MARVIS: Motion & Geometry Aware Real and Virtual Image Segmentation

DEPARTMENT OF COMPUTER SCIENCE

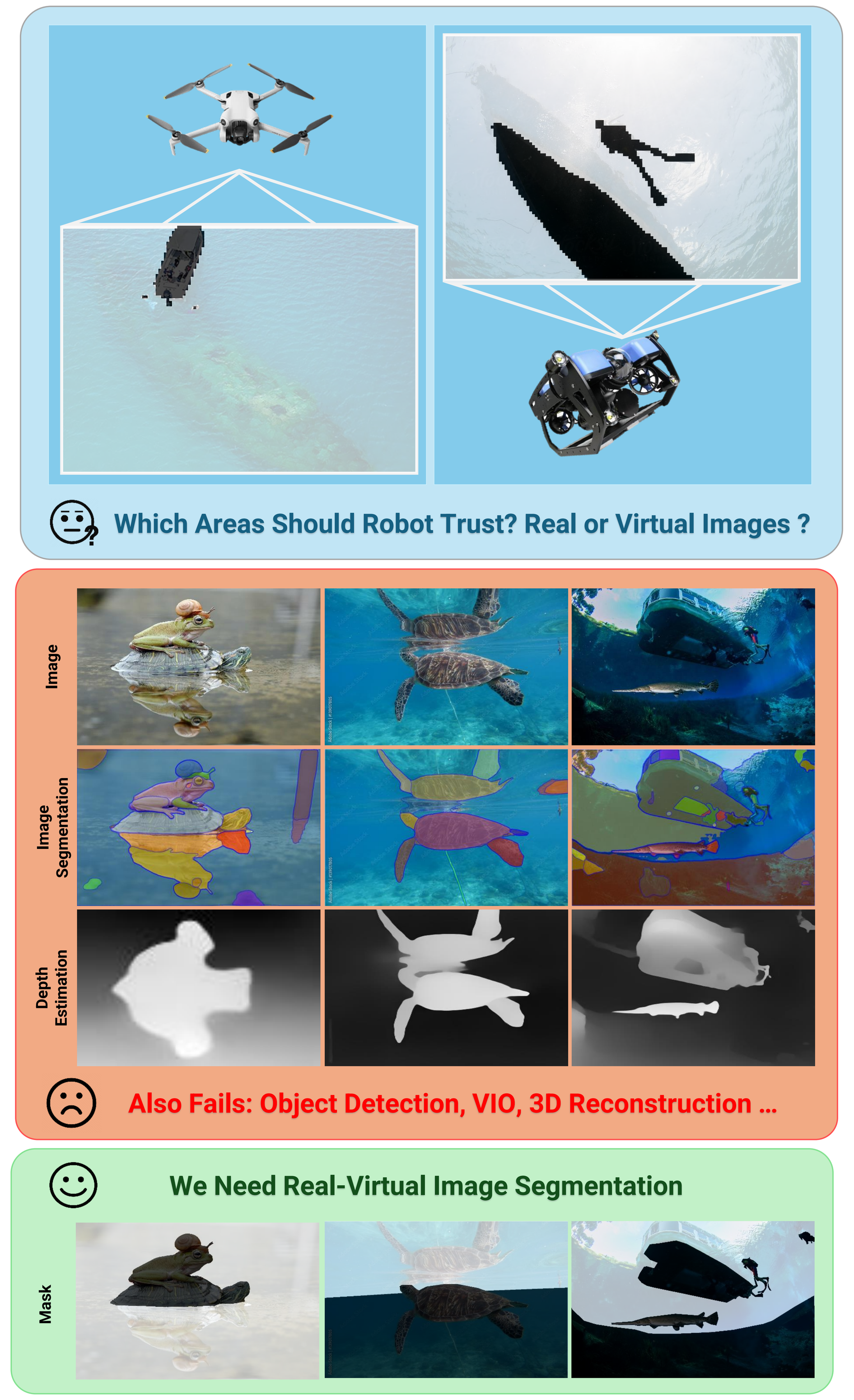
Jiayi Wu[1]*, Xiaomin Lin[1]*, Shahriar Negahdaripour[1,2], Cornelia Fermuller[1], Yiannis Aloimonos[1]

[1]*University of Maryland*, [2]*University of Miami*

PRG

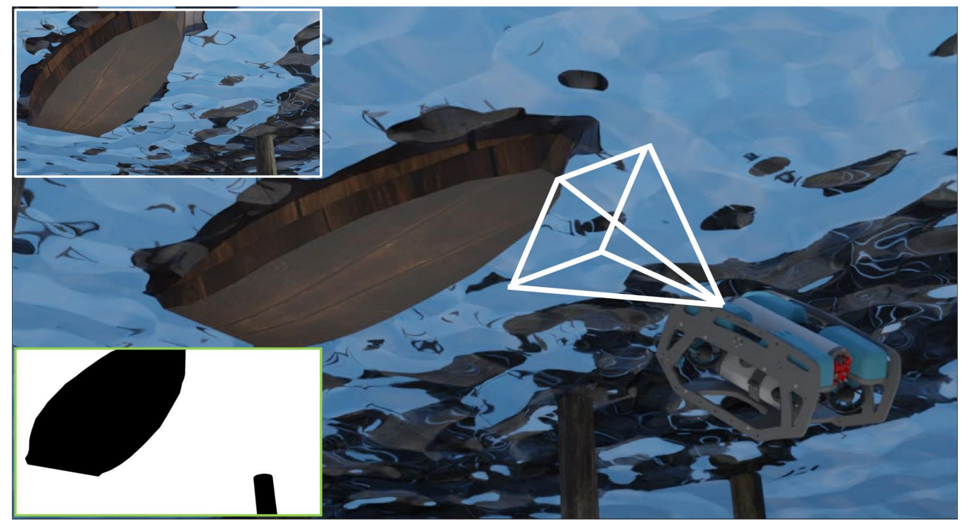
USDA

Motivation



Proposed Simulator: *AquaSim*

Simulator	Quality	Rendering	Surface	GT-Mask
UUV [8]	Low	Gazebo	×	×
URSim [29]	Moderate	Unity3D	×	×
UWRS [30]	Moderate	Unity3D	×	×
HoloOcean [11]	Moderate	UE4	×	×
DAVE [10]	Low	Gazebo	×	×
MARUS [12]	Moderate	Unity3D	✓	×
UNav-Sim [13]	High	UE5	×	×
***AquaSim* (Ours)**	Highest	Blender	✓	✓



- A novel simulator includes the intricate modeling of water-air interface imaging
- Supports the adjustment of various media attribute parameters, such as color, wave characteristics, reflection properties, etc.
- Efficiently generate tailored datasets alongside corresponding ground-truth image masks

Local Motion Entropy (LME)

- Air-water interfaces with highly variable dynamics and turbulence lead to virtual images with chaotic motions.
- Real objects within the same medium as the camera tend to exhibit relatively smooth motions.

$$H(\mathbf{M},\mathbf{A}) = -\alpha \cdot \sum_{m\in\mathcal{M}} p(m)\log_2 p(m) - \beta \cdot \sum_{a\in\mathcal{A}} p(a)\log_2 p(a),$$

$\mathbf{M}$ – *motion vector's magnitude*

$\mathbf{A}$ – *motion vector's angle*

Epipolar Geometric Consistency (EGC)

- The refracted rays form a virtual image that no longer satisfies the epipolar geometry under different camera viewpoints.
- We formulate the *EGC* as a weak supervision to improve the prediction accuracy.

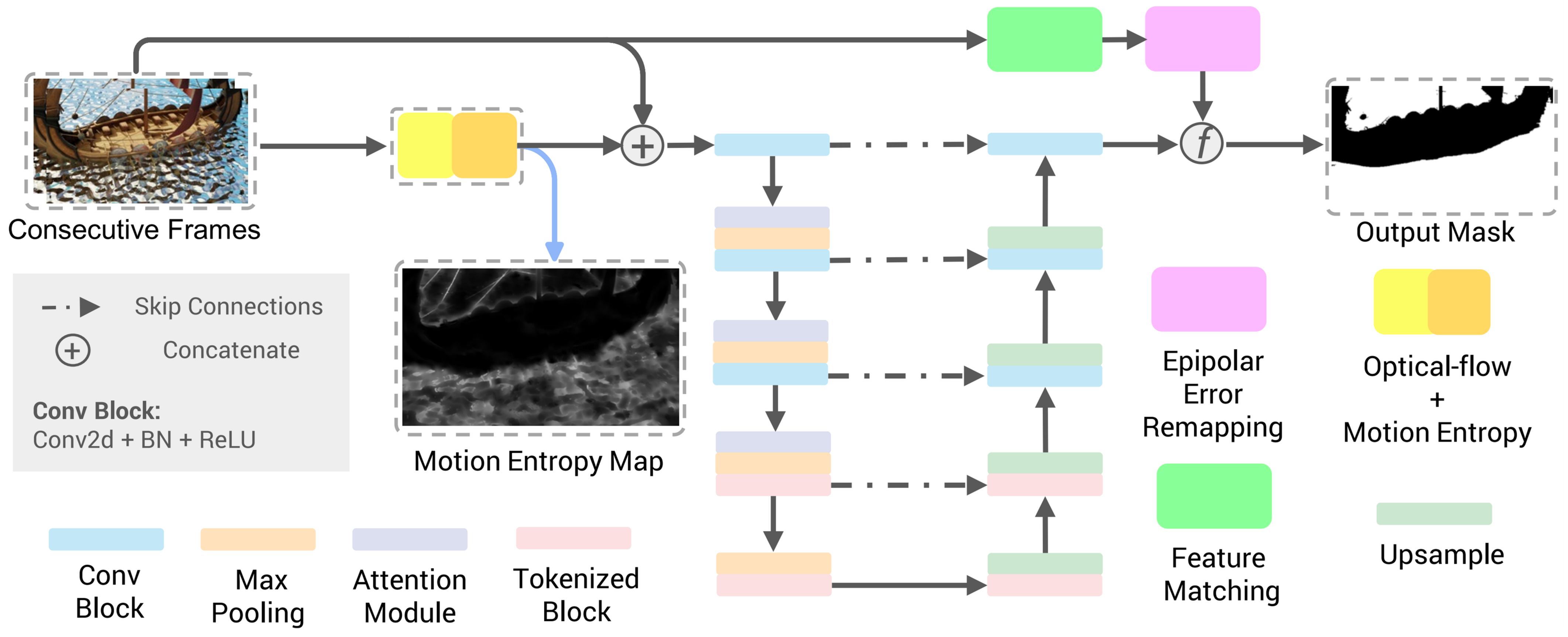
$$\mathcal{L}_{EGC} = \frac{\sum[-(\hat{\mathbf{y}}-1)\cdot \mathbf{E}_{EGC}]}{Count_{(-(\hat{\mathbf{y}}-1)\cdot \mathbf{E}_{EGC}\neq 0)}},$$

$\hat{\mathbf{y}}$ – *predicted binary mask*

$\mathbf{E}_{EGC}$ – *epipolar error map*

Network Architecture: *MARVIS*

- A novel approach for segmentation, exploiting synthetic images combined with domain-invariant information
- A new network layer for extracting *Local Motion Entropy (LME)* features
- A novel *Epipolar Geometric Consistency (EGC)* loss as weak supervision to embed geometric priors in training



Quantitative Comparison

- Trained solely in the synthetic domain without extensive labeling, maintaining performance in unseen real world.
- Significantly improved generalization ability on unseen domains by incorporating domain-invariant priors. (*LME* and *EGC*)
- Leveraging *LME* and *EGC*, enabling a lightweight network to rapidly converge with robust latent feature representation.
- Much lower computational overhead, making it feasible to deploy on computationally constrained drones and AUVs

Model	Params ↓	IoU ↑		F1 ↑	
		Real	Syn	Real	Syn
WASR [17]	$71M$	13.24	29.10	21.78	44.37
WaterNet [23]	$22M$	41.08	53.86	49.63	59.45
PSPNet [42]	$11.32M$	61.35	87.48	74.63	90.51
Deeplabv3 [38]	$5.63M$	68.68	89.02	79.26	93.01
PAN [39]	$4.10M$	61.65	89.69	74.45	93.56
UNet [40]	$31.04M$	51.11	91.91	66.85	94.96
FPN [41]	$13.05M$	66.84	91.91	78.65	94.94
MARVIS(Ours)	**2.56M**	**78.56**	**94.08**	**86.47**	**96.35**

Inference rates:

$43.48\ FPS$ – $NVIDIA^{TM}\ RTX\ 4070\ GPU$

$8.06\ FPS$ – $Intel^{TM}\ Core\ i9 - 4.10GHz\ CPU$

Qualitative Comparison

Experimental Analysis:

- Training on spatial pixel information alone confuses reflections and refractions with real images due to similar RGB colors and textures.
- The domain gap in the RGB space between various domains results in poor generalization to unseen domains.

Advantages of MARVIS:

- Task-tailored feature representation uses motion and geometric cues, enhancing the network's ability to distinguish between virtual and real images in the latent space.
- The domain-invariant priors allow MARVIS to maintain stable and robust segmentation across various environments without retraining.