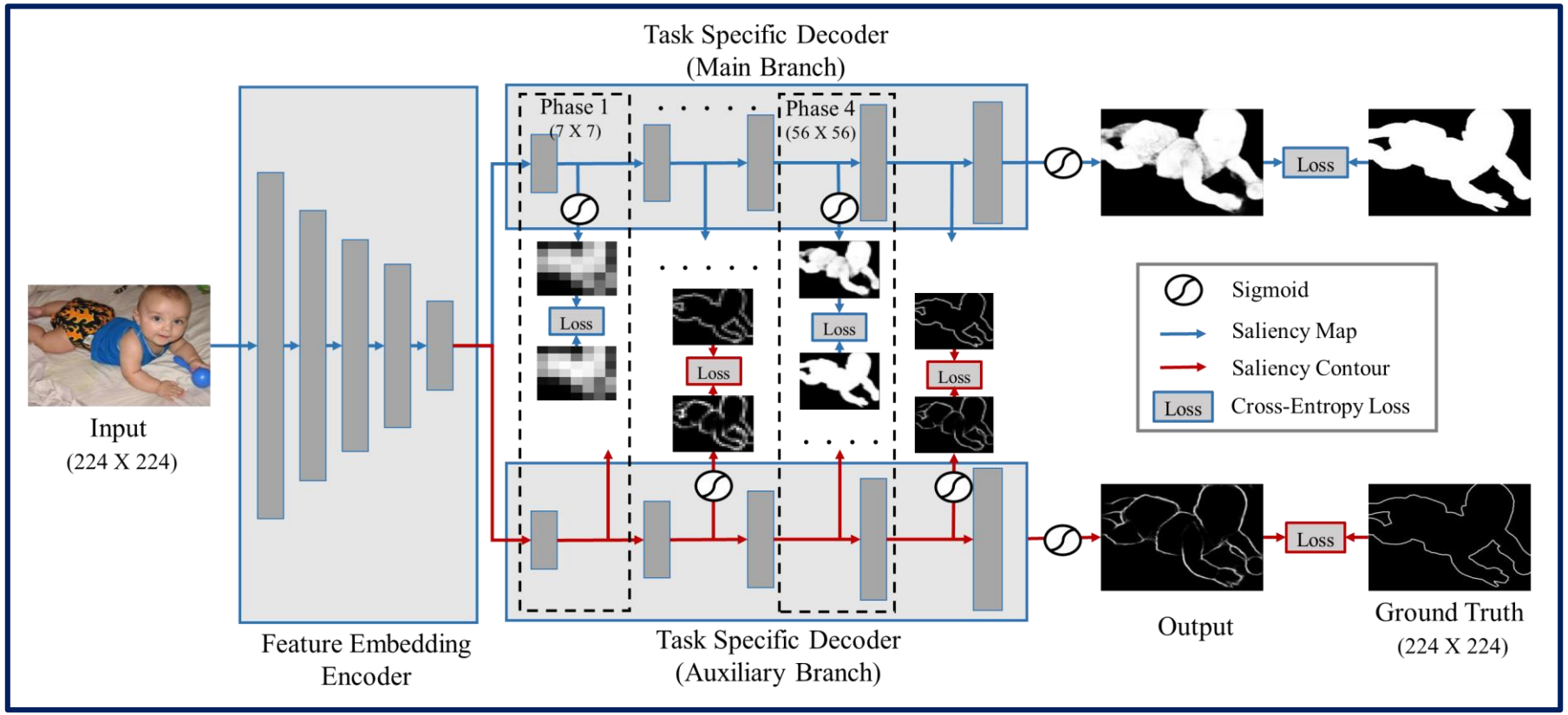
Coarse to Fine:

Progressive and Multi-Task Learning for Salient Object Detection



Donggoo Kang, Sangwoo Park, and Joonki Paik*

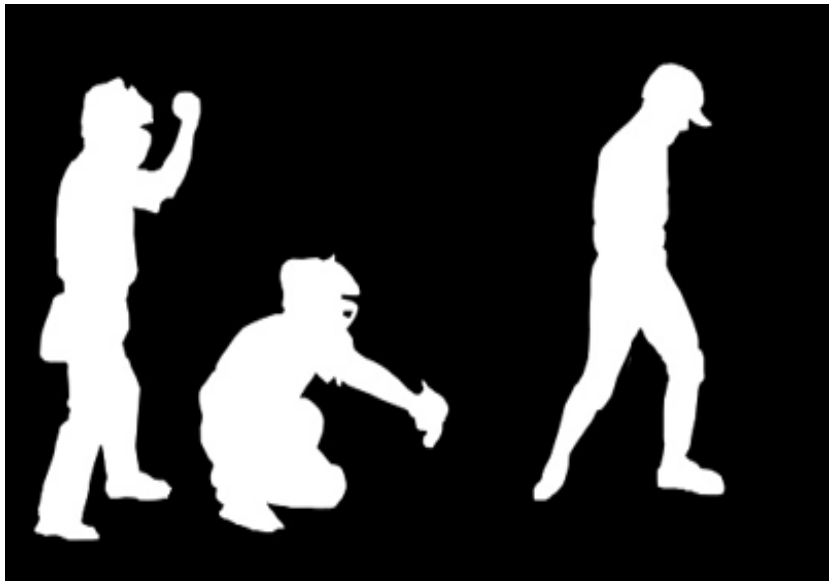
Image Processing and Intelligent System Laboratory

Chung-Ang University, Seoul, Korea

Email: dgkang@ipis.cau.ac.kr, swpark@ipis.cau.ac.kr, paikj@ipis.cau.ac.kr

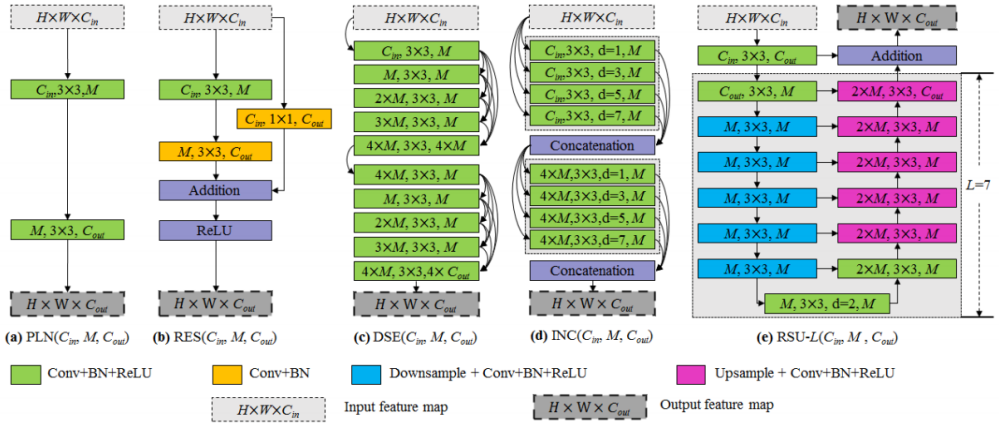
Salient Object Detection

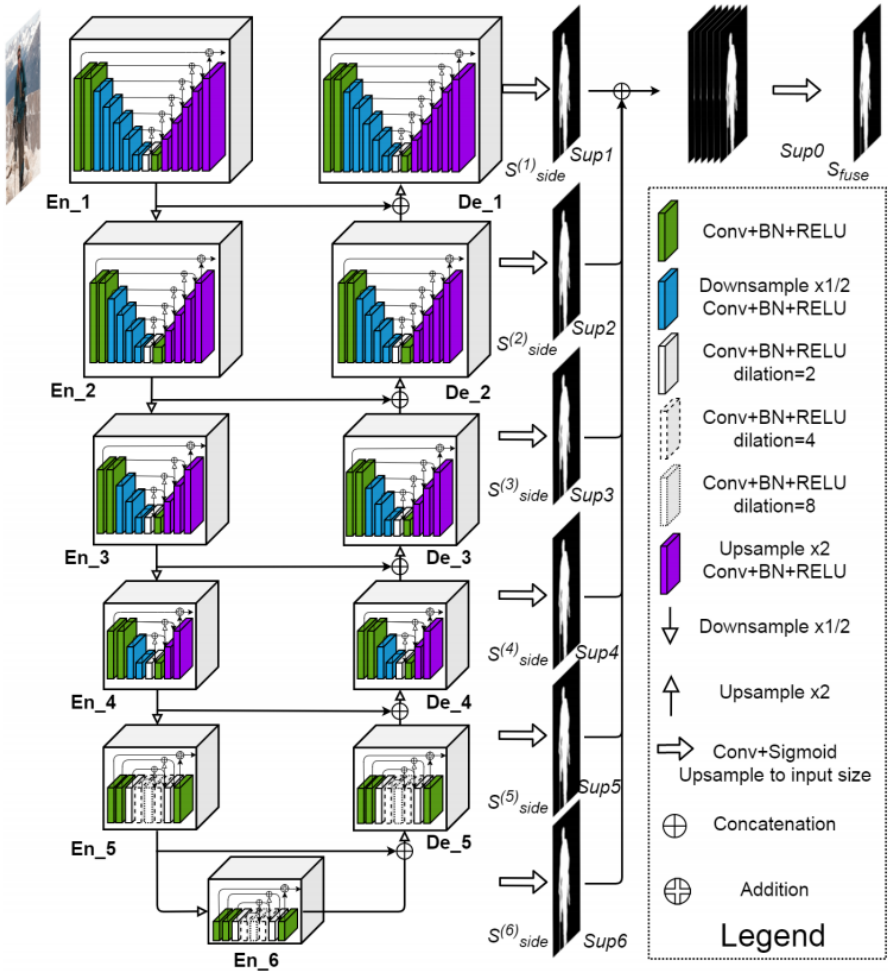
- Salient object detection is a task based on a visual attention mechanism, in which algorithms aim to explore objects or regions more attentive than the surrounding areas on the scene or images.



Related Work – Convolution Block Manipulation

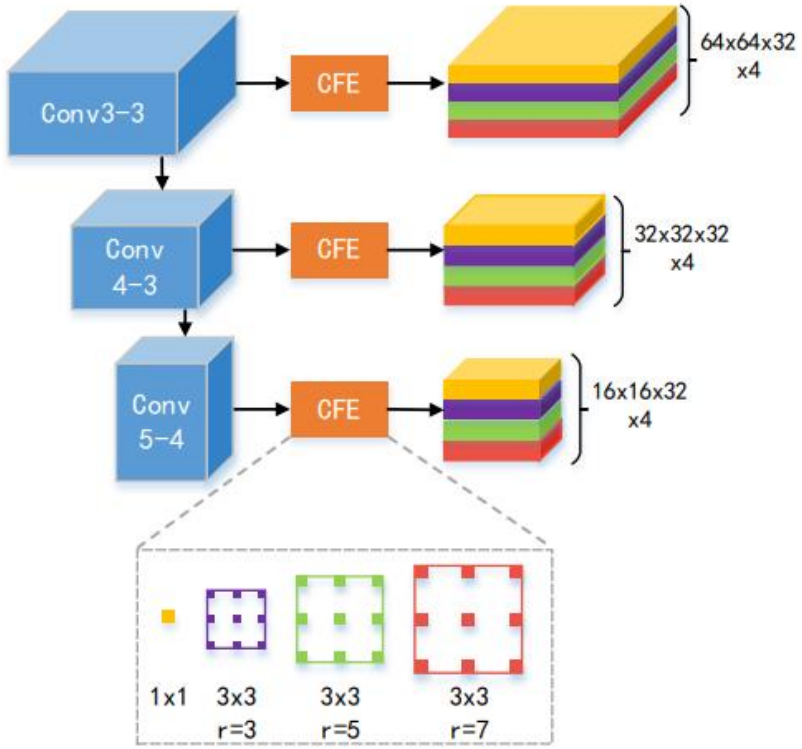
- U2-Net(2020)

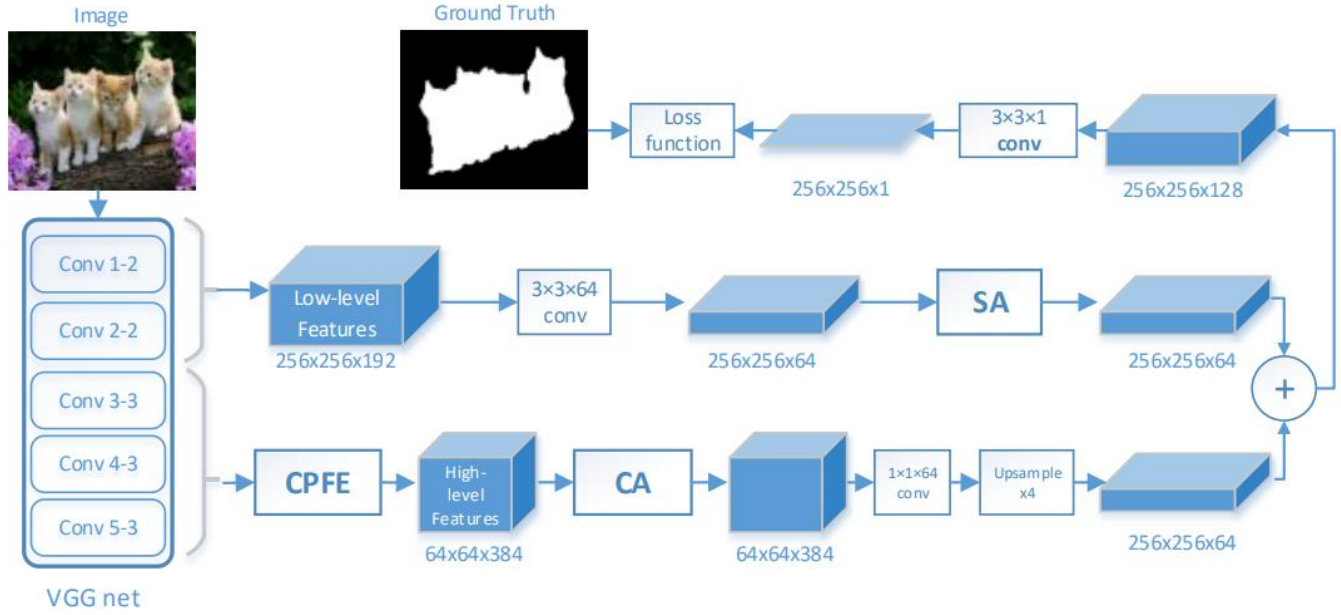




Related Work – Convolution Block Manipulation

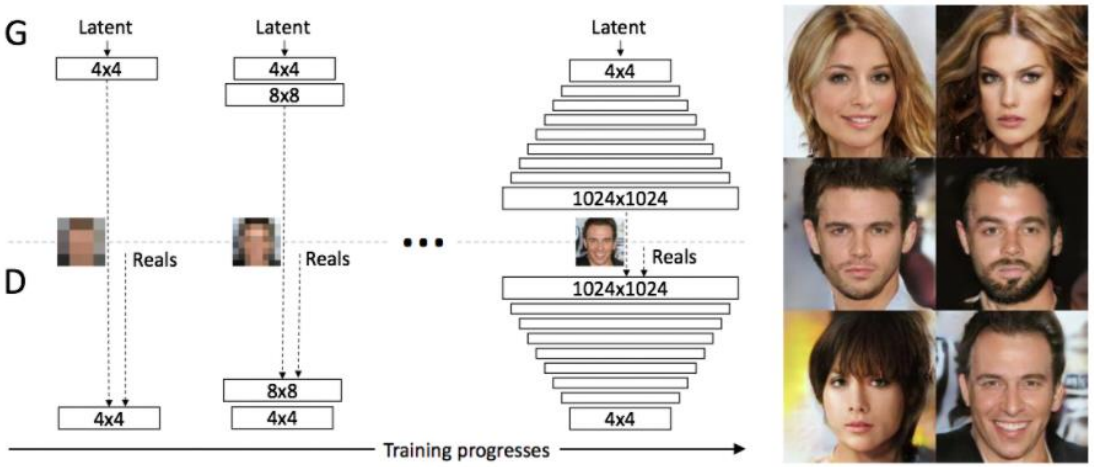
- Pyramid Feature Attention Network for Saliency detection(2020)

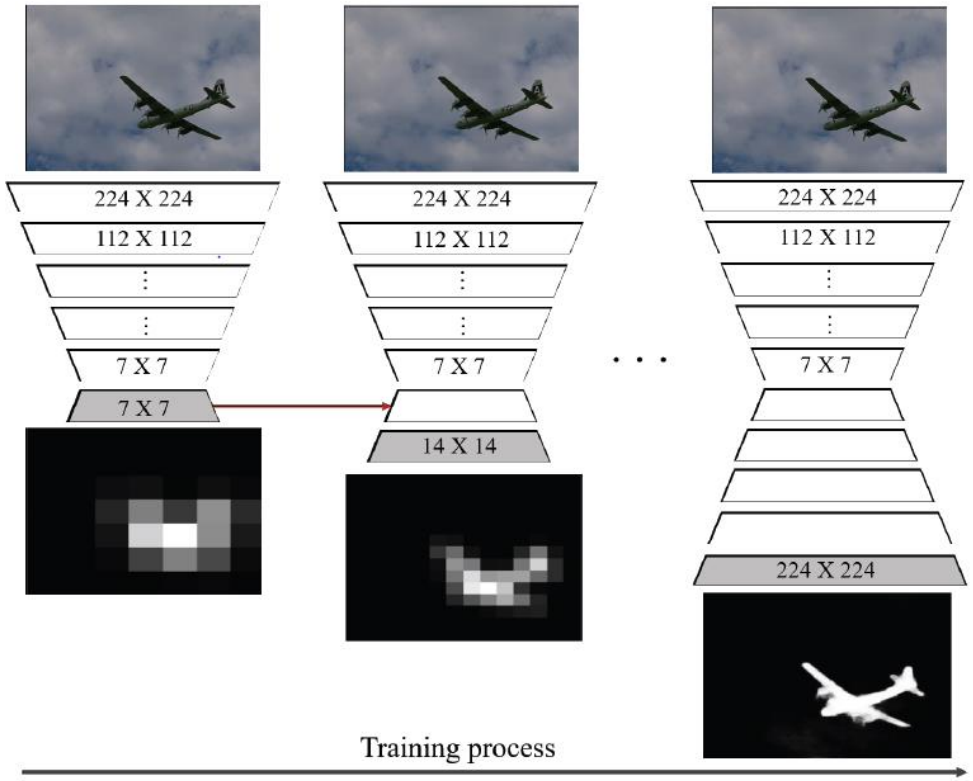




Progressive Learning

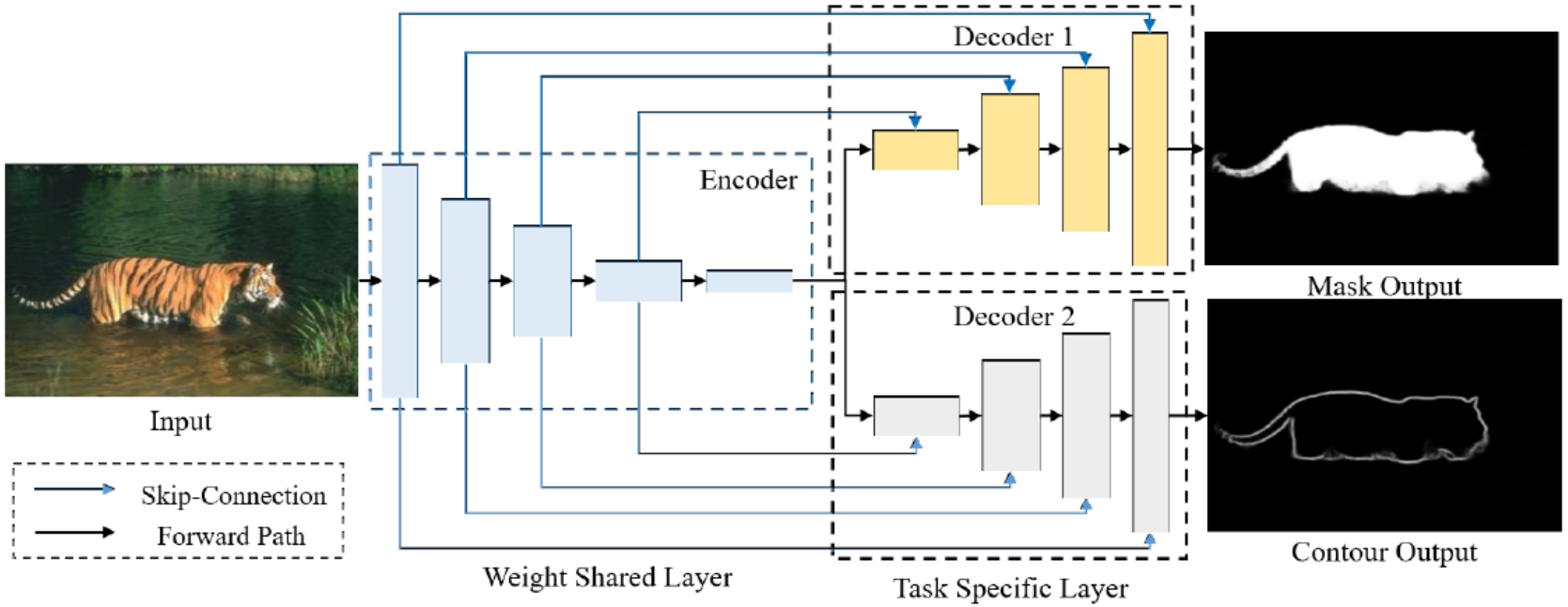
- Inspired by pggan(2018) approach
- Progressive learning scheme progressively grows decoder in the training phase.
- In other words, it starts from easier low-resolution layers, and adds new higher-resolution layers.
- This can weakly localize candidates in row-resolution and improve stability in high-resolution layers.





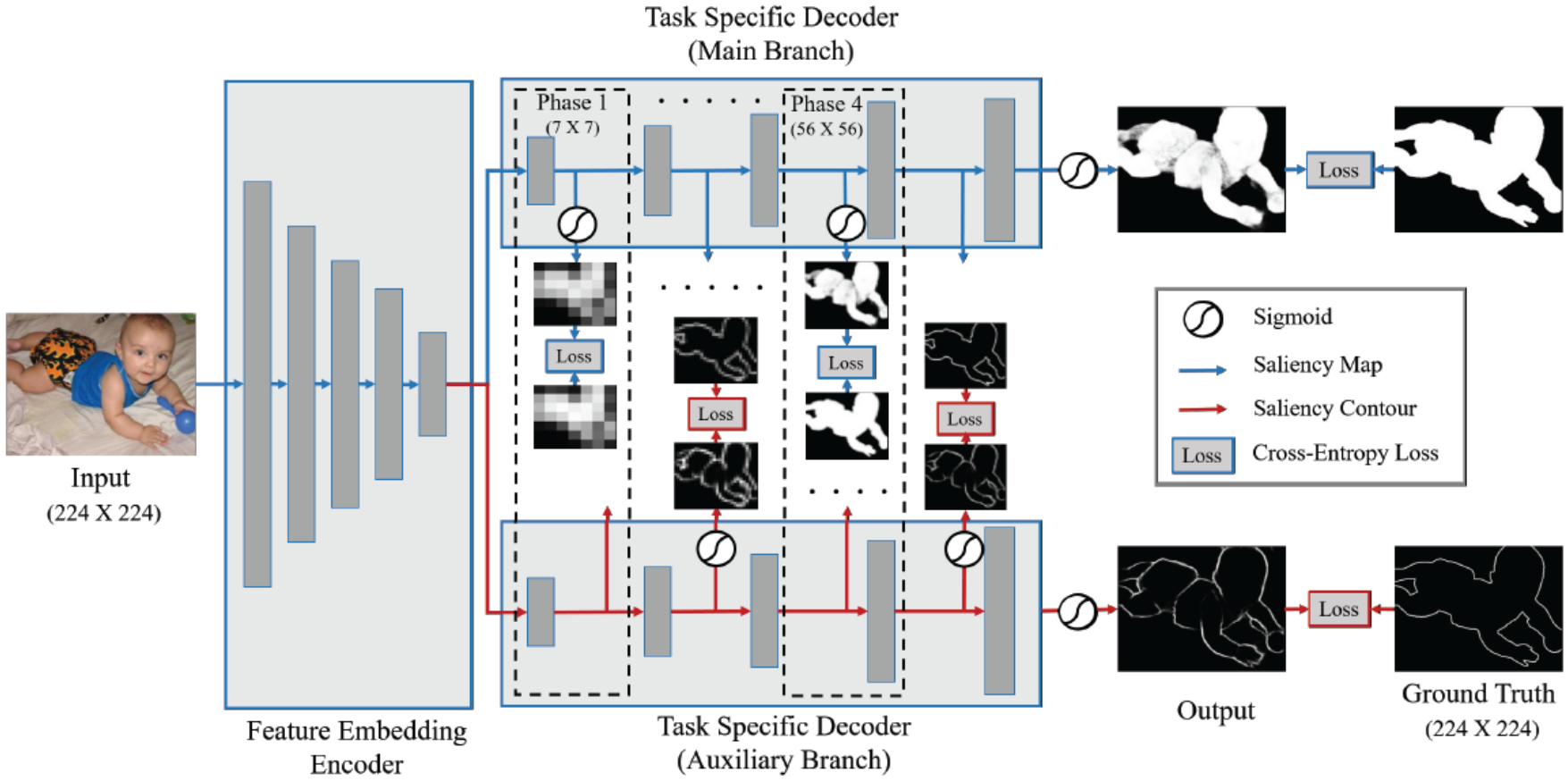
Multi-task Learning

- Multi-task learning is to learn multiple tasks jointly by weight shared encoder and task specific decoders to improve generalization and mitigate undesired artifacts

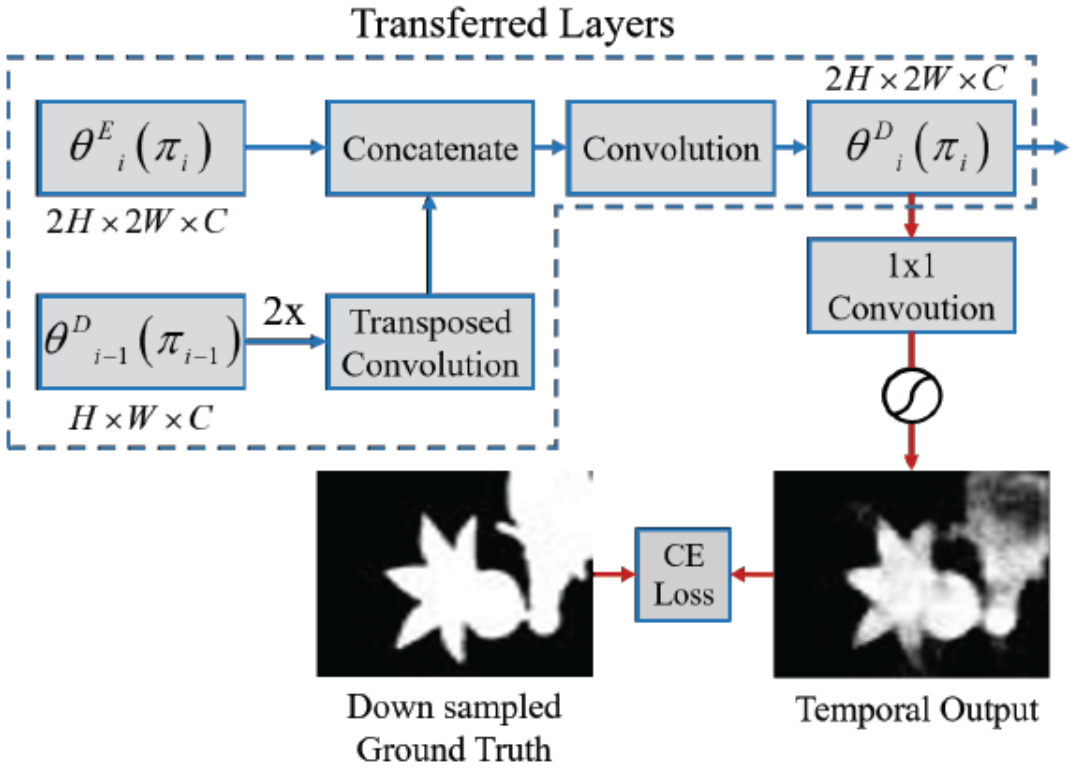


Progressive and Multi-task Learning

- Most deep learning based methods tried to manipulate the convolution block
- Our approach only manipulating the learning scheme without changing the network architecture



Progressive and Multi-task Learning



$$\theta_i^D(\pi_i), i \in \{1, \dots, 6\}$$

- The lower phases $(\theta_1, \theta_2, \theta_3)$ focus on the context of the object, so it can solve the problem of insufficient receptive field
- The higher phases $(\theta_4, \theta_5, \theta_6)$ learn the fine detail of the saliency map

Loss Function

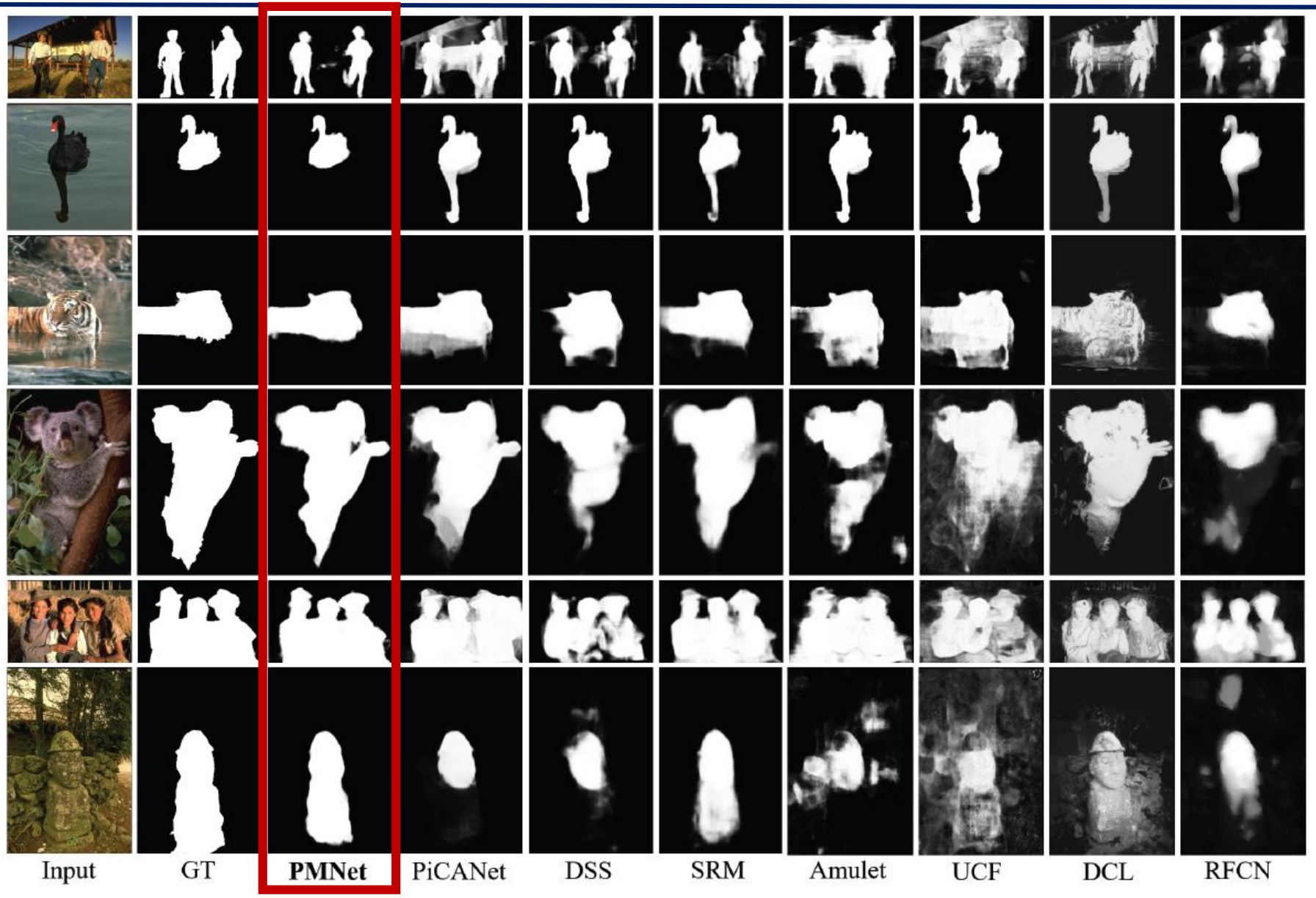
$$L(y, \hat{y}) = -\frac{1}{N}\sum_{j}^{N}\left(y_j \log\left(\hat{y}_j\right) + \left(1 - y_j\right)\log\left(1 - \hat{y}_j\right)\right)$$

❖ Where y is the pixel value of ground truth image and $\hat{y}$ is the pixel value of the network output. $i \in \{0,1,\ldots,N\}$. N is the number of pixel of input image.

$$L_i^{final} = L_i\left(S_i, \hat{S}_i^{E_1D_1}\right) + L_i\left(C_i, \hat{C}_i^{E_1D_2}\right)$$

❖Where S and C represent the saliency map and contour goround truth, respectively. $\hat{C}^{E_1D_1}$ and $\hat{C}^{E_1D_2}$ represent output of each branch. $i \in \{0,\ldots,M\}$. M is the number of training phase.

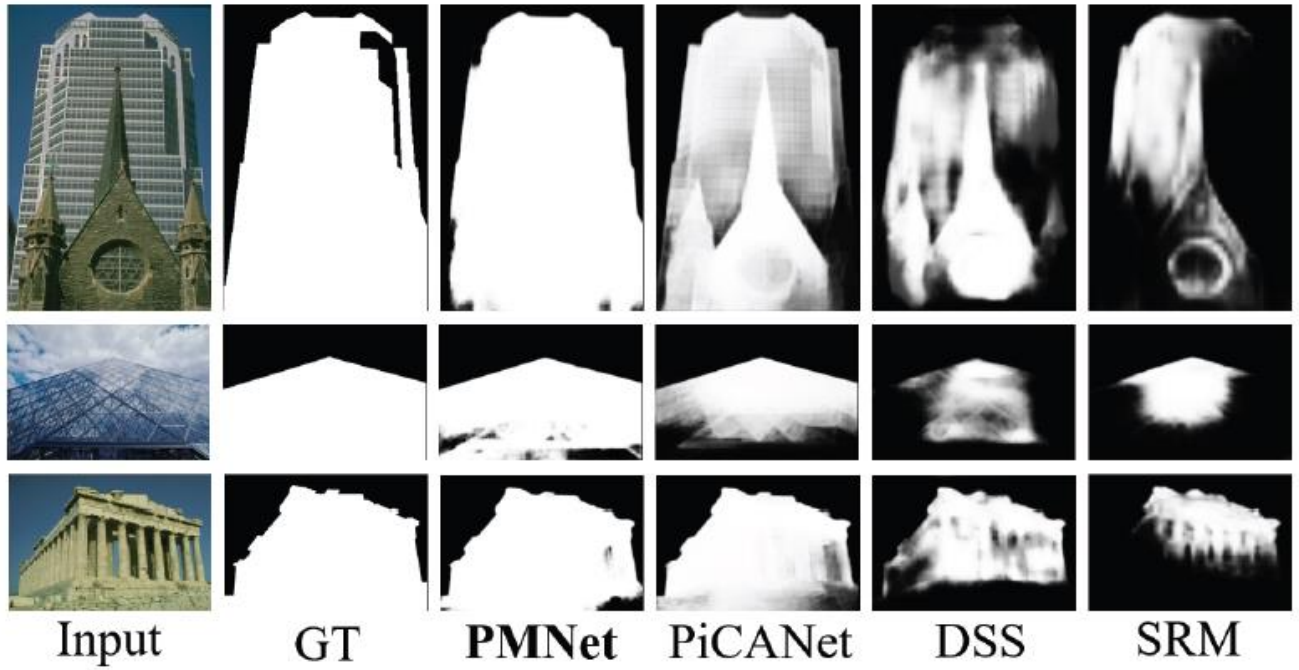
Experiment Result



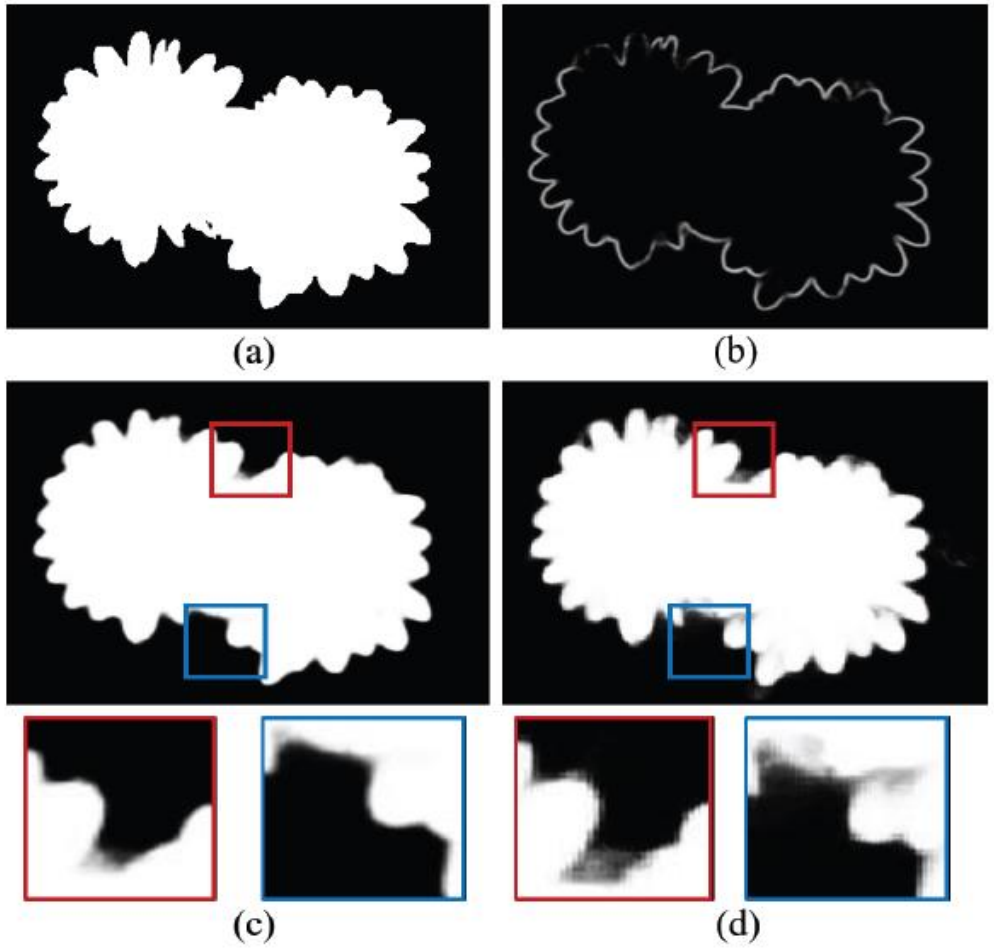
Experiment Result

Dataset	SOD [36]			HKU-IS [37]			DUT-O [38]			DUTS-TE [39]			ECSSD [40]		
Metric	*max* F_β	*mean* F_β	MAE	*max* F_β	*mean* F_β	MAE	*max* F_β	*mean* F_β	MAE	*max* F_β	*mean* F_β	MAE	*max* F_β	*mean* F_β	MAE
PMNet	0.862	0.816	0.094	0.922	0.909	0.036	0.790	0.749	0.060	0.850	0.834	0.051	0.936	0.912	0.041
PiCANet [43]	0.853	0.791	0.102	0.919	0.870	0.043	0.794	0.710	0.068	0.845	0.755	0.054	0.931	0.884	0.047
RAS [44]	0.850	0.799	0.124	0.913	0.871	0.045	0.786	0.713	0.062	0.831	0.755	0.060	0.921	0.889	0.056
PAGR [45]	-	-	-	0.918	0.886	0.048	0.771	0.711	0.071	0.855	0.788	0.056	0.927	0.894	0.061
DSS [27]	0.844	0.795	0.121	0.910	0.895	0.041	0.771	0.729	0.066	0.825	0.791	0.057	0.916	0.901	0.052
SRM [16]	0.843	0.800	0.127	0.906	0.874	0.046	0.769	0.707	0.069	0.827	0.757	0.059	0.917	0.892	0.054
Amulet [17]	0.806	0.755	0.141	0.895	0.839	0.052	0.742	0.647	0.098	0.778	0.676	0.085	0.915	0.870	0.059
DHS [26]	0.823	0.774	0.128	0.892	0.855	0.052	-	-	-	0.815	0.724	0.065	0.905	0.872	0.062
UCF [18]	0.803	0.699	0.164	0.886	0.808	0.074	0.734	0.613	0.132	0.771	0.629	0.117	0.911	0.840	0.078
DCL [46]	0.823	0.741	0.141	0.885	0.853	0.072	0.739	0.684	0.097	0.782	0.714	0.088	0.890	0.829	0.088
RFCN [19]	0.805	0.751	0.161	0.895	0.835	0.079	0.747	0.627	0.094	0.786	0.712	0.090	0.898	0.834	0.097
MDF [37]	0.787	0.721	0.159	0.861	0.784	0.129	0.694	0.644	0.092	0.730	0.673	0.094	0.832	0.807	0.105
DS [47]	0.784	0.698	0.190	0.865	0.788	0.080	0.745	0.603	0.120	0.777	0.633	0.090	0.882	0.826	0.122

Ablation Study – Large Object



Ablation Study – Effect of Multi-task Learning



Ablation Study – Effect of each component

Dataset	SOD [36]	HKU-IS [37]	DUT-O [38]
Method	MAE		
Baseline [20]	0.1398	0.0502	0.0850
w/o Progressive Learning	0.1306	0.0447	0.0760
w/o Multi-task Learning	0.0969	0.0395	0.0639
Ours(PGL + MTL)	**0.0945**	**0.0363**	**0.0603**

Progressive and Multi-task Learning scheme