ADNet: An Asymmetric Dual-Stream Network for RGB-T Salient Object Detection

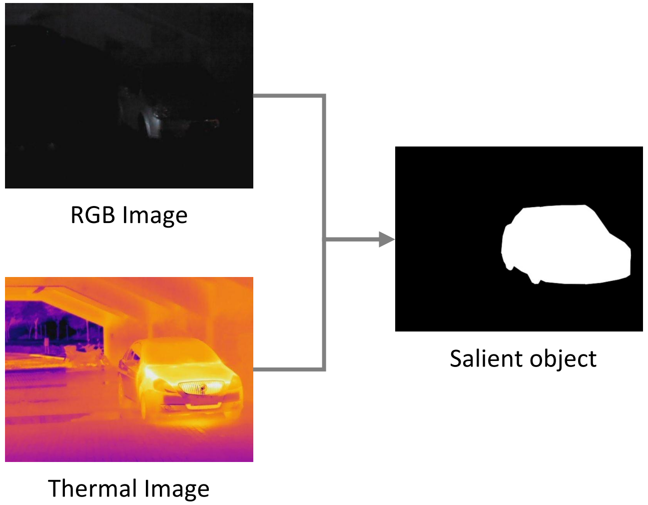
Reporter：Yaqun Fang

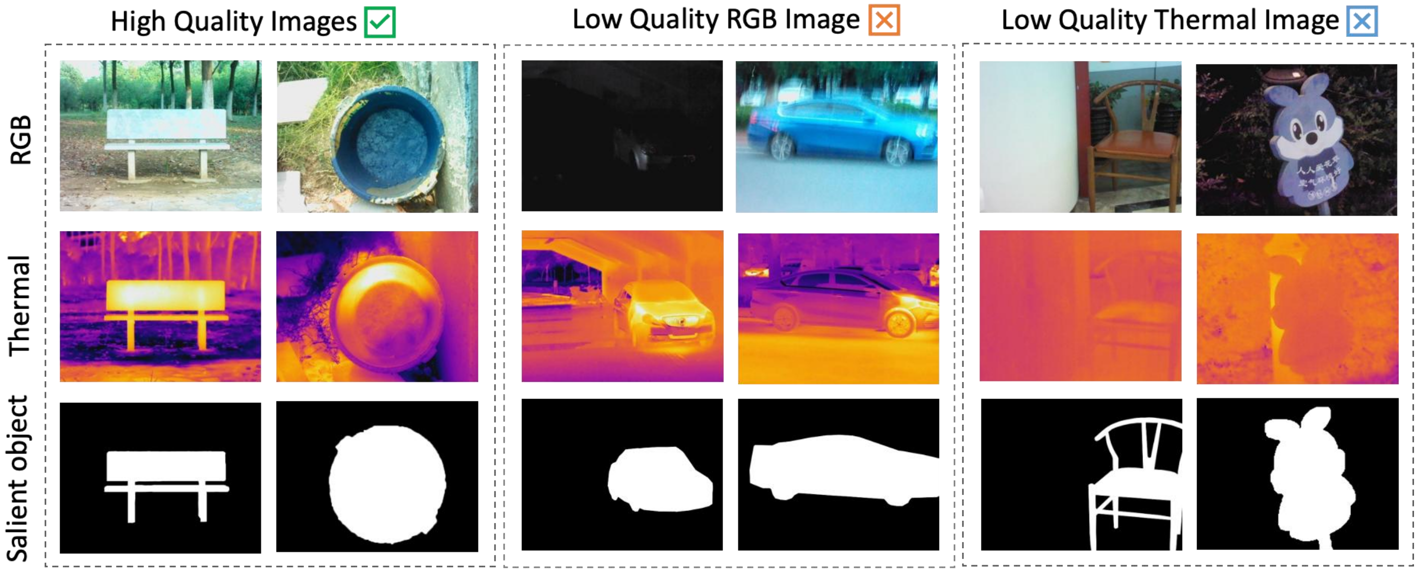
Dec.6, 2023

ACM multimedia Asia 2023

MAGUS
MediA recoGnition and UnderStanding
南京大學
NANJING UNIVERSITY

Introduction





- RGB-Thermal salient object detection (RGB-T SOD) aims to locate salient objects in images that include both RGB and thermal information.
- Traditional approaches often used symmetric dual-stream structures, which did not effectively handle the disparities in information density between RGB and thermal modalities.

Method

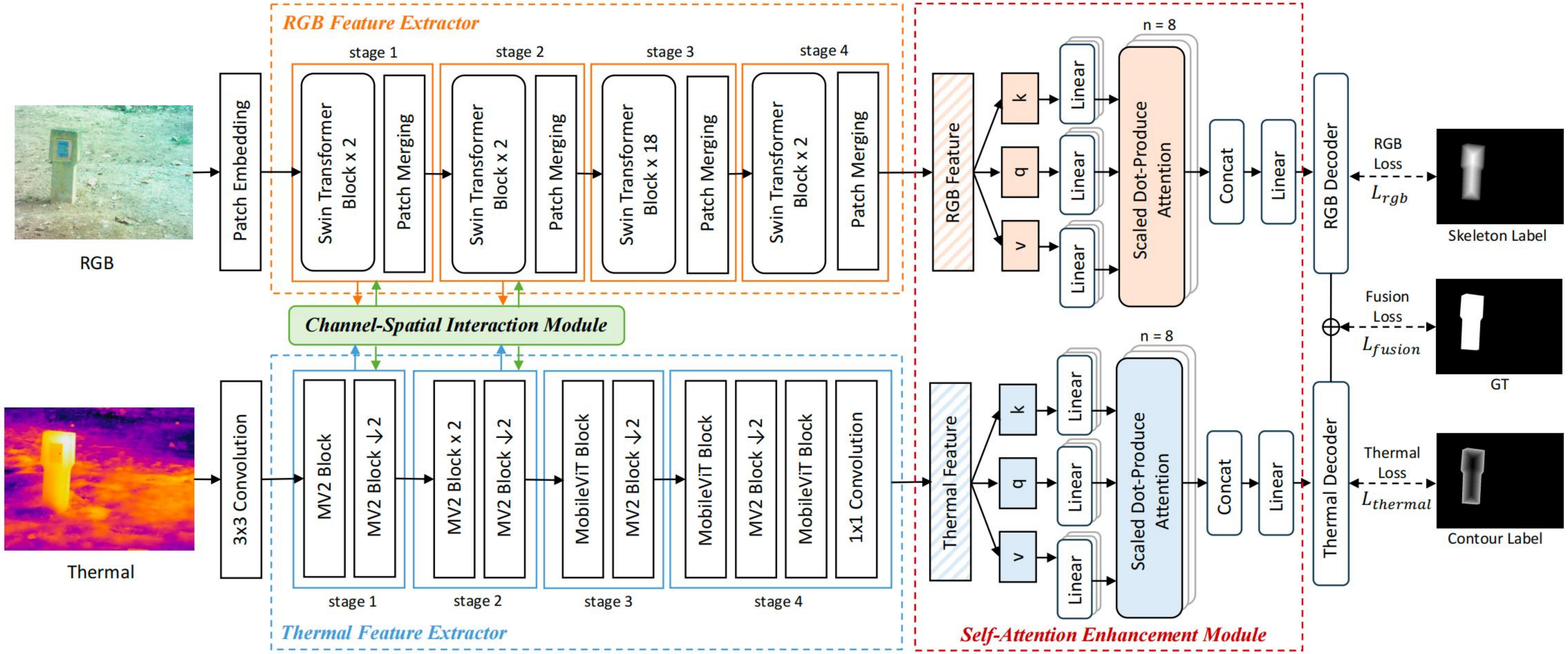
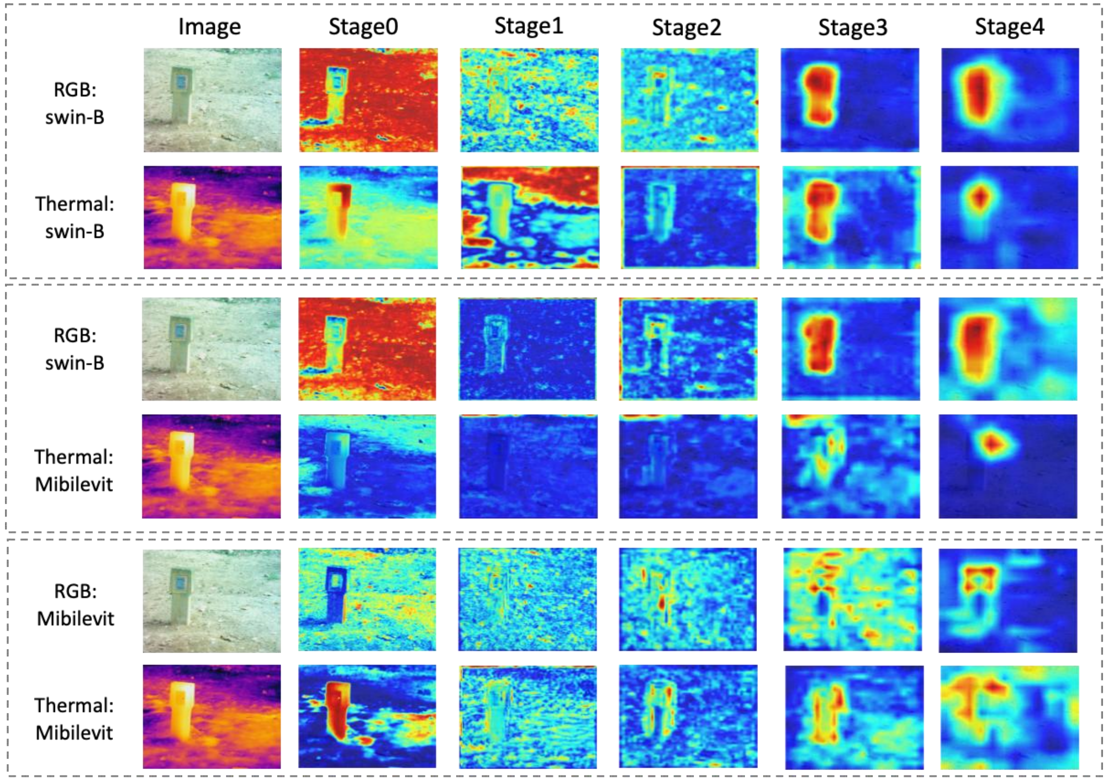


Figure 2: The framework of the proposed ADNet, including RGB feature extractor, Thermal feature extractor, Channel-Spatial Interaction module and Self-Attention Enhancement module.

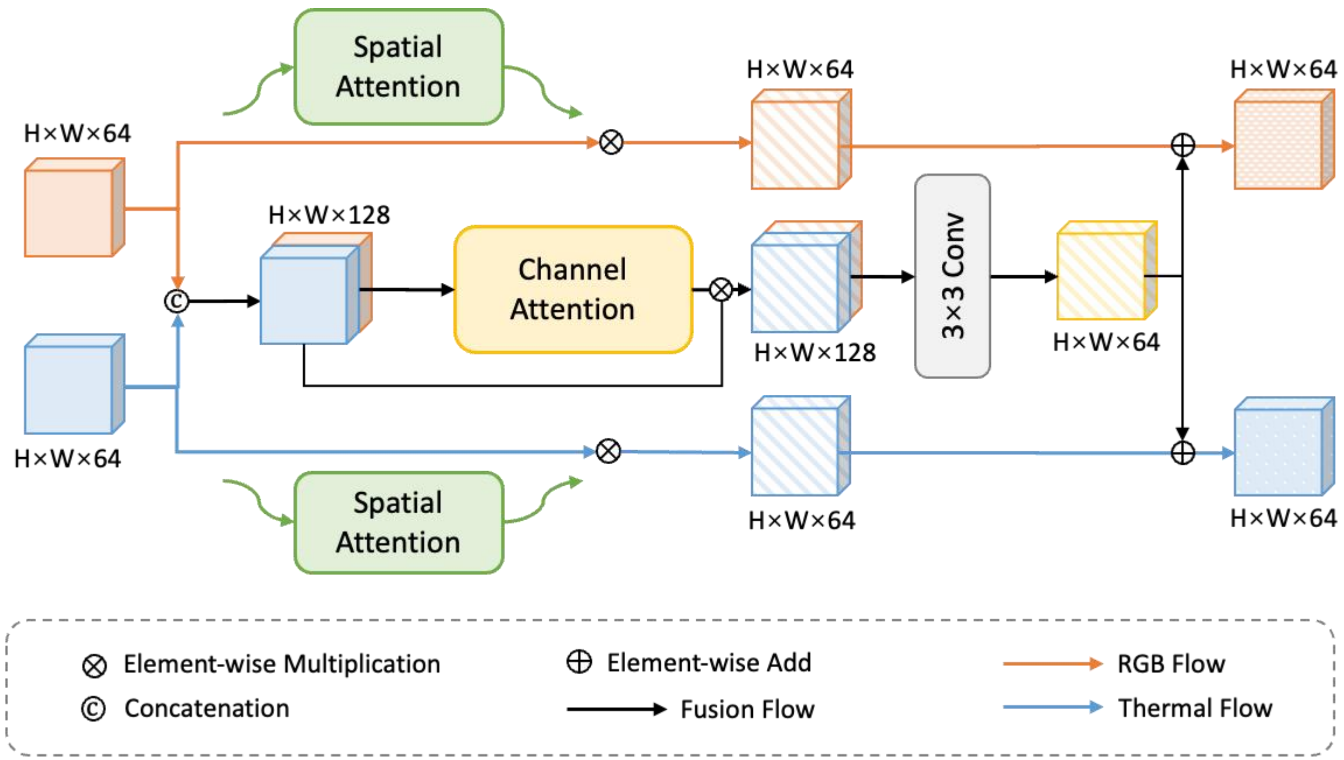
Asymmetric Dual-stream Backbone



Stage	Size of F_{rgb}	Size of $F_{thermal}$	Size of F_{fusion}
stage 0	[128, 96, 96]	[16, 192, 192]	/
stage 1	[128, 96, 96]	[64, 96, 96]	[64, 96, 96]
stage 2	[256, 48, 48]	[96, 48, 48]	[64, 48, 48]
stage 3	[512, 24, 24]	[128, 24, 24]	[64, 24, 24]
stage 4	[1024, 12, 12]	[640, 12, 12]	[64, 12, 12]

The output feature map sizes of different stages for RGB and thermal modalities, where stage 0 represents the feature map before the feature extractor.

Channel-Spatial Interaction Module



$$F_{rgb} = \{F^i_{rgb} | i = 1, 2, 3, 4\} \text{ and } F_t = \{F^i_t | i = 1, 2, 3, 4\}.$$

$$F^i_{fusion} = Concat(F^i_{rgb}, F^i_t).$$

$$CA^i = Sigmoid(MLP(P_{avg}(F^i_{fusion})) + MLP(P_{max}(F^i_{fusion}))),$$

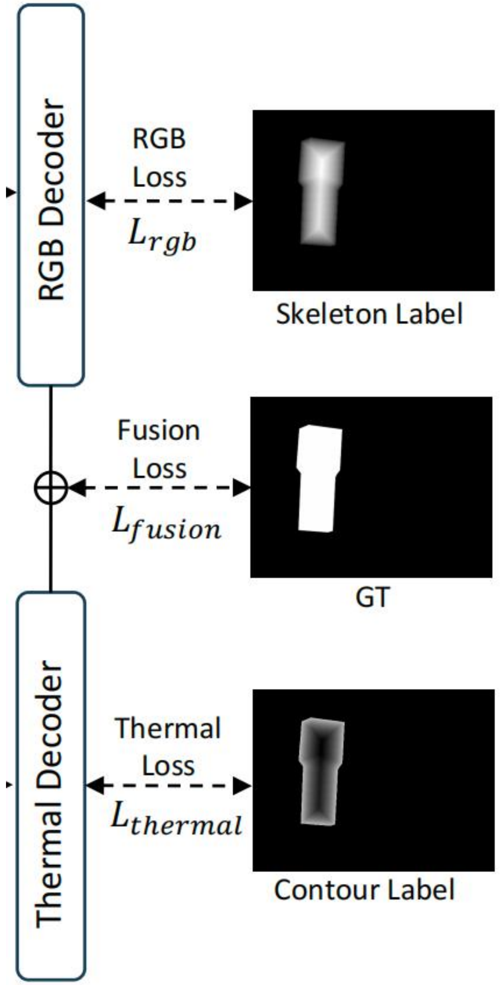
$$SA^i_{rgb} = Sigmoid(Conv^{7\times7}(Concat(P_{avg}(F^i_{rgb}), P_{max}(F^i_{rgb})))), \tag{3}$$

$$SA^i_t = Sigmoid(Conv^{7\times7}(Concat(P_{avg}(F^i_t), P_{max}(F^i_t)))), \tag{4}$$

$$Att^i_{rgb} = SA^i_t + Conv^{3\times3}(F^i_{fusion} \times CA^i),$$

$$Att^i_t = SA^i_{rgb} + Conv^{3\times3}(F^i_{fusion} \times CA^i),$$

Decoder and Loss Function



$$\mathcal{L}_{rgb} = \mathcal{L}_{BCE} + \mathcal{L}_{SSIM},$$
$$\mathcal{L}_{thermal} = \mathcal{L}_{BCE} + \mathcal{L}_{SSIM},$$
$$\mathcal{L}_{fusion} = \mathcal{L}_{BCE} + \mathcal{L}_{SSIM} + \mathcal{L}_{IoU},$$

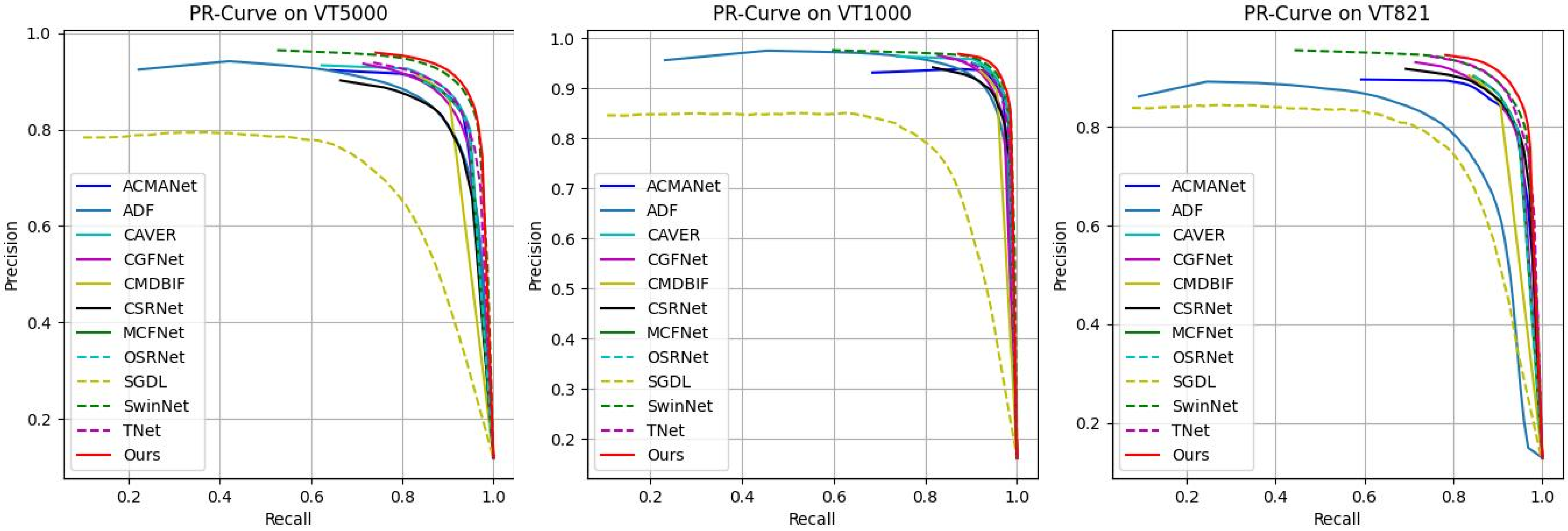
$$\mathcal{L} = \mathcal{L}_{rgb} + \mathcal{L}_{thermal} + \mathcal{L}_{fusion},$$

Experiments and Results

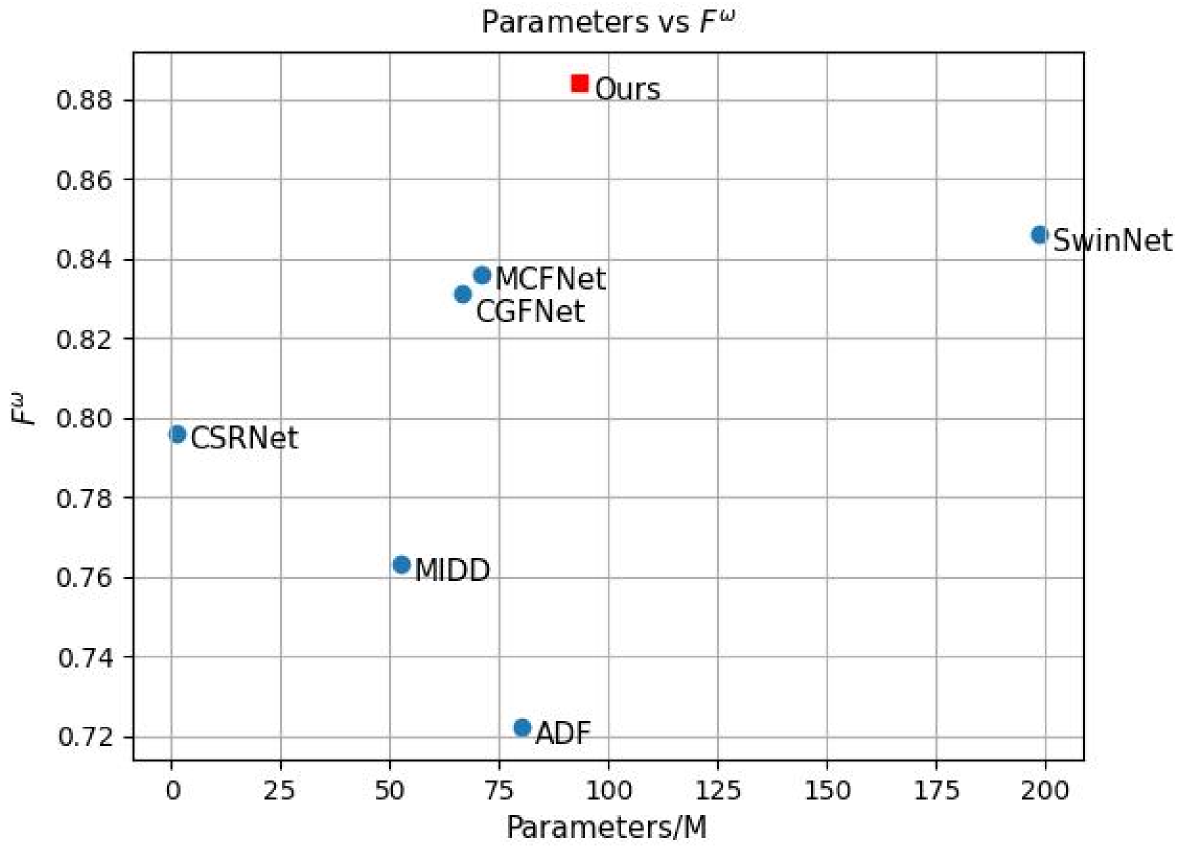
Table 2: Performance compared with SOTA

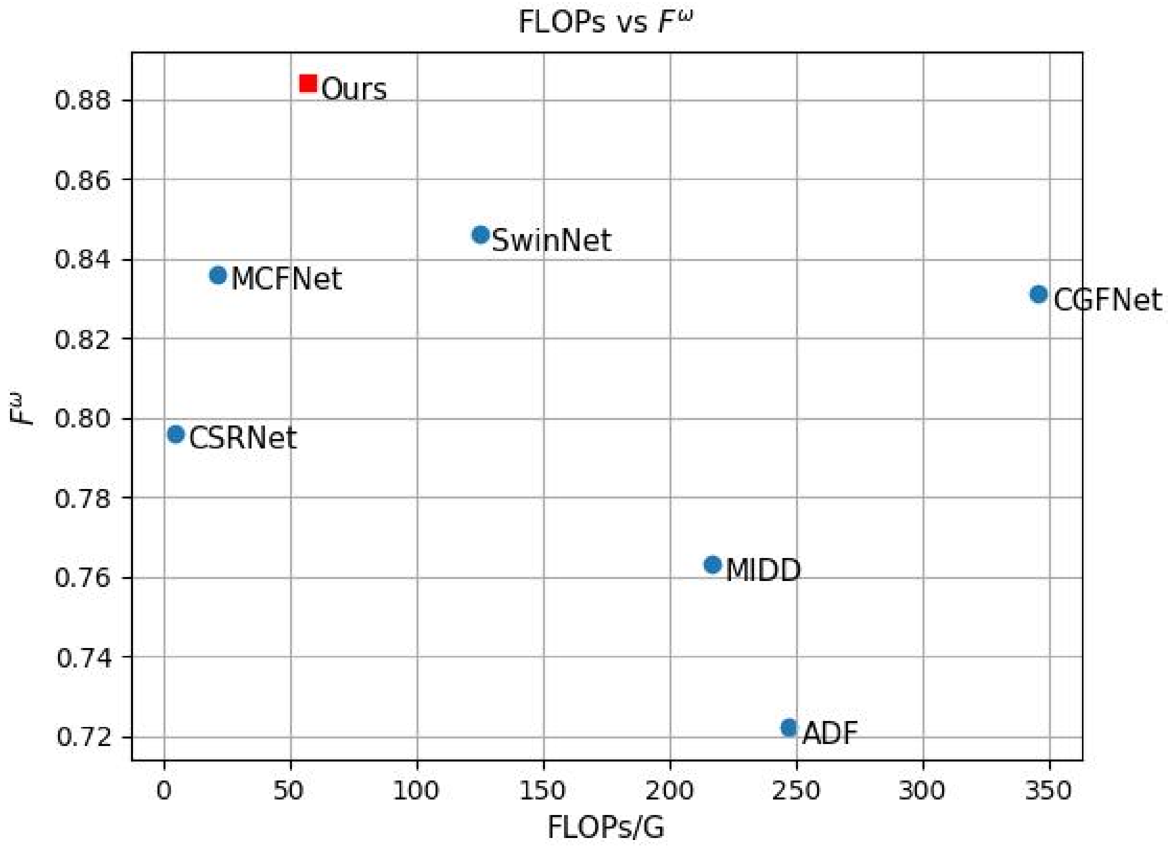
	VT5000						VT1000						VT821					
Method	F_{avg}	F_{max}	F	MAE	E_m	S_m	F_{avg}	F_{max}	F	MAE	E_m	S_m	F_{avg}	F_{max}	F	MAE	E_m	S_m
M3S-NIR	0.575	0.644	0.327	0.168	0.780	0.652	0.717	0.769	0.463	0.145	0.827	0.726	0.734	0.780	0.407	0.140	0.859	0.723
MTMR	0.595	0.662	0.397	0.114	0.795	0.680	0.715	0.755	0.485	0.119	0.836	0.706	0.662	0.747	0.462	0.108	0.815	0.725
SGDL	0.672	0.737	0.558	0.089	0.824	0.750	0.764	0.807	0.652	0.090	0.856	0.787	0.731	0.780	0.583	0.085	0.846	0.764
ADF	0.778	0.863	0.722	0.048	0.891	0.864	0.847	0.923	0.804	0.034	0.921	0.910	0.717	0.804	0.627	0.077	0.843	0.810
MIDD	0.801	0.871	0.763	0.043	0.897	0.867	0.882	0.926	0.856	0.027	0.933	0.915	0.805	0.874	0.760	0.045	0.895	0.871
CSRNet	0.811	0.857	0.796	0.042	0.905	0.868	0.877	0.918	0.878	0.024	0.925	0.918	0.831	0.88	0.821	0.038	0.909	0.885
OSRNet	0.823	0.866	0.807	0.040	0.908	0.875	0.892	0.929	0.891	0.022	0.935	0.926	0.814	0.862	0.801	0.043	0.896	0.875
TNet	0.846	0.895	0.84	0.033	0.927	0.895	0.889	0.937	0.895	0.021	0.937	0.929	0.842	0.904	0.841	0.03	0.919	0.899
mcfnet	0.848	0.886	0.836	0.033	0.924	0.887	0.902	0.939	0.906	0.019	0.944	0.932	0.844	0.889	0.835	0.029	0.918	0.891
CGFNet	0.851	0.887	0.831	0.035	0.922	0.883	0.906	0.936	0.900	0.023	0.944	0.923	0.845	0.885	0.829	0.038	0.912	0.881
CAVER	0.856	0.897	0.849	0.028	0.935	0.899	0.906	0.945	0.912	0.016	0.949	0.938	0.854	0.897	0.846	0.026	0.928	0.897
ACMANet	0.858	0.89	0.823	0.033	0.932	0.887	0.904	0.933	0.889	0.021	0.945	0.927	0.837	0.873	0.807	0.035	0.914	0.883
SwinNet	0.865	0.915	0.846	0.026	0.942	0.912	0.896	0.948	0.894	0.018	0.947	0.938	0.847	0.903	0.818	0.03	0.926	0.904
CMDBIF	0.868	0.892	0.846	0.032	0.933	0.886	0.914	0.931	0.909	0.019	0.952	0.927	0.856	0.887	0.837	0.032	0.923	0.882
Ours	**0.893**	**0.924**	**0.884**	**0.022**	**0.953**	**0.922**	**0.916**	**0.952**	**0.920**	**0.015**	**0.952**	**0.944**	**0.869**	**0.915**	**0.860**	**0.024**	**0.930**	**0.915**

PR Curve


PR-Curve on VT5000
PR-Curve on VT1000
PR-Curve on VT821
Precision
Recall
ACMANet
ADF
CAVER
CGFNet
CMDBIF
CSRNet
MCFNet
OSRNet
SGDL
SwinNet
TNet
Ours

FLOPs and Params





Ablation Studies

Table 3: Ablation results of different backbone, S represents Swin-B, M represents Mobilevit, the results are test on VT5000.

Backbone	F_{avg}	F_{max}	F^{ω}	MAE	E_m	S_m
S + S	0.878	0.915	0.873	0.024	0.947	0.914
M + M	0.766	0.836	0.730	0.049	0.883	0.839
S + M	**0.879**	**0.918**	**0.877**	**0.023**	**0.950**	**0.917**

Table 4: Ablation results of different feature operation.I represents feature interaction module, MH represents multi-head self-attention module.

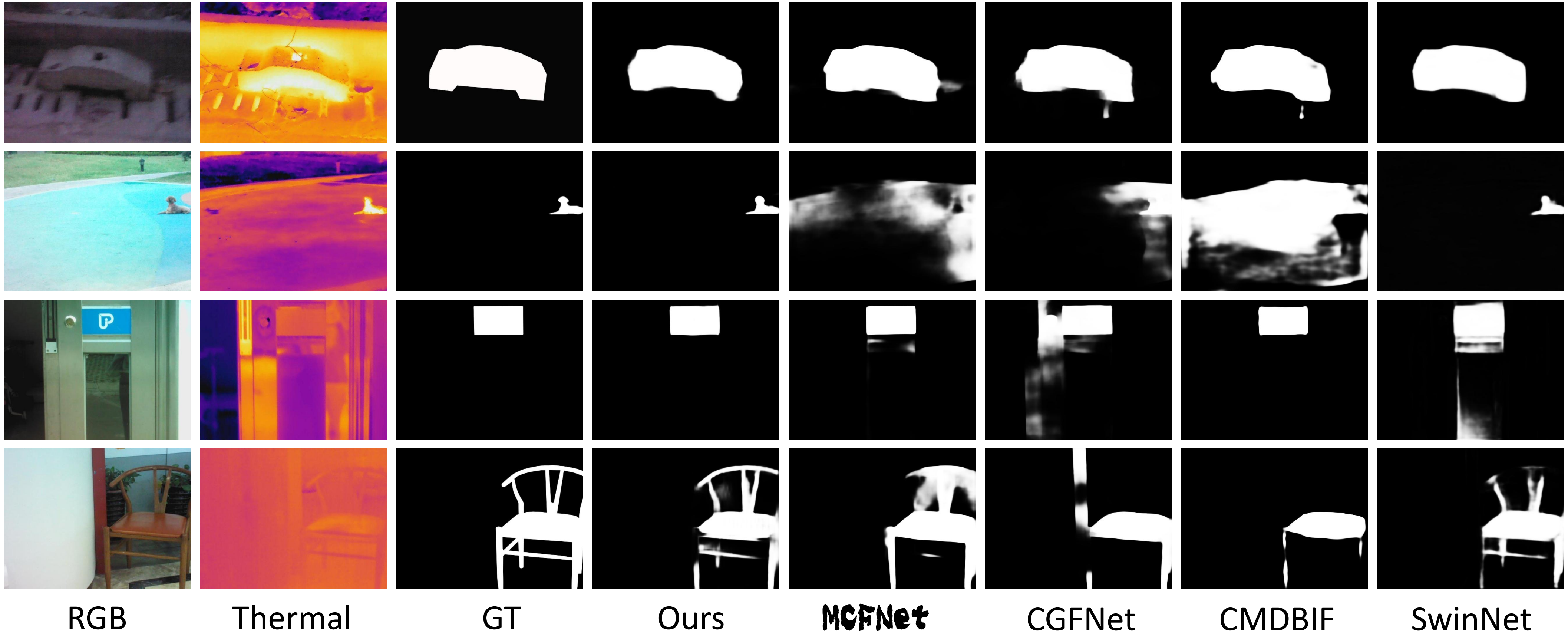
Operation	F_{avg}	F_{max}	F^{ω}	MAE	E_m	S_m
IN	0.887	0.918	0.879	0.023	0.949	0.917
MH	0.889	0.920	0.878	0.023	0.950	0.918
IN + MH	**0.893**	**0.924**	**0.884**	**0.022**	**0.953**	**0.922**

- Asymmetric Backbone
- Feature Fusion
- Model Complexity

Table 5: Ablation of model size and cost of conputation.

Methods	$Paramters(M)$	$FLOPs(G)$
S+S	11.107	13.396
M+M	174.600	94.795
S+M	92.854	54.095
S+M+IN	93.302	56.659
S+M+IN+MH	93.319	56.683

Visual comparison



Conclusion

- We introduce the first asymmetric network for RGB-T salient object detection. Experimental results demonstrate that our method achieves superior performance, reducing the number of parameters by approximately 46% and the computational load by around 40%.

- We introduce a CSI module for low-level features, enabling the model to better leverage the CNN’ s capability to emphasize local features. Additionally, we present an SAE module for enhancing deep features, improving attention on salient regions by enhancing global features in both the RGB branch and the thermal branch.

Thanks for your listening!

MAGUS

MediA recoGnition
and UnderStanding