PMSNet: Propagated Masks Selection Network for Video Object Segmentation

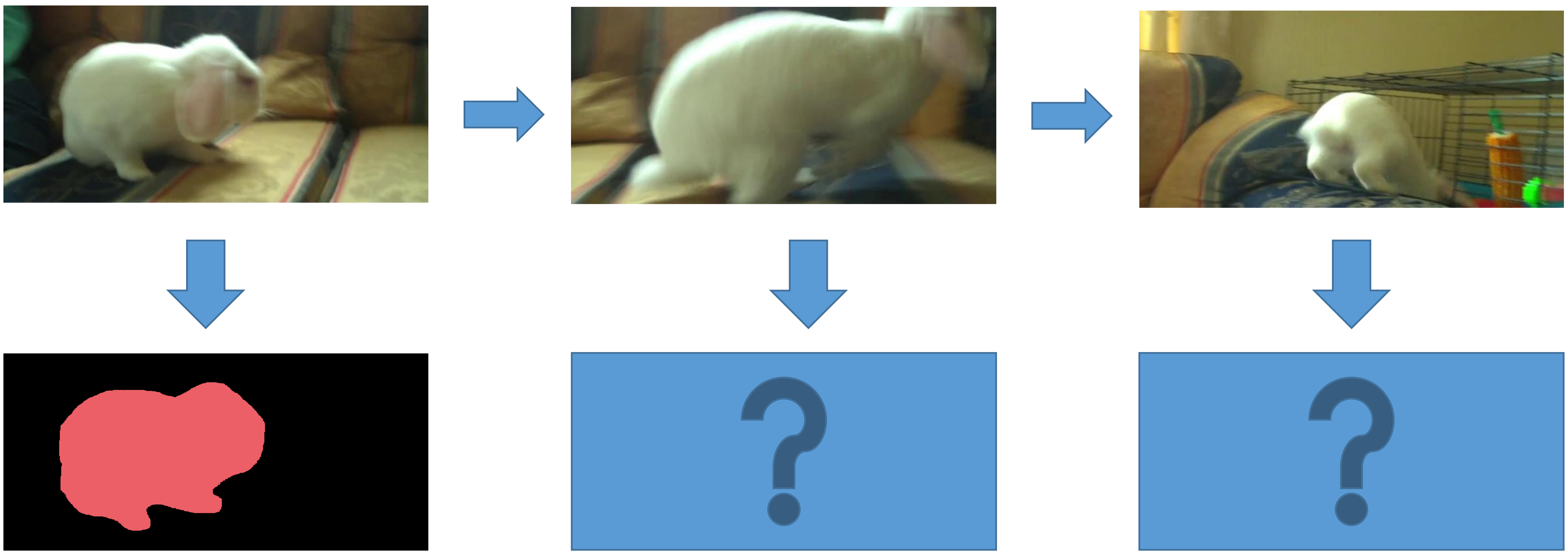
Huaijia Lin[1*], Ruizheng Wu[1,2*], Xiaogang Xu[1], Xiaojuan Qi[1], Jiaya Jia[1,2]

[1]The Chinese University of Hong Kong [2]Tencent YouTu Lab

*indicates equal contribution

Problem Definition

- Separating an object from the background in a video, given the mask of the first frame.

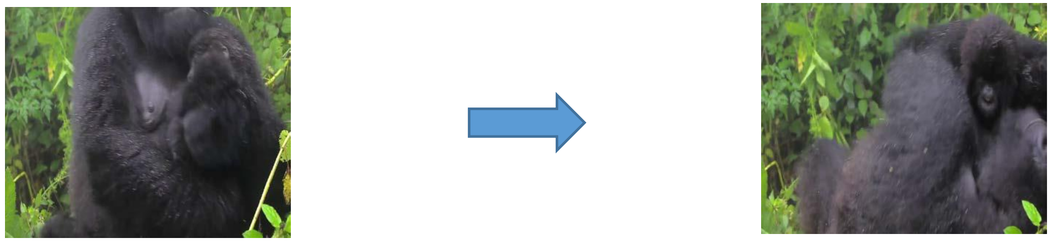


Challenges

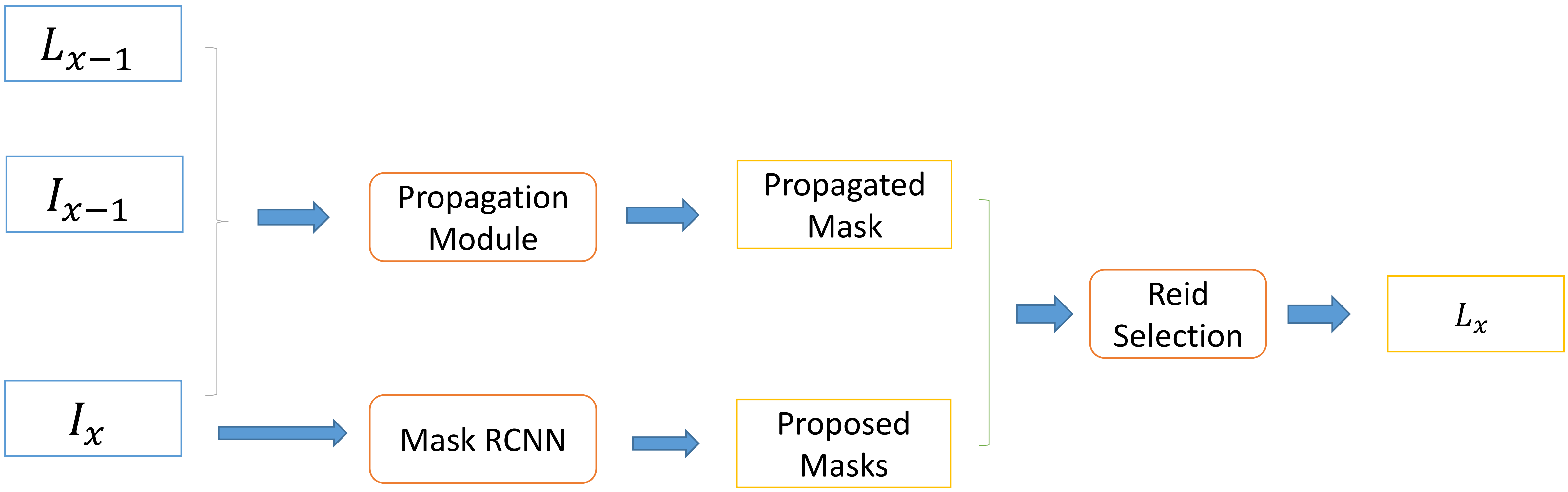
- Missing Objects Reappear

- Substantial Appearance Variations

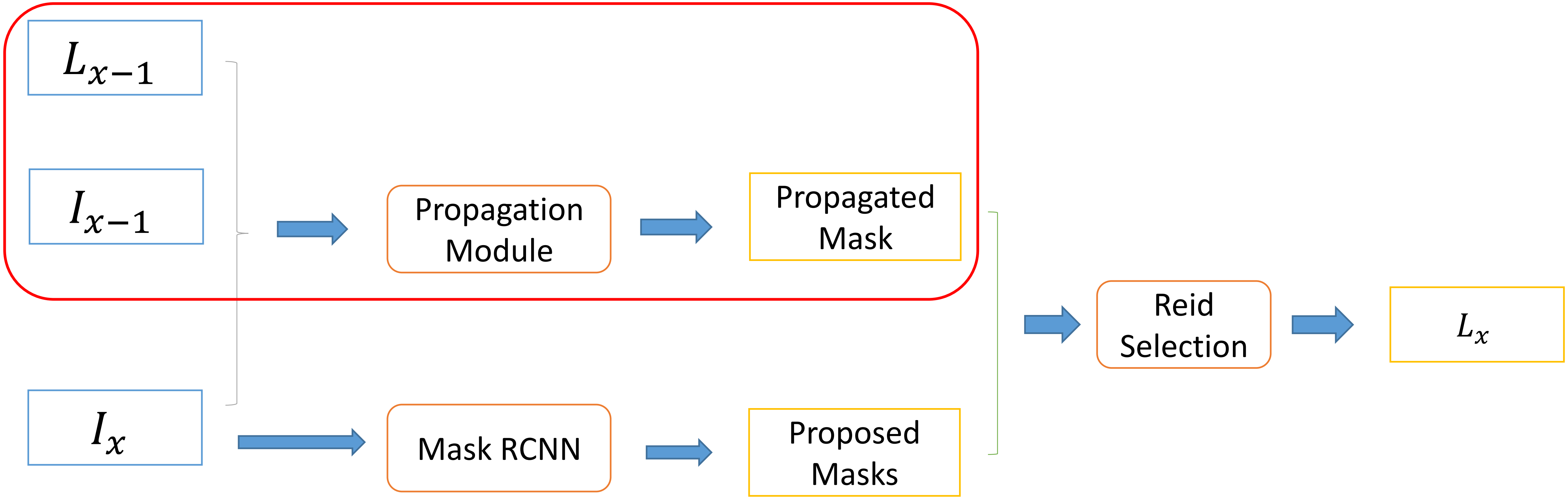
- Multiple Similar Objects Occluding



Framework – Inference Pipeline

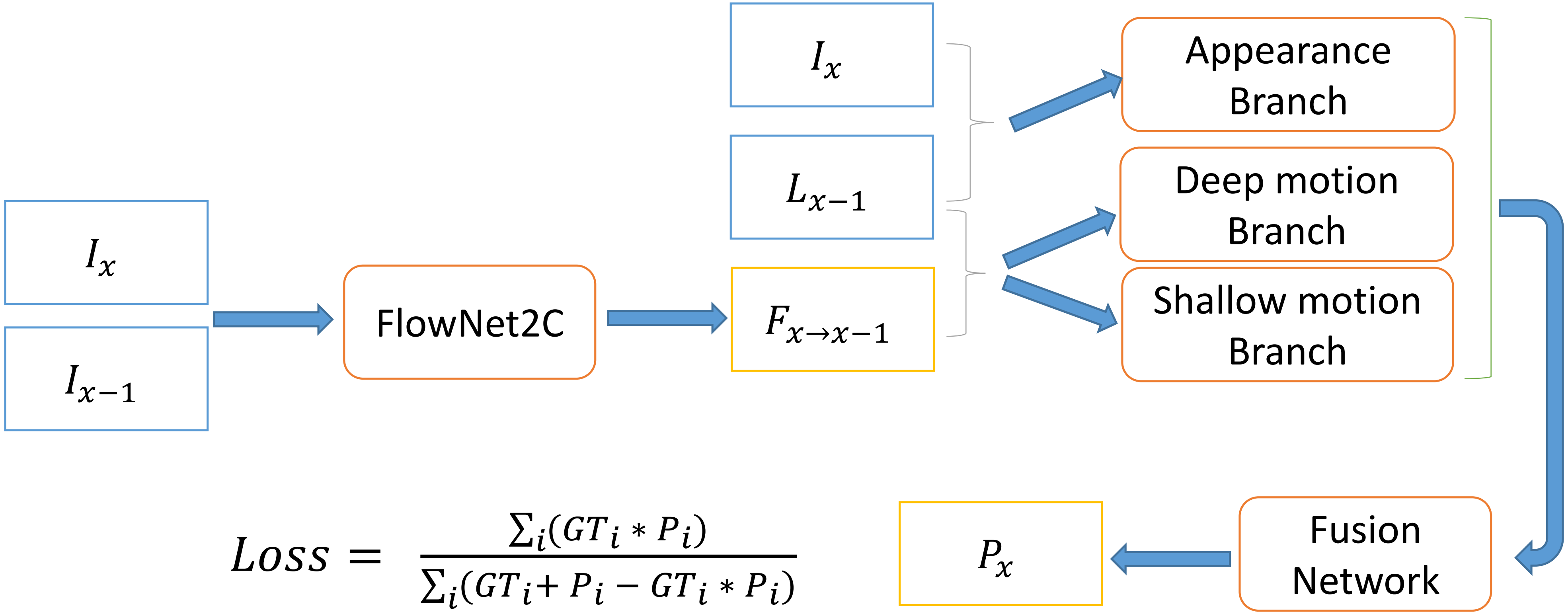


Framework – Inference Pipeline



Framework – Propagation Module

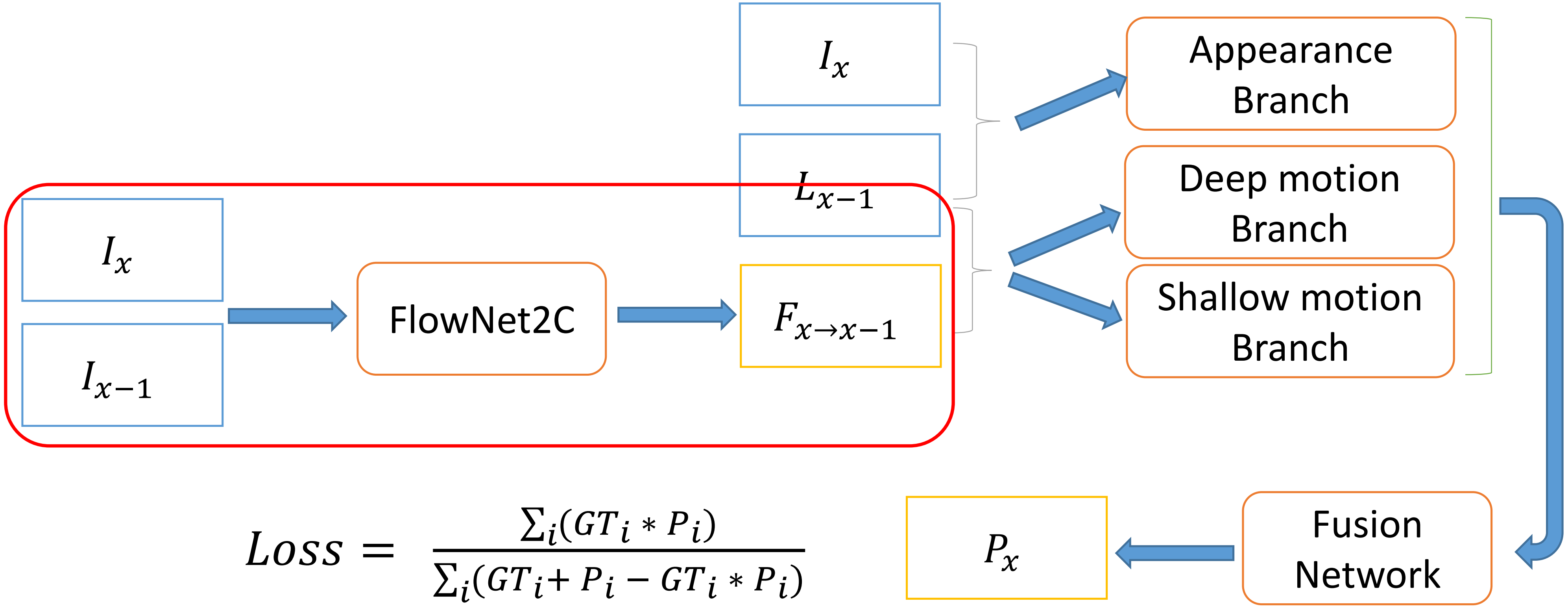
- Overview of Propagation Module



$$Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}$$

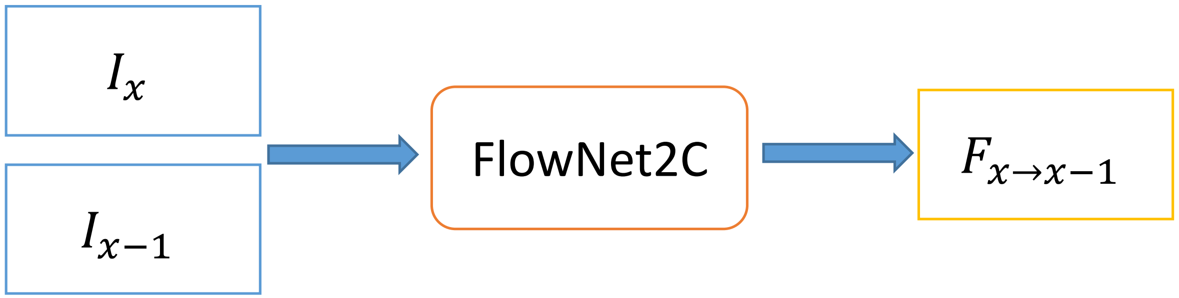
Framework – Propagation Module

- Overview of Propagation Module



Framework – Propagation Module

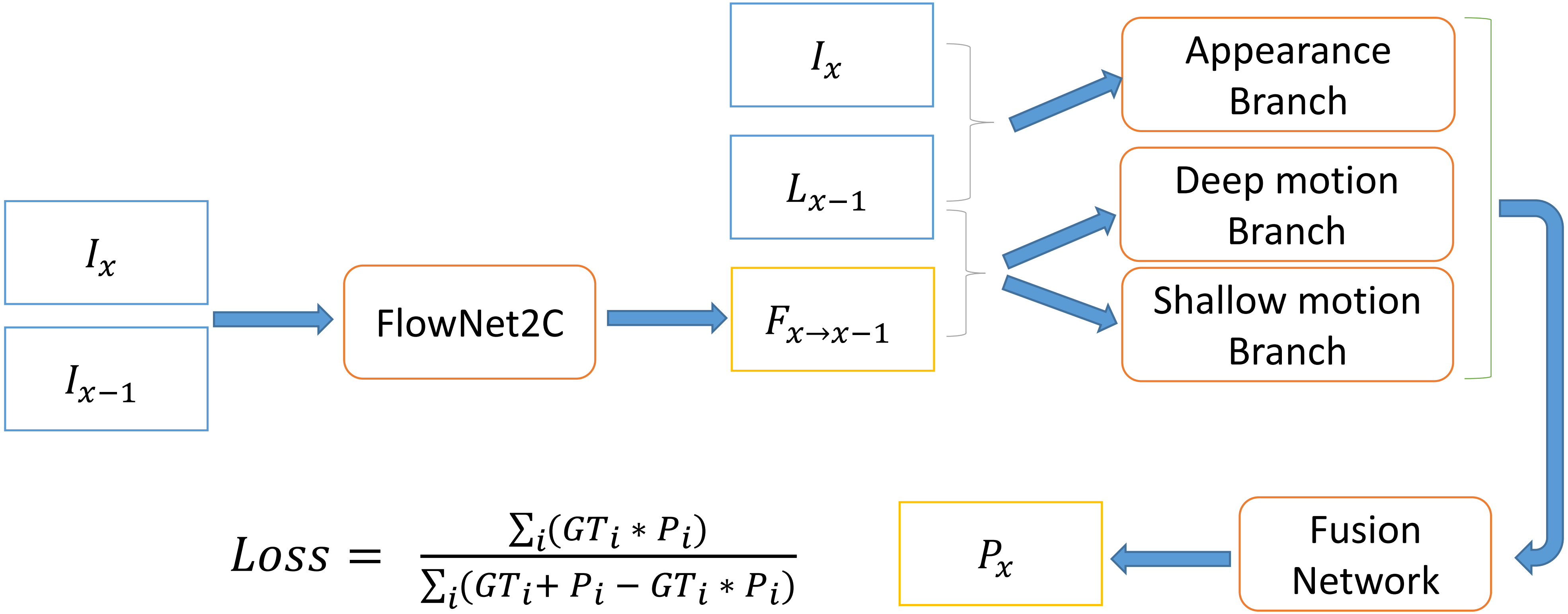
- **Motion Feature extract**
  - Adopt FlowNet2C[1] structure.
  - Load Flownet2C pre-trained weight.
  - The magnitude of optical flow $\|F_{x\rightarrow x-1}\|_2^2$ will be the motion features for subsequent network.
  - Learn motion features end-to-end.



[1] Dosovitskiy, Alexey, et al. "FlowNet: Learning Optical Flow with Convolutional Networks." *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2015.

Framework – Propagation Module

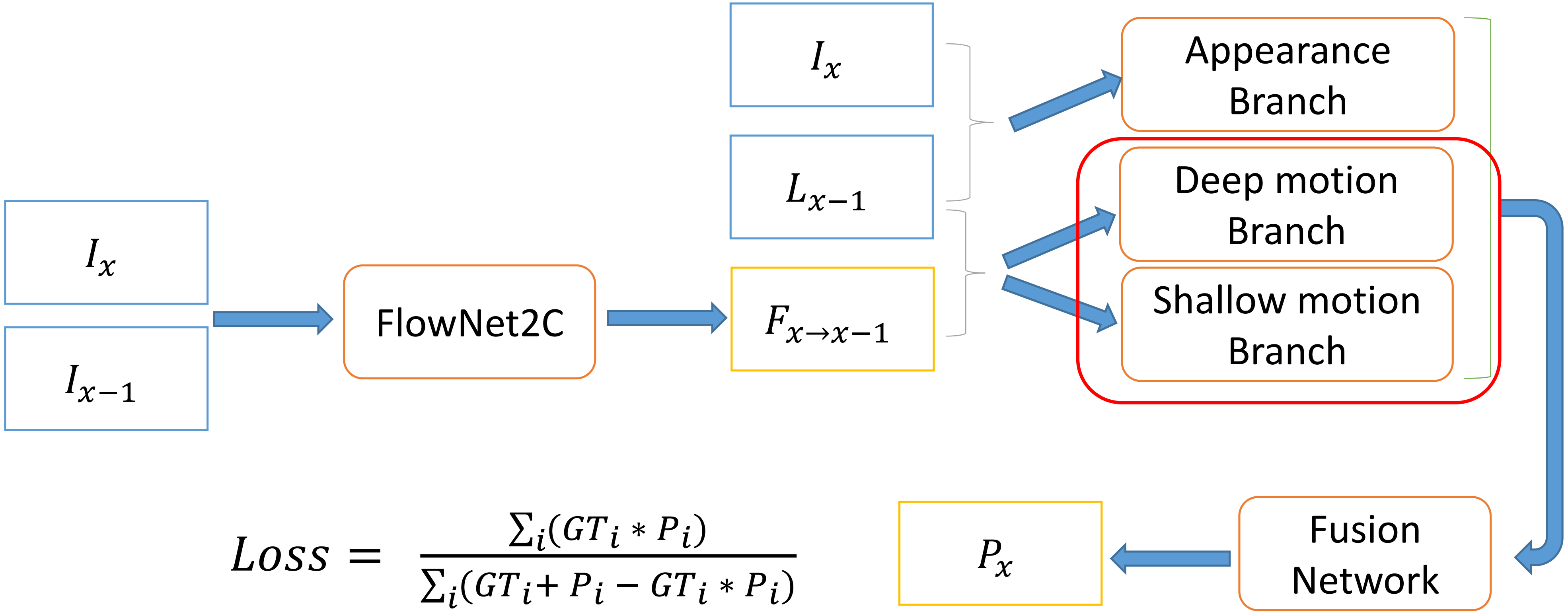
- Overview of Propagation Module



$$Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}$$

Framework – Propagation Module

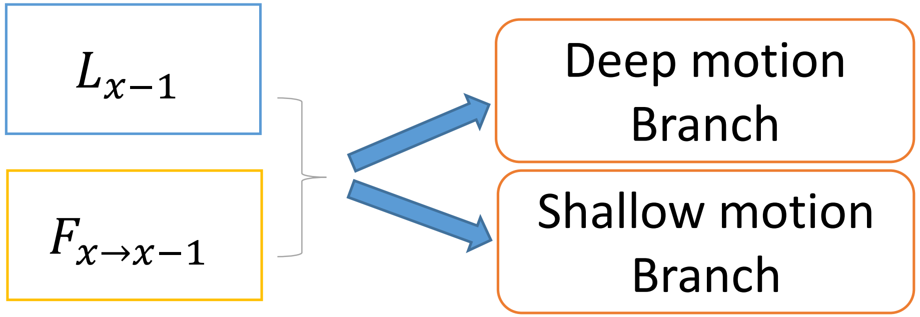
- Overview of Propagation Module



$$Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}$$

Framework – Propagation Module

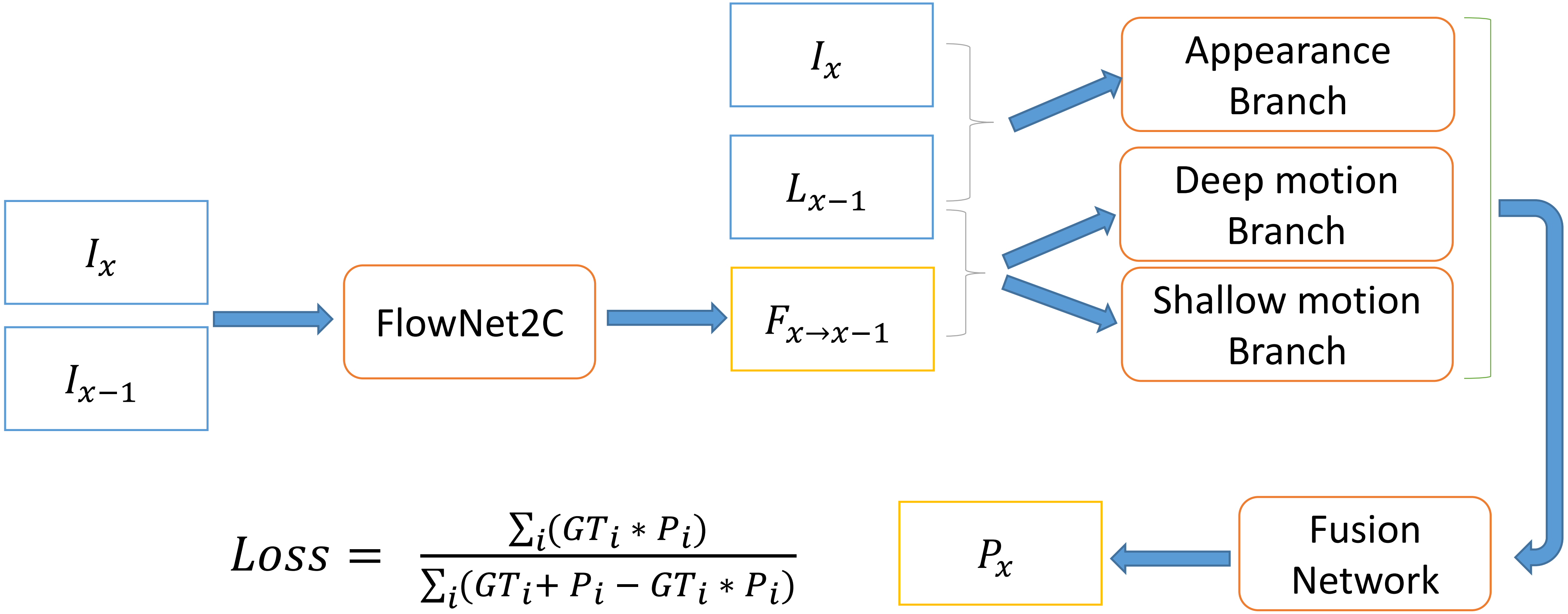
- **Motion branches**
  - **Input:**
    - Last frame label L_{x-1}.
    - Motion features from current frame to last frame $F_{x \rightarrow x-1}$.
  - **Deep Motion Branch**
    - Adopt OSVOS[1] network structure.
    - Load with VGG16 pre-trained weight.
  - **Shallow Motion Branch**
    - Several Convolution-Relu blocks.
    - No down-sample operation.
    - It improves **10.56** in validation set overall score.



[1] S. Caelles*, K.K. Maninis*, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool
One-Shot Video Object Segmentation, Computer Vision and Pattern Recognition (CVPR), 2017.

Framework – Propagation Module

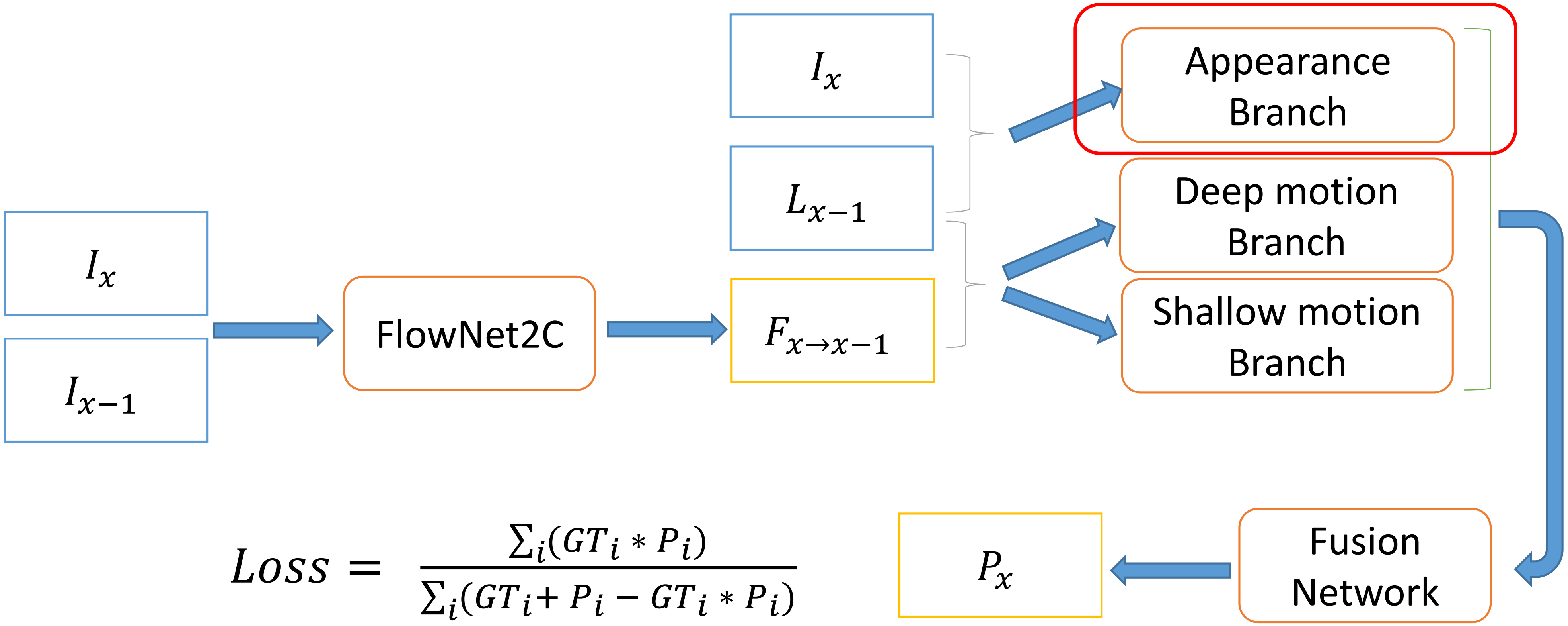
- Overview of Propagation Module



$$Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}$$

Framework – Propagation Module

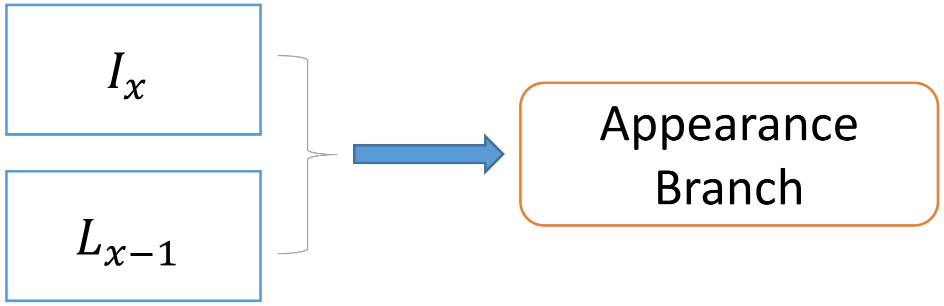
- Overview of Propagation Module



$$Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}$$

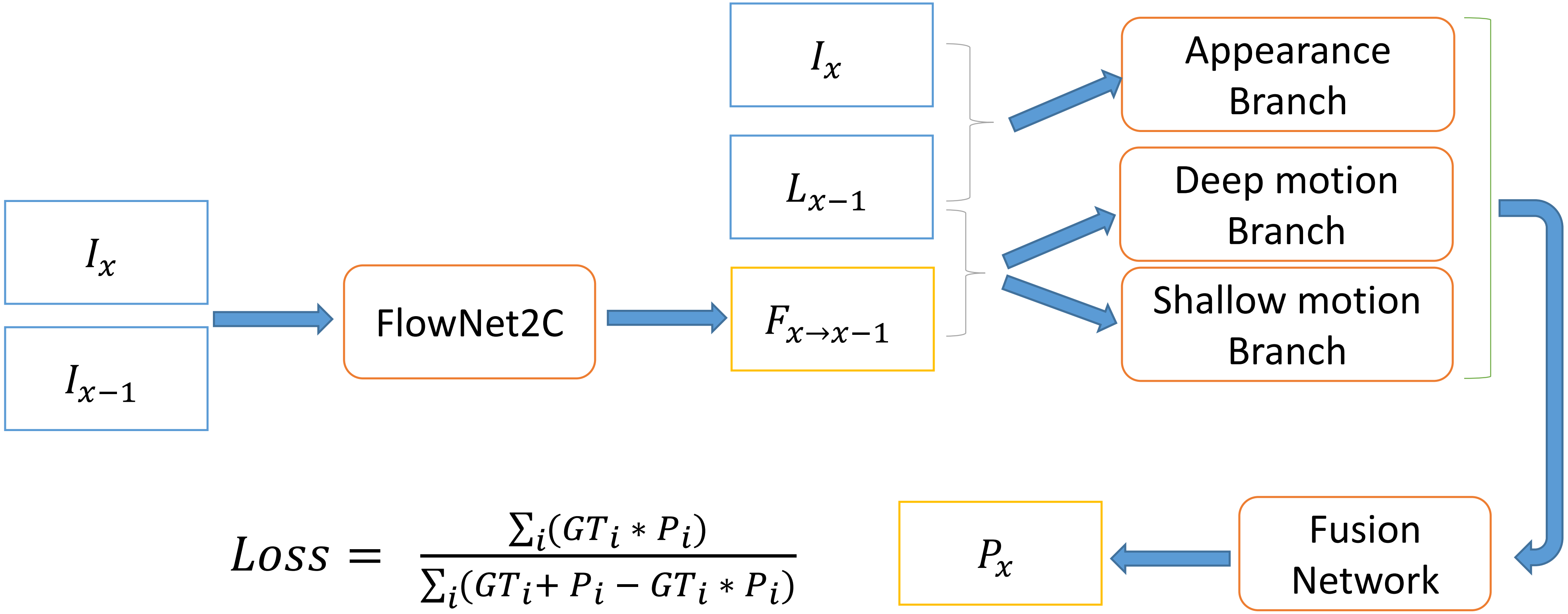
Framework – Propagation Module

- **Appearance Branch**
  - **Input:**
    - Current frame RGB image I_x
    - Last frame label L_{x-1}
  - **Network setting**
    - Adopt OSVOS network structure.
    - Load with VGG16 pre-trained weight.



Framework – Propagation Module

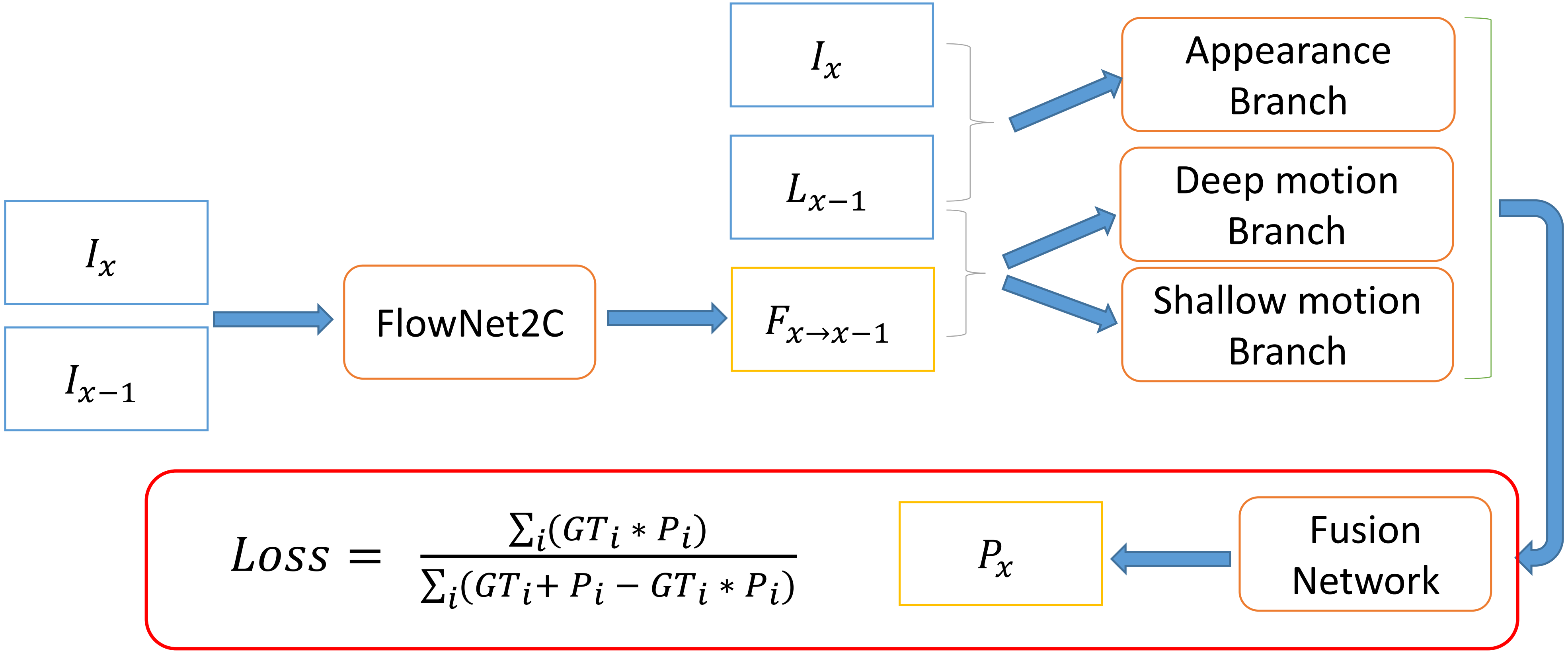
- Overview of Propagation Module



$$Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}$$

Framework – Propagation Module

- Overview of Propagation Module

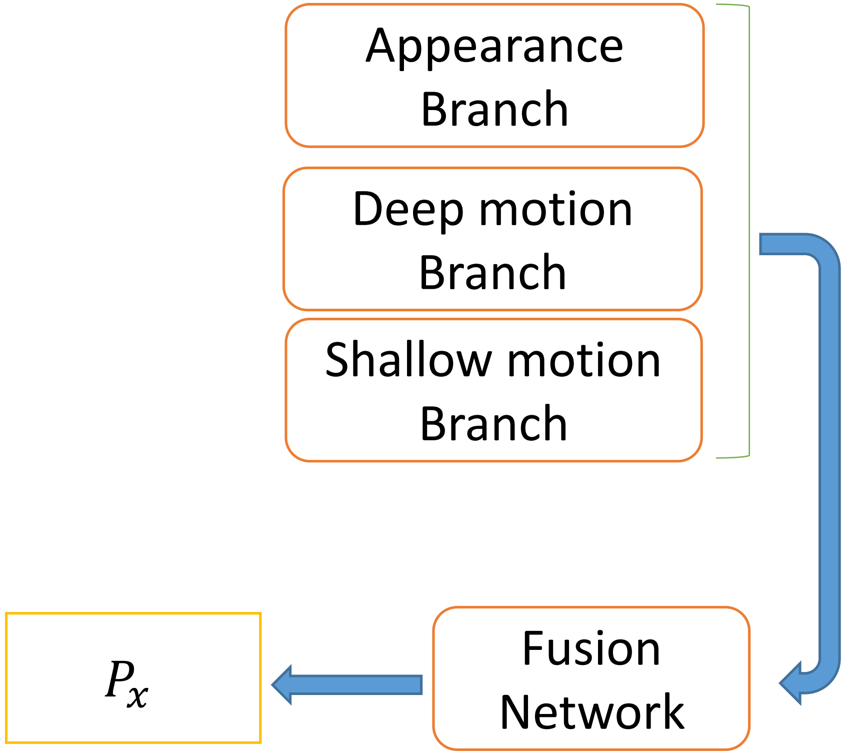


Framework – Propagation Module

- **Fusion & Loss function**
  - We fuse three branch prediction.
  - IoU loss is set as our loss function.

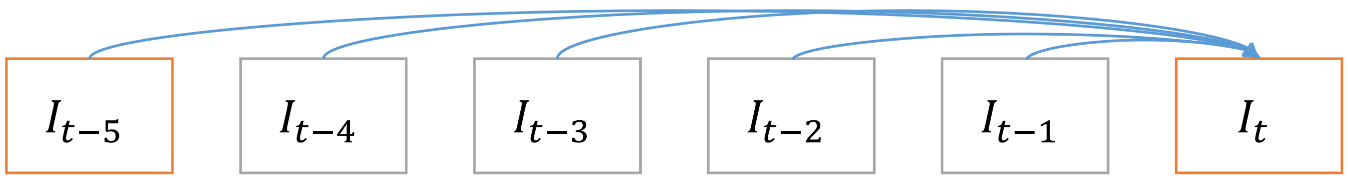
$$Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}$$

  - The propagated mask is utilized for subsequent selection.



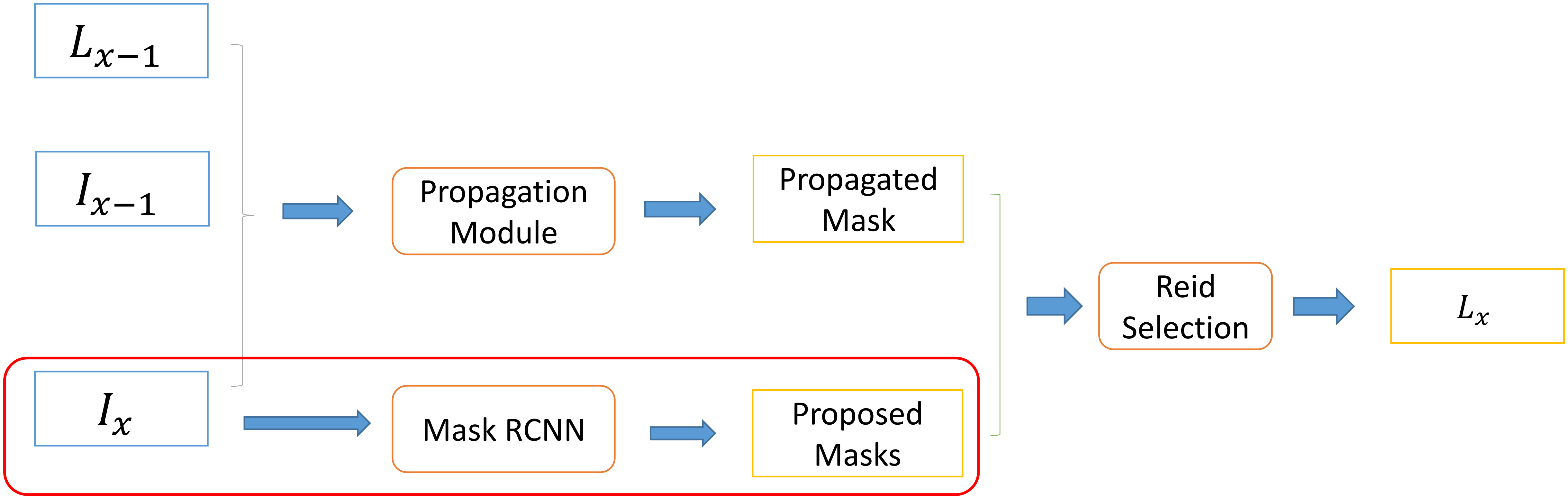
Framework – Propagation Module

- Inference strategy
  - Multi-frame ensemble



- We ensemble the prediction from previous 5 frames.
- Only the results of I_{t-5} and I_t will be saved for validation, all frame results will be saved for testing.
- It improves **2.52** in validation set overall score.

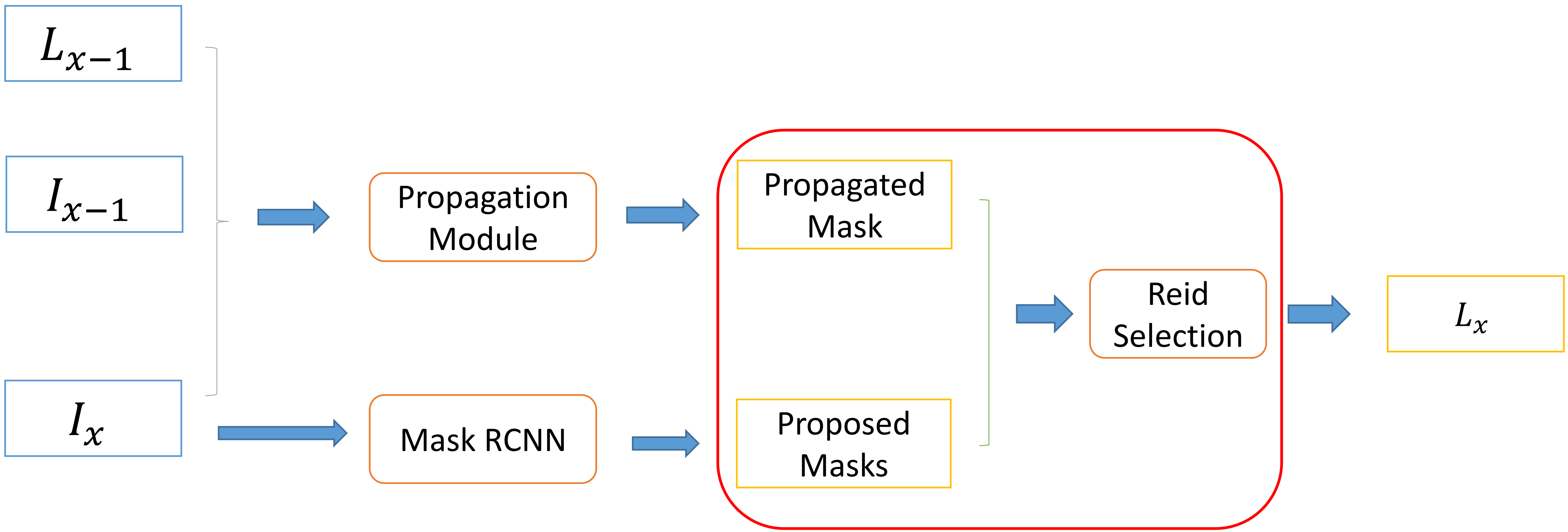
Framework – Inference Pipeline



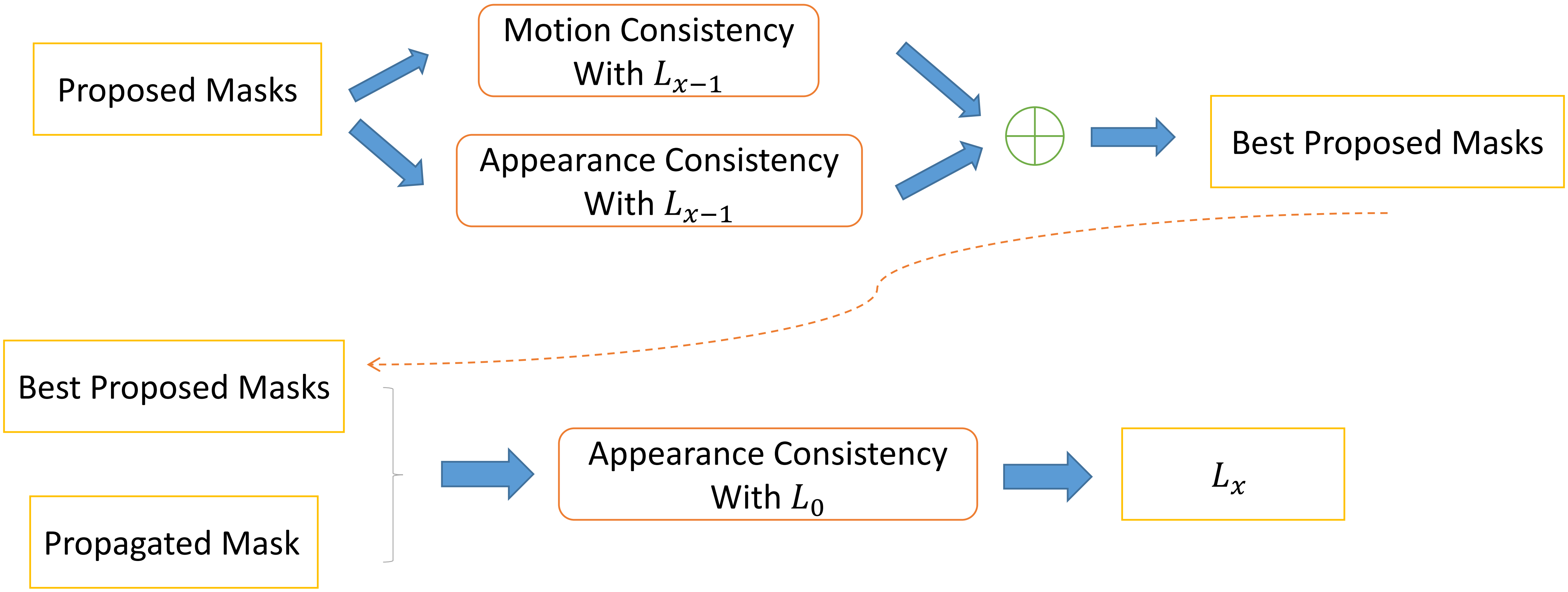
We directly use the pre-trained model[1] of coco dataset.

[1] He, Kaiming, et al. "Mask R-CNN." *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

Framework – Inference Pipeline

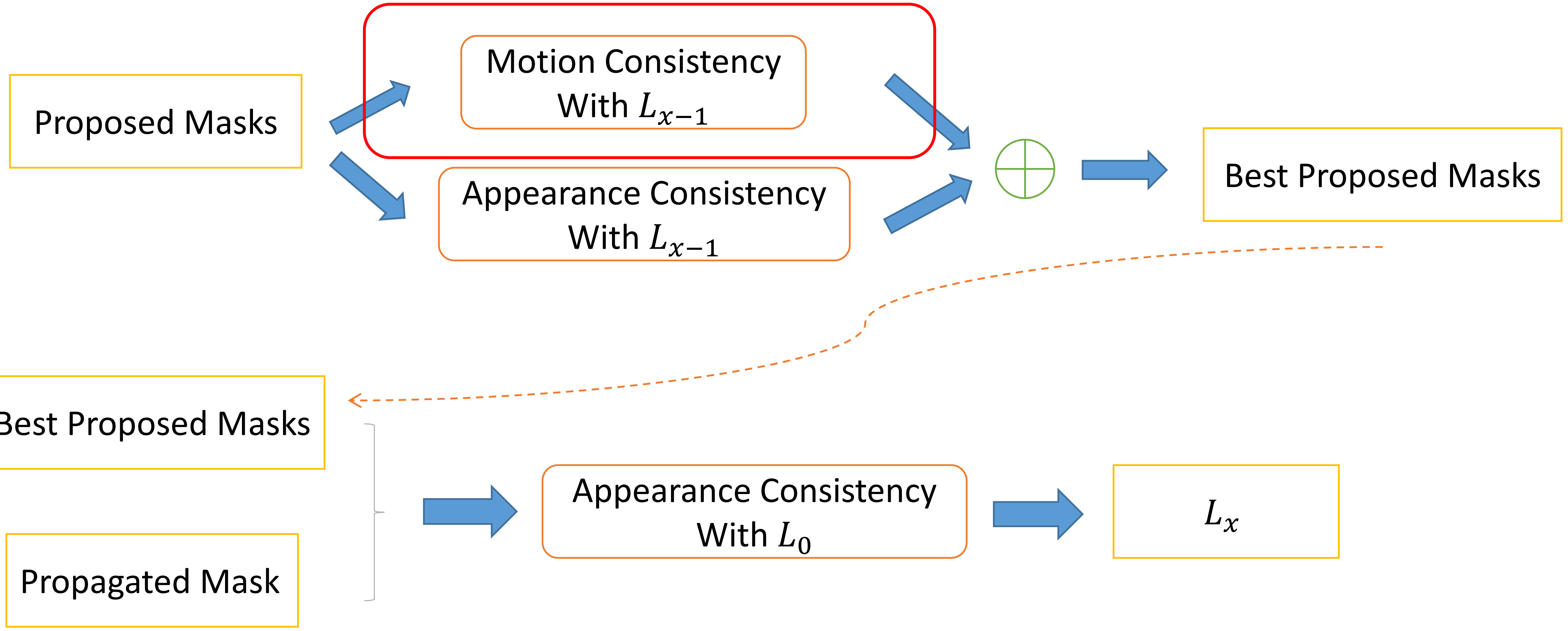


Framework – Reid Selection



L_0 denotes the given mask.

Framework – Reid Selection



L_0 denotes the given mask.

Reid Selection – Motion Consistency

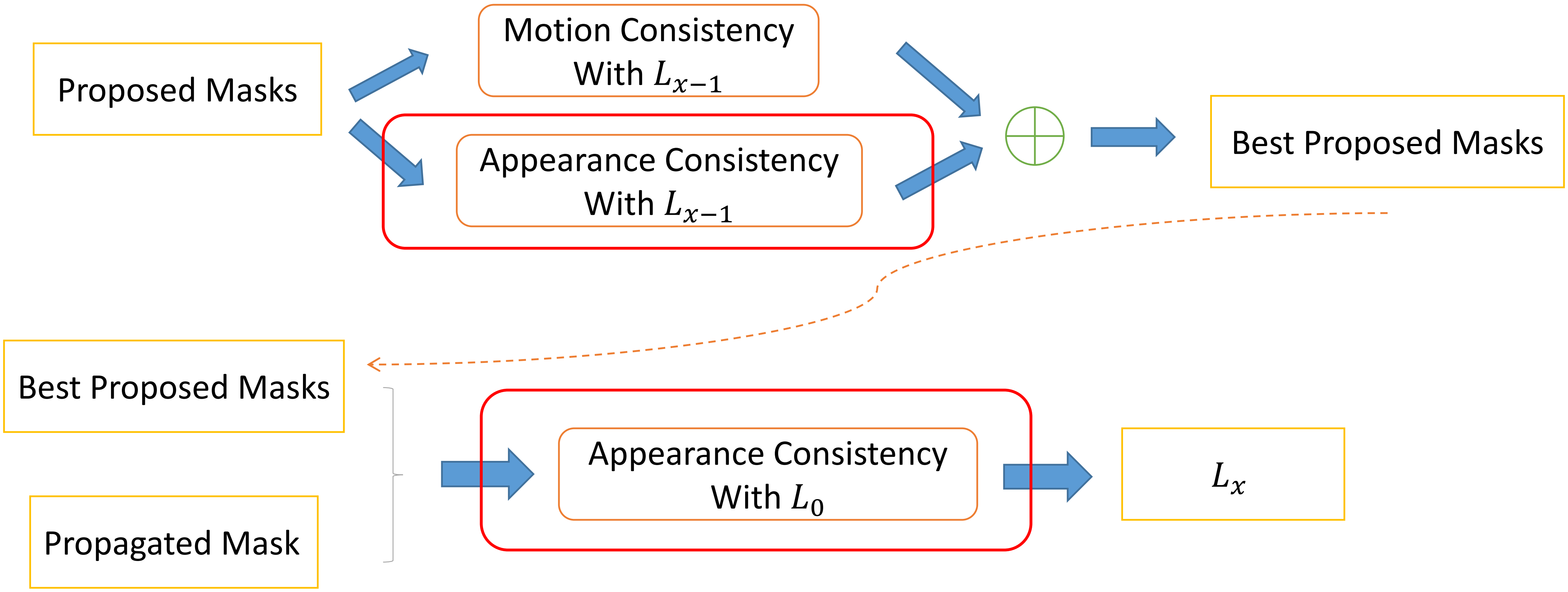
Motion Consistency With L_{x-1}

- **Motion Consistency**

$$S_{MC}(Mask, L_{x-1}) = \frac{1}{2}(IoU(F_{x \to x-1}(Mask), L_{x-1}) + IoU(F_{x-1 \to x}(L_{x-1})), \mathrm{Mask}))$$

  - $F_{x \to x-1}$ and $F_{x-1 \to x}$ is the warp operation with the optical flow
  - For saving computation, the masks with $S_{MC}(Mask, L_{x-1})$ smaller than a threshold (=0.2 in the experiment) are abandoned.

Framework – Reid Selection

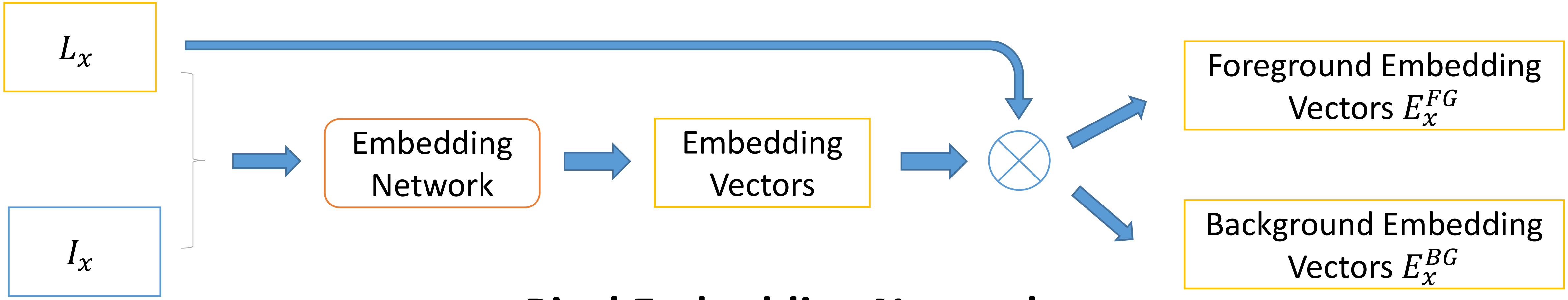


L_0 denotes the given mask.

Reid Selection – Appearance Consistency

- **Appearance Consistency**
  - Realize it by Pixel Embedding Network, inspired by [1]
  - For saving computation, the size of E_x^{FG}/E_x^{BG} is not more than a fixed number (=512 in the experiment).
  - For sampling evenly, down-sample the FG/BG mask to the corresponding area.

Appearance Consistency
With L_{x-1}/L_0



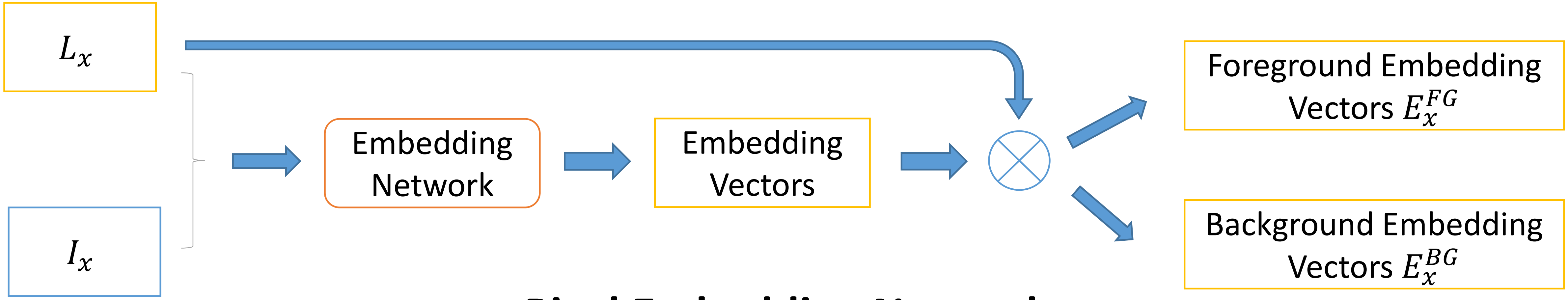
Pixel Embedding Network

[1] Chen, Yuhua, et al. "Blazingly Fast Video Object Segmentation with Pixel-Wise Metric Learning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Reid Selection – Appearance Consistency

- **Appearance Consistency**
  - Use the Embedding Vectors to calculate the number of valid vectors with function $\varphi(\cdot,(\cdot,\cdot))$

$$\varphi(V,(F,B)) = \sum_{f^V \in V} \boldsymbol{I}(min_{f^F \in F}||f^V - f^F|| - min_{f^B \in B}||f^V - f^B|| < 0)$$



Pixel Embedding Network

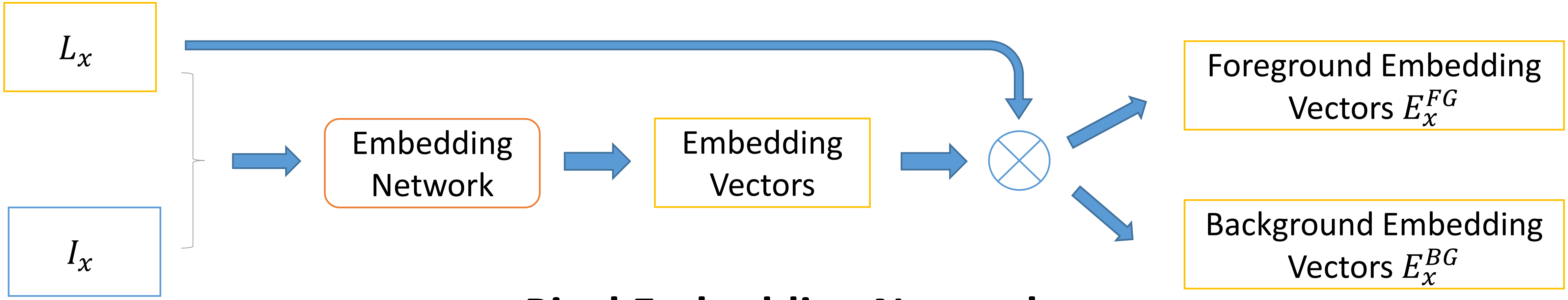
Reid Selection – Appearance Consistency

- **Appearance Consistency**
  - Use the Embedding Vectors to calculate the number of valid vectors with function $\varphi(\cdot,(\cdot,\cdot))$

$$\varphi\big(V,(F,B)\big) = \sum_{f^V \in V} \boldsymbol{I}(min_{f^F \in F}||f^V - f^F|| - min_{f^B \in B}||f^V - f^B|| < 0)$$

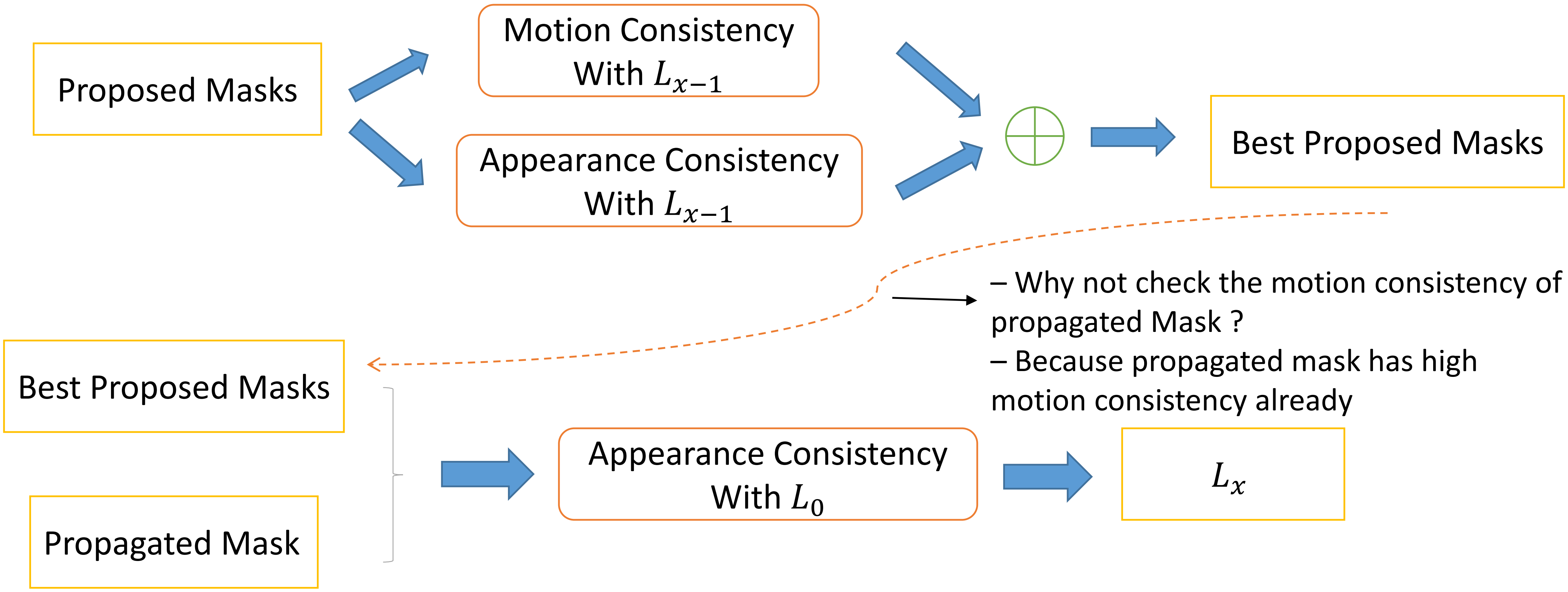
  - Use the valid number to calculate the appearance consistency with L_{x-1}/L_0

$$S_{AC}(Mask, L_{x-1}) = \frac{\varphi\left(E_x^{FG},(E_{x-1}^{FG},E_{x-1}^{BG})\right) + \varphi\left(E_{x-1}^{FG},(E_x^{FG},E_x^{BG})\right)}{|E_x^{FG}| + |E_{x-1}^{FG}|}$$



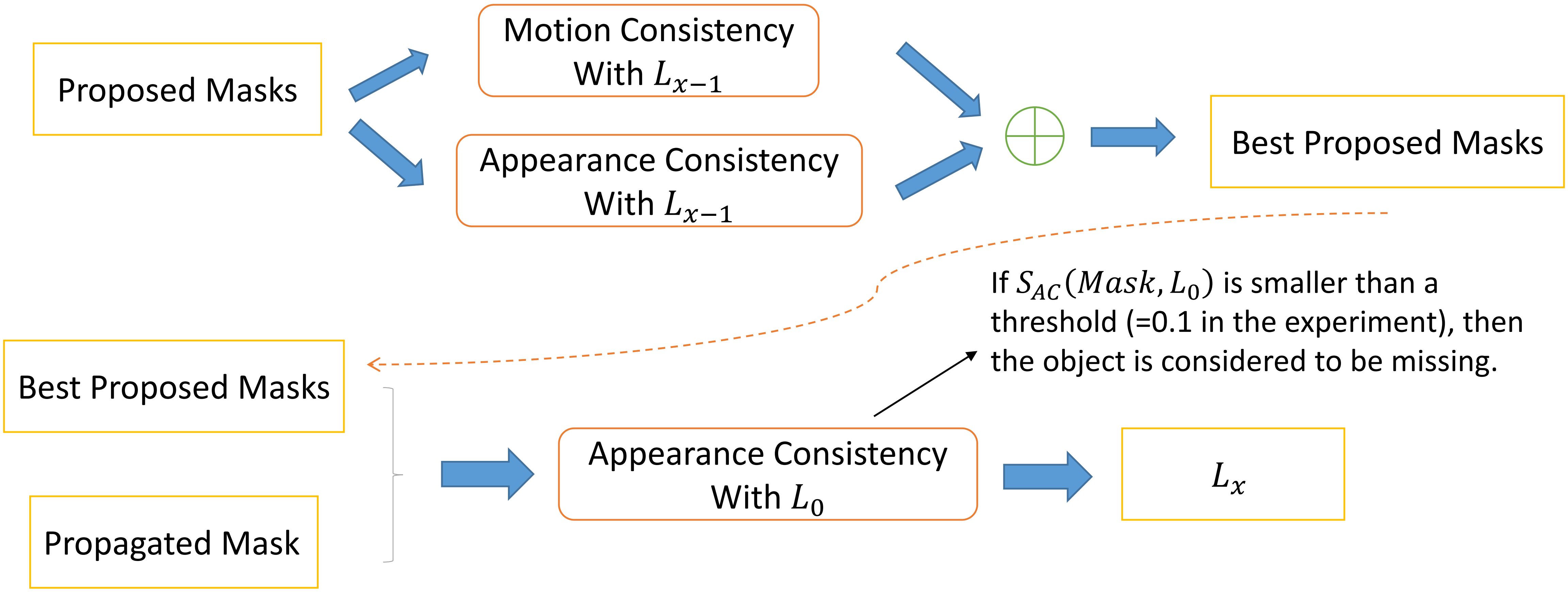
Pixel Embedding Network

Framework – Reid Selection



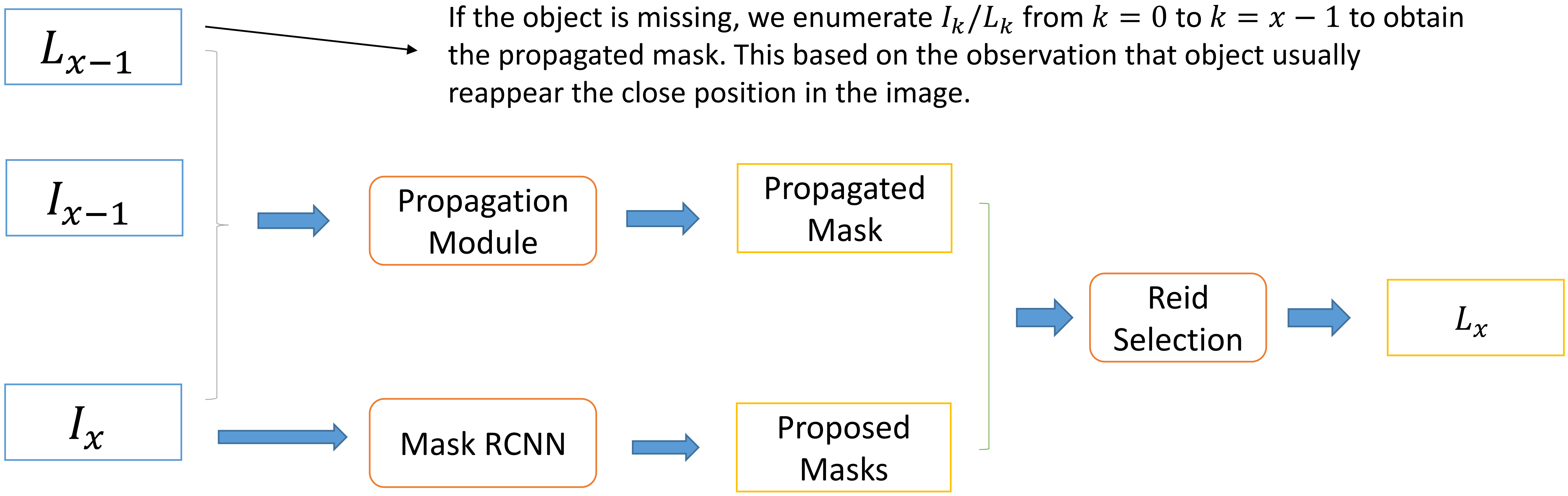
L_0 denotes the given mask.

Framework – Reid Selection



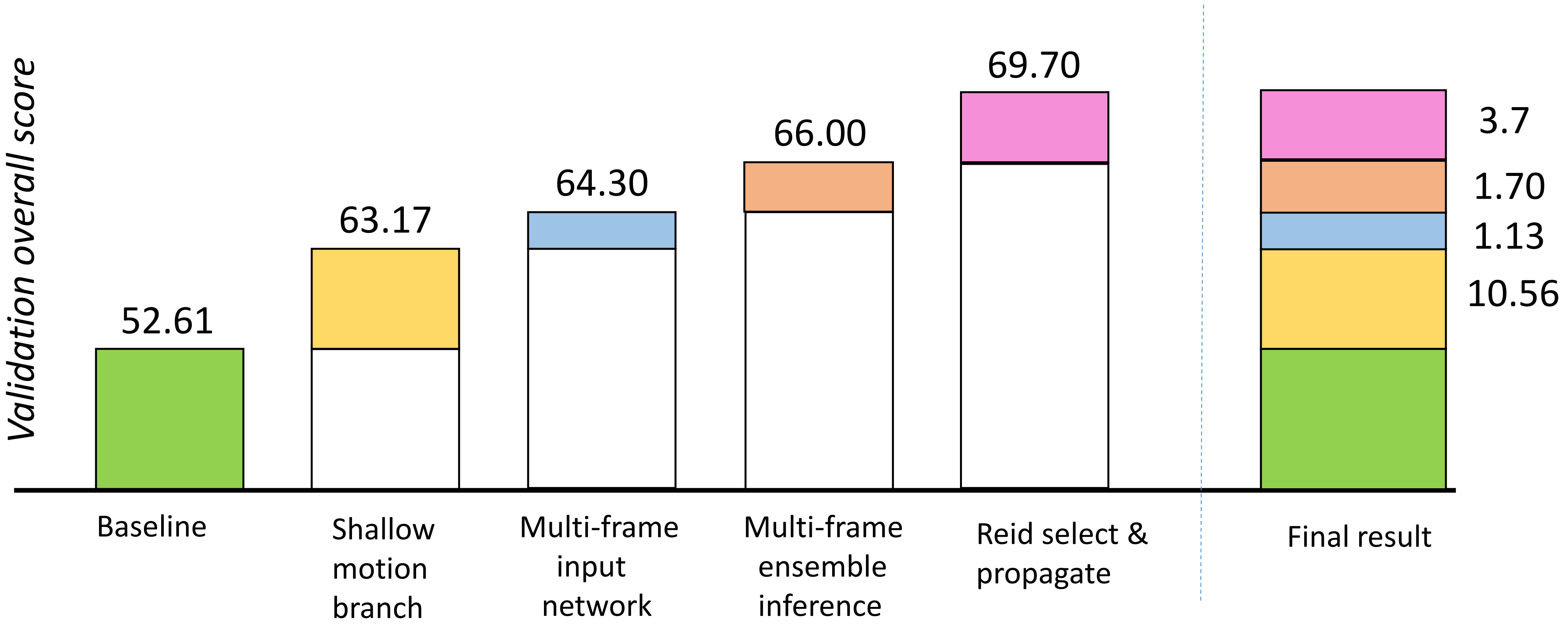
L_0 denotes the given mask.

Framework – Inference Pipeline



Results

- Summary of performance with different components



Results

- Our final results (rank 3rd)
- Validation set:

#	User	Entries	Date of Last Entry	Overall ▲	J_seen ▲	J_unseen ▲	F_seen ▲	F_unseen ▲
1	speeding_zZ	62	08/28/18	0.710 (1)	0.725 (2)	0.644 (1)	0.761 (2)	0.711 (1)
2	Jono	8	08/27/18	0.703 (2)	0.744 (1)	0.606 (4)	0.789 (1)	0.675 (3)
3	linhj	26	08/29/18	0.697 (3)	0.723 (3)	0.631 (2)	0.736 (4)	0.698 (2)

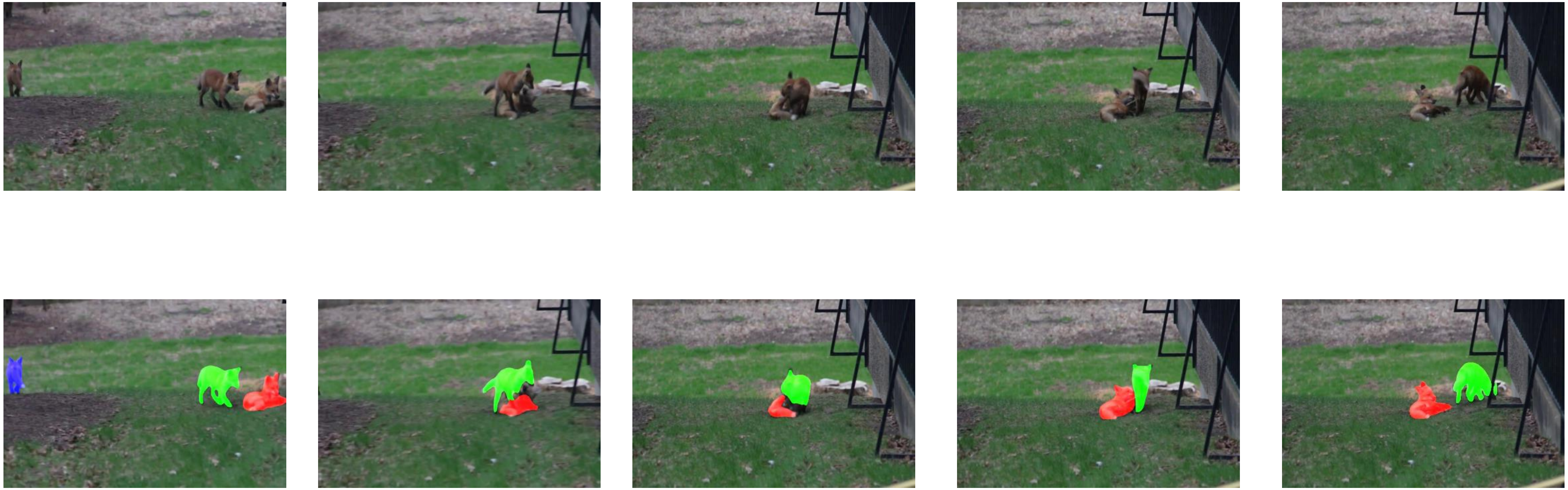
- Test set:

#	User	Entries	Date of Last Entry	Overall ▲	J_seen ▲	J_unseen ▲	F_seen ▲	F_unseen ▲
1	Jono	4	09/01/18	0.722 (1)	0.737 (1)	0.648 (2)	0.778 (1)	0.725 (2)
2	speeding_zZ	8	09/01/18	0.720 (2)	0.725 (3)	0.663 (1)	0.752 (3)	0.741 (1)
3	mikirui	8	09/01/18	0.699 (3)	0.736 (2)	0.621 (4)	0.755 (2)	0.684 (4)

Visual Results

I_x

L_x



Visual Results

I_x

L_x

Future Direction

- Small object
- Learned parameters instead of fixed threshold
- Key multiple previous frames
- Long term understanding for retrieving missing object
- Speedup
- …

Thanks & Questions