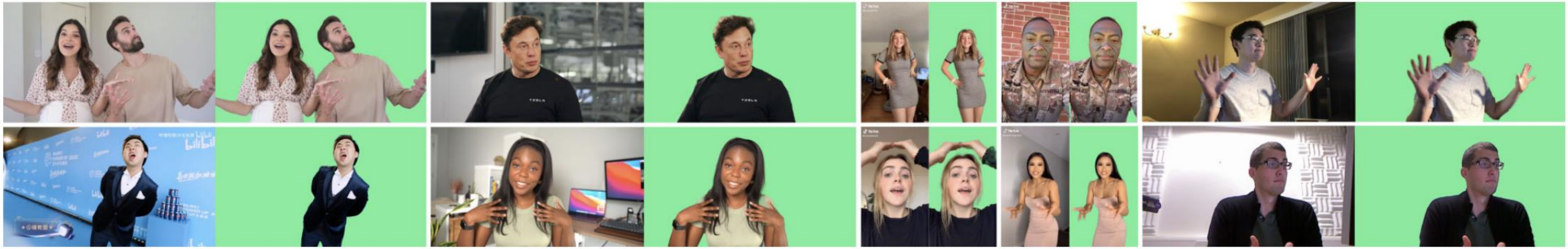
Robust High-Resolution Video Matting with Temporal Guidance

Shanchuan Lin[1*] Linjie Yang[2] Imran Saleemi[2] Soumyadip Sengupta[1]
[1]University of Washington [2]ByteDance Inc.
{linsh,soumya91}@cs.washington.edu {linjie.yang,imran.saleemi}@bytedance.com



Background Matting: The World is Your Green Screen

Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman

University of Washington

CVPR 2020

Real-Time High-Resolution Background Matting

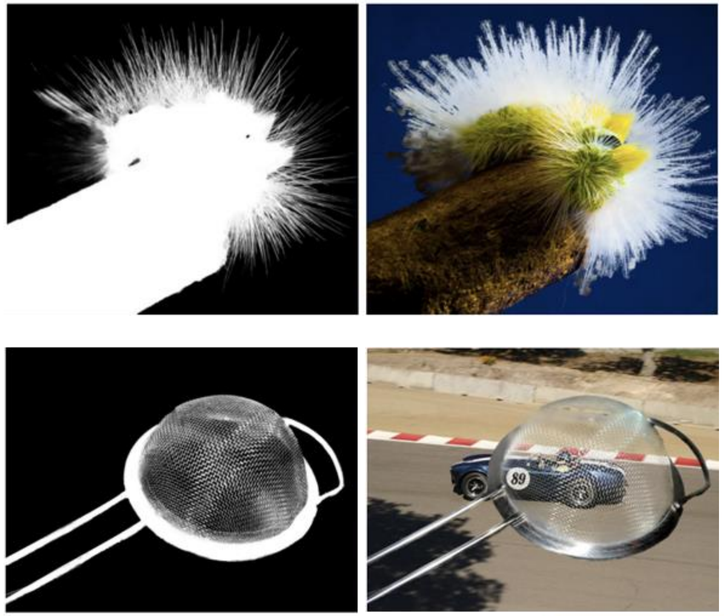
Shanchuan Lin* Andrey Ryabtsev* Soumyadip Sengupta
Brian Curless Steve Seitz Ira Kemelmacher-Shlizerman
University of Washington
{linsh,ryabtsev,soumya91,curless,seitz,kemelmi}@cs.washington.edu

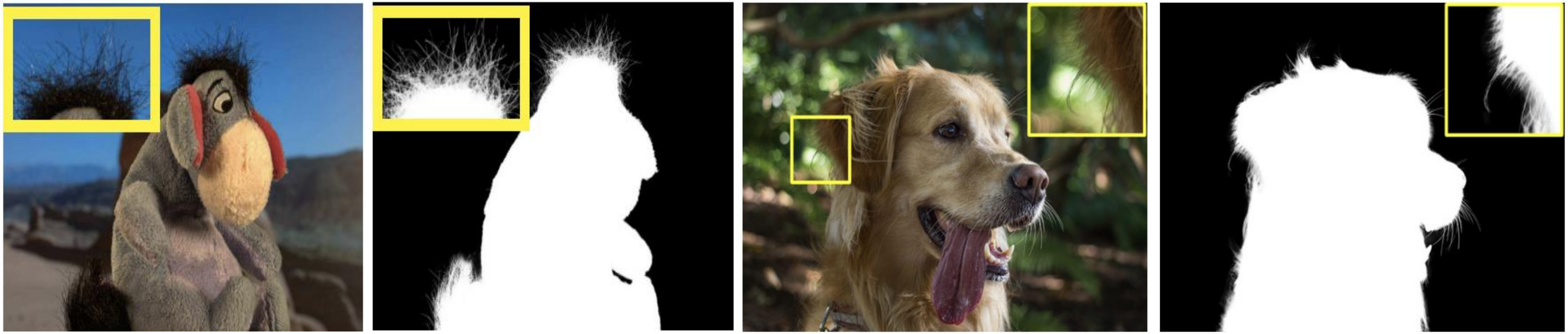
Zoom input and background shot | Zoom with new background | Our Zoom plugin with new background

CVPR 2021 (oral)

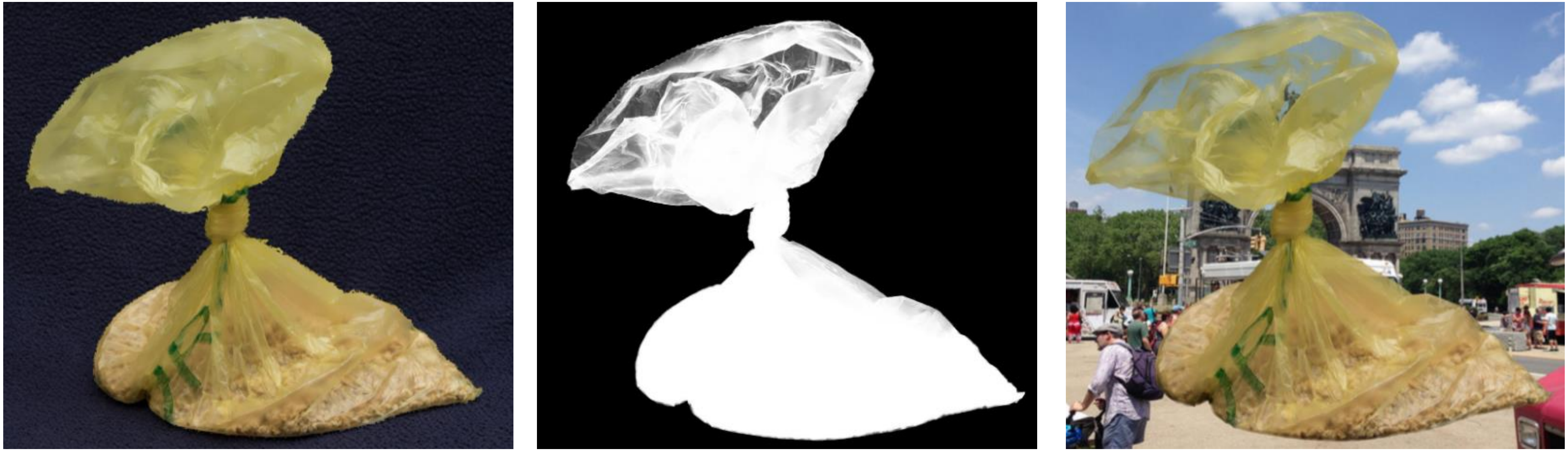
Image Matting

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \quad \alpha_i \in [0, 1].$$





$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \quad \alpha_i \in [0, 1].$$



Guidance

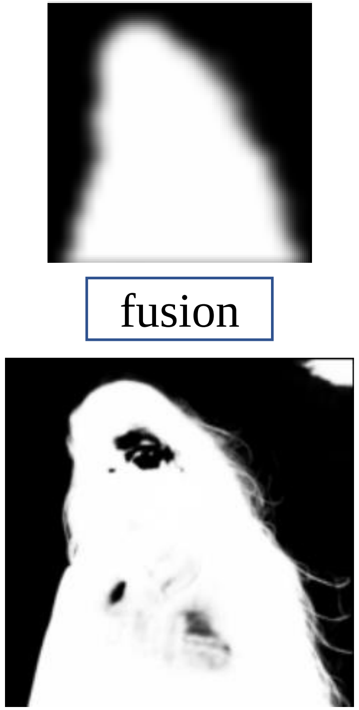
- Trimap

- Background Image

- Mask

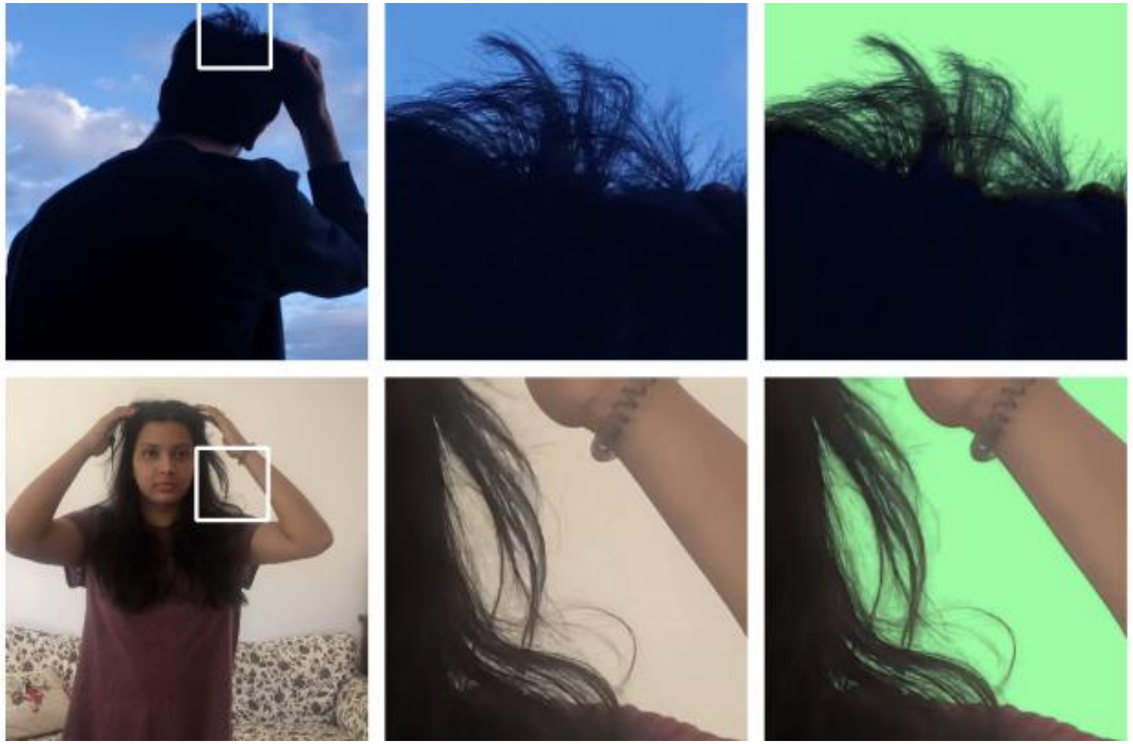
Method without guidance

- For specific objects
  - Semantic information
  - Detail edge information



Video Matting

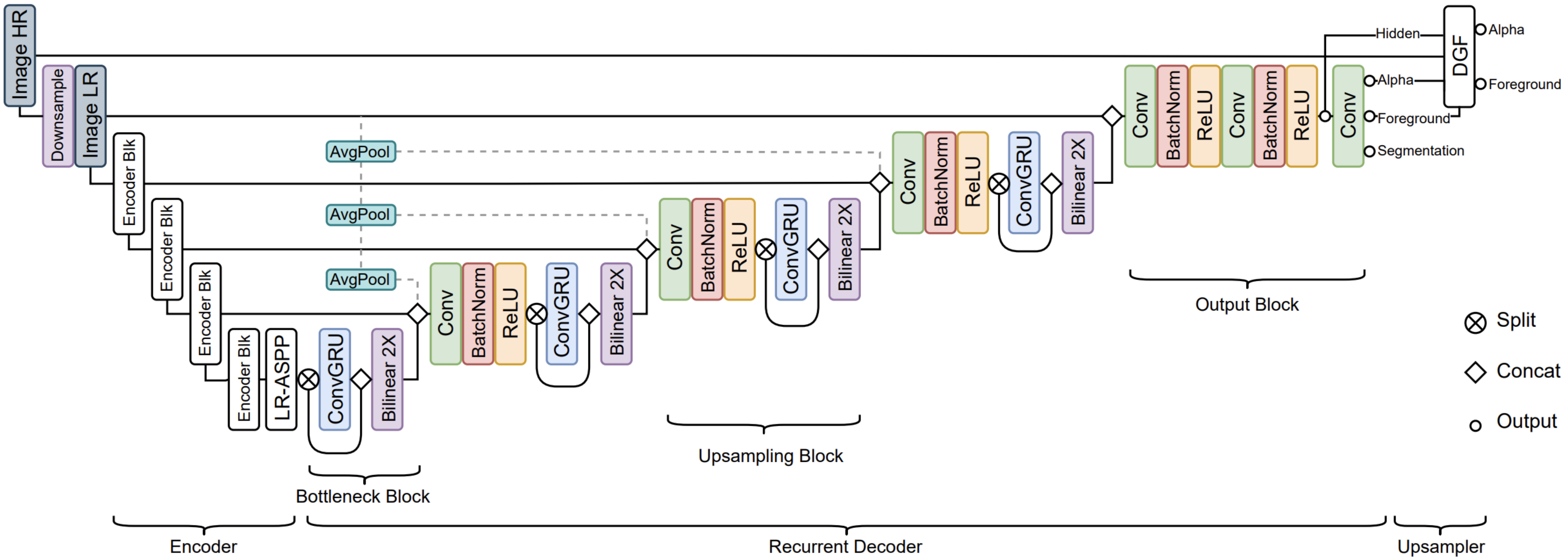
- Faster
  - End-to-end
  - More applicable scenarios
  - Low complexity of objects
- Less monitoring information
  - Static background image
  - First frame trimap
  - No guidance
- Adapt to higher resolution



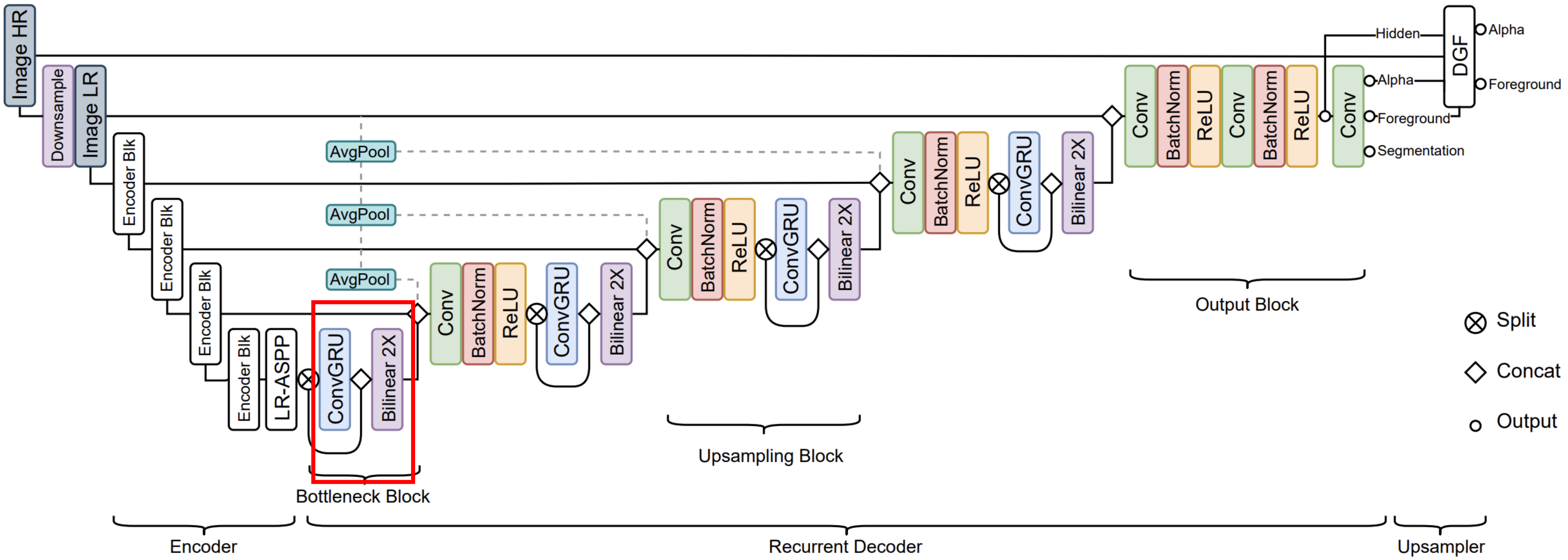
Main Difficulties

- **How to aggregate timing information between video frames?**
- **How to achieve end-to-end real-time matting?**
- **How to meet the high-resolution needs of matting?**

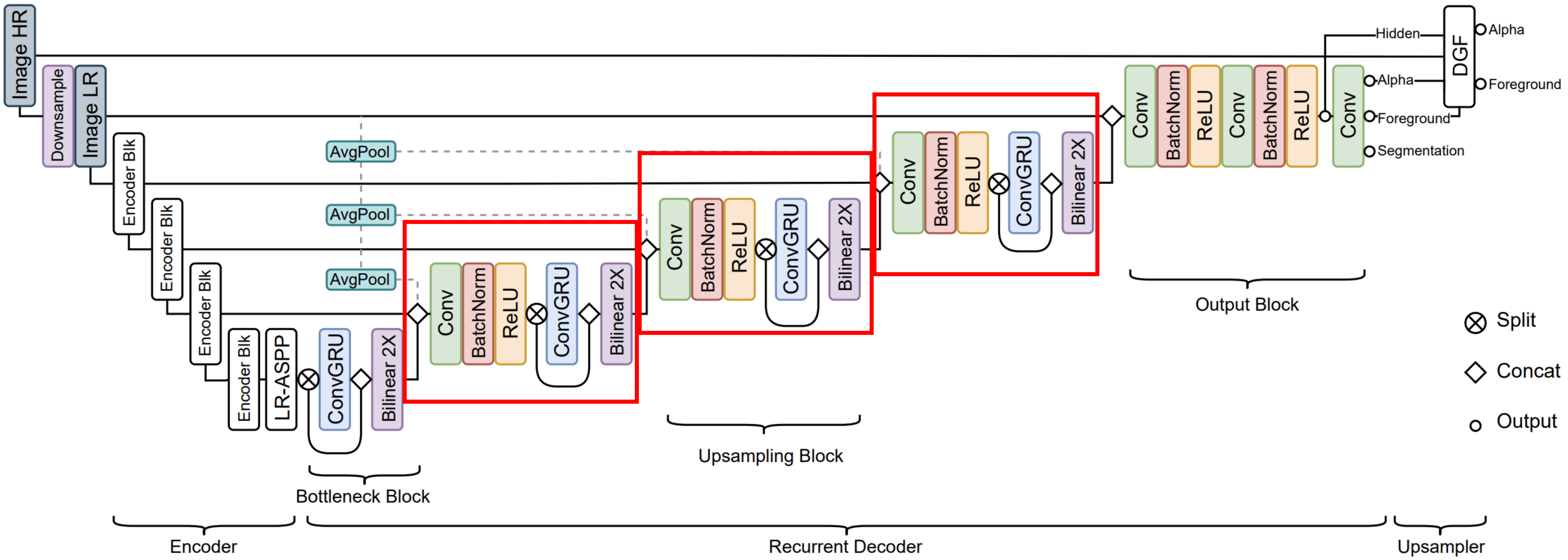
Network Structure



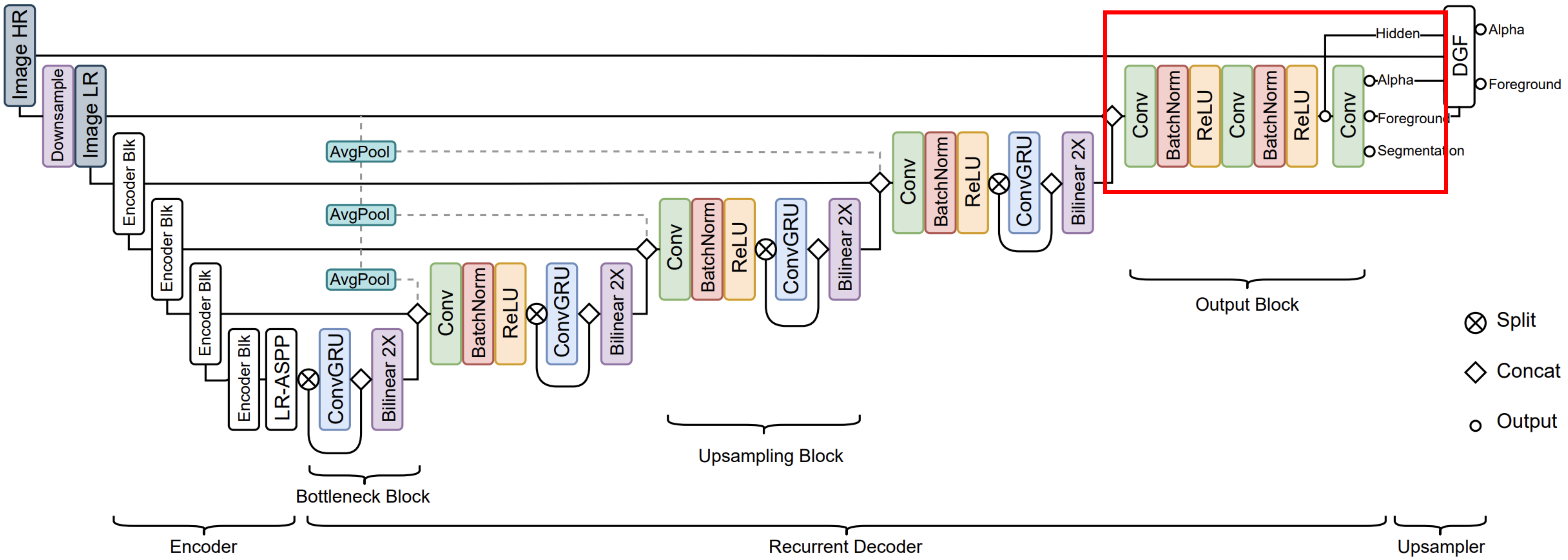
Bottleneck Block



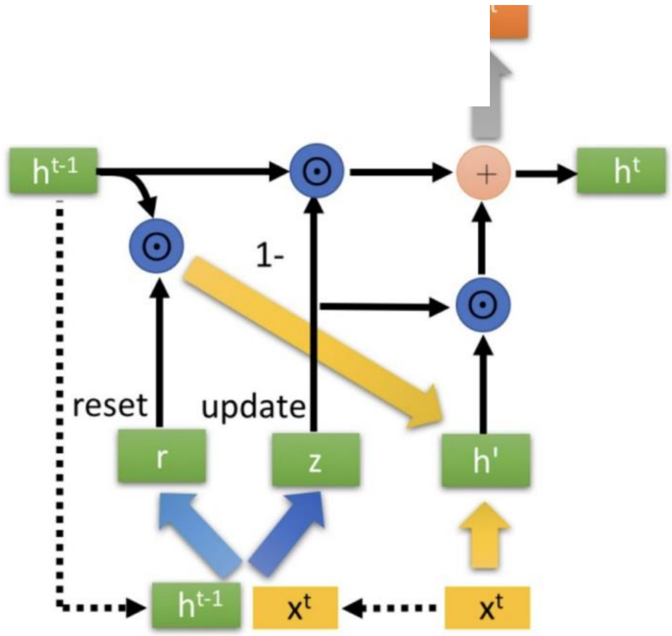
Upsampling Block



Output Block



ConvGRU

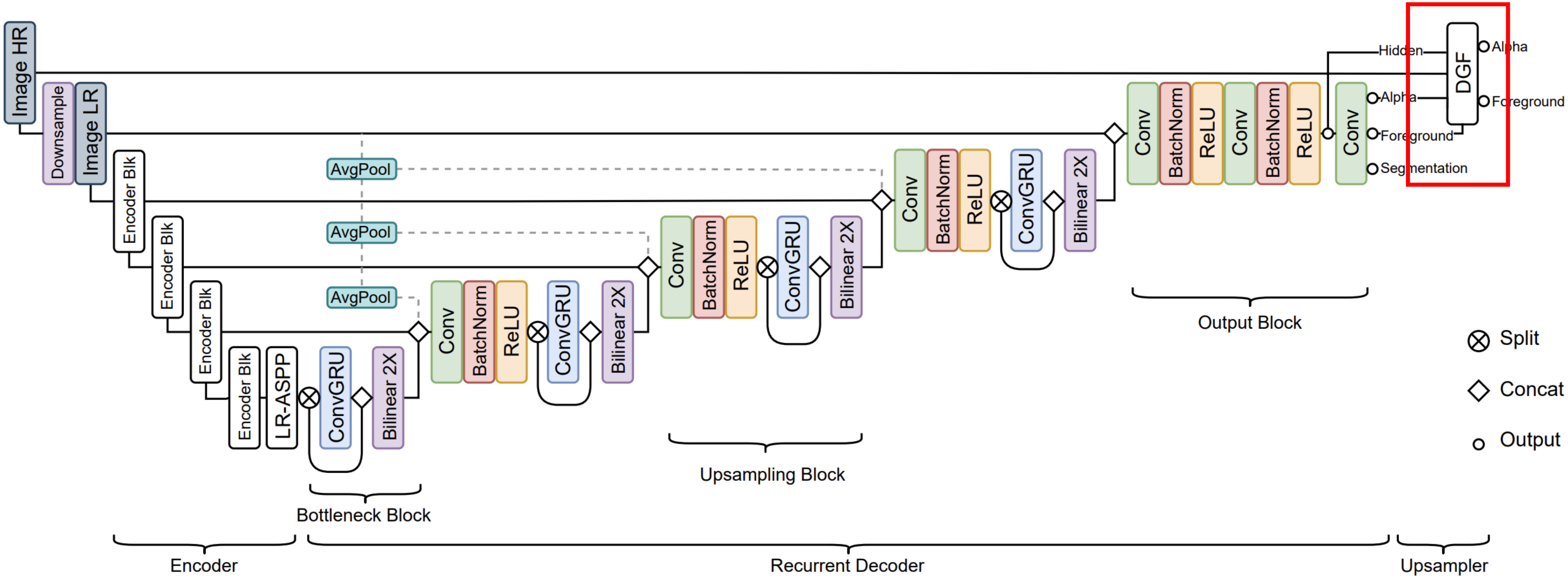


$$z_t = \sigma(w_{zx} * x_t + w_{zh} * h_{t-1} + b_z)$$
$$r_t = \sigma(w_{rx} * x_t + w_{rh} * h_{t-1} + b_r)$$
$$o_t = tanh(w_{ox} * x_t + w_{oh} * (r_t \odot h_{t-1}) + b_o)$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot o_t$$

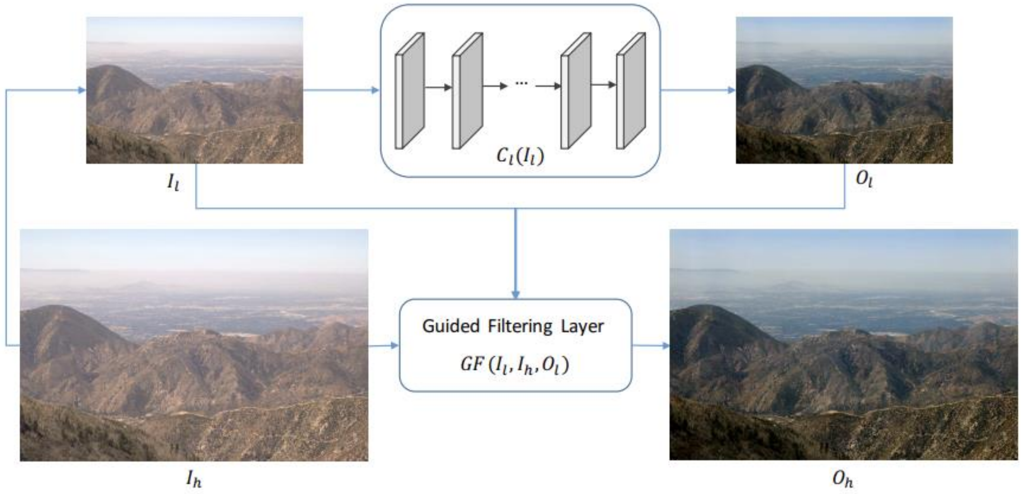
- GRU
  - RNN
  - LSTM

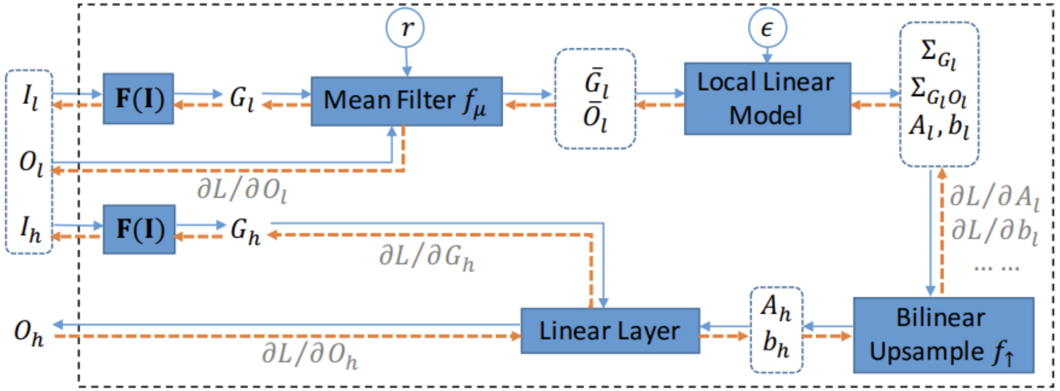
- ConvGRU
  - Replace the linear layer with convolution layer

Deep Guided Filter



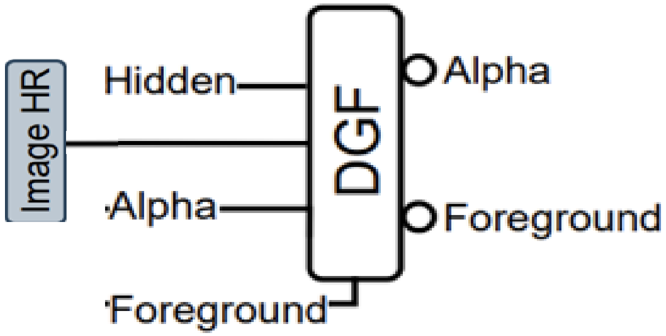
Deep Guided Filter





Guided Filtering Layer

Fast End-to-End Trainable Guided Filter



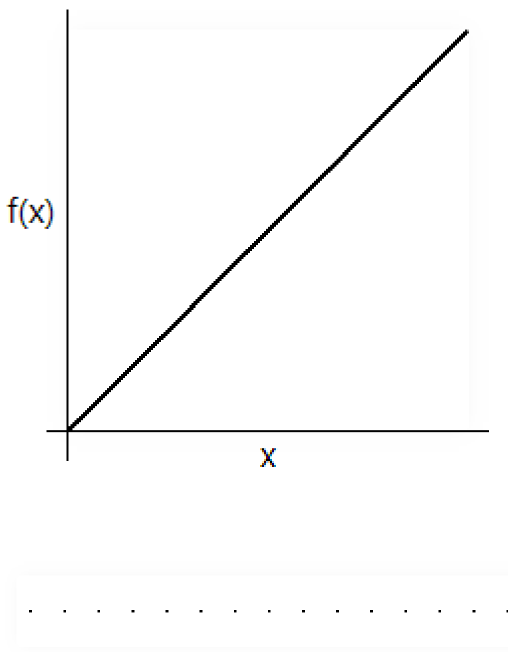

```
mean_x = self.box_filter(base_x)
mean_y = self.box_filter(base_y)
cov_xy = self.box_filter(base_x * base_y) - mean_x * mean_y
var_x  = self.box_filter(base_x * base_x) - mean_x * mean_x

A = self.conv(torch.cat([cov_xy, var_x, base_hid], dim=1))
b = mean_y - A * mean_x

H, W = fine_src.shape[2:]
A = F.interpolate(A, (H, W), mode='bilinear', align_corners=False)
b = F.interpolate(b, (H, W), mode='bilinear', align_corners=False)

out = A * fine_x + b
```

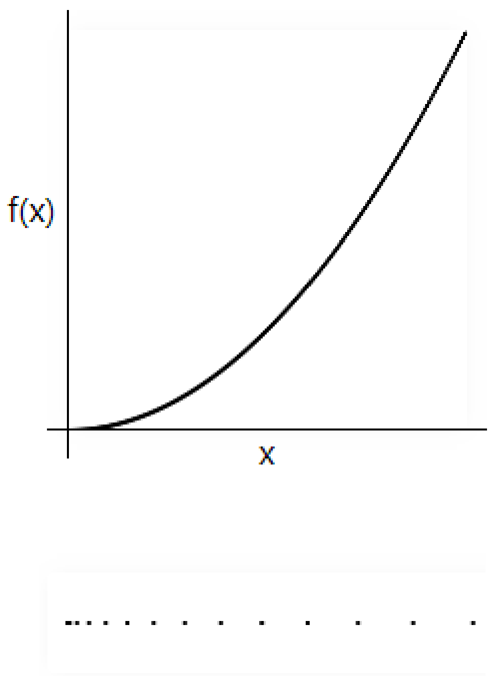
Training Datasets

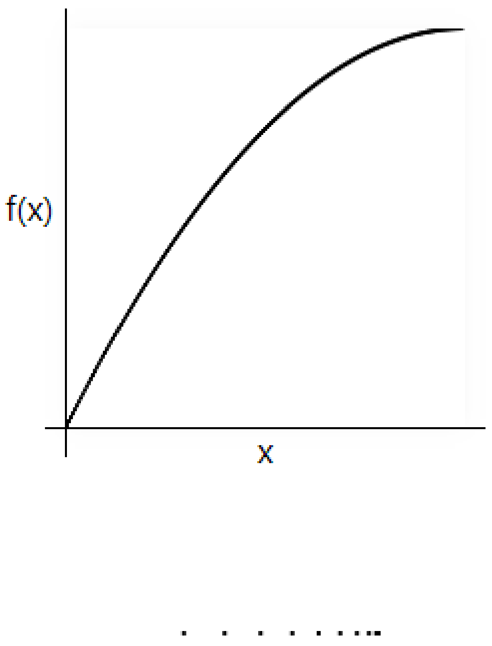
- Matting Datasets
  - VideoMatte240K   484 video， 475/4/5
  - Distinction646 human part
  - Composition1K human part   420/15/21
- Semantic Segmentation Datasets
  - YoutubeVIS 2985 human containing
  - COCO 64,111 human images
  - SPD 5711 human images
- Data Augmentation
  - Matting data augmentation
  - BGM V2 settings
  - Temporal augmentation
  - Motion augmentation

Motion Augmentation

- Easing Function







Training Procedures

- Stage 1
  - Without DGF
  - 15 Epochs
  - T=15（data size is like （B, T, C, H, W））
  - LR backbone is 1e-4, rest is 2e-4
  - Train on VM training set (low resolution (512, 256))
- Stage 2
  - 2 more Epochs
  - T=50（data size is like （B, T, C, H, W））
  - LR backbone is 5e-5, rest is 1e-4
  - Others follows Stage 1
- Stage 3
  - With DGF
  - 1 epoch
  - T=40 for low resolution and T=6 for high resolution(2048, 1024)
  - LR DGF is 2e-4, rest is 1e-5
- Stage 4
  - With DGF
  - 5 epochs
  - Image datasets

Loss Function

- Alpha matte loss
  - L1 loss
  - Pyramid Laplacian Loss
  - Temporal coherence loss
- Foreground loss
  - L1 loss
  - Temporal coherence loss
- Semantic loss
  - BCE loss

$$\mathcal{L}_{l1}^{\alpha} = ||\alpha_t - \alpha_t^*||_1$$

$$\mathcal{L}_{lap}^{\alpha} = \sum_{s=1}^{5} \frac{2^{s-1}}{5} ||L_{pyr}^{s}(\alpha_t) - L_{pyr}^{s}(\alpha_t^*)||_1$$

$$\mathcal{L}_{tc}^{\alpha} = ||\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} - \frac{\mathrm{d}\alpha_t^*}{\mathrm{d}t}||_2$$

$$\mathcal{L}_{l1}^{F} = ||(a_t^* > 0) * (F_t - F_t^*)||_1$$

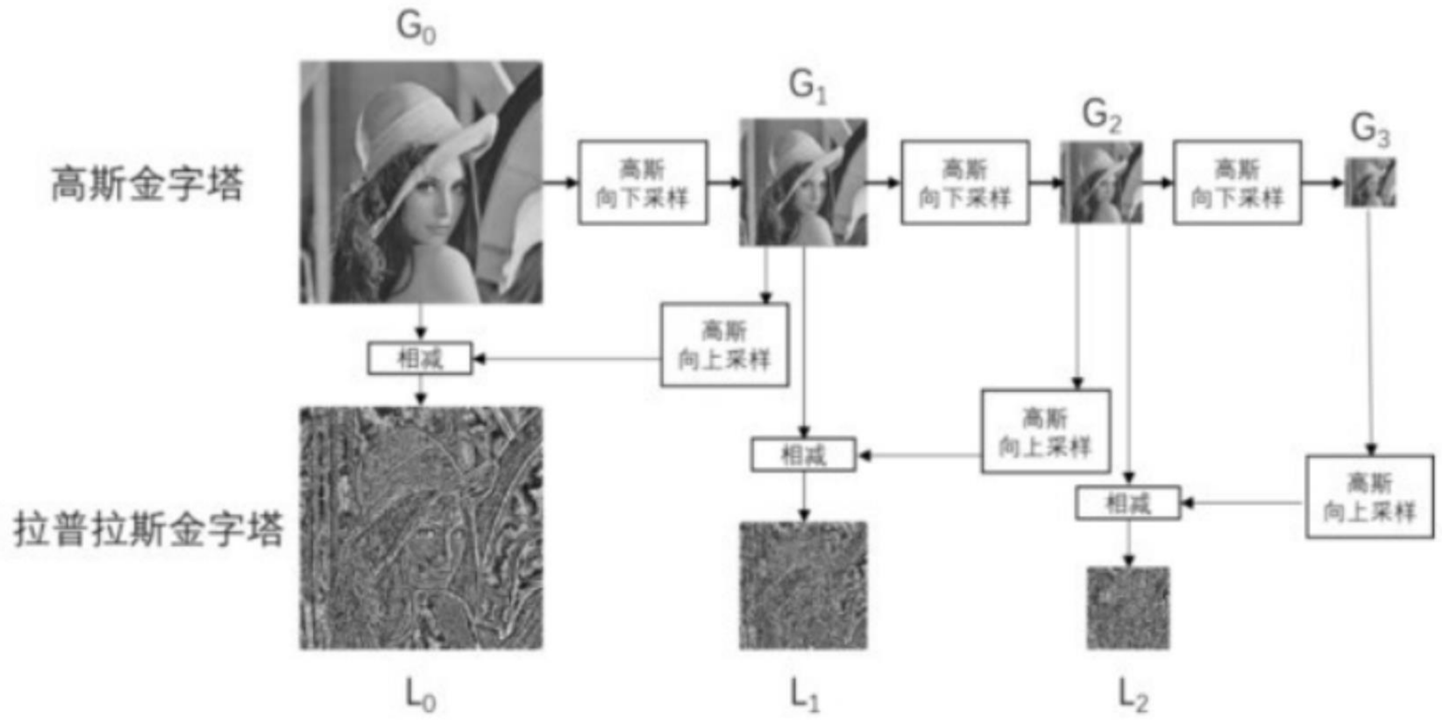
$$\mathcal{L}_{tc}^{F} = ||(a_t^* > 0) * (\frac{\mathrm{d}F_t}{\mathrm{d}t} - \frac{\mathrm{d}F_t^*}{\mathrm{d}t})||_2$$

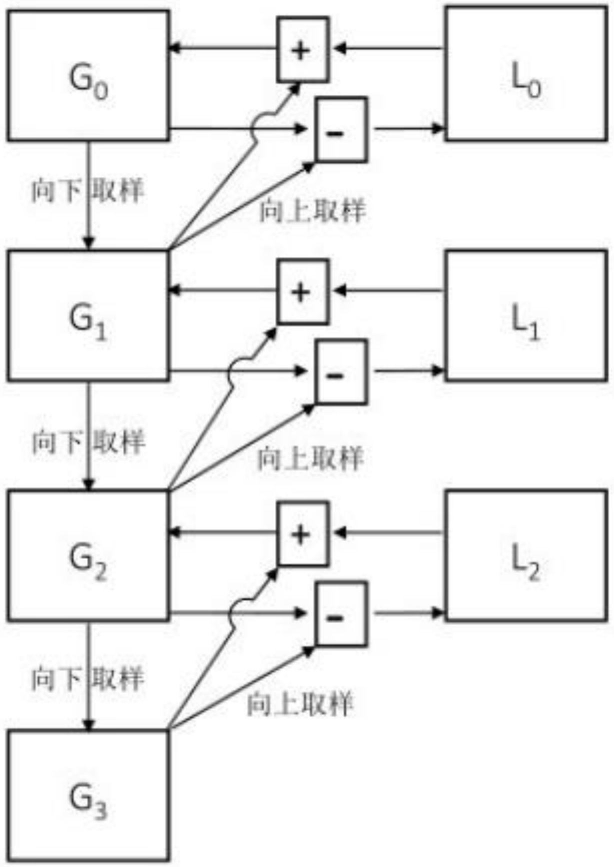
$$\mathcal{L}^{M} = \mathcal{L}_{l1}^{\alpha} + \mathcal{L}_{lap}^{\alpha} + 5\mathcal{L}_{tc}^{\alpha} + \mathcal{L}_{l1}^{F} + 5\mathcal{L}_{tc}^{F}$$

$$\mathcal{L}^{S} = S_t^*(-\log(S_t)) + (1 - S_t^*)(-\log(1 - S_t))$$

Pyramid Laplacian Loss

$$\mathcal{L}_{lap}^{\alpha} = \sum_{s=1}^{5} \frac{2^{s-1}}{5} ||L_{pyr}^{s}(\alpha_t) - L_{pyr}^{s}(\alpha_t^*)||_1$$





Evaluation

Dataset	Method	SAD	MSE	Grad	dtSSD
VM 1920 × 1080	MODNet + FGF	11.13	5,54	15.30	3.08
	Ours	**6.57**	**1.93**	**10.55**	**1.90**
D646 2048 × 2048	MODNet + FGF	11.27	6.13	30.78	2.19
	Ours	**8.67**	**4.28**	**30.06**	**1.64**
AIM 2048 × 2048	MODNet + FGF	17.29	10.10	35.52	2.60
	Ours	**14.89**	**9.01**	**34.97**	**1.71**

Table 2: High-resolution alpha comparison. Ours is better than MODNet with Fast Guided Filter (FGF).

Dataset	Method	Alpha					FG
		MAD	MSE	Grad	Conn	dtSSD	MSE
VM 512 × 288	DeepLabV3	14.47	9.67	8.55	1.69	5.18	
	FBA	8.36	3.37	2.09	0.75	2.09	
	BGMv2	25.19	19.63	2.28	3.26	2.74	
	MODNet	9.41	4.30	1.89	0.81	2.23	
	Ours	**6.08**	**1.47**	**0.88**	**0.41**	**1.36**	
D646 512 × 512	DeepLabV3	24.50	20.1	20.30	6.41	4.51	
	FBA	17.98	13.40	7.74	4.65	2.36	5.84
	BGMv2	43.62	38.84	5.41	11.32	3.08	**2.60**
	MODNet	10.62	5.71	3.35	2.45	1.57	6.31
	Ours	**7.28**	**3.01**	**2.81**	**1.83**	**1.01**	2.93
AIM 512 × 512	DeepLabV3	29.64	23.78	20.17	7.71	4.32	
	FBA	23.45	17.66	9.05	6.05	2.29	6.32
	BGMv2	44.61	39.08	5.54	11.60	2.69	**3.31**
	MODNet	21.66	14.27	5.37	5.23	1.76	9.51
	Ours	**14.84**	**8.93**	**4.35**	**3.83**	**1.01**	5.01

Table 1: Low-resolution comparison. Our alpha prediction is better than all others. Our foreground prediction is behind BGMv2 but outperforms FBA and MODNet. Note that FBA uses synthetic trimap from DeepLabV3; BGMv2 only sees ground-truth background from the first frame; MODNet does not predict foreground so it is evaluated on the input image.

Speed & Size

Method	Parameters (Million)	Size (MB)
DeepLabV3	60.996	233.3
FBA	34.693	138.8
BGMv2	5.007	19.4
MODNet	6.487	25.0
Ours	**3.749**	**14.5**

Table 3: Ours is lighter than all compared methods. Size is measured on FP32 weights.

Resolution	s	Method	FPS	GMACs*
512×288	1	DeepLabV3 + FBA	12.3	205.77
		BGMv2	**152.5**	8.46
		MODNet	104.9	8.80
		Ours	131.9	**4.57**
1920×1080	0.25	BGMv2	70.6	9.86
		MODNet + FGF	100.3	7.78
		Ours	**104.2**	**4.15**
3840×2160	0.125	BGMv2	26.5	17.04
		MODNet + FGF	**88.6**	7.78
		Ours	76.5	**4.15**

Table 4: Model performance comparison. s denotes the downsample scale. Models are converted to TorchScript and optimized before testing (BatchNorm fusion *etc.*). FPS is measured as FP32 tensor throughput on an Nvidia GTX 1080Ti GPU. GMACs is a rough approximation.

Ablation Study

- Temporal Information
- Role of Segmentation Training Objective
- Role of Deep Guided Filter
- Role of dynamic background

Ablation Study

- Temporal Information

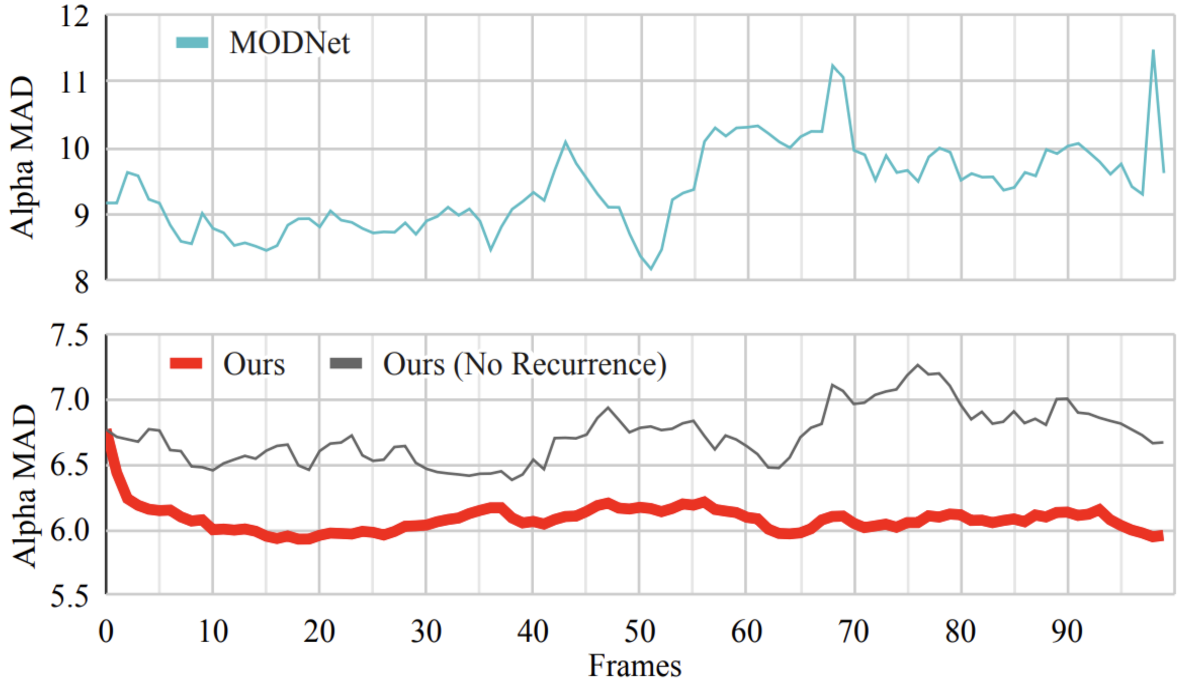


Figure 4: Average alpha MAD over time on VM without DGF. Our metric improves over time and is stable, showing that temporal information improves quality and consistency.

Ablation Study

- Role of Segmentation Training Objective

Method	mIOU
DeepLabV3	68.93
MobileNetV3 + LR-ASPP	58.58
Ours (alpha output, no seg objective)	38.24
Ours (alpha output)	**60.88**
Ours (segmentation output)	**61.50**

Table 5: Segmentation performance on COCO validation set. Training with segmentation objective makes our method robust while training only with pre-trained weights regresses.

Ablation Study

- Role of Deep Guided Filter

Method	Params	FPS	MAD	MSE	Grad	dtSSD
Ours (FGF)	**3.748**	**109.4**	8.70	**4.13**	31.44	1.89
Ours	3.749	104.2	**8.67**	4.28	**30.06**	**1.64**

Table 6: Comparing switching DGF to FGF on D646. Parameters are measured in millions. FPS is measured in HD.

Ablation Study

- Static vs. Dynamic Backgrounds

Background	Method	MAD	MSE	Grad	dtSSD
Static	BGMv2*	**4.33**	**0.32**	**4.19**	**1.33**
	MODNet + FGF	11.04	5.42	15.80	3.10
	Ours	5.64	1.07	9.80	1.84
Dynamic	BGMv2	42.45	37.05	17.30	4.61
	MODNet + FGF	11.23	5.65	14.79	3.06
	Ours	**7.50**	**2.80**	**11.30**	**1.96**

Table 7: Comparing VM samples on static and dynamic backgrounds. Ours does better on static backgrounds but can handle both cases. Note that BGMv2 receives ground-truth static backgrounds, but in reality the backgrounds have misalignment.