EECS 542 Final Project: Unguided DAVIS

Group member: Baikai Li, Zhenzhen Dai, Tianhang Gao, Yu Liu
Group number: 19

1 Problem Description

Public benchmarks and challenges have been an important driving force in the computer vision field. Several state-of-the-art segmentation has achieved outstanding results with examples such as Imagenet for scene classification and object detection, PASCAL for semantic and object instance segmentation. Video object segmentation is a binary labeling problem which separate foreground objects from the background. Despite remarkable progress in recent years, video object segmentation still remains a challenging. In this project, we challenged the Densely-Annotated Video Segmentation (DAVIS) dataset which consists of 50 high-definition sequences with all their frames annotated with object masks at pixel-level accuracy and implemented Generative Adversarial Nets (GAN) to solve this problem.

2 Network Description

We implemented two GANs to solve the problem. The 1st GAN is trained to perform the single frame segmentation, through which we can roughly get "what we may be interested in". The 2nd GAN is trained perform object tracking, where the input is the current frame masked with output of the last frame generated by GAN1. This is a process of fine tuning using the information we get from the last frame. After this we can get "it is the object we want at this particular moment".

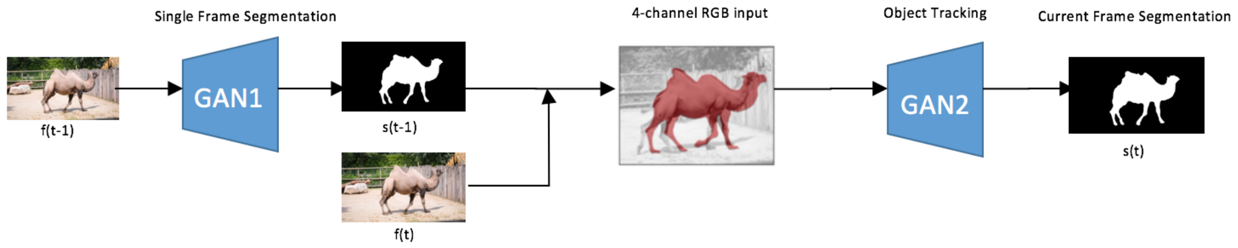


Fig 2.1 Overall Network Architecture

2.1 The 1st GAN architecture

Within the generator, we add skip connection between each encoder and decoder (both of the same size), for sharing some low-level features. For an instance, both inputs and outputs share the location of prominent edges. For this part, we want GAN to generate coarse segmentation of the object for the 2nd part.

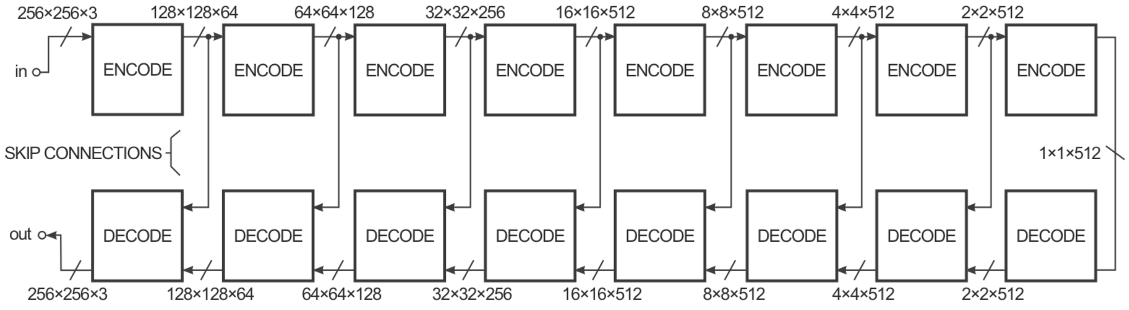


Fig 2.2 Architecture of Generator

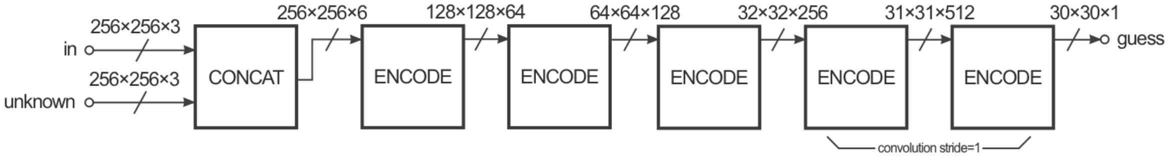


Fig 2.3 Architecture of Discriminator

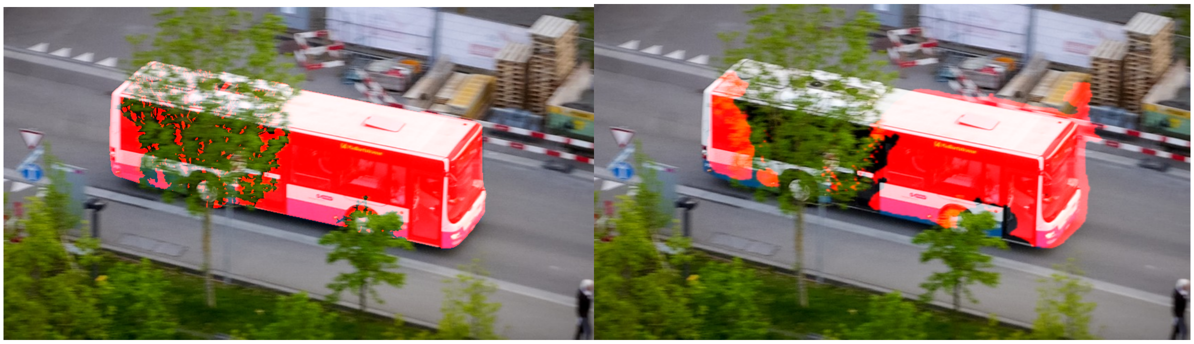
2.2 The GAN2 structure

For the 2nd part, we also use the generator structure like the 1st part, but we concatenate the estimated mask of the (i-1)th frame produced by the 1st generator with the input image of the ith frame as the input of our 2nd ConvNet. We want the generator to capture some motion features based on the mask of last frame, which gives the estimation of shape and position of the object.

3 Detailed Training Process

For GAN1 Training:
We use ADAM optimizers with fixed learning rate 2e-4 for 70k iterations based on DAVIS 2016 dataset. And we use the trained model to produce estimated mask of all the training images for the 2nd part.

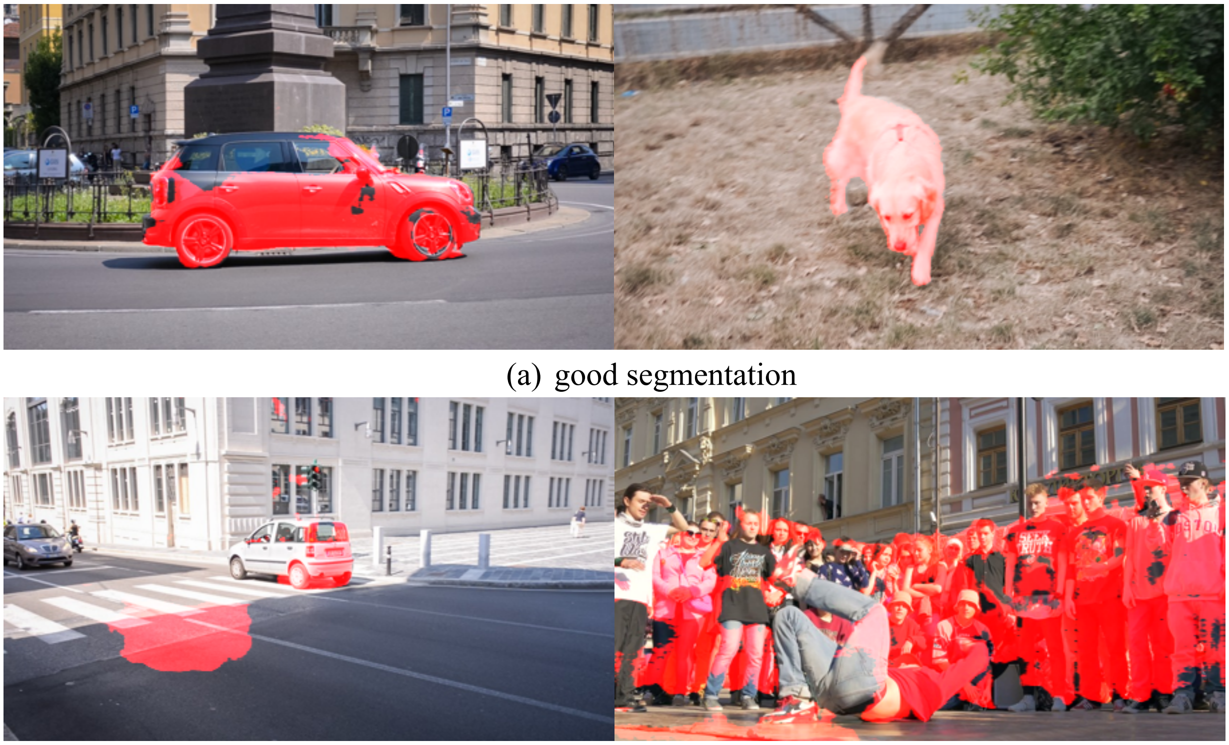
For GAN2 Training:
We begin with expanding the convnet input from RGB to RGB+mask channel (4 channels), and use ADAM optimizers with fixed learning rate 2e-5 for 100k iterations.
The example of the RGB+mask image shown below:



(a) RGB + Annotated image (b) Example RGB+mask image

Fig 3.1 4-channel Masked Image

4 Result



(a) good segmentation

(b)bad segmentation

Fig 4.1 Segmentation Results

	GAN	FST	SAL	KEY	MSG	TRC	CVOS	NLC
J Mean	41.05	57.5	42.6	56.9	54.6	50.1	51.4	**64.1**
J Recall	36.59	65.2	38.6	67.1	63.6	56.0	58.1	**73.1**
J Decay	**1.0**	4.4	8.4	7.5	2.8	5.0	12.7	8.6
F Mean	44.13	53.6	38.3	50.3	52.5	47.8	49.0	**59.3**

F Recall	40.26	57.9	26.4	53.4	61.3	51.9	57.8	**65.8**
F Decay	**1.4**	6.5	7.2	7.9	5.7	6.6	13.8	8.6

Table 4.1 Quantitative Segmentation Result

	Average Training Rate	Training Time
GAN1	5.3 images/second	13:18:07
GAN2	4.6 images/second	10:11:35

Table 4.2 Running Time

5 Difficulties

1. We spent large amount of time on finding the model. We have tried both FCN and GAN, and find GAN gets a better performance.

2. One of the challenges of this project is that the training data has relatively few images. We then tried to external dataset (e.g. ImageNet, COCO, etc.) to pretrain the model, but the effect turned out to be bad, because we had to write code to find the “groundtruth” by ourselves. However, the groundtruth is coarse and there are images to have multiple objects, therefored the trained model by these datasets would not work well. We finally gave up this idea and used only the officially given DAVIS datasets.

3. In the second stage, we did not know at first how to implement the network to let it “track” the object. Compared with common image segmentation, video segmentation has a relationship between the neighboring frames. We tried to use RNN with LSTM at first, but we were confused on how to set value to the input gate, output gate and forget gate. We finally decide to train another GAN by inputting a learned mask and its next video frame.

4. Our model performs well on the slowly-moved, single object, but we find it difficult to segment certain videos. For example, the video with dancer in the center and audiences around. The reason is that the model extracts the features of human, but fails to discriminate which person should be the foreground.

6 Alternative ways to approach the project

1. Use a CNN (or FCN) pretrained on image recognition on ImageNet as the “base network”.
2. Use RNN with LSTM to “fine tune” the learned mask. We need to feed proper values to the input gate, output gate and forget gate.

Reference

[1] Perazzi, Federico, et al. "A benchmark dataset and evaluation methodology for video object segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[2] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

[3] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *arXiv preprint arXiv:1611.07004* (2016).

[4] Caelles, Sergi, et al. "One-Shot Video Object Segmentation." *arXiv preprint arXiv:1611.05198* (2016).