Joint Multiview Segmentation and Localization of RGB-D Images using Depth-Induced Silhouette Consistency

Chi Zhang[1], Zhiwei Li[2], Rui Cai[2], Hongyang Chao[1], Yong Rui[2]

*1 Sun Yat-Sen University,
Guangzhou, P.R. China*

*2 Microsoft Research,
Beijing, P.R. China*

Motivation

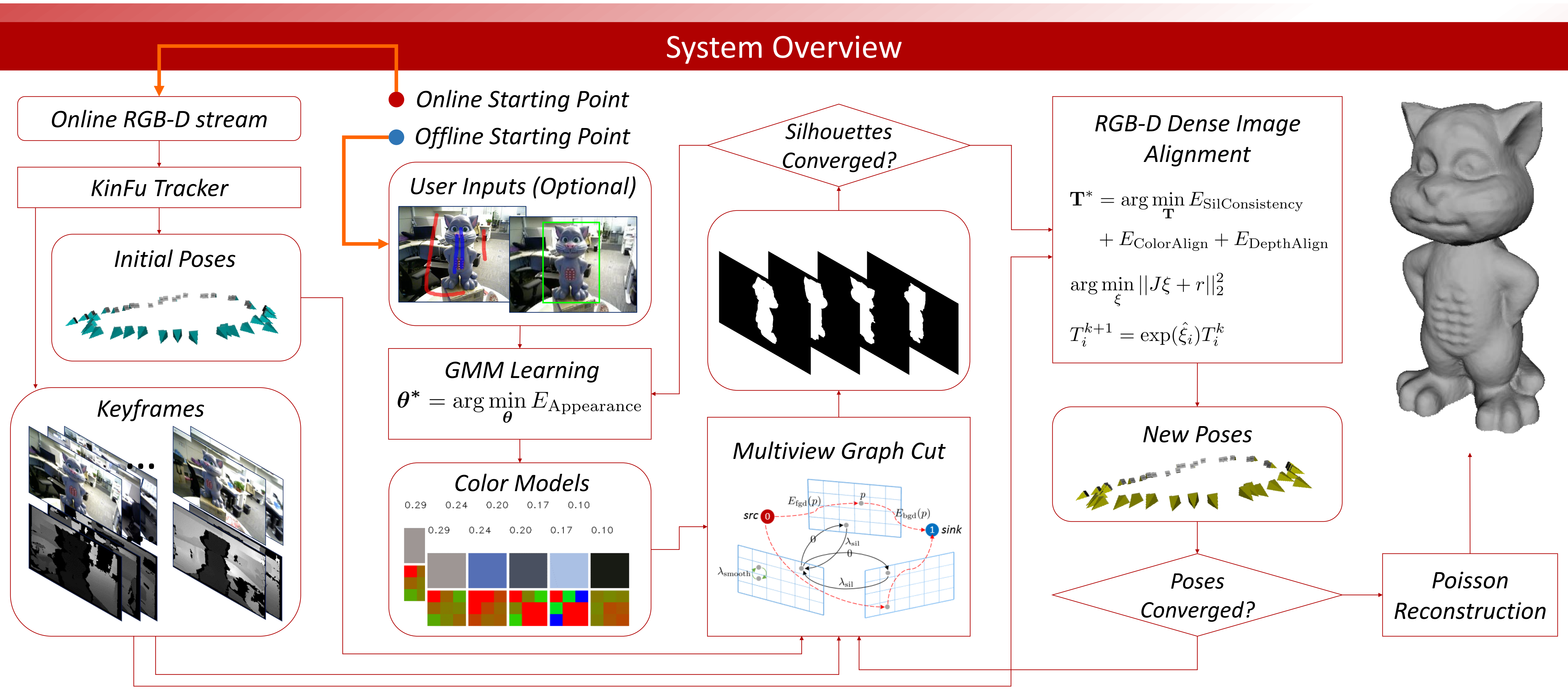
- *Aim at RGB-D SLAM for object scanning.*
    - *Input: An RGB-D stream.*
    - *Output: foreground masks and camera poses of the extracted keyframes.*
- *Use silhouettes to improve pose estimation.*
- *Silhouettes are difficult to obtain in practice. Make it practical by generating silhouettes on-the-fly.*

Variables

For each view i, we optimize:

$\mathbf{S} \equiv \{S_i\}_{i=1}^N$	*A binary foreground mask.*
$\boldsymbol{\theta} \equiv \{\theta^{\text{fgd}}, \theta^{\text{bgd}}\}_{i=1}^N$	*Two GMM color models.*
$\mathbf{T} \equiv \{T_i\}_{i=1}^N$	*A local-to-world camera pose.*

System Overview



$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} E_{\text{Appearance}}$$

$$\mathbf{T}^* = \arg\min_{\mathbf{T}} E_{\text{SilConsistency}} + E_{\text{ColorAlign}} + E_{\text{DepthAlign}}$$

$$\arg\min_{\xi} ||J\xi + r||_2^2$$

$$T_i^{k+1} = \exp(\hat{\xi}_i)T_i^k$$

Formulation

Five Sub-energies:

Detailed definitions:

$$E_{\text{All}}(\mathbf{S}, \boldsymbol{\theta}, \mathbf{T}) = E_{\text{Appearence}}(\mathbf{S}, \boldsymbol{\theta}) \longrightarrow \sum_i \sum_{p\in\Omega_i} -\text{Prob}(I_i(p) \mid S_i(p), \theta_i^{\text{bgd}}, \theta_i^{\text{fgd}})$$

$$+ E_{\text{MaskSmooth}}(\mathbf{S}) \longrightarrow \sum_i \sum_{p,r\in\mathcal{N}_4} w_{pr}||S_i(p) - S_i(r)||^2$$

$$+ E_{\text{SilConsitency}}(\mathbf{S}, \mathbf{T}) \longrightarrow \sum_i \sum_{p\in\tilde{\Omega}_i} \sum_{j\neq i} S_i(p) \cdot ||S_i(p) - S_j(q)||^2$$

$$+ E_{\text{ColorAlign}}(\mathbf{T}) \longrightarrow \sum_i \sum_{p\in\tilde{\Omega}_i} \sum_{j\in\mathcal{N}_i} ||I_i(p) - I_j(q)||^2$$

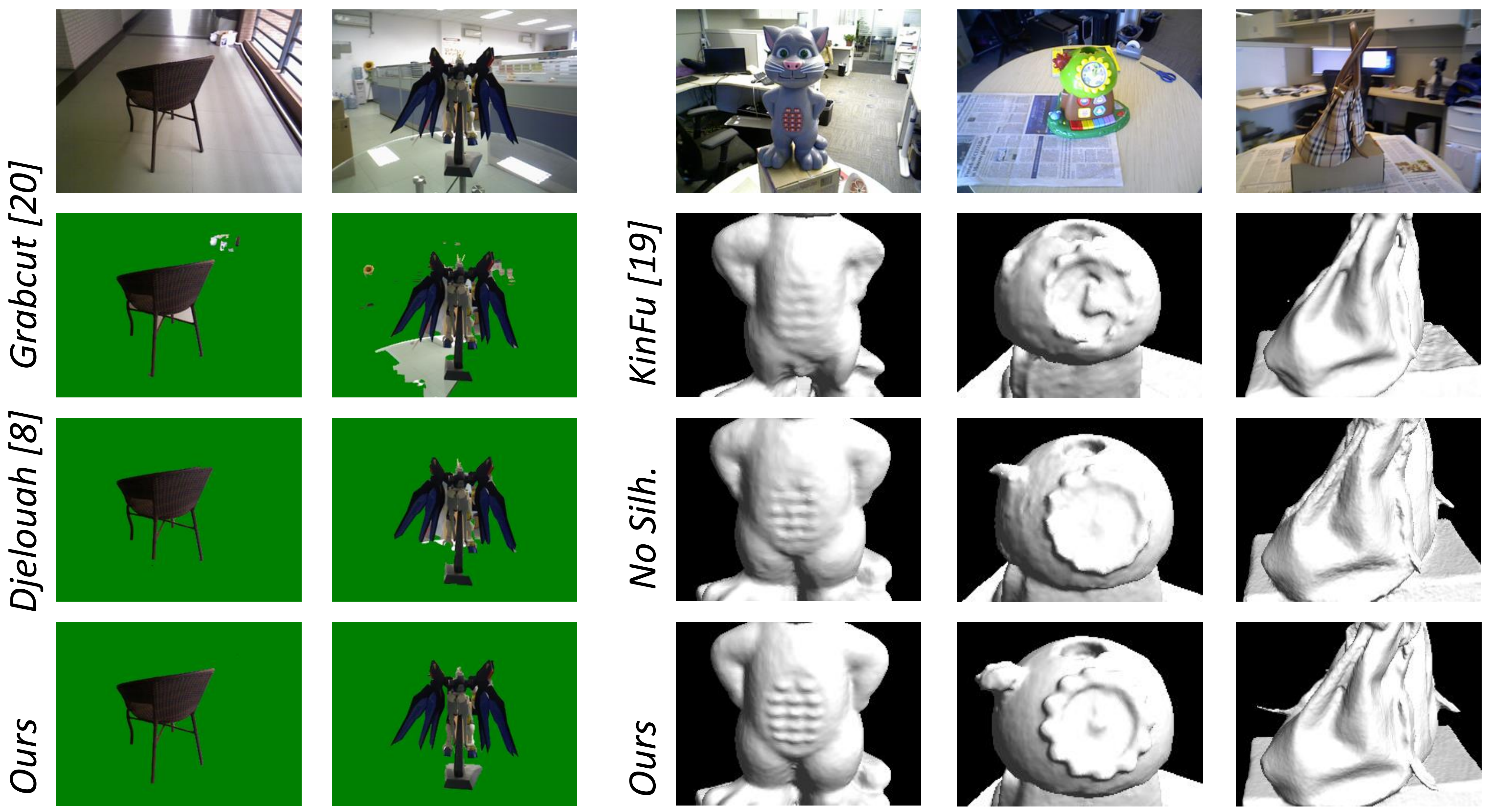
$$+ E_{\text{DepthAlign}}(\mathbf{T}) \longrightarrow \sum_i \sum_{p\in\tilde{\Omega}_i} \sum_{j\in\mathcal{N}_i} ||D_i(p) - D_j(q)||^2$$

- ***Optimize w.r.t.*** $\mathbf{T}$ ***by Gauss-Newton Method.***
- ***Optimize w.r.t.*** $\mathbf{S}, \boldsymbol{\theta}$ ***by Multiview Graph Cut.***

$$q = \pi_j(T_j^{-1}T_i\pi_i^{-1}(p, D_i(p)))$$

$\tilde{\Omega}_i$: *Pixels with depths in view i*

Results


Grabcut [20]
Djelouah [8]
Ours
KinFu [19]
No Silh.
Ours