TempSeg-GAN: Segmenting Objects in Videos Adversarially using Temporal Information

Saptakatha Adak[1] and Sukhendu Das[2]
Visualization and Perception Lab,
Indian Institute of Technology Madras, Chennai - 600036, India

Keywords: Video Object Segmentation, Generative Adversarial Network (GAN), Deep Learning.

Abstract: This paper studies the problem of Video Object Segmentation which aims at segmenting objects of interest throughout entire videos, when provided with initial ground truth annotation. Although, variety of works in this field have been done utilizing Convolutional Neural Networks (CNNs), adversarial training techniques have not been used in spite of their effectiveness as a holistic approach. Our proposed architecture consists of a Generative Adversarial framework for the purpose of foreground object segmentation in videos coupled with Intersection-over-union and temporal information based loss functions for training the network. The main contribution of the paper lies in formulation of the two novel loss functions: (i) Inter-frame Temporal Symmetric Difference Loss (ITSDL) and (ii) Intra-frame Temporal Loss (IFTL), which not only enhance the segmentation quality of the predicted mask but also maintain the temporal consistency between the subsequent generated frames. Our end-to-end trainable network exhibits impressive performance gain compared to the state-of-the-art model when evaluated on three popular real-world Video Object Segmentation datasets *viz.* DAVIS 2016, SegTrack-v2 and YouTube-Objects dataset.

1 INTRODUCTION

Video Object Segmentation has emerged as a popular field of research in Computer Vision in the recent decade. The popularity of this domain mainly lies in its profound impact in the domains of bio-medical research, self-driving cars, video editing, robotics, etc. The recent years have noticed a dramatic advance in Autonomous Driving where instance segmentation in videos have found a huge scope through tasks like segmenting lanes in roads, detecting obstacles in the path of motion, segmentation of street signs, etc. With the recent advancement of deep learning techniques, there has been many works based on Convolutional Neural Networks (CNNs) which not only have improved the performance for problems like image classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), object detection (Ren et al., 2015; Redmon et al., 2016), etc., but also in the field of image segmentation (Maninis et al., 2016; Caelles et al., 2017; Voigtlaender and Leibe, 2017), using pre-trained weights of image recognition models on ImageNet (Deng et al., 2009). The major disadvantage of these CNNs are their hunger for large training data. Recently, approaches based on Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) have also been used for the task of image segmentation (Souly et al., 2017; Luc et al., 2016).

In Video Object Segmentation (VOS), with the introduction of pixel-level annotated DAVIS 2016 (Perazzi et al., 2016) dataset, a shift from segmentation using bounding box tracking (Li et al., 2016; Wang and Yeung, 2013; Perazzi et al., 2017) to pixel-level tracking (Caelles et al., 2017; Voigtlaender and Leibe, 2017) has been noticed. Deep CNNs pre-trained on large image classification datasets show decent performance in VOS (Caelles et al., 2017; Perazzi et al., 2017) and other related tasks like single-object tracking (Bertinetto et al., 2016; Nam and Han, 2016) and back-ground modeling (Braham and Van Droogenbroeck, 2016; Wang et al., 2017). Among these, Caelles *et al.* (Caelles et al., 2017) shows promising result in VOS, by fine-tuning the pre-trained CNN with the initial frame of the target video. But, this method fails to adapt to the drastic appearance changes in the subsequent frames of the video sequences. To cope up with this problem, some recent methods have solved the task in the temporal domain using optical flow (Tsai et al., 2016; Khoreva et al., 2017), spatio-temporal MRFs (Bao et al., 2018), guided mask propagation (Wug Oh et al., 2018; Cheng et al., 2018), etc. In our work, we have followed

Adak, S. and Das, S.
TempSeg-GAN: Segmenting Objects in Videos Adversarially using Temporal Information.
DOI: 10.5220/0007254302210232
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 221-232
ISBN: 978-989-758-354-4

a GAN based approach for semi-supervised instance segmentation of an object of interest throughout the video along with the usage of two temporal information based objective functions for end-to-end training of the network.

To summarize, the novelty of this paper lies in

- an adversarial training based framework (TempSeg-GAN) to generate segmentation masks using the initial ground-truth and generated masks in videos (refer to section 2),
- use of a encoder-decoder model with skip connections as generator network in the proposed architecture,
- variation in discriminator training by adding encoder module between generator and discriminator, such that the encoded output of the generator is used as an input for the discriminator,
- incorporation of two temporal information based loss functions which not only enhances the segmentation quality of predicted masks but also stabilizes the motion features between them.

The proposed Inter-frame Temporal Symmetric Difference Loss (ITSDL) is calculated between predicted and optical flow warped target mask, formed from the ground-truth mask of previous time step. Thus, it not only captures motion features between consecutive frames but also enhances the segmentation quality by minimizing the erroneously identified pixels in the predicted masks (section 3). The Intra-frame Temporal loss (IFTL) along with its long-range variant (L-IFTL) generally preserve the temporal relationship between the generated masks (section 4).

2 TEMPORALLY AIDED SEGMENTATION NETWORK

The proposed network for Video Object Segmentation consists of two sub-networks: (i) the Generator (G) and (ii) the Discriminator (D). The generator G generates images close to the ground-truth by extracting features from the true data distribution p_{data}, thereby making it difficult for the discriminator to differentiate between generated and real images. Whereas, the discriminator D, is optimized to predict whether the generated output is synthetic or real. This process of alternate learning of the two sub-networks in this framework is similar to the two player min-max games (Goodfellow et al., 2014). The overall objective function for simultaneous minimizing the loss at G and maximizing the distinguisher D is as follows:

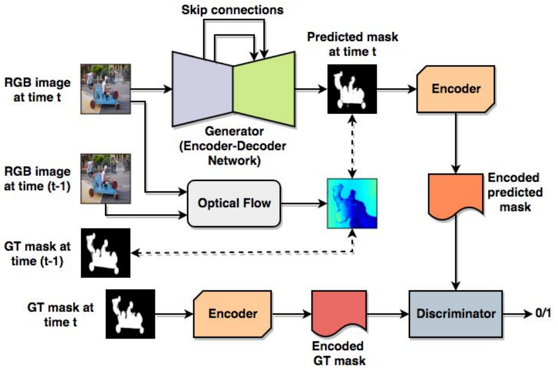

Figure 1: Proposed TempSeg-GAN architecture. GT denotes the ground-truth and $\leftarrow\!-\!-\!-\!\rightarrow$ refers to the Inter-frame Temporal Symmetric Difference Loss ($\mathcal{L}_{ITSDL}$) estimation using GT mask at time t, predicted mask at time $(t-1)$ and optical flow vectors between RGB input images at time $(t-1)$ and t.

$$\min_G \max_D u(G,D) = \mathbb{E}_{x\sim p_{data}}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))] \quad (1)$$

where, x is a real image from the true distribution p_{data} and z is a vector sampled from an uniform or Gaussian random distribution p_z. Since our work is based on videos, a sequence of video frames are provided as an input to the network and is trained using two novel loss functions which provides temporal assistance. The adversarial loss employed in this paper is a variant of that in equation 1, as an additional encoder module is implemented between the generator and the discriminator.

The proposed architecture of Temporally aided Segmentation GAN (TempSeg-GAN) is illustrated in figure 1. The generator of the model is an encoder-decoder structure consisting of convolutional modules coupled with pooling and unpooling layers (Zeiler and Fergus, 2014) with ReLU non-linearity. Batch-normalization (Ioffe and Szegedy, 2015) and dropout have also been incorporated in this network. The contracting path of the encoder captures the context, while the symmetric expanding path of the decoder localizes the information precisely. Following “U-net” (Ronneberger et al., 2015), skip connections are added between each layer i and $(n-i)$, where n denotes the total number of layers. The skip connection concatenates all feature maps of layer i with those at the $(n-i)^{th}$ layer and aids in sharing low-level information like prominent edge details, etc. between the initial and final layers of the generator network. The input to the encoder-decoder like generator are frames of videos of dimension $(W_0 \times H_0 \times 3)$ passed independently in a sequential manner, while the output is a segmentation map of dimension $(W_0 \times H_0 \times 1)$, corresponding to each frame of input. The predicted output and the ground-truth mask are then encoded with two separate encoders of same configuration (see figure 1). Each encoder is formed of convolutional layers combined with pooling modules to down-sample the input. The encoded predicted and ground-truth mask obtained as output from the encoder form the input for the discriminator, which is a collection of convolutional modules with fully-connected layers at the end to generate 0 or 1, denoting synthetic or real data respectively.

The main difference between the proposed TempSeg-GAN model and the traditional one (Goodfellow et al., 2014) is that the encoder modules have been used to encode the output of the generator and the ground-truth before passing it into the discriminator, instead of using the generator output directly. The motivation behind the encoding mask lies in the fact that, the discriminator will be able to distinguish between the real and synthetic outputs more efficiently in the projected feature space than in the RGB image space. The details of the network architecture is mentioned in section A of the Appendix.

2.1 TempSeg-GAN training

Our proposed GAN framework follows the training of the conventional generative adversarial networks with a few variations. The generator G in the adversarial network is a segmentation model that predicts the mask considering the joint data distribution of the input video frame (I) and its corresponding ground-truth mask (Y). On the other hand, the discriminator D distinguishes between predicted and the original mask, thus facilitating the training procedure by minimizing the dissimilarity between the prediction and the ground-truth. The objective function of the adversarial training is as follows:

$$\begin{aligned}\min_{\theta_G}\max_{\theta_D}\sum_I &\mathcal{L}_{bce}(Y, O_G(I;\theta_G)) \\ &-\lambda\big[\mathcal{L}_{bce}(1, O_D(O_E(Y);\theta_D)) \\ &+\mathcal{L}_{bce}(0, O_D(O_E(O_G(I;\theta_G));\theta_D))\big]\end{aligned} \tag{2}$$

where θ_G and θ_D are generator and discriminator parameters respectively; $\mathcal{L}_{bce}$ represents the binary cross-entropy loss; O_G and O_D denote the generator and discriminator output respectively; $O_E(X)$ refers to the encoded segmented mask output of the input X; 1 and 0 denotes the discriminator labels when the input is from the ground-truth Y and the generator $O_G(I;\theta_G)$ respectively and λ is a regularization parameter.

Thus, the adversarial objective function for the generator, obtained by minimizing equation 2 w.r.t. θ_G, is as follows

$$\begin{aligned}\mathcal{L}^{G}_{adv}(I) = \min_{\theta_G}\sum_I &\mathcal{L}_{bce}(Y, O_G(I;\theta_G)) \\ &+\lambda\mathcal{L}_{bce}(1, O_D(O_E(O_G(I;\theta_G));\theta_D))\end{aligned} \tag{3}$$

where, the first term deals with the consistency of the predicted segmentation with the target mask at each position, while the unfitting structure between output and ground-truth is penalized with the help of the second term.

Again, equation 2 is minimized w.r.t. θ_D and the adversarial discriminator loss function is achieved as follows

$$\begin{aligned}\mathcal{L}^{D}_{adv} = \min_{\theta_D}\sum_I &\big[\mathcal{L}_{bce}(1, O_D(O_E(Y);\theta_D)) \\ &+\mathcal{L}_{bce}(0, O_D(O_E(O_G(I;\theta_G));\theta_D))\big]\end{aligned} \tag{4}$$

where, $O_E(O_G(I;\theta_G))$ and $O_E(Y)$ are the encoded predicted and target masks fed into the discriminator, which is trained in such a way that it classifies $(O_E(I), O_E(Y))$ into class 1 and class 0 for $(O_E(O_G(I;\theta_G)), O_E(Y))$.

Though the theoretical foundation of this alternate optimization process of generator and discriminator is logically firm and well-established, in reality it is susceptible to mode collapse leading to instabilities in training. To overcome this implicit instability and produce better predicted segmented masks, two novel loss functions based on Intersection-over-Union (IoU) and optical flow vectors (refer to section 3, 4) have been formulated and are used in addition to the existing conventional adversarial losses described in eqns. 3 and 4. The flow vectors are obtained by passing consecutive RGB input frames through the *FlowNet 2.0* (Ilg et al., 2017) optical flow generation module during training.

3 INTER-FRAME TEMPORAL SYMMETRIC DIFFERENCE LOSS

Unlike images, the advantage of the videos lie in the fact that it provide a latent space of data distribution by combining the temporal information with the spatial one. CNNs are capable of capturing short-range consistencies in the spatial domain, which only forms a small part of the rich input data. Thus, to maintain the temporal coherency between the masks along with enhancing the segmentation quality, an Intersection-over-Union (IoU) based temporal objective function has been incorporated. It measures the region of similarity between the predicted mask ($\hat{Y}$) and the ground-truth mask (Y), by computing the number of overlapping pixels. In other words, it gives an idea of the

Algorithm 1: $optical\_flow\_warp(P_{t-1}, W^*_{t-1}, \hat{P}_t)$

```
Input: Ground-truth mask patch ($P_{t-1}$) at time $t-1$, Predicted mask patch ($\hat{P}_t$) at time $t$, Optical flow
       vector map patch ($W^*_{t-1}$).
Output: Optical flow warped ground-truth mask patch ($P^*_t$) at time $t$ of dimension same as $\hat{P}_t$.
// $s =$ height and width of $\hat{P}_t$
// $s+4 =$ height and width of $P_{t-1}$ and $W^*_{t-1}$
Initialize: patch $P^*_t$ with each pixel value equal to 0.;
for $u = 0$ to $s+4$, $i \leftarrow u+1$ do
    for $v = 0$ to $s+4$, $j \leftarrow v+1$ do
        /* $(u,v)$ is the spatial location of pixel at time $(t-1)$.                */
        $u' \leftarrow u + V_u \Delta t$;
        $v' \leftarrow v + V_v \Delta t$;
        /* $V_u$ and $V_v$ are horizontal and vertical flow vectors of $(u,v)$ obtained from
           $W^*_{t-1}$, $(u',v')$ is the new spatial location of $(u,v)$ at time $t$.  Here, $\Delta t = 1$.
           */
        Label$(u',v') \leftarrow$ Label$(u,v)$;
        /* The label of $(u',v')$ is updated with that of $(u,v)$.                  */
        if $(u',v')$ lies within the patch $P^*_t$ then
            Update corresponding pixel value of $P^*_t$ with Label$(u',v')$.
        end
    end
end
```

number of mispredicted pixels present in the estimated mask compared to the ground-truth. To formulate the function in a patch-wise manner, we calculate the number of mispredicted pixels ($|M_p|$) in a patch of the segmented mask. The set of mispredicted pixels (M_p) for class $c \in \{0,1\}$ obtained by using symmetric difference of the two vectors $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$, is as follows:

$$\boldsymbol{M}_p(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \{\boldsymbol{p}_k = c, \hat{\boldsymbol{p}}_k \neq c\} \cup \{\boldsymbol{p}_k \neq c, \hat{\boldsymbol{p}}_k = c\} \quad \forall k = 1, \cdots, s^2 \tag{5}$$

where, $\hat{\boldsymbol{p}}$ denotes the vector of predicted labels in the patch $\hat{P}$, with top-left pixel index (i,j), of the estimated mask ($\hat{Y}$) and $\boldsymbol{p}$ is the vector of ground-truth labels in the corresponding patch P of the target mask (Y); $\boldsymbol{p}, \hat{\boldsymbol{p}} \in \{0,1\}^{s^2}$; s is the height and width of the patches P and $\hat{P}$.

In Video Object Segmentation, we implement the symmetric difference by extracting non-overlapping patches of dimension $s \times s$ $(1 < s \leq 4)$, represented by $\hat{P}_t\{i,j,s\}$, where (i,j) is the top-left index of the patch, from the predicted mask at time t and then evaluating the number of mislabeled pixels with the corresponding target patch $P^*_t\{i,j,s\}$ at same time t. The mechanism of formation of the target patch from the ground-truth patch at time $(t-1)$, denoted by $P_{t-1}\{i-2, j-2, s+4\}$, by warping with optical flow vector patch $W^*_{t-1}\{i-2, j-2, s+4\}$, at time $(t-1)$, is explained step-wise in Algorithm 1.

In simpler terms, we calculate the symmetric difference score between small portions of the predicted mask and the corresponding optical flow warped ground-truth mask. According to our assumption, the motion features are effectively transfered from the ground-truth mask of the previous time step to the warped target mask of current time step using optical flow vectors. Thus the motion related features can be well approximated with the low resolution patches both in the spatial as well as temporal domains. The smoothness of the features are also guaranteed, unless there is a sudden change of scene or rapid movement in the videos. Computing the loss function in a patch-wise manner enhances the attention over small disjoint sections formed as a result of occlusions, in the segmented masks, which are often ignored when estimated globally.

The Inter-frame Temporal Symmetric Difference score for each patch ($ITSD_{patch}$) of a predicted mask with top-left index (i,j) is denoted as:

$$ITSD_{patch} = \frac{|\boldsymbol{M}_{p^*}|}{|\{\hat{\boldsymbol{p}}_k = 1\} \cup \boldsymbol{M}_{p^*}|} \tag{6}$$

which estimates a ratio of mispredicted pixels to the total number of pixels for the patch having top-left pixel index (i,j) (refer to line 11 in algorithm 2).

The step-by-step procedure of evaluating the ITSD score by matching the corresponding local patches of the estimated and optical flow warped masks is explained in algorithm 2. The objective function modeled by calculating the ITSD score for the generator (G) network not only minimizes the score for

Algorithm 2: Inter-frame Temporal Symmetric Difference score to estimate the similarity between optical flow warped ground-truth mask(s) and predicted mask(s).
$extract\_patch(X_t, a, b, d)$ is used to extract a patch of dimension $d \times d$ starting from the top-left pixel index (a,b) of the frame X_t at time t.

```
Input: Ground-truth masks ($Y$), Predicted masks ($\hat{Y}$), Optical flow vector maps ($W$)
Output: Inter-frame Temporal Symmetric Difference score ($Score_{ITSD}$)
// $s =$ height and width of an patch on the mask
// $S =$ height and width of the masks
// $t =$ current time
// $T =$ Number of masks predicted
Initialize: $Score_{ISD} = 0$;
for $t = 1$ to $T$ do
    for $i = 0$ to $S$, $i \leftarrow i + s$ do
        for $j = 0$ to $S$, $j \leftarrow j + s$ do
            $\hat{P}_t \leftarrow extract\_patch(\hat{Y}_t, i, j, s)$;
            $P_{t-1} \leftarrow extract\_patch(Y_{t-1}, i-2, j-2, s+4)$;
            $W^*_{t-1} \leftarrow extract\_patch(W_{t-1}, i-2, j-2, s+4)$;
            $P^*_t \leftarrow optical\_flow\_warp(P_{t-1}, W^*_{t-1}, \hat{P}_t)$ (refer to Algo. 1);
            $\hat{\boldsymbol{p}} \leftarrow$ vector of predicted labels of pixels obtained from the patch $\hat{P}_t$;
            $\boldsymbol{p}^* \leftarrow$ vector of labels of pixels obtained from the patch $P^*_t$;
            $\boldsymbol{M}_{p^*} \leftarrow \{\boldsymbol{p}^*_k = 1, \hat{\boldsymbol{p}}_k = 0\} \cup \{\boldsymbol{p}^*_k = 0, \hat{\boldsymbol{p}}_k = 1\}, \quad \forall k = 1, \cdots, s^2$ (see eqn. 5);
            /* $\boldsymbol{M}_{p^*}$ denotes the set of mislabeled pixels in the $\hat{\boldsymbol{p}}$ compared with $\boldsymbol{p}^*$ */
            $Score_{ITSD} \leftarrow Score_{ITSD} + \frac{|\boldsymbol{M}_{p^*}|}{|\{\hat{\boldsymbol{p}}=1\} \cup \boldsymbol{M}_{p^*}|}$ (see eqn. 6);
        end
    end
    $Score_{ITSD} \leftarrow {}^{Score_{ITSD}}/_{\lfloor S/s \rfloor^2}$ ;          // Average over all the patches
end
$Score_{ITSD} \leftarrow {}^{Score_{ITSD}}/_{T}$;          // Average over all the masks
```

batch inputs but also maintains the temporal data distribution by stabilizing the motion features generated by the network with the help of flow vectors. The loss function, $\mathcal{L}_{ITSDL}$ is defined as:

$$\mathcal{L}_{ITSDL}(Y, \hat{Y}) = Score_{ITSD}(Y, \hat{Y}) \tag{7}$$

where, $\hat{Y}$ and Y represents the predicted and optical flow warped mask, and $Score_{ITSD}$ denotes the mean symmetric difference score over all the masks, obtained using the process mentioned in algorithm 2.

4 INTRA-FRAME TEMPORAL LOSS

The ITSD Loss, mentioned in section 3, estimates the motion features that change slowly with respect to time using the local symmetric difference measures which in turn also enhances the segmentation quality. Thus to maintain the temporal relationship between the frames globally, we introduce the idea of Intra-frame Temporal Loss over the network output masks. A few works (Goroshin et al., 2015; Mobahi et al., 2009) in the recent past exploits the idea of the temporal coherence to learn the motion features. Assuming slow variation of motion features over time, we consider 2 consecutive frames $\hat{Y}_t$ and $\hat{Y}_{t+1}$ as temporal pair, where $\hat{Y}_t$ and $\hat{Y}_{t+1}$ are TempSeg-GAN generated output masks at time t and $t+1$ respectively with value of the discriminator (D) outputs, q_t and q_{t+1} being equal to 1 for both the masks. The slow variation of motion features is modeled through an objective function as

$$\begin{aligned} &\mathcal{L}_{IFTL}(\hat{Y}, \overrightarrow{q}) \\ &= \sum_{t=0}^{T-1} d_\delta(\hat{Y}_t, \hat{Y}_{t+1}, q_t \times q_{t+1}) \\ &= \sum_{t=0}^{T-1} \Big(q_t \times q_{t+1} \times d(\hat{Y}_t, \hat{Y}_{t+1}) \\ &\quad + (1 - q_t \times q_{t+1}) \times \max(0, \delta - d(\hat{Y}_t, \hat{Y}_{t+1})) \Big) \end{aligned} \tag{8}$$

where, T is the total time duration of the masks generated by the network, $q_t \in \{0, 1\}$ gives the value of the discriminator output, $d(x, y)$ is the measure for eucledian distance and δ is a positive constant. Thus speaking in simpler terms, equation 8 minimizes the intra-frame distance between the predicted masks which have been generated correctly while penalizing the disparity between the incorrectly predicted frames with a positive margin δ.

Long-range Intra-frame Temporal Loss: Though the IFT Loss maintains the temporal consistency between the consecutive frames, it does not guarantee the same for the long term frames. Thus, to keep the stability intact in the spatio-temporal feature space a Long-range Intra-frame Loss (L-IFTL) is incorporated by extending the IFT Loss as an estimation of the distance between initial predicted mask ($\hat{Y}_0$) and all other predicted masks ($\hat{Y}_t$) at time $t(>0)$. The proposed loss is defined as

$$\begin{aligned}
&\mathcal{L}_{L-IFTL}(\hat{Y},\overrightarrow{q}) \\
&= \sum_{t=1}^{T} d_\delta(\hat{Y}_0,\hat{Y}_t,q_0\times q_t) \\
&= \sum_{t=1}^{T} \Big(q_0\times q_t\times d(\hat{Y}_0,\hat{Y}_t) \\
&\qquad +(1-q_0\times q_t)\times \max(0,\delta-d(\hat{Y}_0,\hat{Y}_t))\Big)
\end{aligned} \tag{9}$$

where, the symbols have the same meaning as in equation 8.

Thus, L-IFT preserves the temporal coherency among the distant frames by estimating the distance between the initial and rest of the generated frames.

5 MULTI-COMPONENT OBJECTIVE FUNCTION

Finally, the overall objective function is formed by combining the loss functions given in eqns. 7 - 9 with the adversarial loss (refer to eqn. 3) and the traditional $\mathcal{L}_1$ objective with respective weights as follows

$$\begin{aligned}
&\mathcal{L}_{combined} \\
&= \alpha_{adv}\mathcal{L}^{G}_{adv}(I)+\alpha_{L_1}\mathcal{L}_{L_1}(Y,\hat{Y})+\alpha_{ITSDL}\mathcal{L}_{ITSDL}(Y,\hat{Y}) \\
&\quad +\alpha_{IFTL}\mathcal{L}_{IFTL}(\hat{Y},\overrightarrow{q})+\alpha_{L-IFTL}\mathcal{L}_{L-IFTL}(\hat{Y},\overrightarrow{q})
\end{aligned} \tag{10}$$

where, the weights *viz.* α_{L_1}, α_{ITSDL}, α_{IFTL} and α_{L-IFTL} are set to 0.25 while α_{adv} is kept at 0.1. This combined loss is minimized during the training of TempSeg-GAN using Adam optimizer (Kingma and Ba, 2014).

6 EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we discuss the performance analysis of our proposed model for video object segmentation on three popular benchmark real-world datasets. Sequences of video frames with corresponding ground-truth masks are used to train the model. During validation, the network generates segmented mask of the object of interest when provided with frames of video sequences as input. To specify the particular object to be segmented the model is fine-tuned with first two annotated frames of the videos. The network generated segmented masks are again used by our proposed architecture as reference masks to produce the predicted masks of the next time steps. Optical flow vectors are also calculated using *FlowNet 2.0* (Ilg et al., 2017) between the consecutive frames which helps in formulating the ITSD Loss (described in section 3). Three metrics: Region similarity ($\mathcal{J}$), Contour accuracy ($\mathcal{F}$) and Temporal (in-)stability ($\mathcal{T}$) (Perazzi et al., 2016) have been used to compare the results of our network with the existing state-of-the-art techniques.

6.1 Datasets

Evaluation of the proposed adversarial training based method is done on three benchmark datasets with challenging characteristics like occlusion, motion blur, background clutter, change of appearance, etc.

DAVIS 2016 (Perazzi et al., 2016) consists of 50 high resolution video sequences with 30 being used for training and remaining for validation purposes. Single or multiple connected objects in each of the 3,455 frames of the dataset are provided with pixel-level segmentation.

SegTrack-v2 (Li et al., 2013) contains 14 videos with a total of 947 frames. Sequences with multiple objects are annotated with instance-level segmentation where each annotation is treated as an individual object.

YouTube-Objects Dataset (Prest et al., 2012) includes 126 videos with 10 object classes. The ground-truth segmentation masks with pixel level accuracy of ∼4,250 frames have been obtained from (Jain and Grauman, 2014).

6.2 Evaluation Metric for Segmentation

Three methods (Perazzi et al., 2016) used for quantitative assessment of the predicted masks in comparison with the ground-truth masks are as follows : (a) Region similarity ($\mathcal{J}$), (b) Contour accuracy ($\mathcal{F}$) and

(c) Temporal (in-)stability ($\mathcal{T}$).

Region similarity ($\mathcal{J}$) or *intersection-over-union* (IoU) measures the similarity in segmentation by computing the region overlap between the estimated ($\hat{Y}$) and ground-truth (Y) masks and is defined as: $\mathcal{J} = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}$.

Contour accuracy ($\mathcal{F}$) gives a measure based on the precision and recall of the contour points forming the segmentation boundary.

Temporal (in-)stability ($\mathcal{T}$) evaluates how smooth are the motion features in a video, thus measuring the consistency of sequences of frames over time.

6.3 Performance Analysis of Video Object Segmentation

The proposed TempSeg-GAN model is first pre-trained on the ImageNet (Deng et al., 2009) dataset. During training on DAVIS 2016 (Perazzi et al., 2016), 30 sequences of annotated video frames pre-allocated for training purpose has been used. Augmentation of the frames by random rotation, flipping and zooming is also done for training, in such a way that it does not disturb the temporal consistency among the frames. While validating, first two frames of each of the remaining 20 video sequences along with its ground-truth mask is fed into the GAN to fine-tune the generator before providing the remaining video frames sequentially into the network as input to generate the corresponding segmented masks. Fine-tuning aids in capturing the appearance of the specific object of interest required for foreground segmentation. Also, to maintain the temporal relationship between the predicted masks, the network-output masks of previous time steps are used as reference, during generation of next time step masks, by the framework. A well-tuned Conditional Random Field (CRF) (Krähenbühl and Koltun, 2011) is used as a post-processing module on top of our method for fair comparative study of results among the recent state-of-the-art methods. Augmentation of the frames at fine-tuning stage (Test-time augmentation) is also done to improve the quality of segmentation. TempSeg-GAN++ refers to the modified version of our baseline model with the aforementioned add-ons attached.

Ablation Studies on DAVIS 2016

Variations have been made in the proposed architecture to study its performance and the results obtained in different cases are exhibited in table 1. First, we remove the fine-tuning based on the initial frames of the videos and then the output on the validation set is studied in an unsupervised setup.

Table 1: Ablation study of our proposed method on DAVIS 2016 dataset. Keeping the entire system intact, one variation is made at a time to see contribution of each module. The last row consists of the result after adding test-time augmentation and CRF on the top of our base method. The result of the best configuration is in bold. The right-most column gives the $\mathcal{J}_{Mean}$ difference ($\Delta\mathcal{J}_{Mean}$) of different settings in comparison with the baseline method (in row 5).

Aspect	System variant	$\mathcal{J}_{Mean}$	$\Delta\mathcal{J}_{Mean}$
Training	w/o fine-tune	76.8	-8.3
	w/o DAVIS training	68.7	-16.4
	w/o ImageNet weights	79.3	-5.8
Loss	$\mathcal{L}_1$ loss	81.2	-3.9
	$\mathcal{L}_1 + ITSDL$	84.6	-0.5
	TempSeg-GAN	85.1	-
Add-ons	TempSeg-GAN++	**86.3**	+ 1.2

A substantial decrease in $\mathcal{J}_{Mean}$ was noticed relying only on pre-trained ImageNet (Krizhevsky et al., 2012) weights and DAVIS training data, thus making fine-tuning indispensable for expanding the tracking capabilities in the video sequences. Again, relying on only pre-trained ImageNet weights and fine-tuning, skipping the training on DAVIS 2016 (Perazzi et al., 2016) dataset, shows a drastic drop (68.7 $\mathcal{J}_{Mean}$) in the performance of the model. Removing the pre-trained ImageNet weights results in decrease in $\mathcal{J}_{Mean}$, owing to the loss of scale information. We argue that tracking a specific object in a video requires a reasonable amount of pre-knowledge which comes from pre-training the network on ImageNet (Deng et al., 2009) dataset which consists of ∼10 M objects belonging to 1000 categories. Thus, these pre-trained weights assist the model to learn the general objectness prior, while the training set of DAVIS 2016 provides an advantage in evaluation by aiding the model to adapt to the characteristics of the dataset. The initial frames of the validation set videos guide the network to track the specific object of interest throughout the sequences.

The proposed objective functions play an important role in the generation of segmentation masks in the sequence of video frames. Using only $\mathcal{L}_1$ loss produces holes in the segmented mask and some of them contain small blobs generated outside the region of interest causing inaccurate segmentation. On the other hand, the proposed Inter-frame Temporal Symmetric Difference Loss (ITSDL), combined with $\mathcal{L}_1$ and Intra-frame Temporal Loss (IFTL) produces impressive results (see figure B.2 in the Appendix), where the ITSDL helps in removing the blob like artifacts, thereby improving the contours of the output masks. We also add a well-tuned post-processing CRF on top of our proposed method along with augmentation of initial frames during

Table 2: Quantitative analysis of TempSeg-GAN in comparison with other existing semi-supervised methods on DAVIS 2016, YouTube-Objects and SegTrack-v2 datasets. Other results used for comparison are from the respective papers. Best results are in bold. Values underlined represents the next best results. ↑ = 'higher the value better'; ↓= 'lower the value better'.

Method	DAVIS 2016			YouTube-Objects	SegTrack-v2
	$\mathcal{J}_{Mean}\uparrow$	$\mathcal{F}_{Mean}\uparrow$	$\mathcal{T}_{Mean}\downarrow$	$\mathcal{J}_{Mean}\uparrow$	$\mathcal{J}_{Mean}\uparrow$
BVS (Märki et al., 2016)	60.0	58.8	34.7	59.7	58.4
OFL (Tsai et al., 2016)	68.0	63.4	22.2	70.1	67.5
OSVOS (Caelles et al., 2017)	79.8	80.6	37.8	72.5	65.4
Masktrack (Perazzi et al., 2017)	80.3	75.8	18.6	72.6	70.3
RGMP (Wug Oh et al., 2018)	81.5	82.0	**13.3**	-	71.1
LucidTracker (Khoreva et al., 2017)	80.5	-	-	76.2	77.6
FAVOS (Cheng et al., 2018)	82.4	79.5	26.3	-	-
OnAVOS (Voigtlaender and Leibe, 2017)	85.7	84.2	18.5	77.4	-
CINM (Bao et al., 2018)	83.4	85.0	28.0	**78.4**	77.1
TempSeg-GAN (ours)	85.1	83.3	15.1	77.6	76.8
TempSeg-GAN++ (ours)	**86.3**	**85.2**	14.2	**78.4**	**77.9**

fine-tuning stage to boost the $\mathcal{J}_{Mean}$ value further. It is evident from table 1 that, each of the above factors is important and removing any one of them causes deterioration in terms of quantitative as well as qualitative outputs.

Quantitative Analysis with Existing Methods
The major part of our experiments are performed on the DAVIS 2016 (Perazzi et al., 2016) dataset, which consists of high-resolution video sequences with all of their frames annotated with pixel-level segmentation. For DAVIS, 3 metrics: (i) region similarity in terms of mean Jaccard index ($\mathcal{J}_{Mean}$), (ii) mean contour accuracy ($\mathcal{F}_{Mean}$) and (iii) mean temporal (in-)stability of the segmented masks ($\mathcal{T}_{Mean}$), have been relied upon for evaluation. The validation set of DAVIS 2016 has been used for computation and comparison purposes.

We compare our work with a number of recent and state-of-the-art semi-supervised methods like OnAVOS (Voigtlaender and Leibe, 2017), Masktrack (Perazzi et al., 2017), OSVOS (Caelles et al., 2017), LT (Khoreva et al., 2017), CINM (Bao et al., 2018), RGMP (Wug Oh et al., 2018), FAVOS (Cheng et al., 2018), OFL (Tsai et al., 2016) and BVS (Märki et al., 2016). The quantitative results of our method in comparison with other techniques are shown in table 2. In terms of region similarity $\mathcal{J}_{Mean}$, our baseline *TempSeg-GAN w/o adapt* model beats all other existing techniques except OnAVOS (Voigtlaender and Leibe, 2017) which uses online adaptation, test time augmentation and CRF. On using CRFs and Test-time augmentation on the top of our base network (*TempSeg-GAN++*), the result obtained surpasses OnAVOS. In terms of contour accuracy $\mathcal{F}_{Mean}$, *TempSeg-GAN++* outperforms all other methods, though the base model falls short when compared with CINM (Bao et al., 2018) and OnAVOS. Temporal (in-)stability measure $\mathcal{T}_{Mean}$ of both *Temp-GAN* and *TempGAN++* exhibits dominant performance over all the recent and state-of-the-art semi-supervised methods except RGMP (Wug Oh et al., 2018) which used guided mask propagation as a part of the model (refer column 3 of results under DAVIS 2016, in table 2). Thus, the overall performance of TemSeg-GAN base model along with its modified variant (TemSeg-GAN++) is better than most of the existing methods with small exceptions in few cases. The *Inter-frame Temporal Symmetric Difference Loss (ITSDL)* used in our network can be accounted for the success of our base model which has outperformed majority of the existing state-of-the-art techniques by minimizing the number of mispredicted pixels in segmentation. Thus it has not only increased the $\mathcal{J}_{Mean}$ value, but also has improved the segmented contour by working on small patches. Again, ITSDL along with Intra-Frame Temporal Loss (IFTL) and its long-range variant has contributed to the temporal stability in between the generated masks. The quantitative results in table 2 clarifies the effectiveness of our proposed ITSDL and IFTL objective functions.

For complete evaluation, experimentations are also done on SegTrack-v2 (Li et al., 2013) and YouTube-Objects (Prest et al., 2012) datasets and compared our results (refer table 2) with recent state-of-the-art methods. Due to the lack of proper training set in YouTube-Objects dataset, the same parameters as of DAVIS 2016 has been used and the pre-training step on DAVIS training set is removed to evaluate the generalization capability of our method. While evaluating on this dataset, we have been consistent with (Khoreva et al., 2017) i.e. the frames in which the object of interest are absent has also been included. Both TempSeg-GAN base network (77.6 $\mathcal{J}_{Mean}$) and its mo-

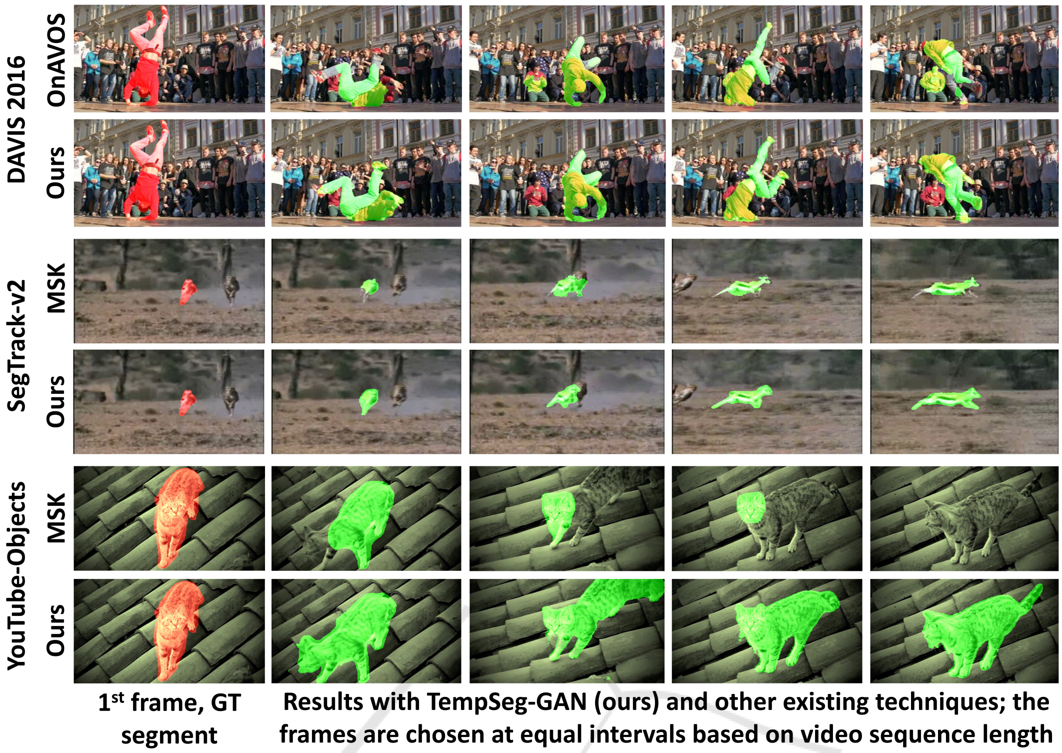
DAVIS 2016
OnAVOS
Ours
SegTrack-v2
MSK
Ours
YouTube-Objects
MSK
Ours
1st frame, GT segment
Results with TempSeg-GAN (ours) and other existing techniques; the frames are chosen at equal intervals based on video sequence length

Figure 2: Qualitative results on three benchmark real-world datasets exhibit that our proposed method gives impressive results in challenging situations like change of appearance, occlusions, background clutter and motion blur, when compared to OnAVOS (Voigtlaender and Leibe, 2017) and MSK (Perazzi et al., 2017) (best viewed in color).

dified variant (78.4 $\mathcal{J}_{Mean}$) give better results than OnAVOS (refer table 2). TempSeg-GAN++ performs at par with the state-of-the-art CINM model.

Assessment on the SegTrack-v2 dataset is performed following the similar choice of setting as done in recent existing methods (Perazzi et al., 2017; Khoreva et al., 2017). Same protocols as of DAVIS 2016 evaluation are carried out by fine-tuning on the initial frame of the test video sequences. Table 2 shows that modified version of our base network (*TempSeg-GAN*) exhibits promising result in comparison with the recent state-of-the-art models. The comparative study of the qualitative results of our proposed network with OnAVOS (Voigtlaender and Leibe, 2017) and MaskTrack (Perazzi et al., 2017) on three popular real-world datasets is shown in Figure 2, where our method performs well in segmenting specific object of interest under difficult conditions like background clutter, viewpoint change, motion blur, occlusions and shape deformation of object. More visual results of TempSeg-GAN on the three real-world datasets are shown in figure B.1 of the Appendix.

7 CONCLUSION

The paper proposes a temporally aided Generative Adversarial Network for the purpose of Video Object Segmentation. The generator of the model is modified by implementing an encoder-decoder type architecture with skip connections, along with a variation in the discriminator training by introducing an additional encoder module. Introduction of Inter-frame Temporal Symmetric Difference Loss (ITSDL) and Intra-frame Temporal Loss (IFTL) not only provides a significant improvement in the segmentation results over the existing state-of-the-art techniques, but also preserves the motion features among the generated masks. Quantitative results on three benchmark datasets reveals the superiority of TempSeg-GAN over other recent state-of-the-art methods. This work can be effectively implemented to segment traffic signs, vehicles and other obstacles in the context of autonomous cars.

REFERENCES

Bao, L., Wu, B., and Liu, W. (2018). Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5977–5986.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision (ECCV)*, pages 850–865. Springer.

Braham, M. and Van Droogenbroeck, M. (2016). Deep background subtraction with scene-specific convolutional neural networks. In *IEEE International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–4. IEEE.

Caelles, S., Maninis, K.-K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., and Van Gool, L. (2017). One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Cheng, J., Tsai, Y.-H., Hung, W.-C., Wang, S., and Yang, M.-H. (2018). Fast and accurate online video object segmentation via tracking parts.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680.

Goroshin, R., Bruna, J., Tompson, J., Eigen, D., and LeCun, Y. (2015). Unsupervised learning of spatiotemporally coherent metrics. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4086–4093.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 6.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456.

Jain, S. D. and Grauman, K. (2014). Supervoxel-consistent foreground propagation in video. In *European Conference on Computer Vision (ECCV)*, pages 656–671. Springer.

Khoreva, A., Benenson, R., Ilg, E., Brox, T., and Schiele, B. (2017). Lucid data dreaming for object tracking. In *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems (NIPS)*, pages 109–117.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105.

Li, F., Kim, T., Humayun, A., Tsai, D., and Rehg, J. M. (2013). Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2199.

Li, H., Li, Y., and Porikli, F. (2016). Deeptrack: Learning discriminative feature representations online for robust visual tracking. *IEEE Transactions on Image Processing (TIP)*, 25(4):1834–1848.

Luc, P., Couprie, C., Chintala, S., and Verbeek, J. (2016). Semantic segmentation using adversarial networks. In *NIPS Workshop on Adversarial Training*.

Maninis, K.-K., Pont-Tuset, J., Arbeláez, P., and Van Gool, L. (2016). Convolutional oriented boundaries. In *European Conference on Computer Vision (ECCV)*, pages 580–596. Springer.

Märki, N., Perazzi, F., Wang, O., and Sorkine-Hornung, A. (2016). Bilateral space video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 743–751.

Mobahi, H., Collobert, R., and Weston, J. (2009). Deep learning from temporal coherence in video. In *International Conference on Machine Learning (ICML)*, pages 737–744. ACM.

Nam, H. and Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4293–4302.

Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., and Sorkine-Hornung, A. (2017). Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., and Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732.

Prest, A., Leistner, C., Civera, J., Schmid, C., and Ferrari, V. (2012). Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3282–3289. IEEE.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Souly, N., Spampinato, C., and Shah, M. (2017). Semi supervised semantic segmentation using generative adversarial network. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5689–5697. IEEE.

Tsai, Y.-H., Yang, M.-H., and Black, M. J. (2016). Video segmentation via object flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3899–3908.

Voigtlaender, P. and Leibe, B. (2017). Online Adaptation of Convolutional Neural Networks for Video Object Segmentation. In *British Machine Vision Conference (BMVC)*.

Wang, N. and Yeung, D.-Y. (2013). Learning a deep compact image representation for visual tracking. In *Advances in Neural Information Processing Systems (NIPS)*, pages 809–817.

Wang, Y., Luo, Z., and Jodoin, P.-M. (2017). Interactive deep learning method for segmenting moving objects. *Pattern Recognition Letters*, 96:66–75.

Wug Oh, S., Lee, J.-Y., Sunkavalli, K., and Joo Kim, S. (2018). Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7376–7385.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer.

APPENDIX

A Network Architecture Details

The details of the generator (G), discriminator (D) and encoder (E) networks of our proposed TempSeg-GAN are presented in table A.1 for experimental analysis . All the convolution layers of G follow ReLU non-linearity, while batch-normalization and dropout at a rate of 50% are also included in the network. G is an encoder-decoder network where skip connections connect layer i with layer $(n-i)$ by concatenating the feature maps of former with that of later. Unpooling layers are also used to upsample the image by a factor of 2 into higher resolution in terms of both width and height. G is initialized with pre-trained weights of ImageNet and the learning rate is fixed to 0.002 for training purpose, which decreases gradually over time upto 0.0004. The learning rate of the discriminator (D) network is set to 0.01 and also uses ReLU non-linearities. For training the entire network, mini-batches of 50 frames of video sequences were used.

Table A.1: Network architecture details; *G*, *D* and *E* are the generator, discriminator and encoder networks respectively.

Network	G	D	E
Number of feature maps	64, 128, 256, 512, 512, 512, 512, 512, 256, 128, 64	256, 512, 512	64, 128, 256
Kernel sizes	5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 5	3, 5, 5	5, 3, 3
Fully connected	N/A	1024, 512	N/A

B Qualitative Results of TempSeg-GAN

More qualitative results of our TempSeg-GAN base network on three benchmark real-world datasets *viz.* DAVIS 2016, SegTrack-v2 and YouTube-Objects datasets, have been shown in figure B.1.

Additional Illustrations

Apart from figures B.1-B.2, we also provide video output using .gif format. We perform qualitative evaluation on two video clips, one from each of DAVIS and SegTrack-v2 respectively. The video outputs on DAVIS 2016 (DAVIS.gif) and SegTrack-v2 (Segtrackv2.gif) contains the comparison of TempSeg-GAN with the existing state-of-the-art techniques like OnAVOS (Voigtlaender and Leibe, 2017) and MaskTrack (Perazzi et al., 2017) respectively. Both the video clips show the superiority of the TempSeg-GAN as the result is quite close to the ground-truth.

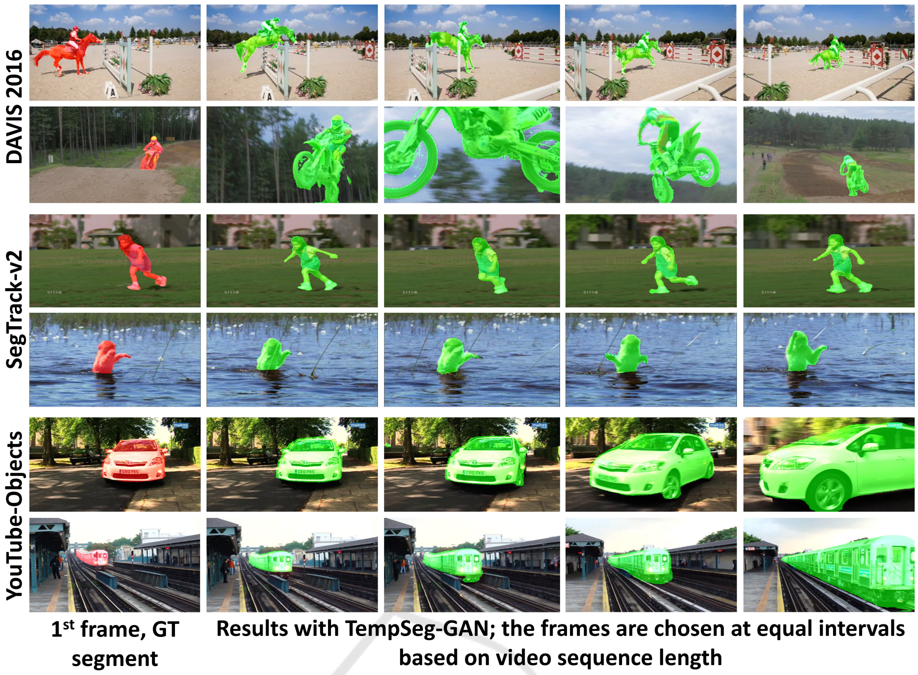
DAVIS 2016
SegTrack-v2
YouTube-Objects
1st frame, GT segment
Results with TempSeg-GAN; the frames are chosen at equal intervals based on video sequence length

Figure B.1: Qualitative results on three benchmark real-world datasets exhibit that our proposed method gives impressive results in challenging situations like change of appearance, occlusions, camera view change, background clutter and motion blur (best viewed in color).

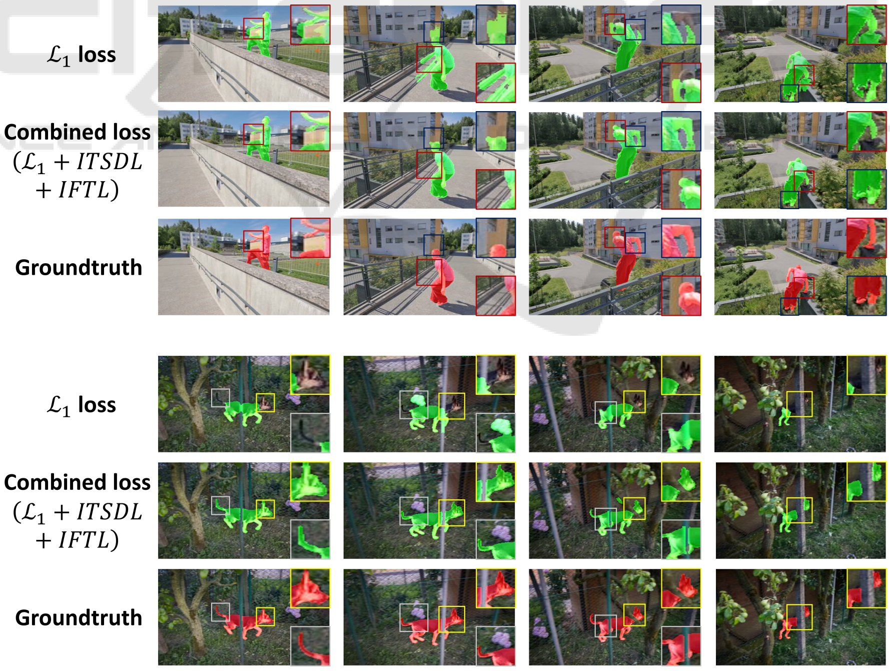
$\mathcal{L}_1$ loss
Combined loss ($\mathcal{L}_1 + ITSDL + IFTL$)
Groundtruth
$\mathcal{L}_1$ loss
Combined loss ($\mathcal{L}_1 + ITSDL + IFTL$)
Groundtruth

Figure B.2: Comparative study of predicted segmentation results on DAVIS 2016 obtained from our TempSeg-GAN models using only $\mathcal{L}_1$ loss and Combined ($\mathcal{L}_1 + ITSDL + IFTL$) loss (refer eqn. 10) respectively. $ITSDL$ and $IFTL$ denotes the proposed Inter-frame Temporal Symmetric Difference Loss (section 3) and Intra-Frame Temporal Loss (section 4) respectively. Figures in insets show zoomed-in patches for better visibility of the estimated segmented masks in areas with background clutter, occlusion and significant motion blur. Using only $\mathcal{L}_1$ loss produces holes in the segmented mask and some of them contain small blobs generated outside the region of interest causing inaccurate segmentation. On the other hand, $ITSDL + \mathcal{L}_1 + IFTL$ produces impressive results, where $ITSDL$ helps in removing the blob like artifacts, thereby improving the contours of the output masks, while $IFTL$ preserves the temporal consistency between the generated masks (best viewed in color).