STESA-GRU: Spatio Temporal Self Attended Gated Recurrent Units for Video Object Segmentation

Diego Valderrama & Mateo Rueda

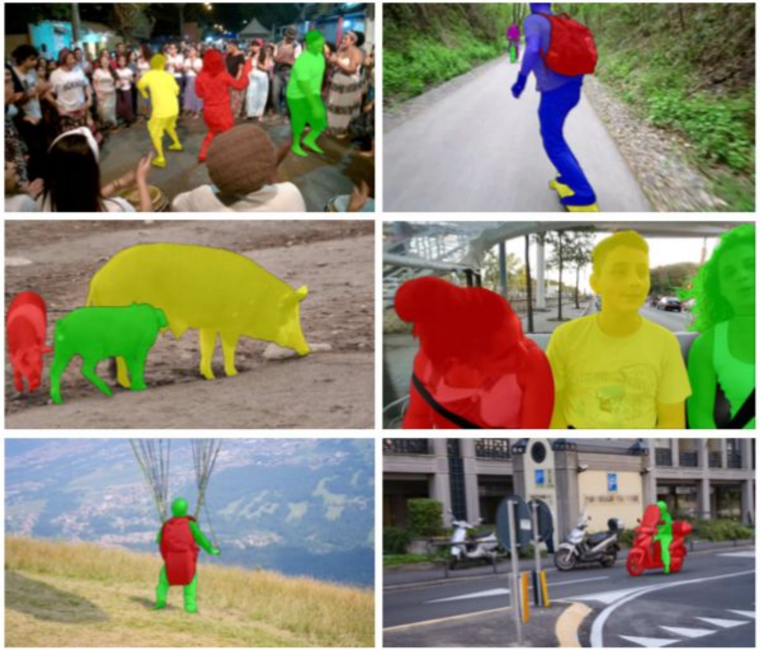
INTRODUCTION

INTRODUCTION

- Separate objects from a video sequence background
- Beneficial applications
  - Video editing
  - Scene understanding
  - Autonomous vehicles

INTRODUCTION

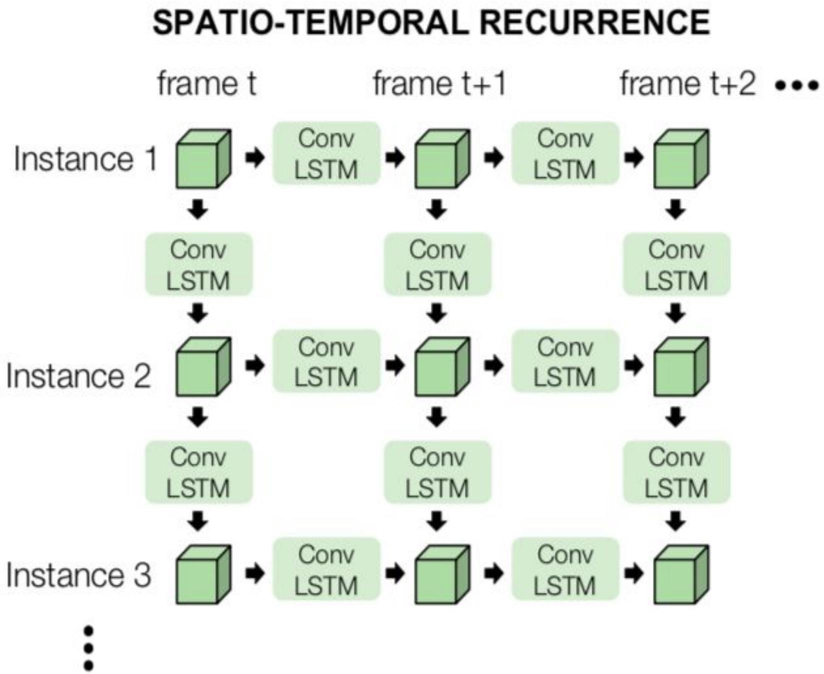
- Semi-supervised task
- Predict masks from the first frame ground-truth
- Motion based or detection based approaches



BASELINE

RVOS

- Encoder - Decoder architecture
- Resnet 101 as backbone
- LSTM spatio-temporal
- Segmentation per instance
- Off-line approach



APPROACH

APPROACH

- Conv GRU spatio-temporal
  - Two hidden states
  - Conv reset gate
  - Conv update gate
  - Conv “current”
  - Reset hidden

$$r_g = \sigma\Big(W_{xr}x_t + W_{hr}h_{t-1} + b_r\Big)$$

$$u_g = \sigma\Big(W_{xu}x_t + W_{hu}h_{t-1} + b_u\Big)$$

$$c = \sigma\Big(W_{xc}x_t + b_c\Big)$$

$$r_h = r_g * \Big(hS_{t-1} + hT_{t-1}\Big)$$

$$\tilde{h_t} = tanh\Big(c + r_h\Big)$$

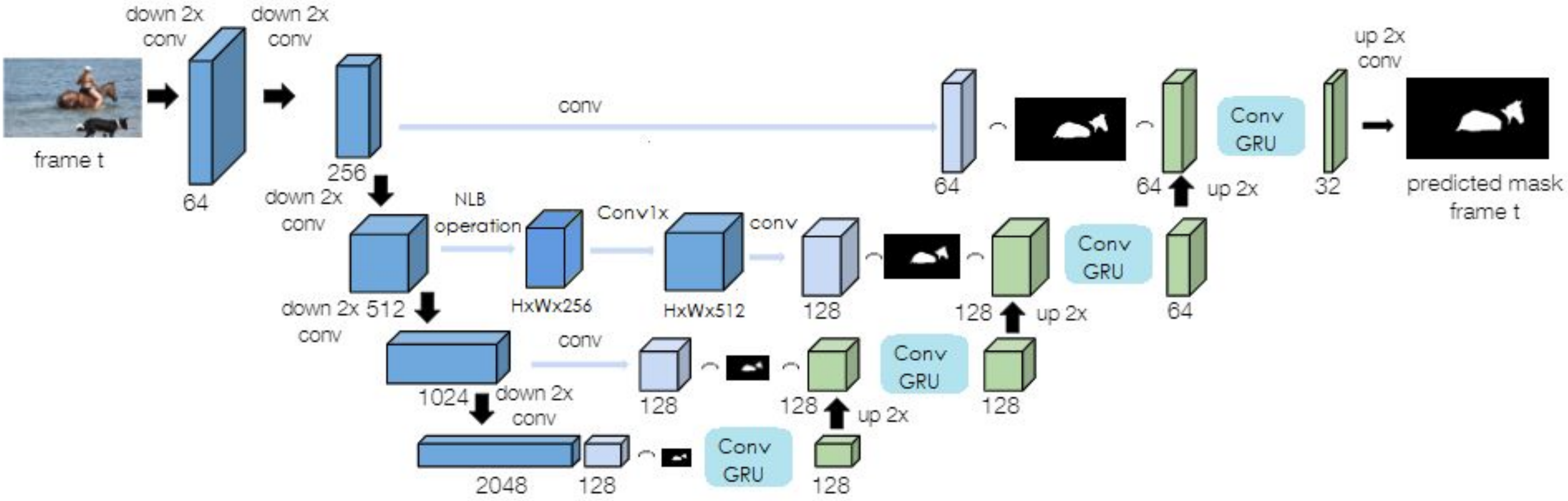
$$h = (u_g * hS_{t-1}) + (1 - u_g) * \tilde{h_t}$$

APPROACH

- Self- Attention
  - 2D Non Local Block Module after *res3*
  - Auxiliary Loss

$$Loss = mIoU_{NLB} + mIoU_P$$

STESA-GRU



down 2x
conv
down 2x
conv
frame t
64
256
conv
64
down 2x
conv
NLB
operation
HxWx256
Conv1x
HxWx512
conv
128
down 2x 512
conv
conv
1024
down 2x
conv
2048
128
Conv
GRU
128
128
up 2x
128
Conv
GRU
128
128
up 2x
Conv
GRU
64
64
up 2x
Conv
GRU
up 2x
conv
32
predicted mask
frame t

EXPERIMENTS

DAVIS 2017

- 60 videos - Train split
- 30 videos - Val split
- 30 videos - Test-dev

YOUTUBE VOS V1

- 3471 videos - Train split

- Region similarity (IoU)
- F-Measure score

TRAINING DETAILS

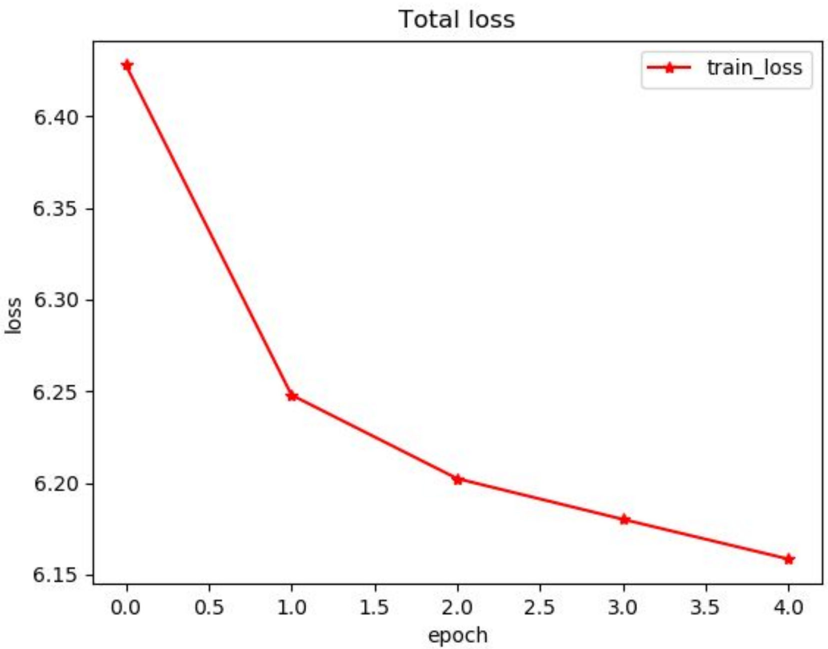
- Frames and annotations have been resized to 256x448
- Each mini-batch is composed of 4/3 videos and 5 consecutive frames
- 5/10 epochs in YouTube-VOS and 40/20 in DAVIS 2017
- Learning rate of 10^{-6} with Adam optimizer
- Single GPU TESLA K40c

RESULTS

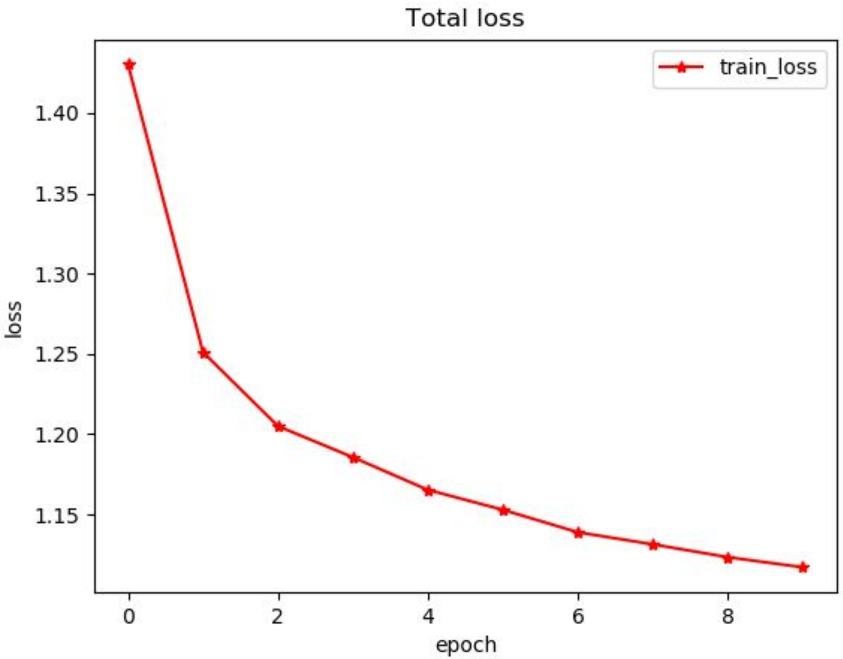
Model	OL	$\mathcal{J}-\mathcal{F}$ mean	$\mathcal{J}$	$\mathcal{F}$
PremVOS [13]	Yes	71,6	67,5	75,7
MRF [2]	Yes	67,5	64,5	70,5
OnAVOS [22]	Yes	56,5	53,4	59,6
FeelVOS [21]	No	57,8	55,2	60,5
RGMP [16]	No	52,9	51,4	54,4
RVOS [20]	No	50,3	47,9	52,6
Ours	No	42.6	40.6	44.6
Ours w/o AL	No	44.3	42.9	45.7
RVOS$_{w/o}$	No	33,6	32,1	35,0
Ours$_{w/o}$	No	36,1	33,9	38,4

Table 1. Comparison against state-of-the-art results for semi-supervised video object segmentation in DAVIS-2017 test-dev. OL refers to online learning. $RVOS_{w/o}$ and $Ours_{w/o}$ are the models without the pre-training on YouTube-VOS dataset. Ours w/o AL refers to our model pretrained without the auxiliary loss

RESULTS

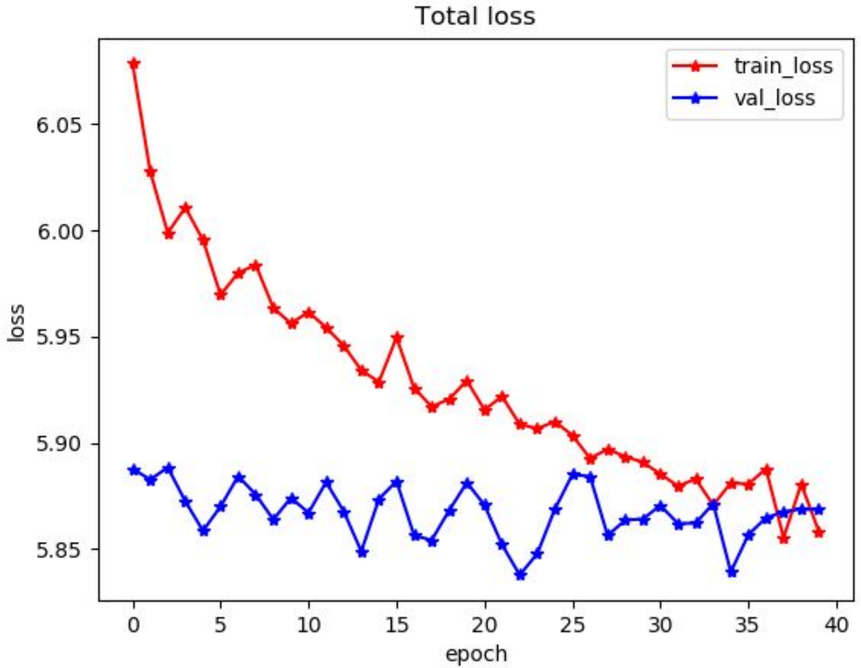


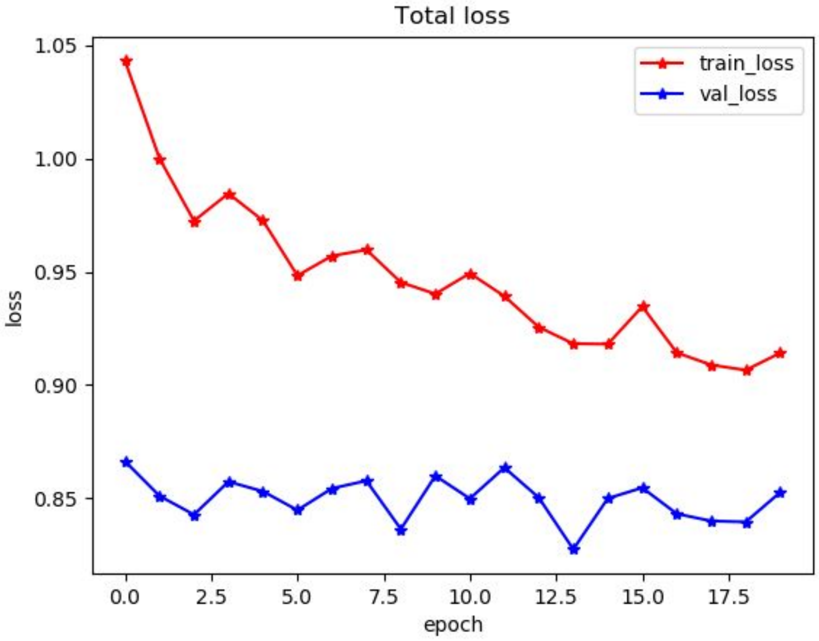
Total loss
train_loss
6.40
6.35
6.30
6.25
6.20
6.15
loss
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
epoch

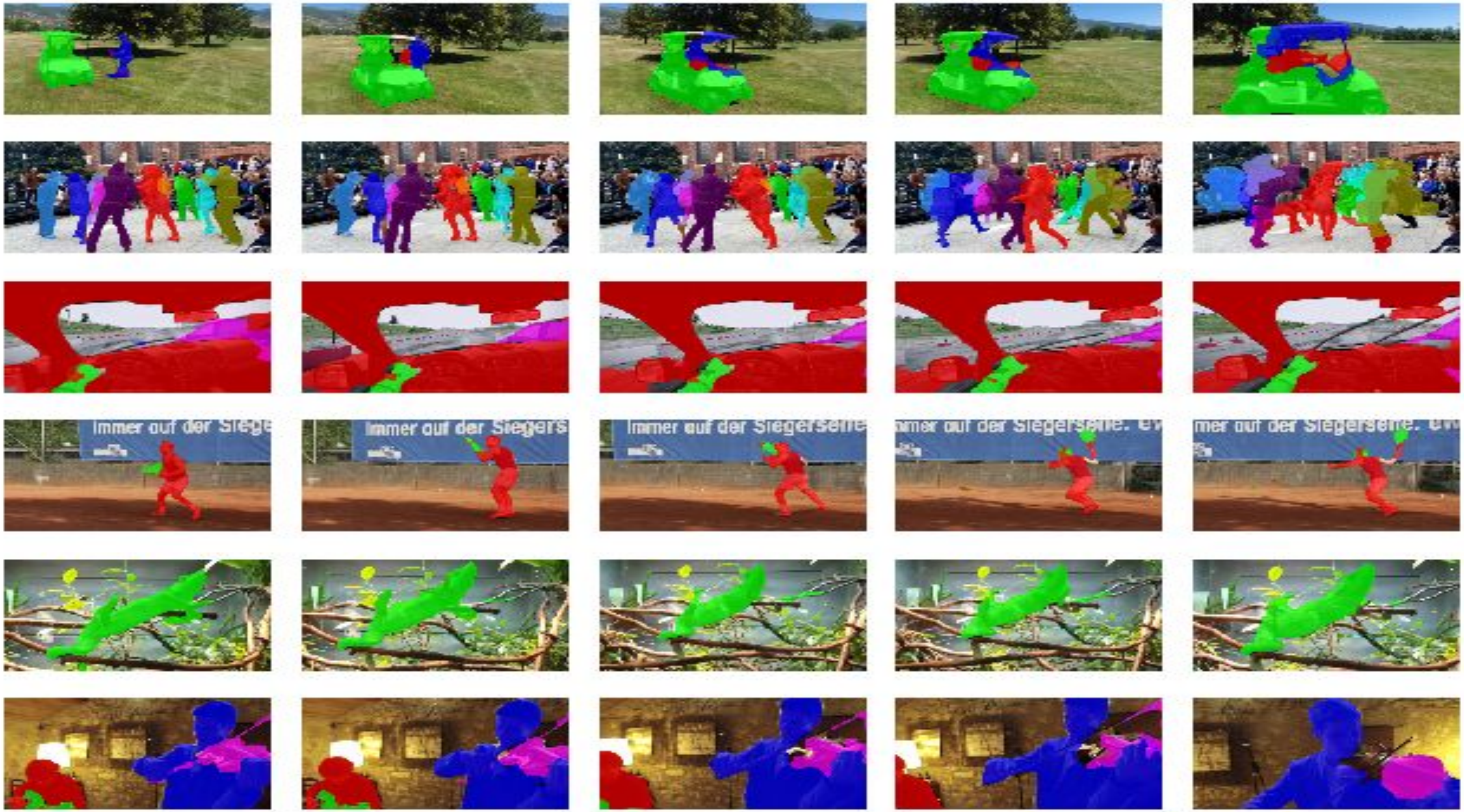


Total loss
train_loss
1.40
1.35
1.30
1.25
1.20
1.15
loss
0
2
4
6
8
epoch

RESULTS


Total loss
train_loss
val_loss
loss
epoch
6.05
6.00
5.95
5.90
5.85
0
5
10
15
20
25
30
35
40


Total loss
train_loss
val_loss
loss
epoch
1.05
1.00
0.95
0.90
0.85
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Immer auf der

ABLATION STUDY

Model	FT	$\mathcal{J}-\mathcal{F}$ mean	$\mathcal{J}$	$\mathcal{F}$
Ours	No	41.5	40.1	42.9
Ours w/o AL	No	42.9	40.8	44.9
Ours	Yes	42.6	40.6	44.6
Ours w/o AL	Yes	44.3	42.9	45.7

Table 2. Comparison between model which are just pretrained and those that are fine-tuned in DAVIS dataset.All the results are on DAVIS-2017 test-dev set. FT refers to fine-tuning on DAVIS dataset.Ours w/o AL refers to our model without the auxiliary loss

ABLATION STUDY

Model	r_g	$\mathcal{J} - \mathcal{F}$ mean	$\mathcal{J}$	$\mathcal{F}$
Ours w/o NLB	1	**35,6**	**34,2**	**37,0**
Ours w/o NLB	2	31,7	30,1	33,3

Table 3. Ablation study about number of reset gates (r_g) on GRU implementation in DAVIS 2017 dataset without pre-training in YouTube-VOS.

ABLATION STUDY

Recurrence	NLB	$\mathcal{J}-\mathcal{F}$ mean	$\mathcal{J}$	$\mathcal{F}$
GRU	Res3	**35,4**	**33,5**	**37,4**
GRU	Res4	34,6	32,4	36,8
GRU	Res5	30,4	28,9	32,0
LSTM	Res3	31,7	30,0	33,4
LSTM	Res4	31,2	29,8	32,5
LSTM	Res5	32,1	30,6	33,7

Table 4. Ablation study about the effect of using the NLB module after different parts of the decoder architecture with different kind of recurrence in DAVIS 2017 dataset without pre-training in YouTube-VOS

ABLATION STUDY

Model	NLB	$\mathcal{J}-\mathcal{F}$ mean	$\mathcal{J}$	$\mathcal{F}$
STESA-GRU	Res3	**36,1**	**33,9**	**38,4**
STESA-GRU$_2$	Res4	33,7	31,9	35,6

Table 5. Ablation study about the effect of using the NLB module with auxiliary loss after different positions in DAVIS 2017 dataset without pre-training in YouTube-VOS

CONCLUSIONS

- End-to-end trainable spatio temporal recurrent network with a self-attention module
- Comparable results to method which not use online learning in DAVIS test-dev set
- It is necessary to prove training with different learning rate values
- Evaluate the model in YouTube-VOS validation set