wu et al. [60] emphasized the significance of temporal coherence in cardiac motion for ultrasound video to enhance model segmentation accuracy and efficiency, in contrast to processing frame-by-frame video input without considering inter-frame coherence. Directly applying methods from natural image segmentation to ultrasound images is ineffective due to the inherent scattering noise in ultrasound images. The authors propose a model that addresses the challenges of ultrasound video segmentation accuracy and real-time segmentation by exploiting the space-time coherence between frames in the presence of scattered noise. Specifically, the model is implemented based on the Mean Teacher semi-supervised framework and primarily consists of: 1) An adaptive space-time semantic calibration module for aligning feature maps of consecutive frames; 2) A temporal context-aware feature extraction module; 3) A weight learning module. During training, the model receives input from three consecutive frames of images and achieves accurate segmentation of the current frame by extracting and calibrating the contextual features of adjacent frames. However, this training method is specific to offline video, and its

effectiveness for online video transmission remains unknown. Ta et al. [61] proposed a semi-supervised joint learning network utilizing a bridging structure with two branches—one for motion tracking and the other for segmentation—while incorporating physiological constraints as shape priors to enforce realistic heart motion.

El Rai et al. [62] viewed VOS as a node classification problem and proposed GraphECV, a semi-supervised VOS method based on graph signal processing, and applied it to echocardiogram video segmentation. Specifically, they initially employed FgSegNet_S as the segmentation network, then extracted optical flow, texture, and statistical features from each instance, connecting them to represent the vertices of the graph. Subsequently, the graph was constructed utilizing the K-nearest neighbor algorithm, which connected the K-nearest neighbors of each node or vertex, followed by graph sampling performed by embedding the graph with a limited quantity of labeled data. Ultimately, all nodes were labeled and reconstructed using semi-supervised Sobolev parametric minimization techniques to classify nodes as either left ventricle or background.

Chen et al. [63] highlighted that backbone models, pre-trained on natural image datasets, experience significant performance degradation when processing medical images. However, due to the scarcity of medical images, publicly available models pre-trained for ultrasound images are limited. To address the data shortage issue, the authors initially constructed an ultrasound video dataset. They then developed a semi-supervised contrastive learning method to train a generic model using ultrasound video for downstream tasks, such as ultrasound image segmentation. This approach demonstrated performance advantages over ImageNet pre-training methods across multiple downstream tasks.

Sirjani et al. [64] developed EchoRCNN, a video object segmentation network for echocardiograms, aimed at extracting cardiac features from echocardiogram sequences. The network is built upon the robust image segmentation architectures of Mask R-CNN, RetinaNet, and the RGMP video object segmentation network. Specifically, the network comprises a siamese encoder, a Feature Pyramid Network (FPN), and three sub-networks for classification, regression, and segmentation. Additionally, the model incorporates a recurrent neural network component for processing time series data. The training process involves a main stream and a reference stream, based on the RGMP network. The main stream receives the current frame and the prediction mask of the previous frame, while the reference stream takes the first image frame and its annotations as input. The input passes through the siamese encoder and FPN before entering the three sub-networks for classification, regression bounding box, and segmentation tasks. Finally, the prediction mask undergoes post-processing to extract the primary parameters of left ventricular (LV) or right ventricular (RV) function, enabling the detection of end-diastolic and end-systolic frames.

Other methods on SUVOS

Song et al. [65] introduced a semi-supervised segmentation method that utilizes continuous minimum cut blocks of superpixels and neighborhood patches. Superpixels are blocks of pixels composed of neighboring pixels with similar intensities. This approach aims to improve segmentation performance by complementarily combining the position information of each superpixel as coarse features with the grayscale information of the neighborhood patches as fine texture features. Shin et al. [66] performed breast ultrasound tumor segmentation by combining bounding box regression and weakly supervised learning methods. Dai et al. [67] integrated target detection and unsupervised learning for the segmentation of prostate ultrasound images, incorporating the YOLOv5 target detection network and the C2Fnet [74] unsupervised segmentation network. Specifically, the detection model first identifies the target, and the region of interest is extracted. The extracted slices are then fed into the segmentation network for segmentation. Judge et al. [68] proposed a semi-supervised segmentation method that optimizes the non-differentiable anatomical prior, with the anatomical prior providing the regularization term. Yang et al. [69] improved segmentation performance by learning image semantic information from various hybrid loss functions, such as uncertainty loss and context-constrained loss, performing uncertainty estimation in both models. [70] suggested a cyclic self-supervised method based on heartbeat cycles to enhance feature similarity and employed teacher-student distillation to improve segmentation using pseudo-labels on unlabeled ultrasound videos.

Wang et al. [71] developed a semi-supervised bone shadow segmentation method for conditional generative adversarial networks based on multi-task learning. The proposed CNN model consists of a shared encoder and two independent multiscale decoders for coarse bone shadow enhancement (BSE) and horizontal bone interval mask (HBIM), respectively. Specifically, the BSE decoder output is an enhanced bone shadow image, and the HBIM decoder output is a one-dimensional row vector representing the horizontal spacing of the bone shadow. Then the outputs of these two tasks are processed by a masking operation to generate the final bone shadow segmentation. By multiplying the pixel

values in the BSE image with the HBIM vector, pixel values in the BSE image not belonging to the bone shadow region can be set to zero, ultimately obtaining the bone shadow mask. Moreover, by introducing a conditional shape discriminator, shape information-based adversarial loss is added to guide the training of the bone shadow segmentation network, further regularizing the output bone shadow shape.

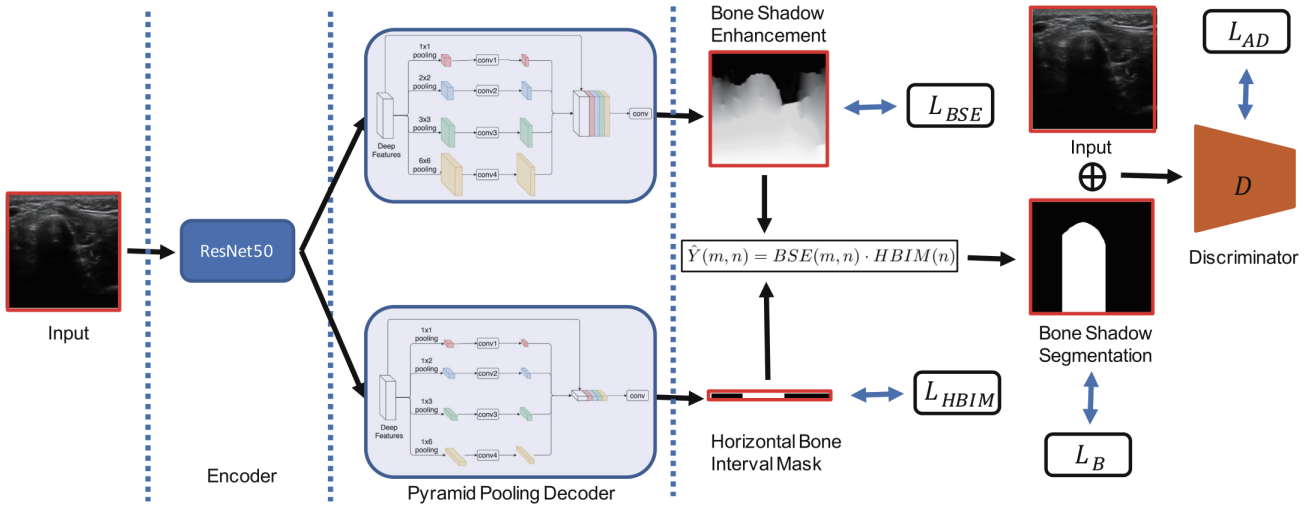


Figure 8 Ultrasonic bone shadow segmentation method based on multi-task learning [71].

In summary, real-time semi-supervised ultrasound video object segmentation methods demonstrate significant advantages in processing medical images. These techniques have achieved remarkable success in handling temporal information and enhancing segmentation accuracy as well as real-time performance. The majority of researchers have tackled this issue by combining SSL with semi-supervised VOS methods, encompassing approaches such as pseudo-labeling-based methods, consistency learning, and video object segmentation-based methods. Additionally, there are solutions integrating other techniques like superpixels, weakly supervised learning, target detection, and unsupervised learning. These methods have been employed in various clinical ultrasound tasks, including thyroid nodule segmentation, breast tumor segmentation, prostate image segmentation, renal parenchyma segmentation, and echocardiogram segmentation.

Evaluation Indicators and Methods

In this section, we discuss the metrics and methods employed to evaluate real-time ultrasound video object segmentation algorithms. Evaluation metrics and methods are crucial for assessing the performance of different approaches and assisting researchers in comprehending the practical performance of various techniques. Generally, these methods should be evaluated across multiple aspects, including segmentation accuracy, inference time, and memory footprint. We provide an overview of commonly used metrics for gauging the performance of segmentation algorithms and elaborate on real-time metrics.

Segmentation accuracy is the core metric for evaluating the performance of image segmentation algorithms. Pixel Accuracy is used to evaluate the accuracy of the predicted pixels. It is calculated by dividing the number of correctly predicted pixels by the total number of pixels. Pixel Accuracy is defined as Eq1:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{1}$$

The Intersection over Union (IoU) or Jaccard index is one of the most commonly used metrics in semantic segmentation. It is used to evaluate the difference between the predicted mask and the true annotations. It is defined as the intersection of the predicted mask and the true annotation divided by the concatenation of the predicted mask and the true annotation, as in Eq2:

$$IoU = J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Where A and B represent the true annotation and the prediction mask respectively. Another metric, Mean-IoU (mIoU), is the average intersection ratio of all classes and is also widely used in segmentation algorithms. It can be expressed in an alternative form, as shown in Eq3:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{3}$$

Where TP is the true positive score, FP is the false positive score, FN is the false negative score and TN is the true negative score. Using the foreground-background dichotomy as an example, as shown in Table 1:

Table 1 Evaluation of accuracy of dichotomous results

True value	Predicted value	
	Prospects	Background
Prospects	TP	FN
Background	FP	TN

Precision, Recall and F1 score are common accuracy evaluation metrics for image segmentation model applications. Precision refers to the proportion of samples that actually belong to a positive class among all samples identified as positive classes. It measures the ability of the model to correctly identify positive classes and is calculated as Eq4:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall refers to the proportion of samples that are correctly identified as positive classes out of all samples that are actually positive classes. It measures the extent to which the model correctly identifies samples in the positive class, as shown in Eq5:

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

The F1 score is the combined mean of Precision and Recall, which combines the performance of Precision and Recall and can be used to assess the overall performance of the model when both metrics are considered. the F1 score is calculated as Eq6:

$$F1\ score = \frac{2\ Prec\ Rec}{Prec + Rec} \tag{6}$$

The Dice Coefficient is a metric used to measure the geometric similarity of two samples. It is often used as a metric to assess model performance in medical image segmentation and is calculated as Eq7:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN} \tag{7}$$

When applied to a binary classification problem, the Dice coefficients and F1 scores are equivalent.

Inference speed and computation are also vital metrics when evaluating the performance of an image segmentation model. They reflect the efficiency and resource consumption of the model in applications and are particularly crucial for real-time scenarios and equipment requirements. Inference speed is the time required for a model to make a single prediction, usually expressed as the prediction time per sample. It is influenced by various factors such as hardware configuration, model complexity, and optimization methods.

Computation amount refers to the computational resources required by a model to perform inference or training. Commonly used computational metrics include Floating Point Operations (FLOPs) and the Number of Parameters (Params). Params refers to the total number of trainable parameters in the model, reflecting the storage requirements and memory usage of the model. More parameters may render the model unworkable on resource-limited devices or affect the speed of inference.

Additionally, in video object segmentation tasks or some tasks with real-time requirements, FPS (Frames Per Second) is a key performance metric that measures the real-time performance of a model when processing video sequences, reflecting how quickly the model can predict and segment between consecutive frames. A high FPS means that the model can complete the processing of video frames in a shorter time, thus meeting the requirements of real-time processing.

When evaluating different real-time segmentation algorithms for semi-supervised ultrasound video objects, metrics such as segmentation accuracy, inference speed and computation, and FPS can be taken into account to identify the best model that balances real-time performance with segmentation accuracy.

Analysis

SUVOS is focused on video segmentation, making continuous image ultrasound video datasets essential to drive this type of research. Currently, there are limited public datasets for ultrasound video, with CAMUS [72] and EchoNet-Dynamic [73] being among the better-known echocardiography video datasets.

The CAMUS dataset is a publicly available dataset for cardiac ultrasound image segmentation and includes sequences of two- and four-chamber views collected from 500 volunteers. Each sequence contains approximately 20 frames of images with two frames of expert annotation. EchoNet-Dynamic is a large dataset of 10,030 annotated echocardiographic videos, including cardiac cases and health samples. Each echocardiogram video is approximately 200 frames.

These two datasets provide rich cardiac ultrasound images and detailed structural annotations, serving as important benchmark datasets in the field of cardiac ultrasound image segmentation. Researchers can use them to evaluate and improve their own segmentation algorithms.

Table 2 Comparison of semi-supervised ultrasound video object segmentation models in real-time

Author	Method (Based on)	Dice/mIou	FPS	Device	Dataset
Zhai D et al. [48]	Generative adversarial networks	0.8319/0.7123	17.48	1080Ti	DBUS (15% labeled)
Pang T et al. [51]	Generative adversarial networks	0.94	33.2	2060Super	Mendeley US
Wang P et al. [71]	Generative adversarial networks	0.962/0.9297	33.33	TiTan Xp	Bone US
Cao X et al. [46]	Consistency regularization	0.7287		2x2080Ti	ABUS (300 labeled)
Wu H et al. [60]	Consistency regularization	0.9379	31.25	2080Ti	CAMUS
Xie X et al. [55]	Consistency regularization	0.7951		V100	BUSI (10% labeled)
Sirjani N et al. [64]	Detection-based	0.7220	12.5	2060-A8G	Echocardiography series
Dai F et al. [67]	Detection-based	0.7594/0.6271			BUS
Wang R et al. [57]	Matching-based	0.8387	Real-time (Not verified)	2080Ti	Renal Parenchyma
B Li Y et al. [47]	Simple linear iterative clustering super-pixel	0.8823/0.8495			BUS (100 labeled)
Song X et al. [65]	Super-Pixel	0.900		CPU	MUS
El Rai M C et al. [62]	Graph Signal Processing	0.9270	Real-time (Not verified)		CAMUS (5% labeled)
Yang H et al. [69]	Uncertainty and contextual constraint loss	0.65	50	1080Ti	Heart US

We have compiled a table comparing the performance of the algorithms discussed earlier, utilizing values from the corresponding papers and indicating the datasets they employ. In Table 2, we categorize and evaluate six main aspects, including the method used, the proportion of unlabeled data, Dice and IoU, the real-time metric FPS, hardware information, and the dataset. As different methods have varying optimization approaches and hardware, and different models focus on different datasets, this information is provided for reference only. In terms of real-time performance, several models [48,50,57,60,62,64,71] have achieved this. Notably, Wang et al. [57] and El Rai et al. [62] do not mention specific real-time metrics in their papers. Yang H et al. [69] and Wang P et al. [71] along with previous researchers, did not provide real-time performance metrics. We have calculated the FPS from the training or testing time and batch size for reference purposes only. Most approaches concentrate solely on segmentation accuracy, which is difficult to deploy for scenarios requiring real-time performance. Generative adversarial networks (GAN) have been favored by many researchers and applied to their own algorithms. Some studies have also employed a combination of consistency learning and video object segmentation to achieve a balance between segmentation accuracy and real-time performance. In Wang et al. [57] 's work, STM was applied directly to an ultrasound kidney parenchymal segmentation scenario, and although the segmentation accuracy of STM on the DAVIS dataset was achieved according to the results given in the paper, the memory and temporal redundancy inherent in STM may adversely affect real-time performance and deployment. Additionally, there are challenges with accurate segmentation of target boundaries in ultrasound images.

Conclusion

This paper reviews representative approaches to semi-supervised learning (SSL) and semi-supervised video object segmentation in recent years, providing an overview of the current research progress in real-time semi-supervised ultrasound video object segmentation. In general, there is relatively limited research in this area, but some notable results have been achieved. The application of SSL methods in the field of real-time intelligent segmentation of ultrasound video objects offers new ideas and possibilities for addressing the issue of limited annotated data.

In this respect, this paper equips the reader with the necessary background knowledge, focusing on pseudo-label-based, consistency learning, and video object segmentation methods, as well as other semi-supervised segmentation approaches. These methods have demonstrated good performance on ultrasound images in various scenarios. However, current research

has limitations, such as the generalization ability, adaptability, and computational efficiency of the models on ultrasound images, which still require further improvement.

We believe that there are promising directions in the field of semi-supervised real-time intelligent segmentation for ultrasound video. SSL can be combined with transfer learning, such as domain adaptation, to further reduce the reliance on labeled data and enhance model generalization performance. The construction of large pre-training datasets for ultrasound video is crucial. Currently, models are primarily pre-trained on ImageNet or other publicly available natural image datasets; however, ultrasound images differ significantly from natural images. Utilizing models pre-trained on such datasets may constrain the potential improvement of model accuracy. Additionally, the well-known Transformer architecture has recently achieved remarkable results in computer vision tasks. Its application in real-time segmentation for semi-supervised ultrasound video object warrants exploration in future research.

Acknowledgment

This study was funded by the National Natural Science Foundation of China (Grant number 52175020), Natural Science Foundation of Shanghai (21ZR1451400) and Shanghai Jiading District Health and Family Planning Commission Fund (2021-KY-10).

Conflict of Interest

The authors have no conflict of interest to declare.

References

[1] Klibanov AL, Hossack JA. Ultrasound in radiology: from anatomic, functional, molecular imaging to drug delivery and image-guided therapy. *Invest Radiol* 2015;50:657-670.

[2] Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med Image Anal* 2020;63:101693.

[3] Chapelle O, Scholkopf B, Zien A. Semi-supervised learning. *IEEE Transactions on Neural Networks* 2009;20:542-542.

[4] Yao Q, Xiao L, Liu P, Zhou SK. Label-free segmentation of COVID-19 lesions in lung CT. *IEEE Trans Med Imaging* 2021;40: 2808-2819.

[5] Zhang L, Wang X, Yang D, Sanford T, Harmon S, Turkbey B, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans Med Imaging* 2020;39:2531-2540.

[6] Jiao R, Zhang Y, Ding L, Cai R, Zhang J. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *arXiv preprint arXiv* 2207.14191, 2022.

[7] Zou Y, Yu Z, Liu X, Kumar B, Wang J. Confidence regularized self-training. *IEEE/CVF International Conference on Computer Vision (ICCV)* 2019;5981-5990.

[8] Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, et al. Semi-supervised learning for network-based cardiac MR image segmentation. *Medical Image Computing and Computer Assisted Intervention-MICCAI* 2017:253-260.

[9] Chaitanya K, Erdil E, Karani N, Konukoglu E. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical Image Analysis* 2023:102792.

[10] He R, Yang J, Qi X. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. *Proceedings of the IEEE/CVF International Conference on Computer Vision(CVPR)* 2021:6930-6940.

[11] Yang L, Zhuo W, Qi L, Shi Y, Gao Y. St++: Make self-training work better for semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)* 2022:4268-4277.

[12] Van Engelen J E, Hoos H H. A survey on semi-supervised learning. *Machine learning* 2020;109:373-440.

[13] Yao H, Hu X, Li X. Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* 2022;36:3099-3107.

[14] Lee D H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning ICML* 2013;3:896.

[15] Xia Y, Yang D, Yu Z, Liu F, Cai J, Yu L, et al. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Med Image Anal* 2020;65:101766.

[16] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 2017,30.

[17] Laine S, Aila T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv* 1610.02242, 2016.

[18] Li X, Yu L, Chen H, Fu CW, Xing L, Heng PA. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Trans Neural Netw Learn Syst* 2020;32:523-534.

[19] Li X, Yu L, Chen H, Fu CW, Xing L, Heng PA. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv* 1808.03887,2018.

[20] Bortsova G, Dubost F, Hogeweg L, Katramados L & De Bruijine M. Semi-supervised medical image segmentation via learning consistency under transformations. *Medical Image Computing and Computer Assisted Intervention-MICCAI* 2019;22:810-818.

[21] French G, Laine S, Aila T, Mackiewicz M, Finlayson G. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv* 1906.01916,2019.

[22] Olsson V, Tranheden W, Pinto J, Svensson L. Classmix: segmentation-based data augmentation for semi-supervised learning. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 2021:1369-1378.

[23] Ouali Y, Hudelot C, Tami M. Semi-supervised semantic segmentation with cross-consistency training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)* 2020:12674-12684.

[24] Chen X, Yuan Y, Zeng G, Wang J. Semi-supervised semantic segmentation with cross pseudo supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)* 2021:2613-2622.

[25] Alonso I, Sabater A, Ferstl D, Montesano L, Murillo, AC. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*

2021:8219-8228.

[26] Lai X, Tian Z, Jiang L, Liu S, Zhao H, Wang L, et al. Semi-supervised semantic segmentation with directional context-aware consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)* 2021:1205-1214.

[27] Caelles S, Maninis K K, Pont-Tuset J, Leal-Taixe L, Cremers D, Van Gool L. One-shot video object segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)* 2017:221-230.

[28] Voigtlaender P, Leibe B. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv* 1706.09364,2017.

[29] Maninis KK, Caelles S, Chen Y, Pont-Tuset J, Leal-Taixe L, Cremers D, et al. Video object segmentation without temporal information. *IEEE Trans Pattern Anal Mach Intell* 2019;41:1515-1530.

[30] Perazzi F, Khoreva A, Benenson R, Schiele B, Sorkine-Hornung A. Learning video object segmentation from static images. *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)* 2017:2663-2672.

[31] Hu Y T, Huang J B, Schwing A. Maskrnn: instance level video object segmentation. *Advances in neural information processing systems* 2017;30.

[32] Luiten J, Voigtlaender P, Leibe B. Premvos: proposal-generation, refinement and merging for video object segmentation. *Computer Vision -ACCV2018: 14th Asian Conference on Computer Vision* 2019:565-580.

[33] Cheng J, Tsai Y H, Hung W C, Wang S, Yang M. Fast and accurate online video object segmentation via tracking parts. *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)* 2018:7415-7424.

[34] Chen X, Li Z, Yuan Y, Yu G, Shen J, Qi D. State-aware tracker for real-time video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)* 2020:9384-9393.

[35] Wang Q, Zhang L, Bertinetto L, Hu W, Torr P H. Fast online object tracking and segmentation: A unifying approach. *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition(CVPR)* 2019:1328-1338.

[36] Li B, Yan J, Wu W, Zhu Z, Hu X. High performance visual tracking with Siamese region proposal network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2018:8971-8980.

[37] Shin Yoon J, Rameau F, Kim J, Lee S, Shin S, Kweon I. Pixel-level matching for video object segmentation using convolutional neural networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* 2017:2167-2176.

[38] Hu Y T, Huang J B, Schwing AG. Videomatch: Matching based video object segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)* 2018:54-70.

[39] Oh S W, Lee J Y, Xu N, Kim S J. Video object segmentation using space-time memory networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 2019:9226-9235.

[40] Lu X, Wang W, Danelljan M, Zhou T, Shen J& Van Gool L. Video object segmentation with episodic graph memory networks. *Computer Vision-ECCV* 2020:661-679.

[41] Li Y, Xu N, Peng J, See J, Lin W. Delving into the cyclic mechanism in semi-supervised video object segmentation. *Advances in Neural Information Processing Systems* 2020;33:1218-1228.

[42] Wang H, Jiang X, Ren H, Hu Y, Bai S. Swiftnet: Real-time video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)* 2021:1296-1305.

[43] Wang Q, Zheng J, Yu H, Zhang J Q, Zhang J. Automatic detection of thyroid nodules with ultrasound images: Basing on semi-supervised learning. *Journal of Physics* 2021;1976:012012.

[44] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer- Assisted Intervention-MICCAI* 2015;18:234-241.

[45] Li Y, Liu Y, Huang L, Wang Z, Luo J. Deep weakly-supervised breast tumor segmentation in ultrasound images with explicit anatomical constraints. *Med Image Anal* 2022;76:102315.

[46] Cao X, Chen H, Li Y, Peng Y, Wang S, Cheng L. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. *IEEE Trans Med Imaging* 2020; 40:431-443.

[47] Li Y, Liu Y, Wang Z, Luo J. Semi-supervised deep learning for breast anatomy decomposition in ultrasound images. *IEEE International Ultrasonics Symposium (IUS)* 2021:1-4.

[48] Zhai D, Hu B, Gong X, Zou H, Luo J. ASS-GAN: Asymmetric semi-supervised GAN for breast ultrasound image segmentation. *Neurocomputing* 2022;493:204-216.

[49] Huang K, Zhang Y, Cheng HD, Xing P. MSF-GAN: Multi-scale fuzzy generative adversarial network for breast ultrasound image segmentation. *Annu Int Conf IEEE Eng Med Biol Soc* 2021:3193-3196.

[50] Iqbal, Ahmed, and Muhammad Sharif. PDF-UNet: a semi-supervised method for segmentation of breast tumor images using a U-shaped pyramid-dilated network. *Expert Systems with Applications* 2023;221.

[51] Pang T, Wong JHD, Ng WL, Chan CS. Semi-supervised GAN-based radiomics model for data augmentation in breast ultrasound mass classification. *Comput Methods Programs Biomed* 2021;203:106018.

[52] Han L, Huang Y, Dou H, Wang S, Ahamad S, Luo H, et al. Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network. *Comput Methods Programs Biomed* 2020;189:105275.

[53] Xu X, Sanford T, Turkbey B, Xu S, Wood BJ, Yan P. Shadow-consistent semi-supervised learning for prostate ultrasound segmentation. *IEEE Trans Med Imaging* 2022;41:1331-1345.

[54] Chen Y, Wang Y, Lai B, Chen Z, Cao X, Ye N, et al. VeniBot: towards autonomous venipuncture with semi-supervised vein segmentation from ultrasound images. *arXiv preprint arXiv* 2105.12945,2021.

[55] Xie X, Niu J, Liu X, Li Q, Wang Y, Tang S. DK-Consistency: a domain knowledge guided consistency regularization method for semi-supervised breast cancer diagnosis. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2021:3435-3442.

[56] Mendel R, Rauber D, de Souza LA Jr, Papa JP, Palm C. Error-correcting mean-teacher: corrections instead of consistency-targets applied to semi-supervised medical image segmentation. *Comput Biol Med* 2023;154:106585.

[57] Wang R. Ultrasound video object segmentation of renal parenchyma. *International Core Journal of Engineering* 2022;8:101-106.

[58] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. *Advances in neural information processing systems* 2017;30.

[59] Chen C H, Lee C S, Ahmad S, Hartley-Brown M. A single center review evaluating daratumumab outcomes in multiple myeloma patients at Monter Cancer Center/Northwell Health Cancer Institute. *Blood* 2020;136:25.

[60] Wu H, Liu J, Xiao F, Wen Z, Cheng L, Qin J. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Med Image Anal* 2022;78:102397.

[61] Ta K, Ahn S S, Lu A, Stendahl J C, Sinusas A J, Duncan J S. A semi-

supervised joint learning approach to left ventricular segmentation and motion tracking in echocardiography. *IEEE 17th International Symposium on Biomedical Imaging (ISBI)* 2020:1734-1737.

[62] El Rai M C, Darweesh M, Al-Saad M. Semi-supervised segmentation of echocardiography videos using graph signal processing. *Electronics* 2022;11:3462.

[63] Chen Y, Zhang C, Liu L, Dong C, Luo Y, Wan X. USCL: pretraining deep ultrasound image diagnosis model through video contrastive representation learning. *Medical Image Computing and Computer Assisted Intervention-MICCAI* 2021;24:627-637.

[64] Sirjani N, Moradi S, Oghli MG, Hosseinsabet A, Alizadehasl A, Yadollahi M, et al. Automatic cardiac evaluations using a deep video object segmentation network. *Insights Imaging* 2022;13:69.

[65] Song XF, Wang YN, Feng QJ, Wang Q. Improved graph cut model with features of superpixels and neighborhood patches for myocardium segmentation from ultrasound image. *Math Biosci Eng* 2019;16:1115-1137.

[66] Shin S Y, Lee S, Yun I D, Kim S M, Lee K M. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE Trans Med Imaging* 2018;38:762-774.

[67] Dai F, Li Y, Zhang G, Shi Q, Xing W, Liu X, et al. An automatic ultrasonic segmentation method by two-stage semi-supervised learning strategy. *IEEE International Ultrasonics Symposium (IUS)* 2022:1-4.

[68] Judge T, Judge A, Jodoin P. Anatomically constrained semi-supervised learning for echocardiography segmentation. *Medical Imaging with Deep Learning* 2022.

[69] Yang H, Shan C, Kolen AF, Peter HN. Deep Q-network-driven catheter segmentation in 3D US by hybrid constrained semi-supervised learning and dual-UNet. *Medical Image Computing and Computer Assisted Intervention-MICCAI* 2020:646-655.

[70] Dai W, Li X, Ding X, Cheng K. Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos. *IEEE Transactions on Medical Imaging* 2022.

[71] Wang P, Vives M, Patel VM, Hacihaliloglu I. Robust bone shadow segmentation from 2D ultrasound through task decomposition. *Medical Image Computing and Computer Assisted Intervention-MICCAI* 2020:805-814.

[72] Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans Med Imaging* 2019;38:2198-2210.

[73] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020;580:252-256.

[74] Sun, Y, Chen G, Zhou T, Zhang Y, Liu N. Context-aware cross-level fusion network for camouflaged object detection. *arXiv preprint* arXiv 2105.12555,2021.