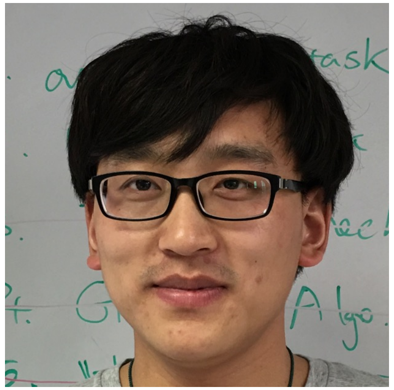
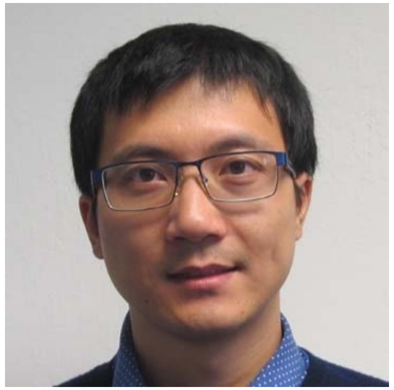
OPPO

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

RIAV-MVS: Recurrent-Indexing an Asymmetric Volume for Multi-View Stereo

Changjiang Cai | Pan Ji | Qingan Yan | Yi Xu

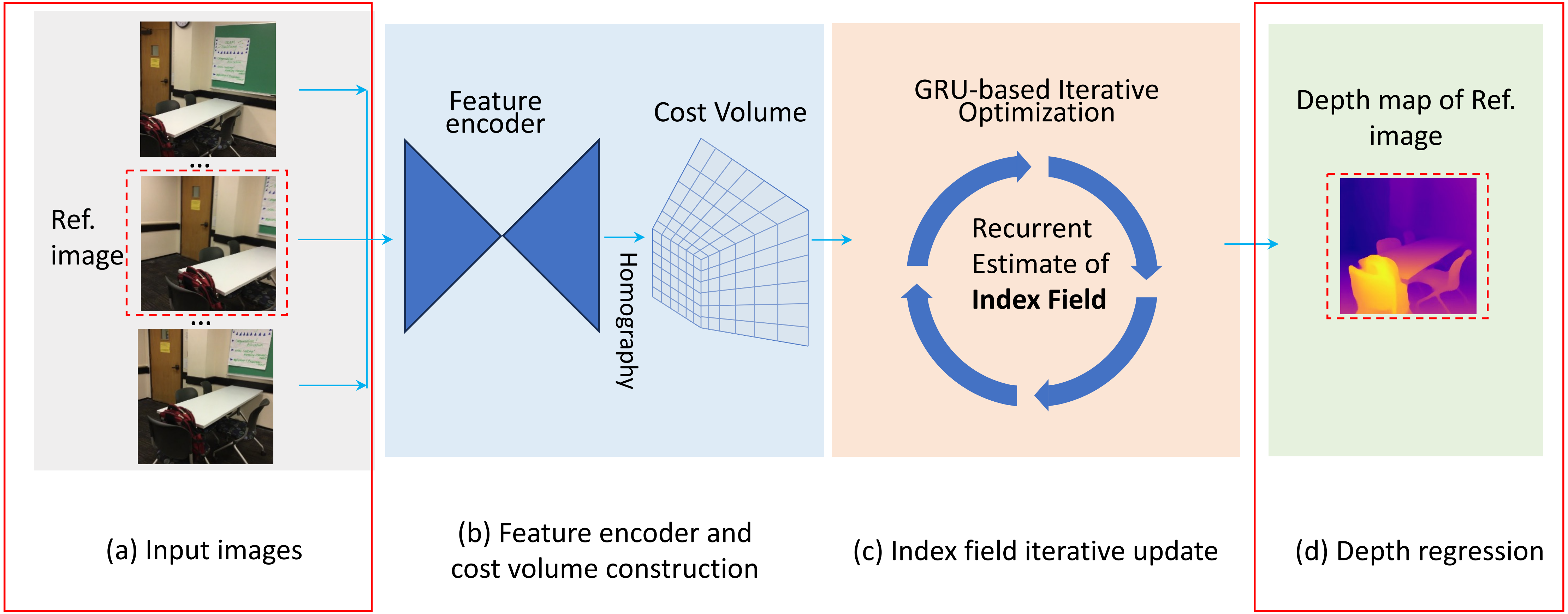
OPPO US Research Center, InnoPeak Technology, Inc.

Poster Tag: TUE-AM-087

github.com/oppo-us-research/riav-mvs

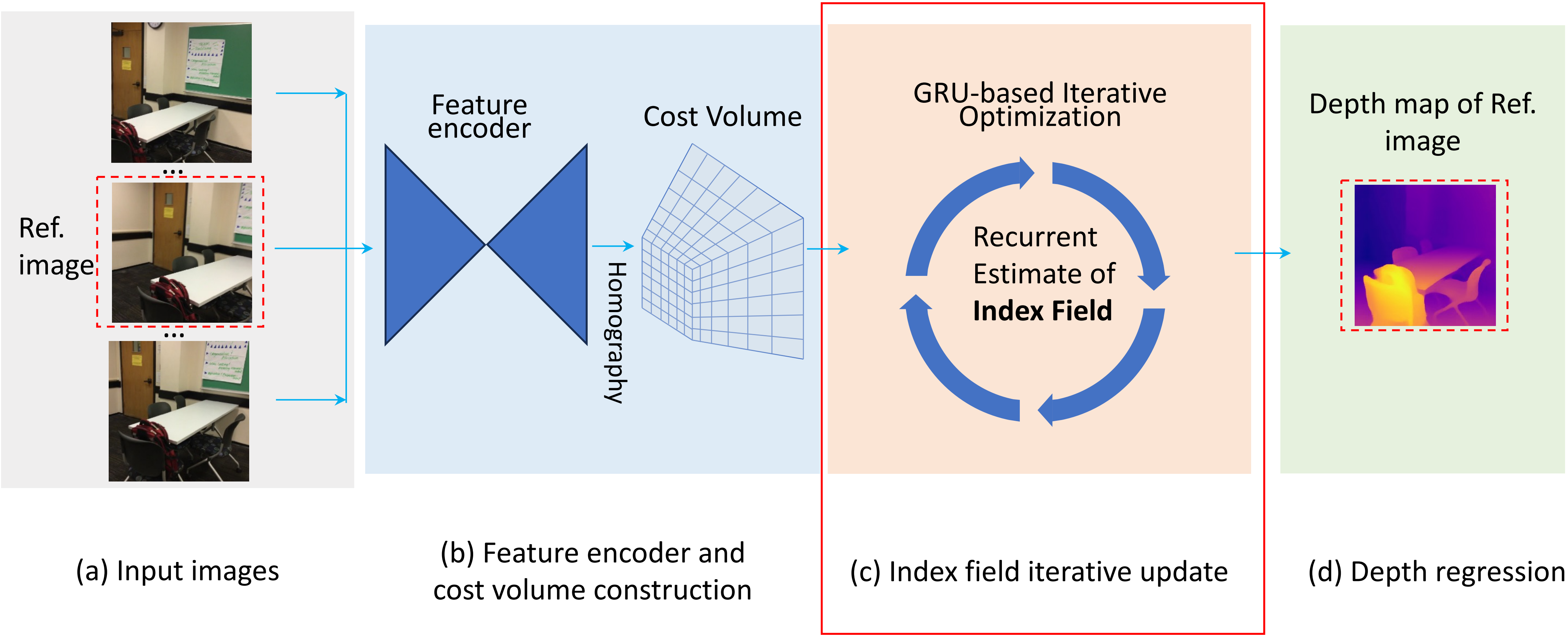
Overview

- Our core idea is a “learning-to-optimize” paradigm that iteratively indexes a plane-sweeping cost volume and regresses the depth map via a convolutional GRU.



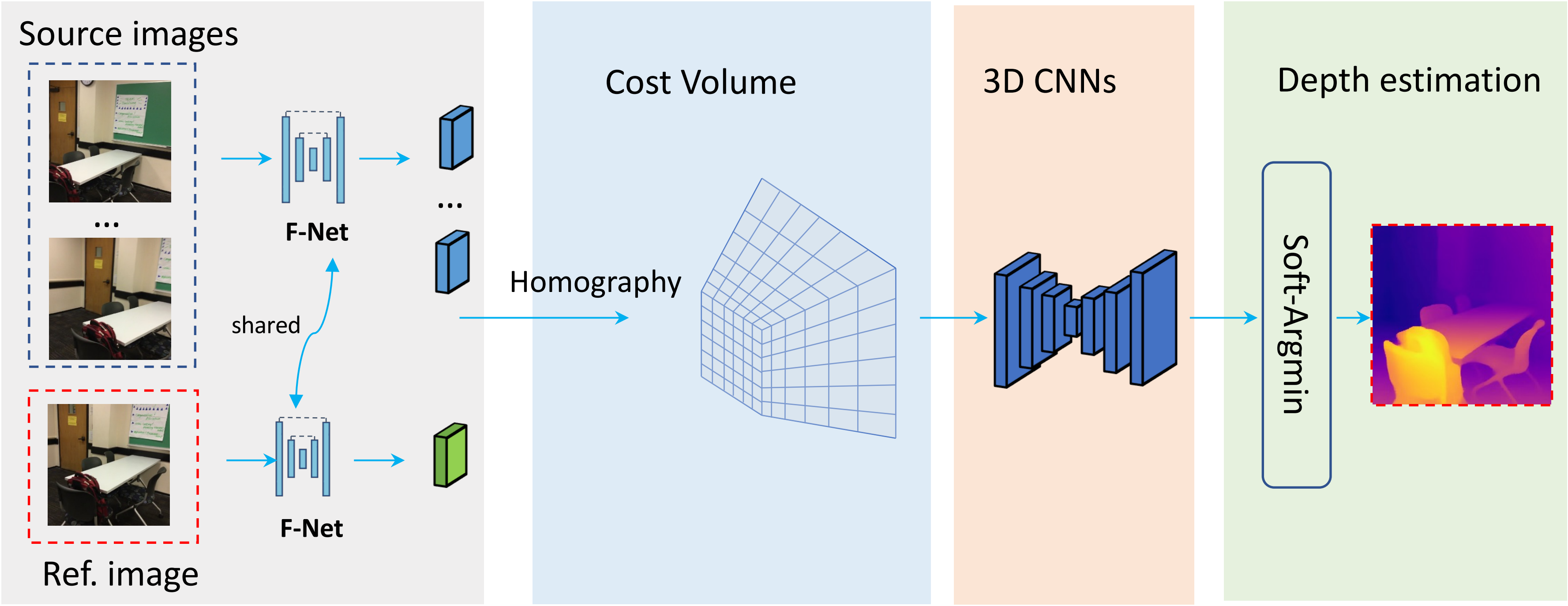
Overview

- Our core idea is a "learning-to-optimize" paradigm that iteratively indexes a plane-sweeping cost volume and regresses the depth map via a convolutional GRU.



Motivation

- Existing CNN-based MVS methods:
  - Concerns in (a), (b), (c) and (d)



(a) Feature extraction

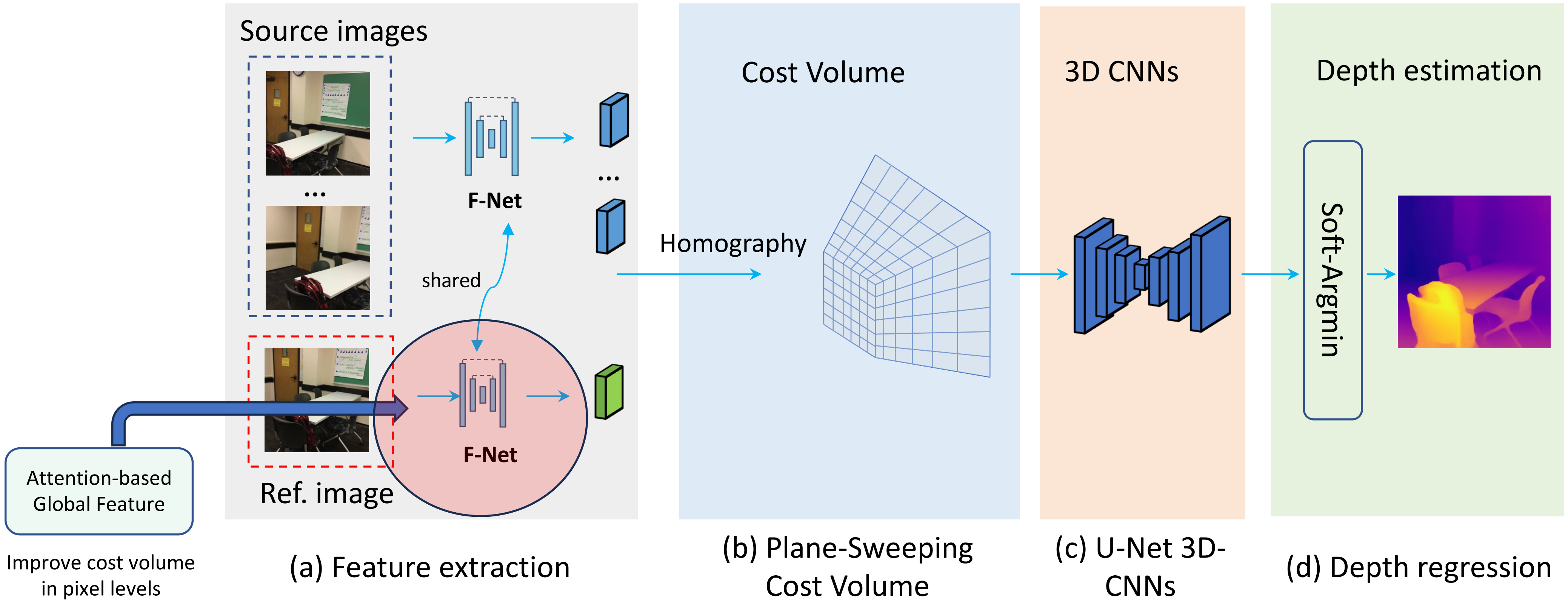
(b) Plane-Sweeping Cost Volume

(c) U-Net 3D-CNNs

(d) Depth regression

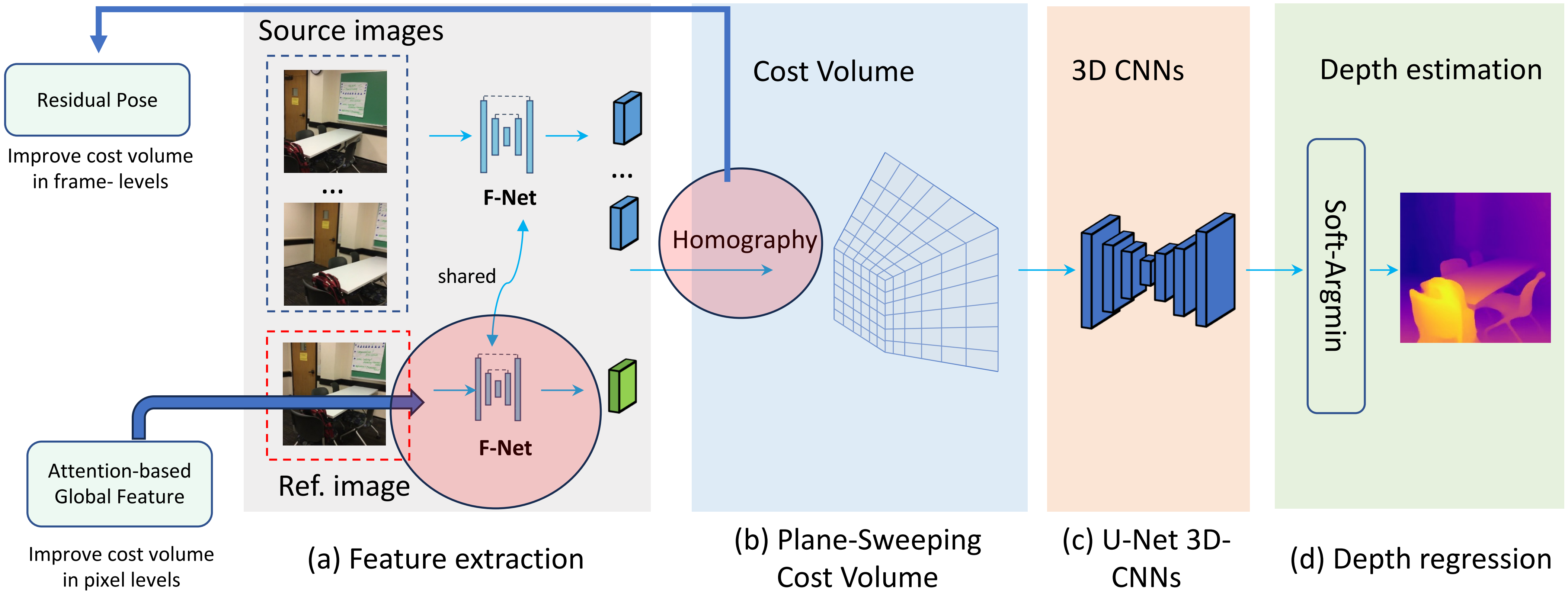
Contributions

- Our contributions: 1) An **asymmetric** cost volume ★★★☆☆



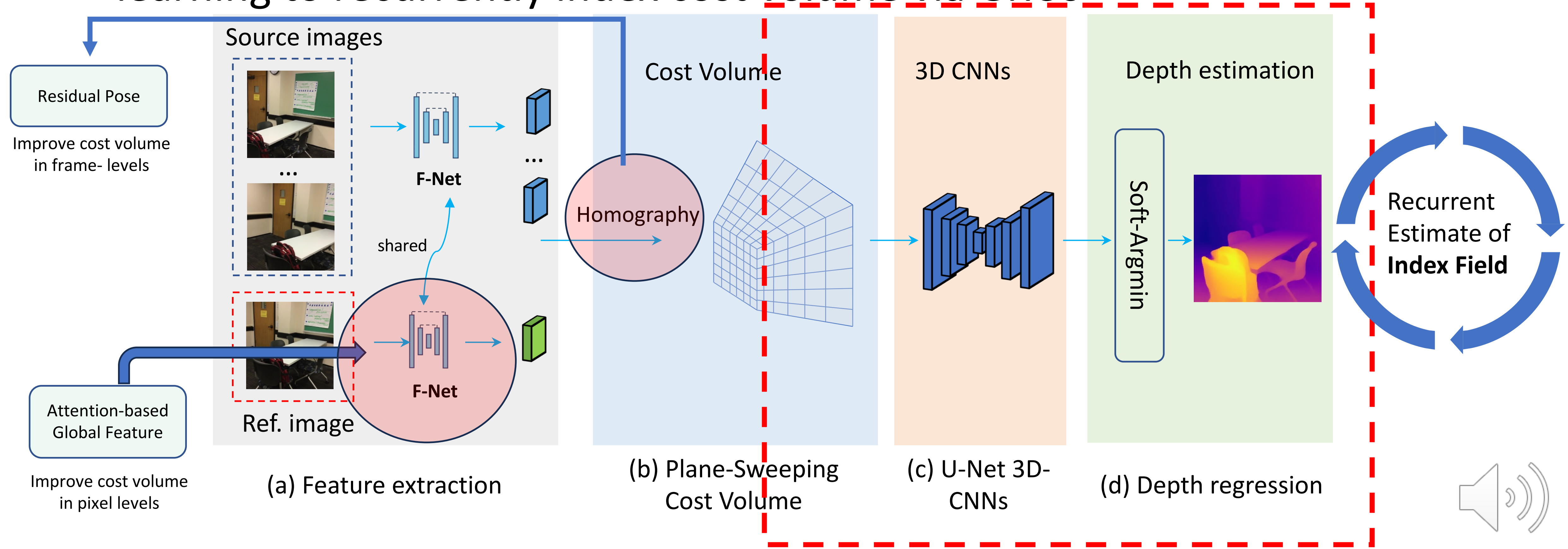
Contributions

- Our contributions: 2) **Residual** pose update ★★★★☆



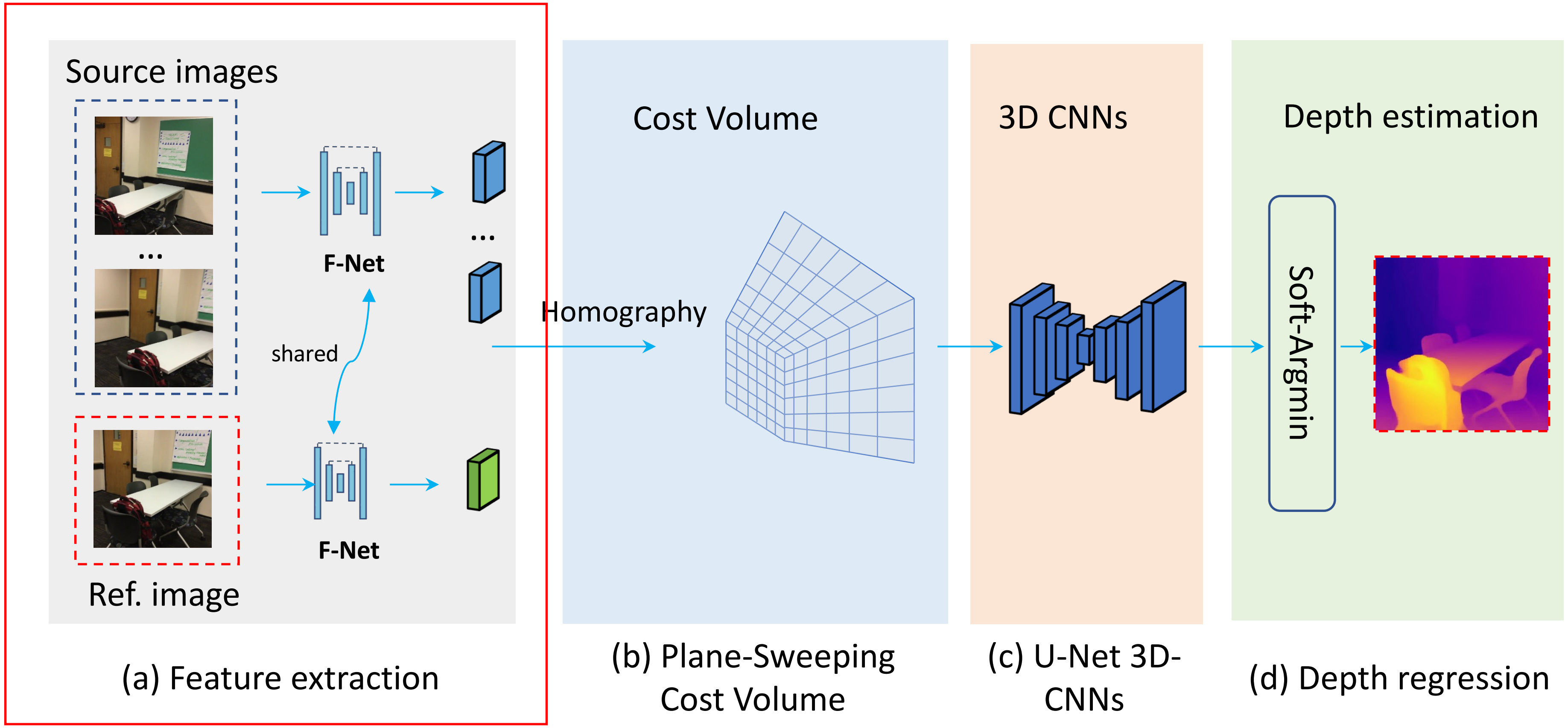
Contributions

- Our contributions: 3) A **new paradigm** to predict the depth by ★★★★★ learning to recurrently index cost volume via GRUs



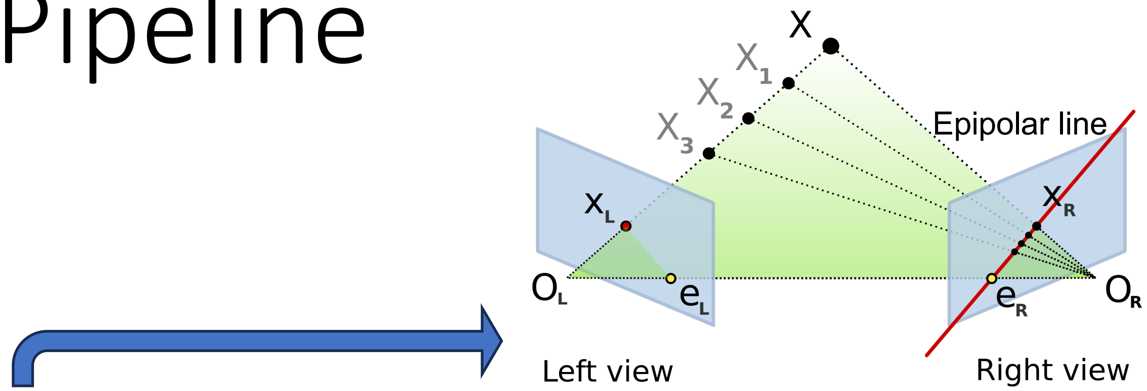
Existing CNN-based MVS Pipeline

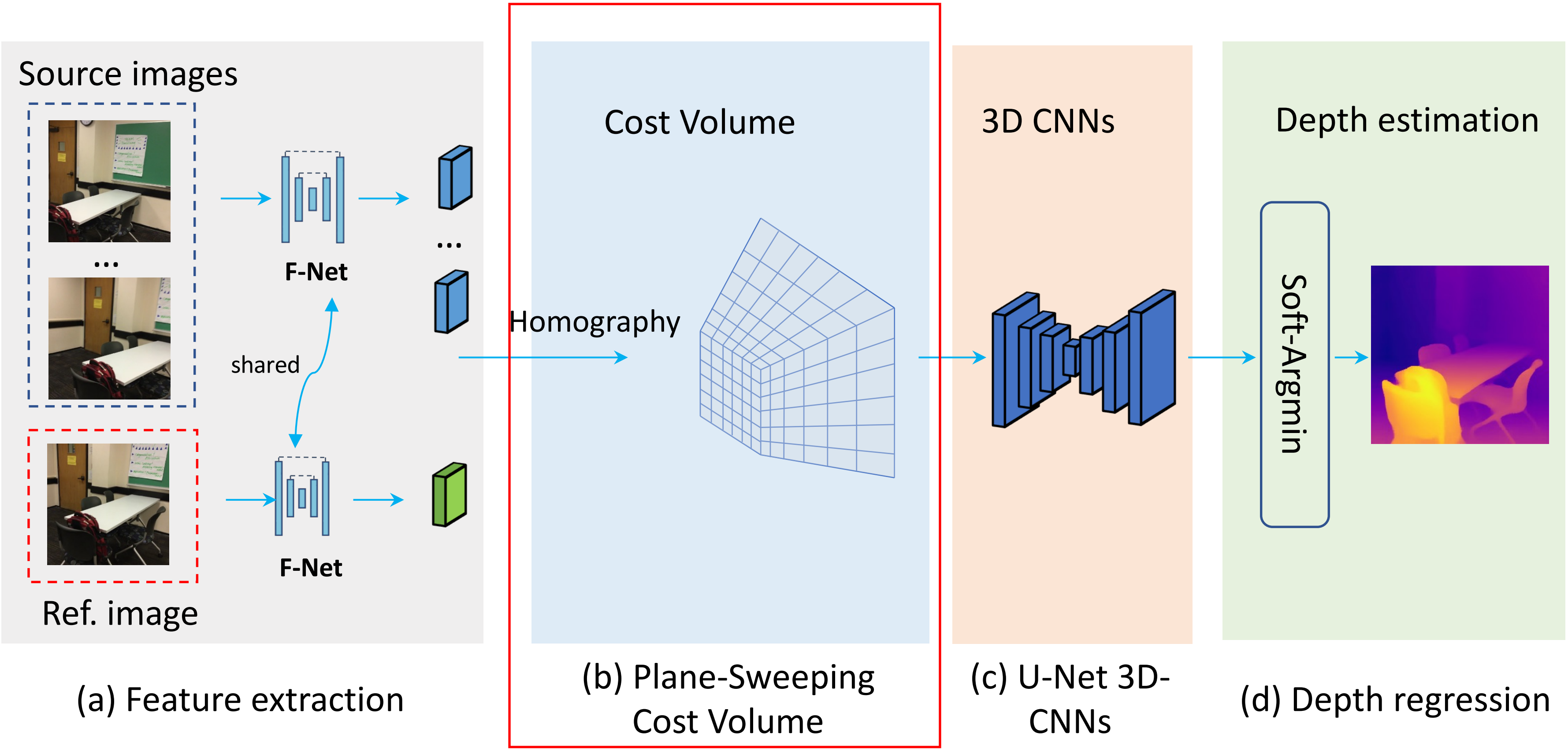
- Existing CNN-based MVS methods:
  - Symmetric features, local context



Existing CNN-based MVS Pipeline

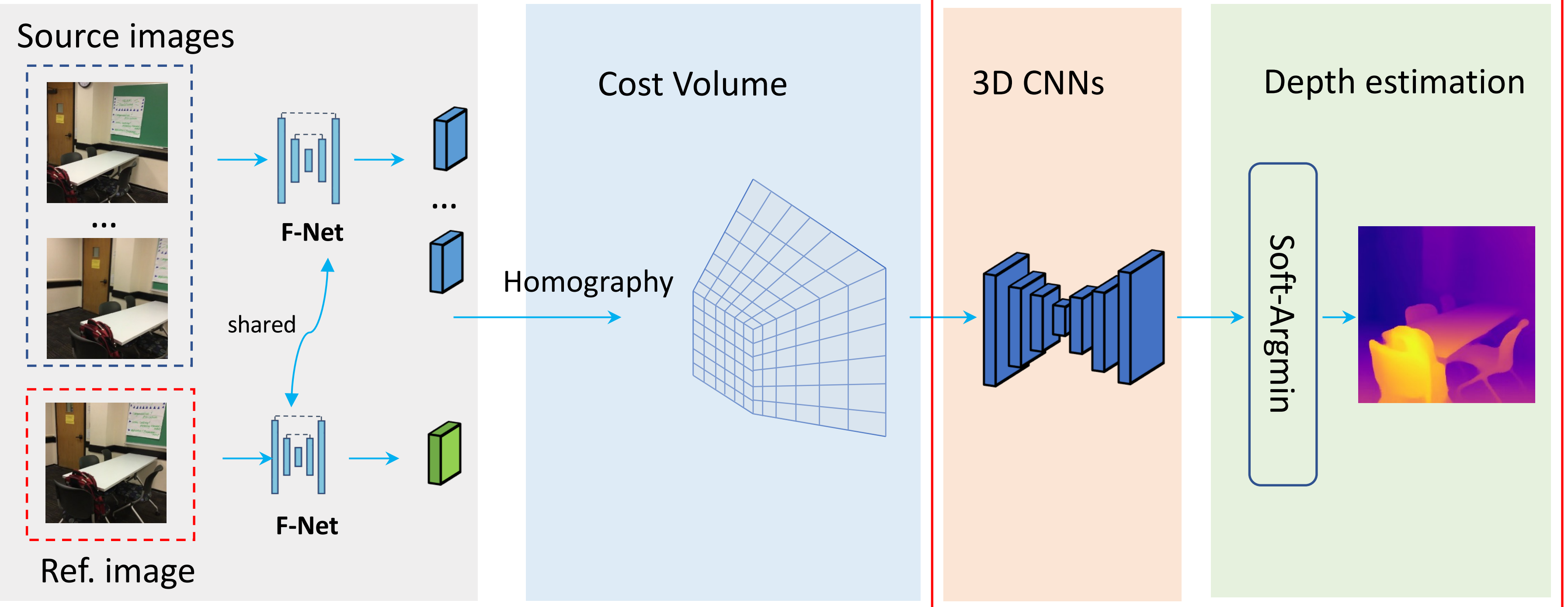
- Existing CNN-based MVS methods:
  - assuming poses being accurate





Existing CNN-based MVS Pipeline

- Existing CNN-based MVS methods:
  - 3D CNNs are time and memory consuming
  - Soft-argmin is not robust to multi-modal distributions



(a) Feature extraction

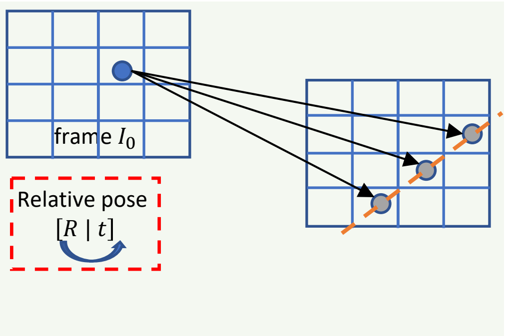
(b) Plane-Sweeping Cost Volume

(c) U-Net 3D-CNNs

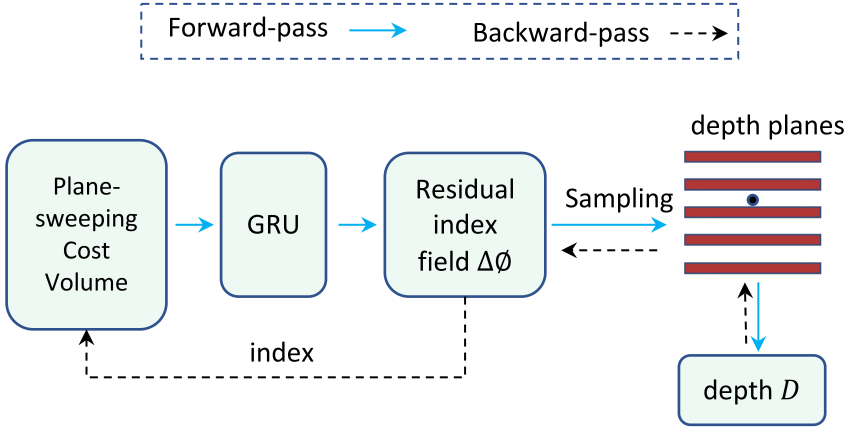
(d) Depth regression

Our Approach

- Constructing a good cost volume:



(a) Reference feature and source feature for matching along Epipolar line

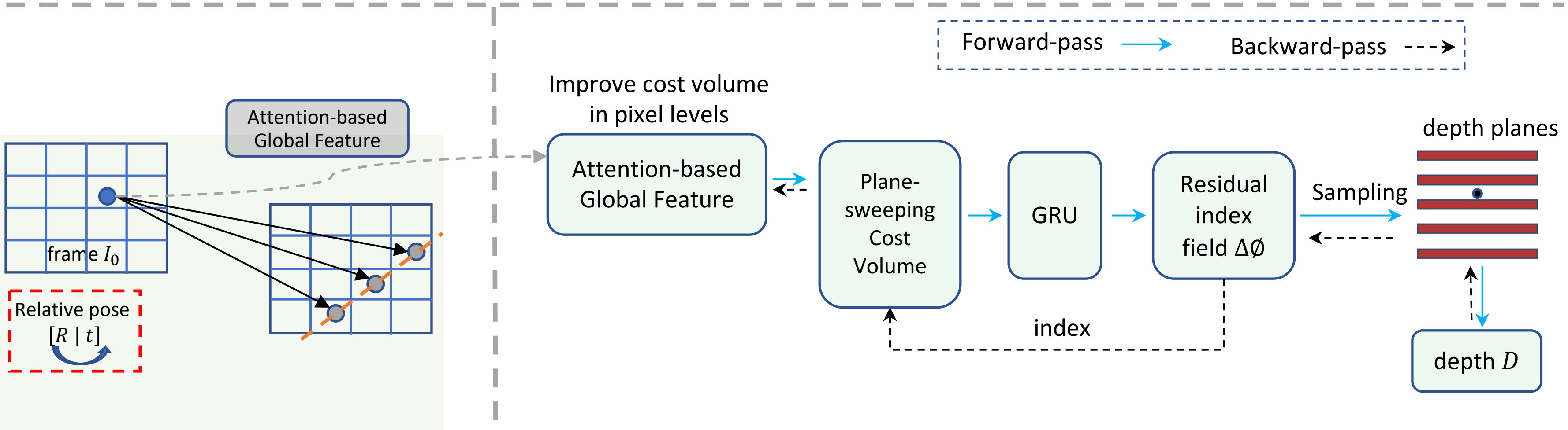


(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

Our Approach

- Constructing a good cost volume:

  1) To break the symmetry of the Siamese network by introducing a transformer block to the reference image (but not to the source images)



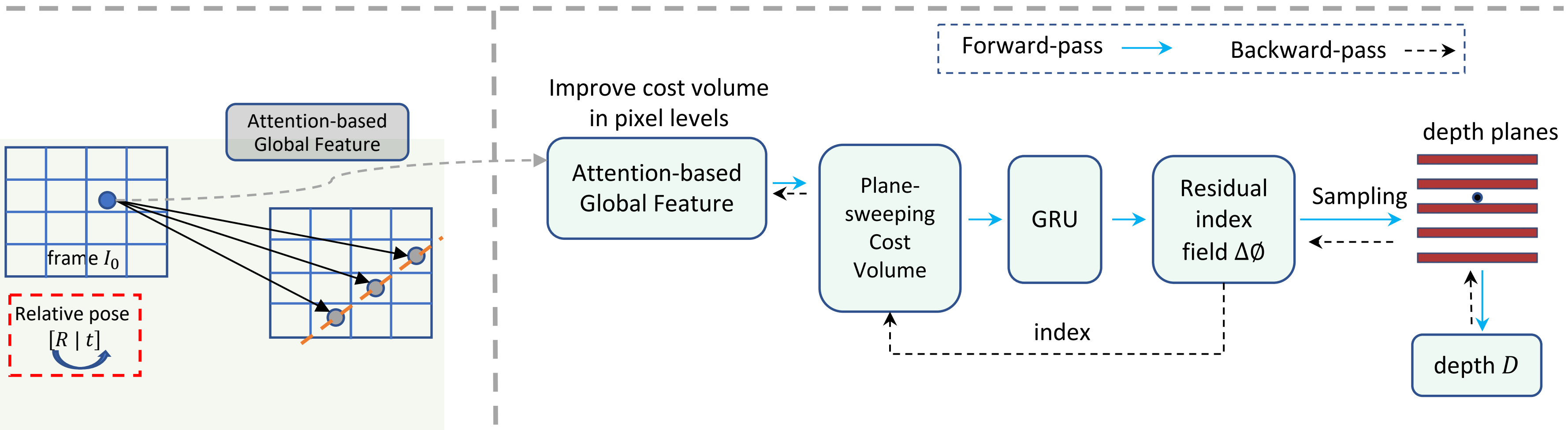
(a) Reference feature and source feature for matching along Epipolar line

(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

Our Approach

- Constructing a good cost volume:

  1) To break the symmetry of the Siamese network by introducing a transformer block to the reference image (but not to the source images)

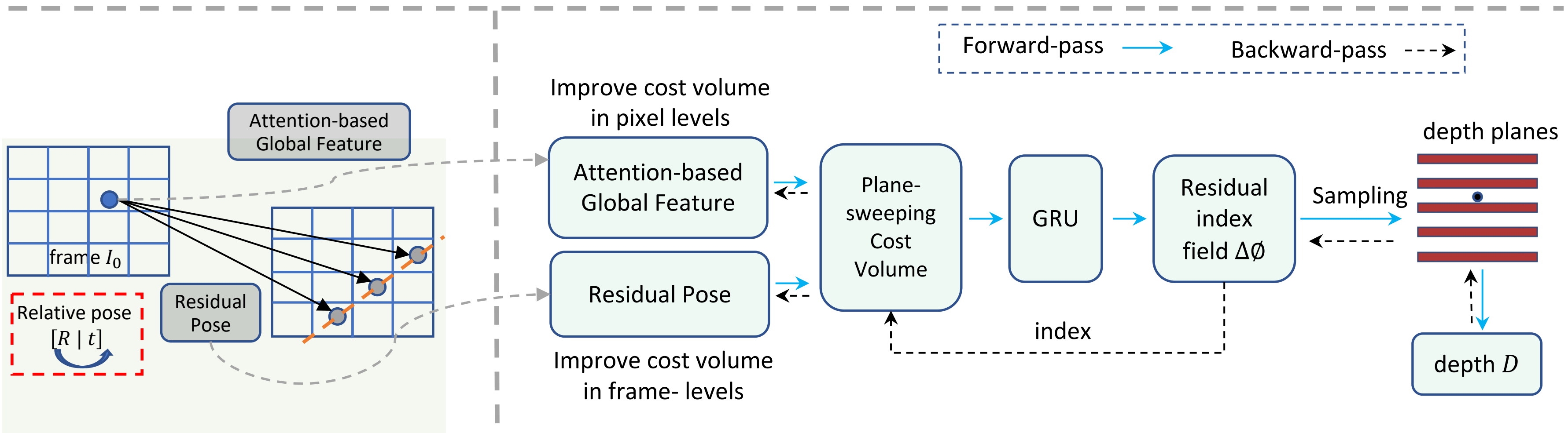


(a) Reference feature and source feature for matching along Epipolar line

(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

Our Approach

- Constructing a good cost volume:
  2) To incorporate a residual pose network to correct the relative poses

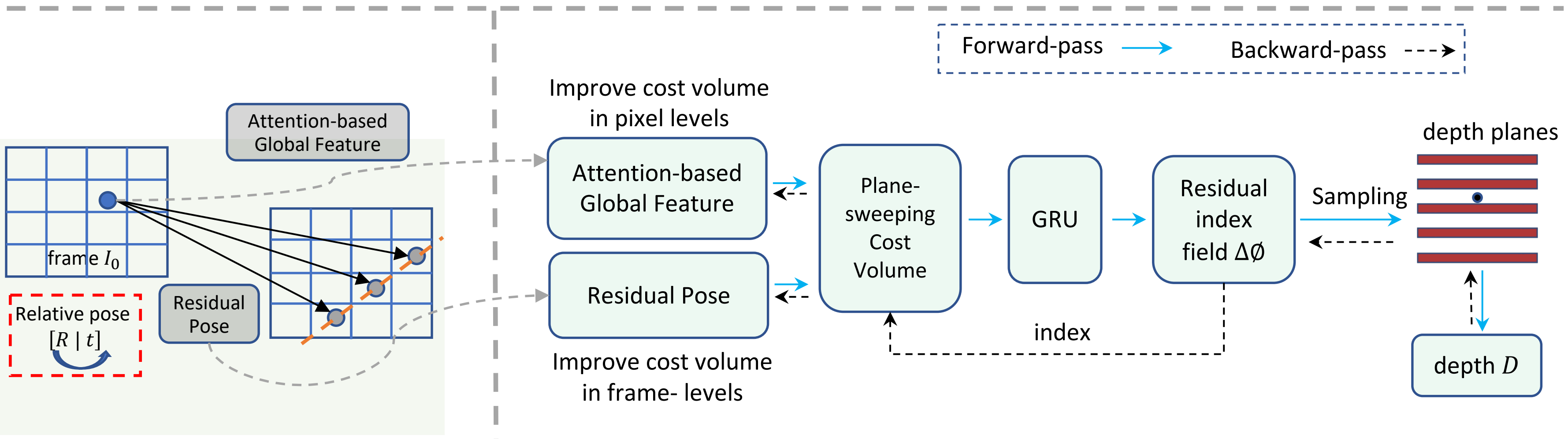


(a) Reference feature and source feature for matching along Epipolar line

(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

Our Approach

- Constructing a good cost volume:
  2) To incorporate a residual pose network to correct the relative poses

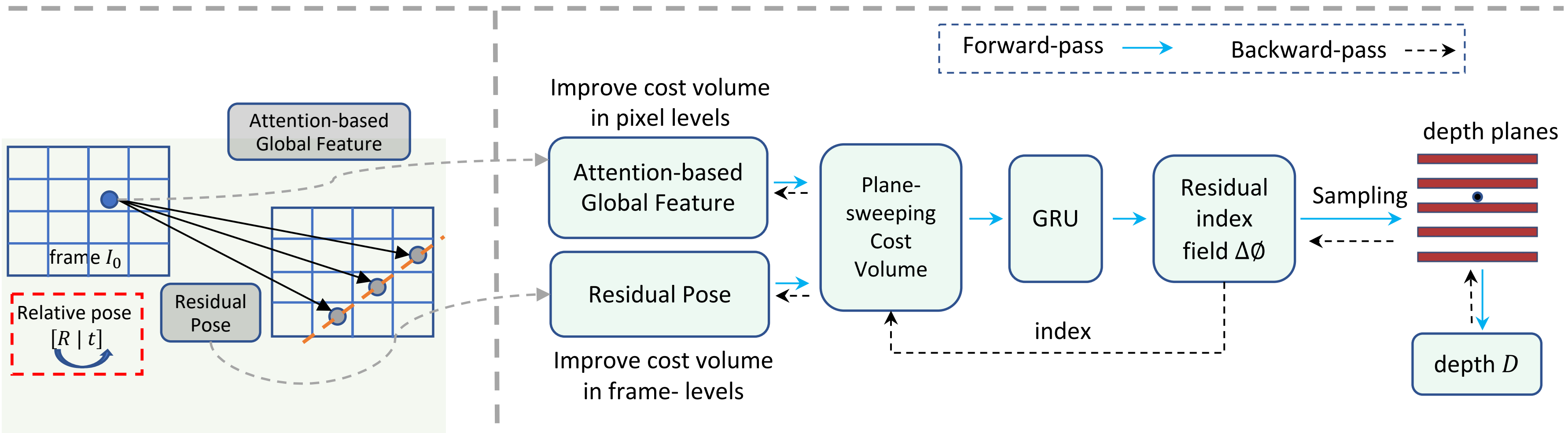


(a) Reference feature and source feature for matching along Epipolar line

(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

Our Approach

- A new paradigm to predict the depth via learning the proposed **index filed** to recurrently index an asymmetric plane-sweeping cost volume via GRUs

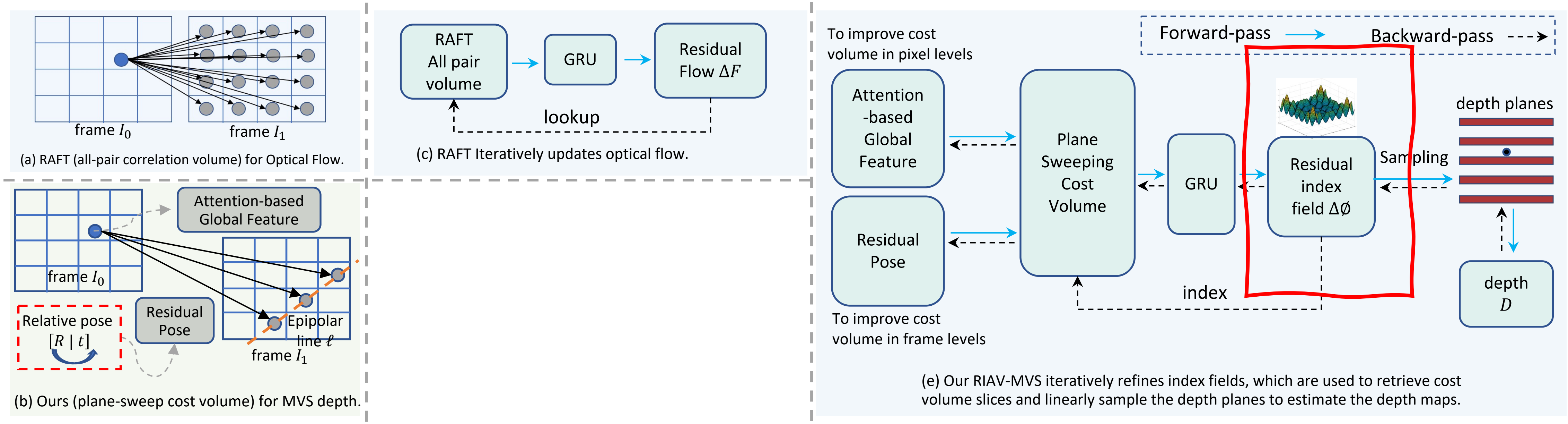


(a) Reference feature and source feature for matching along Epipolar line

(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

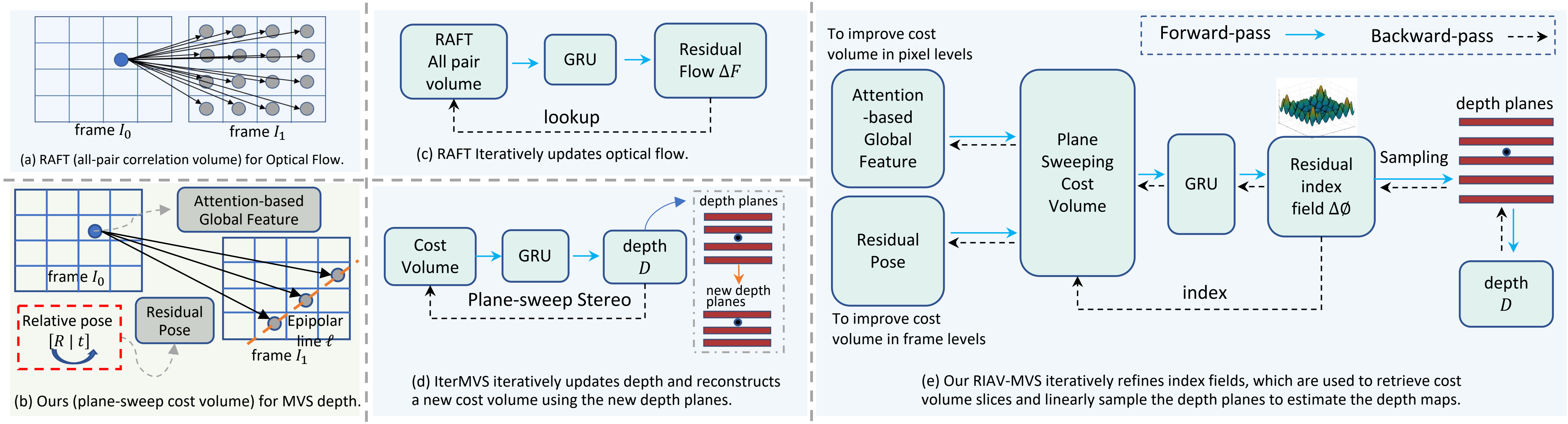
Background - Ours vs RAFT (ECCV'20)

- We borrowed ideas from RAFT for learning to optimize via GRU:
  - RAFT's all-pair correlation for optical flow: NO multi-view geometry constraints → Ours use plane-sweeping cost volume for MVS (Fig. a,b&c)
  - We propose ***index field*** that serves as a new design to bridge cost volume optimization and depth map estimation (Fig. e)



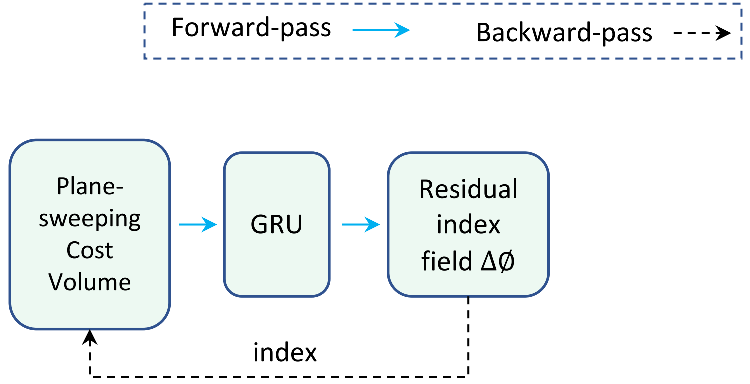
Background - Ours vs IterMVS (CVPR'22)

- IterMVS iteratively predicts a depth and reconstructs a new plane-sweeping cost volume using the updated depth planes (Fig. d)
- Ours learns to index the cost volume by approaching the "correct" depth planes per pixel via an index field (Fig. e)



Our RIAV-MVS

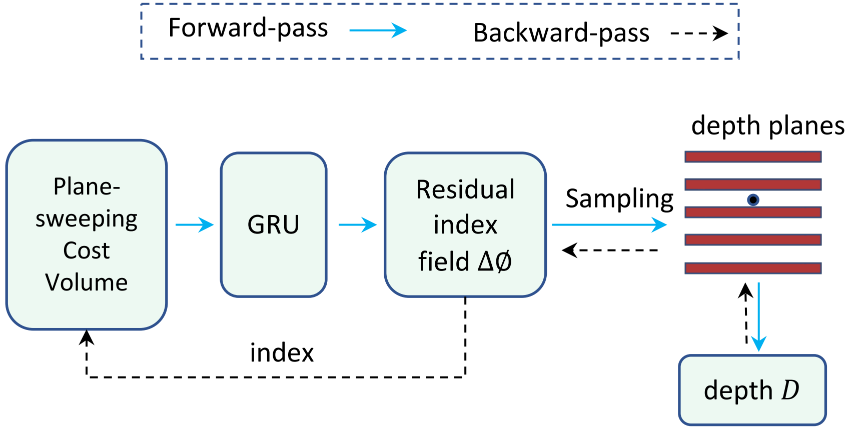
- Our proposed RIAV-Net iteratively refines index fields and retrieve plane-sweeping cost volume slices based on new index fields.
  - a **residual** index field ΔØ is predicted as an update direction for next iteration



(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

Our RIAV-MVS

- Our proposed RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.
  - a depth map is estimated by sampling depth plane hypotheses via index field

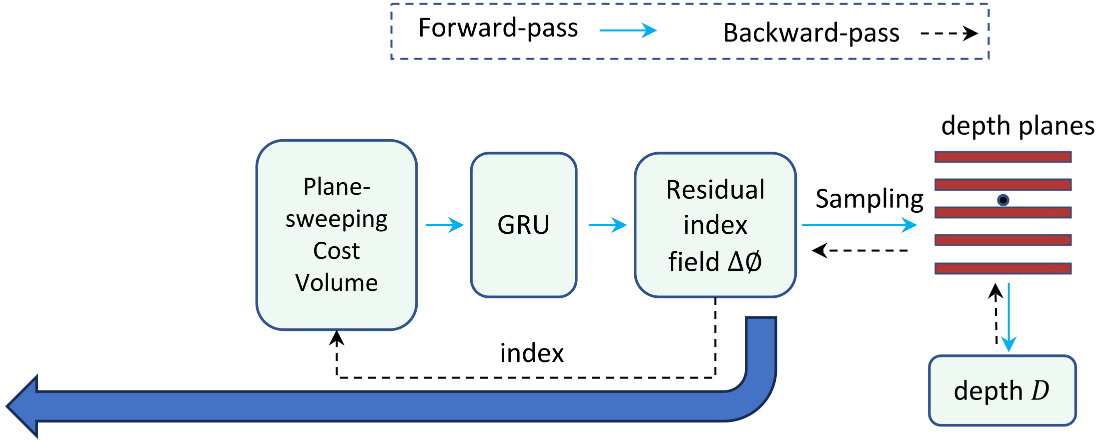


(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

Our RIAV-MVS

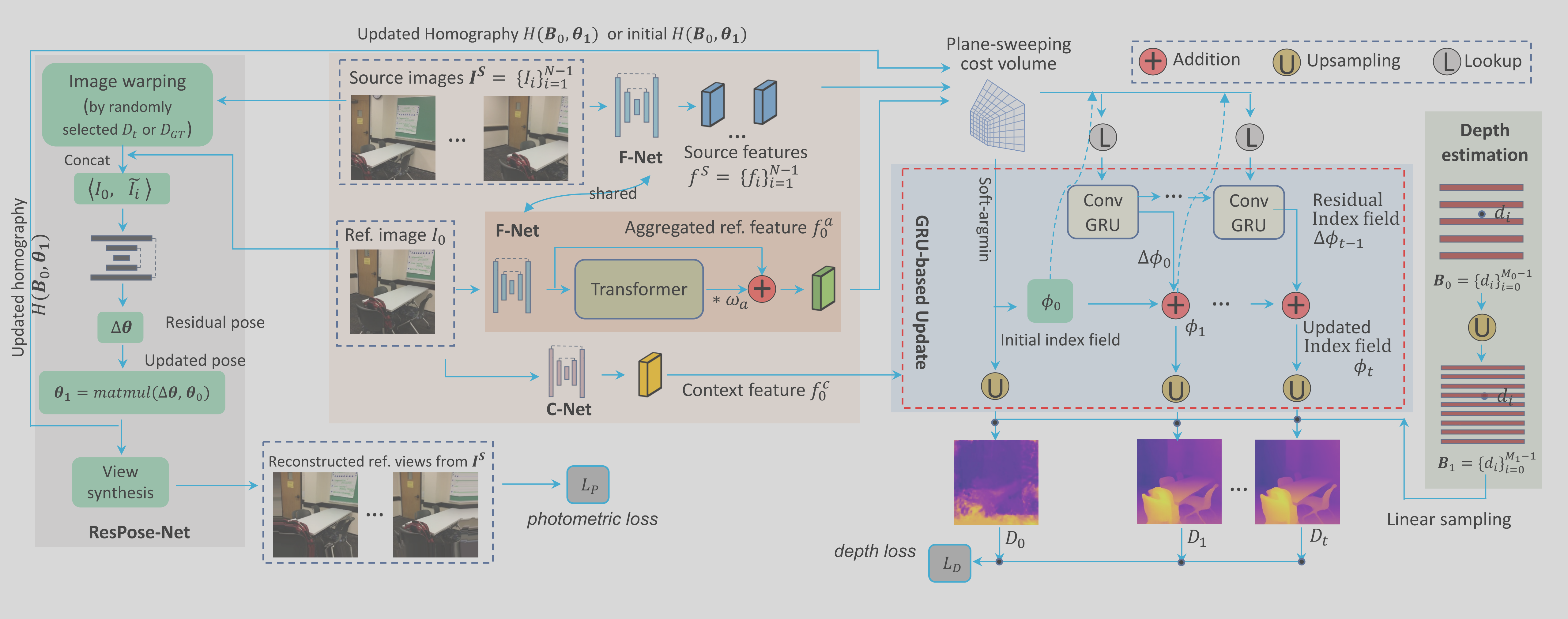
- Our proposed RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.
  - a depth map is estimated by sampling depth plane hypotheses via **index field**

Our ***index field*** serves as a new design to bridge cost volume optimization and depth map estimation

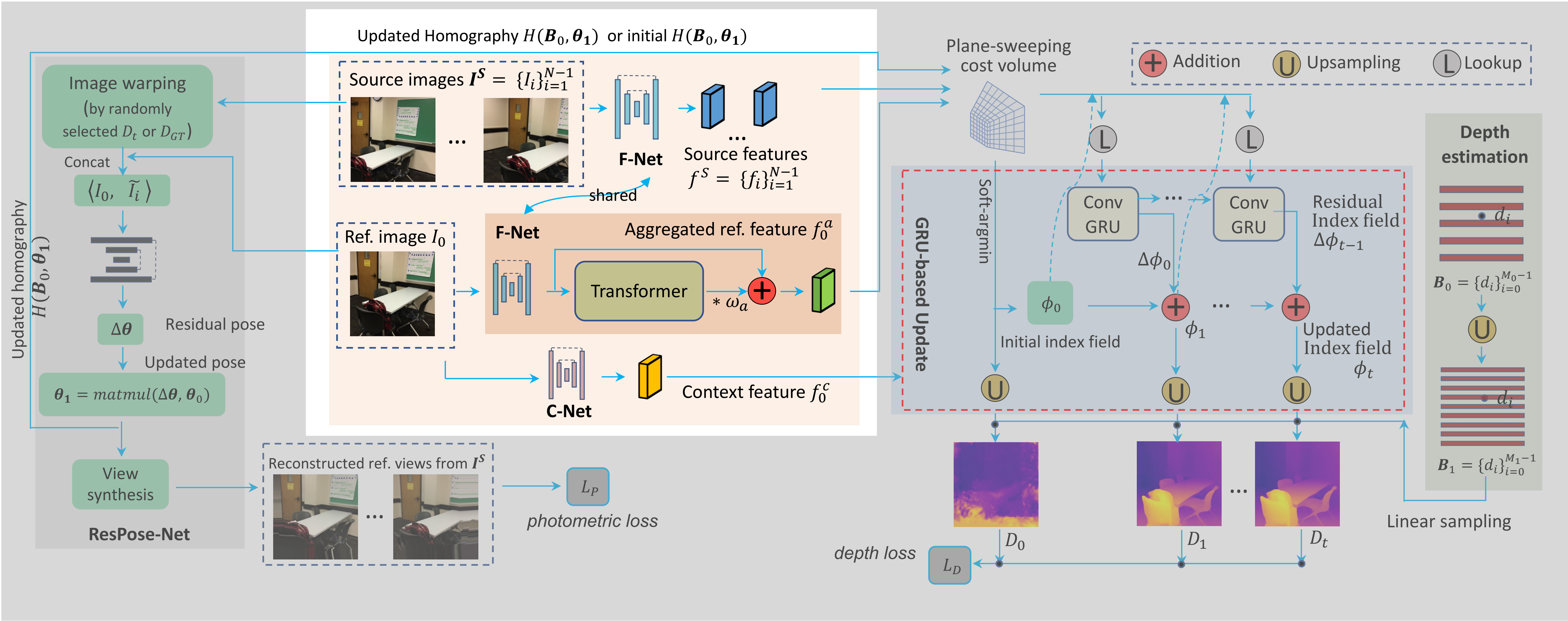


(b) Our RIAV-Net iteratively refines index fields and retrieve cost volume slices based on new index fields.

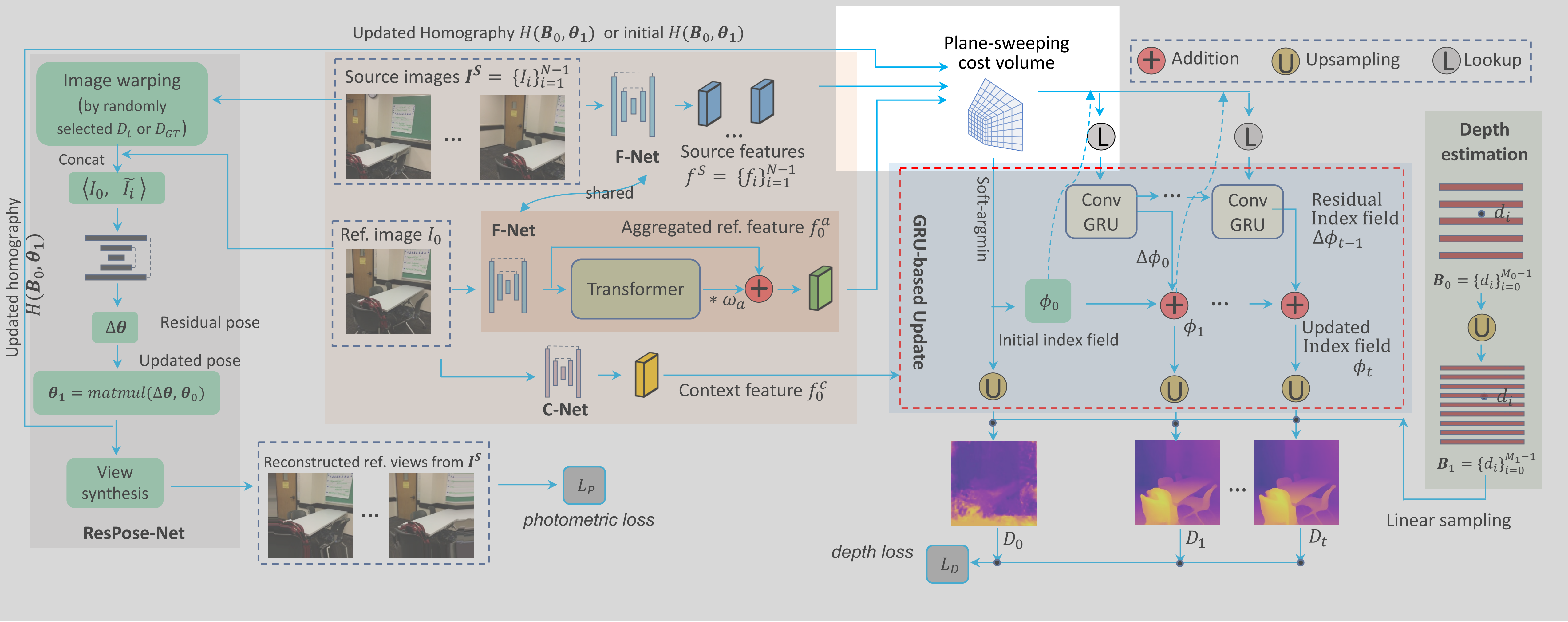
Architecture



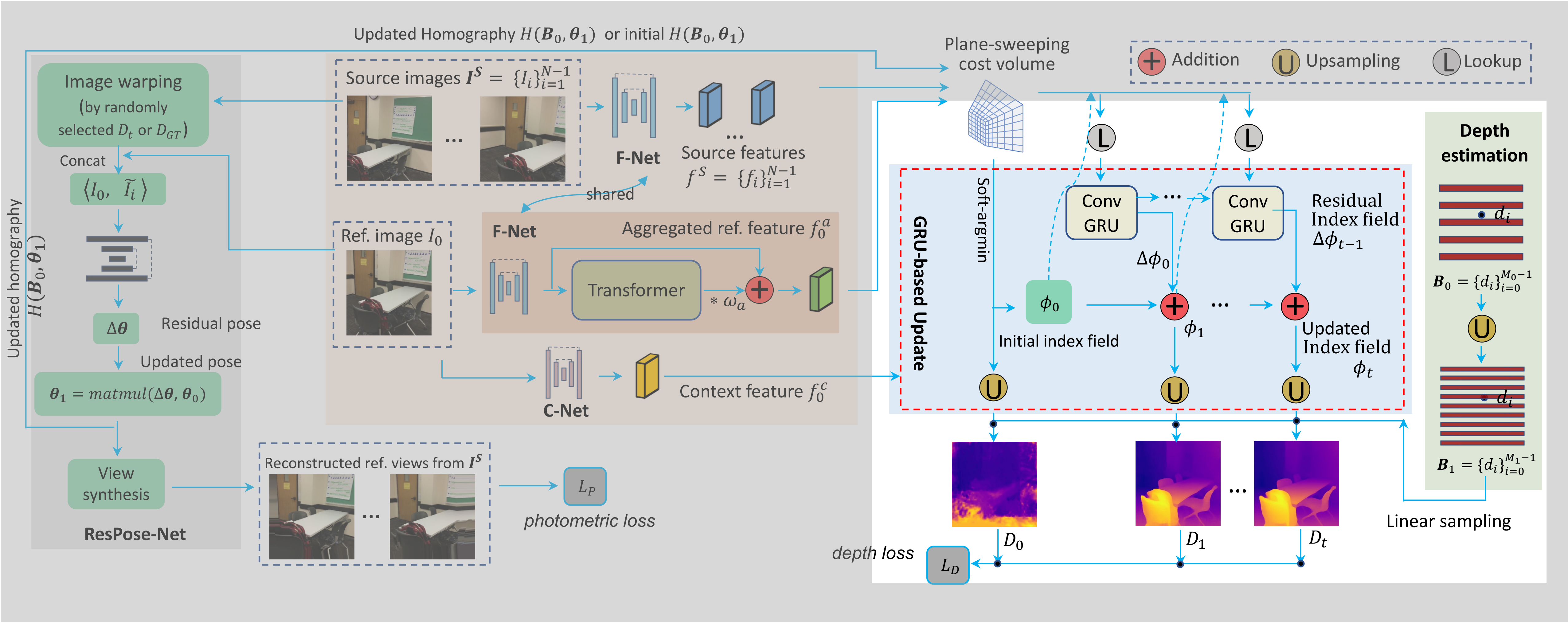
Architecture



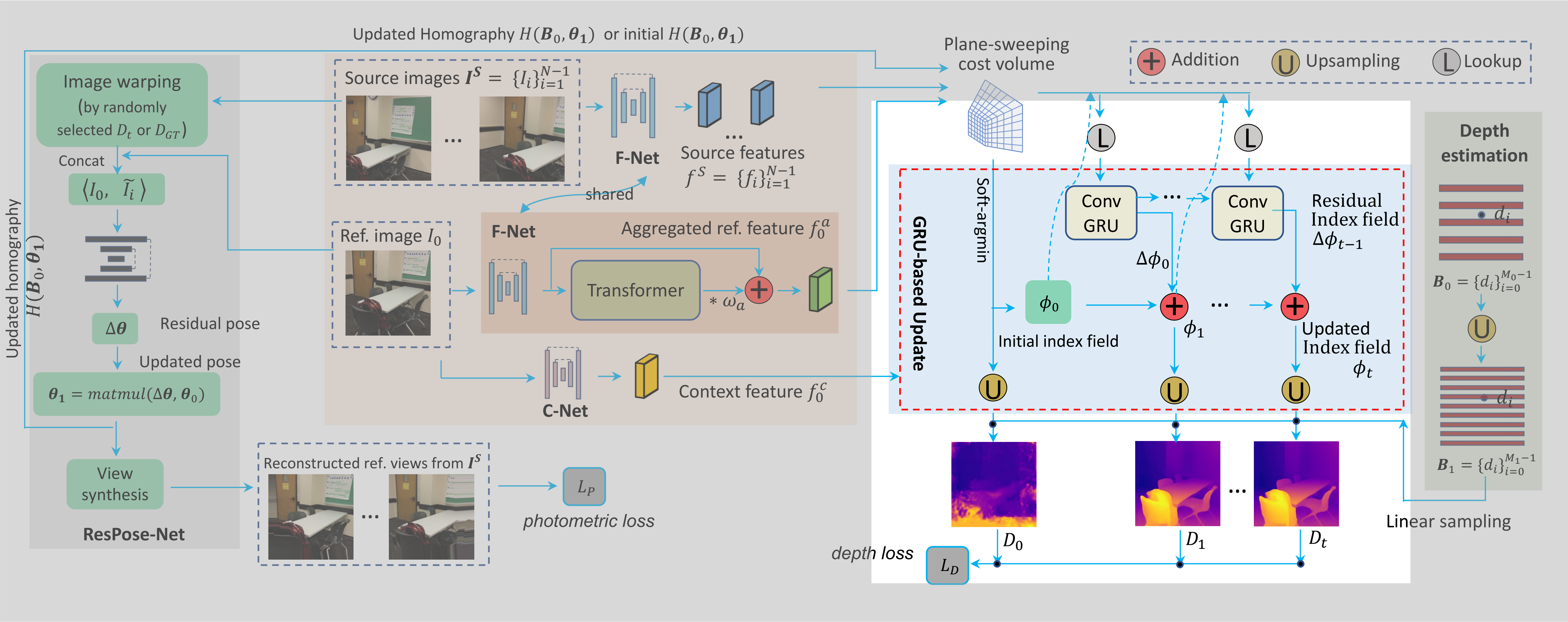
Architecture



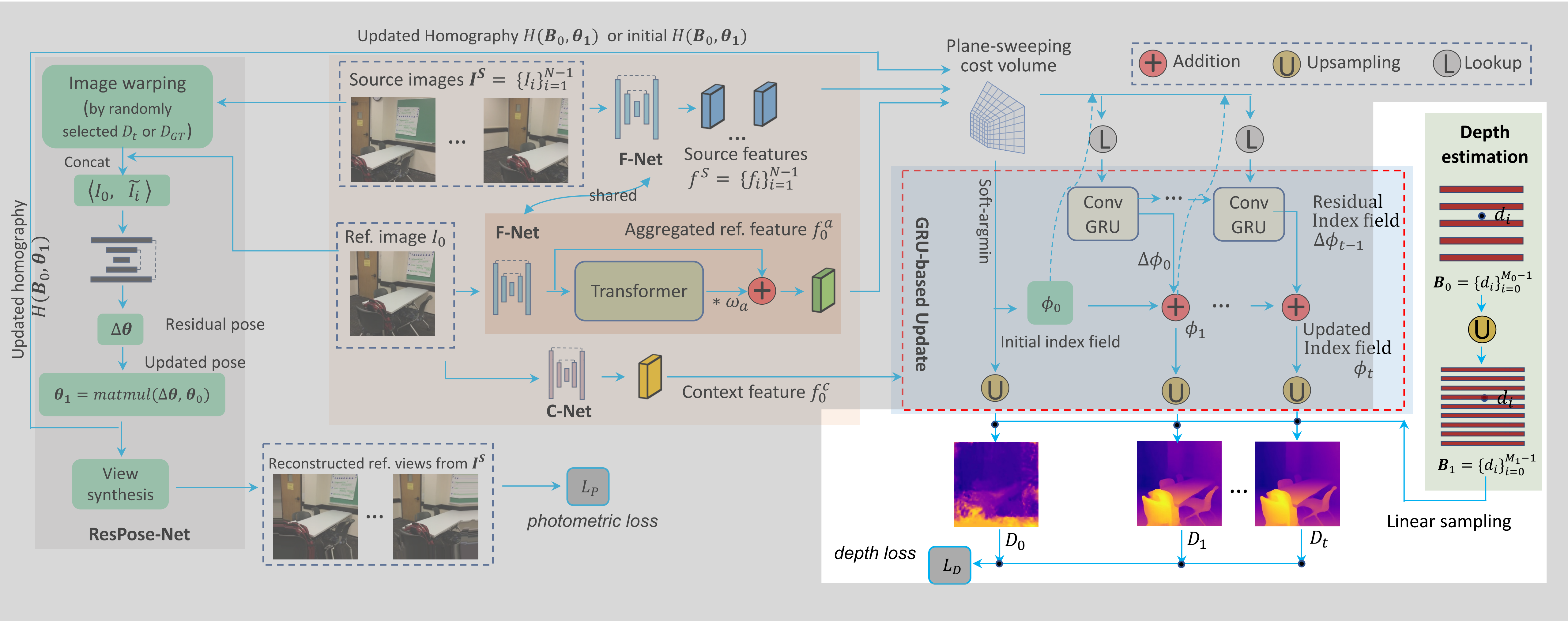
Architecture



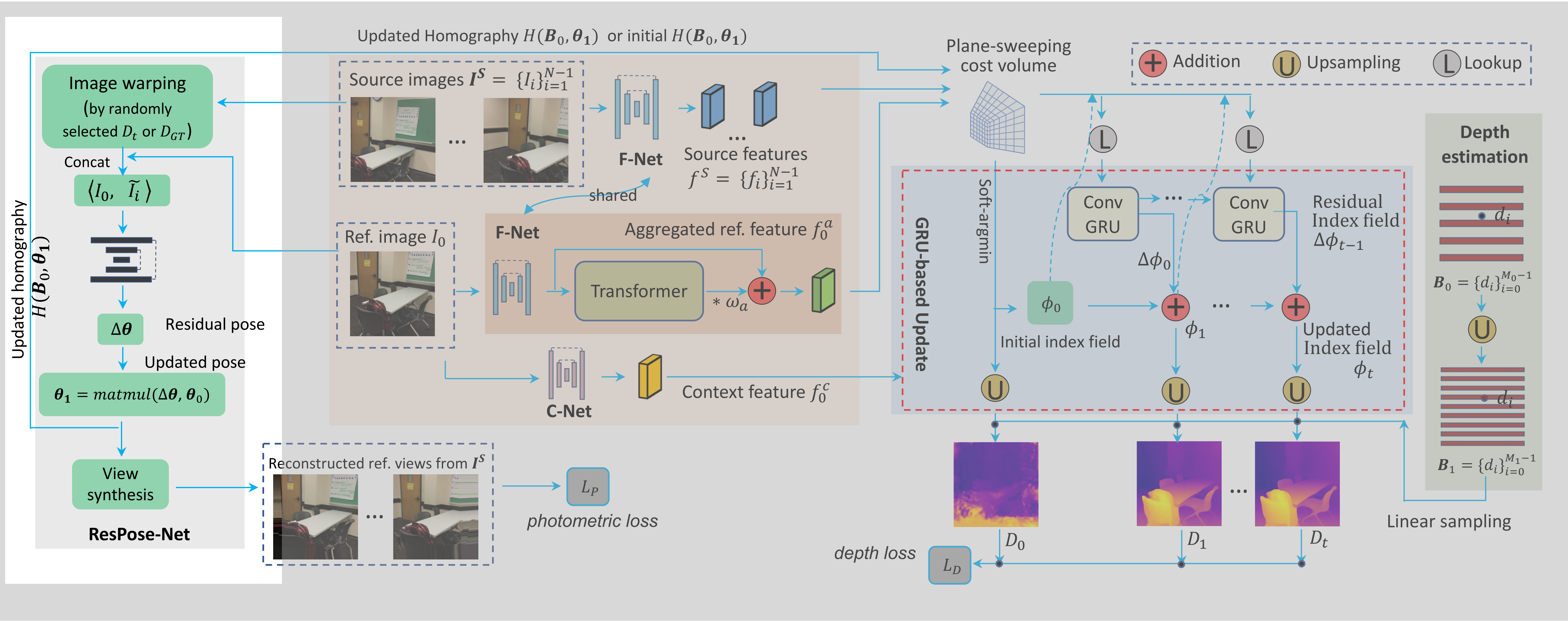
Architecture



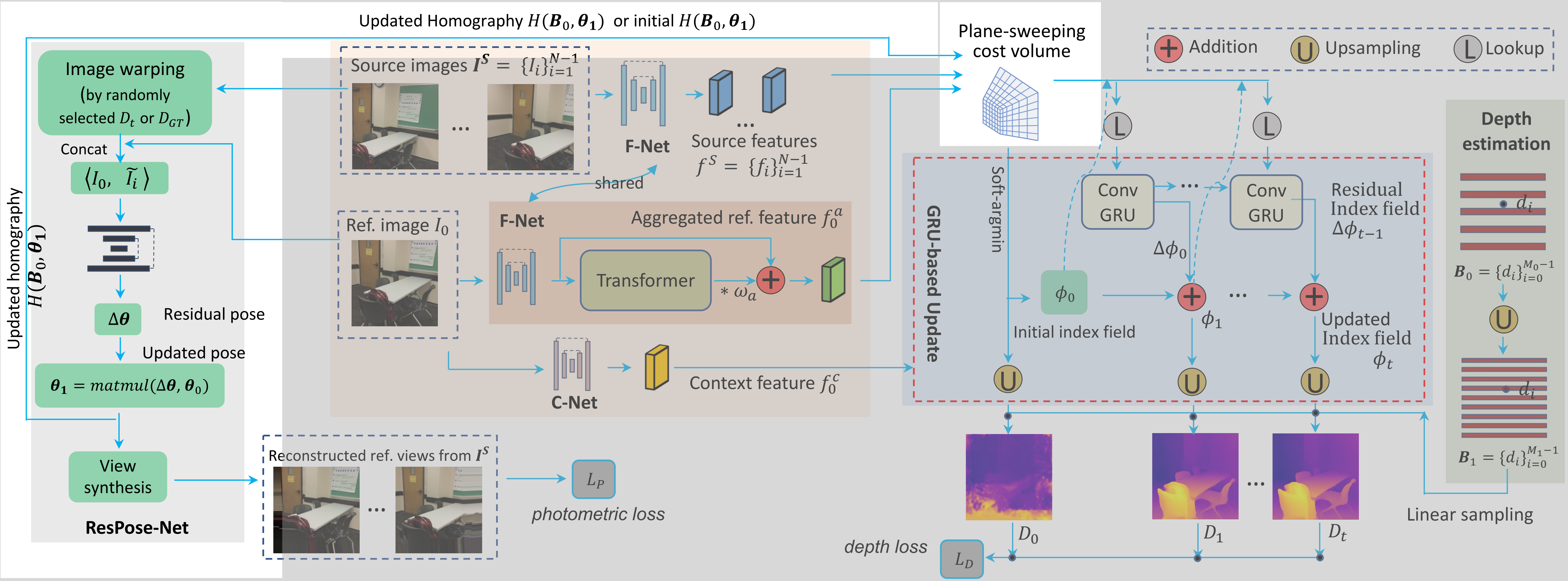
Architecture



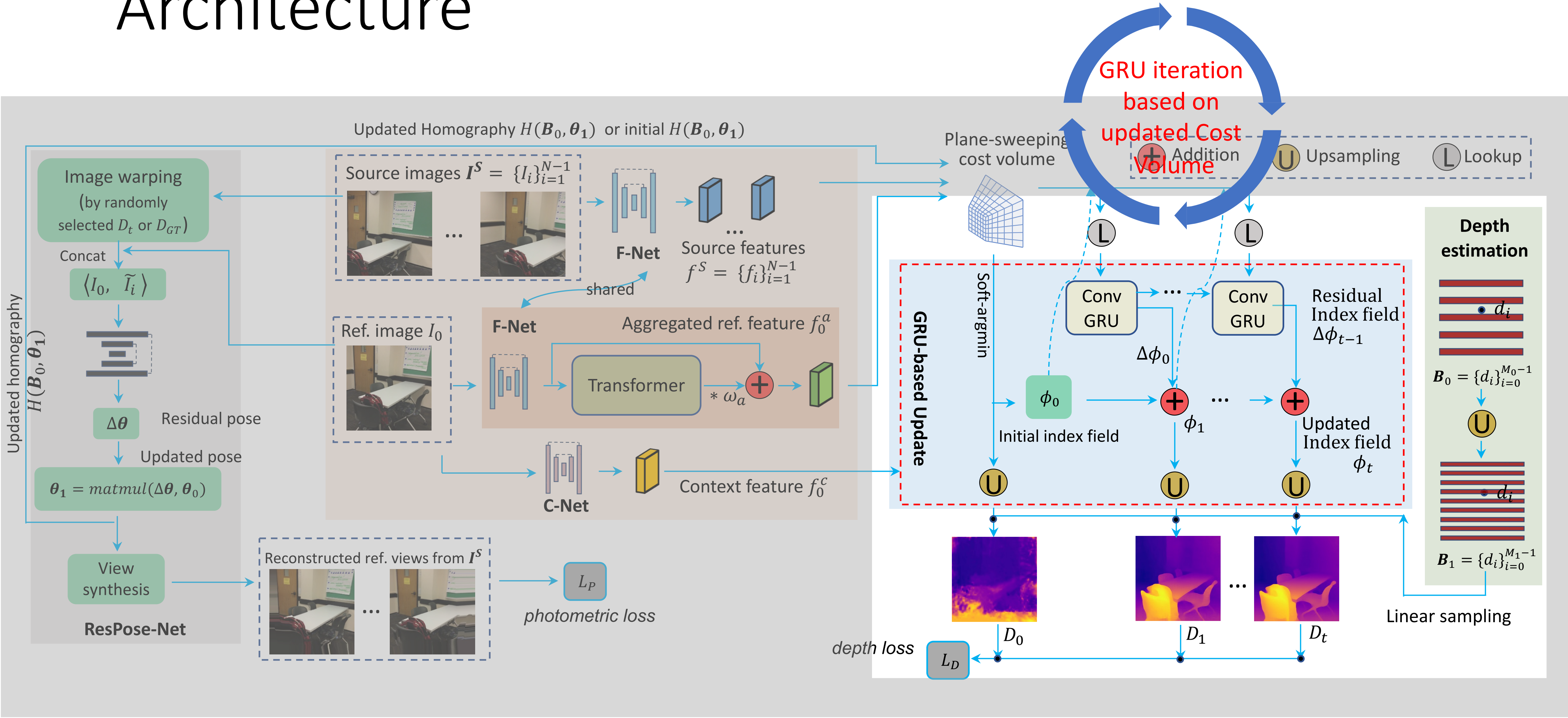
Architecture



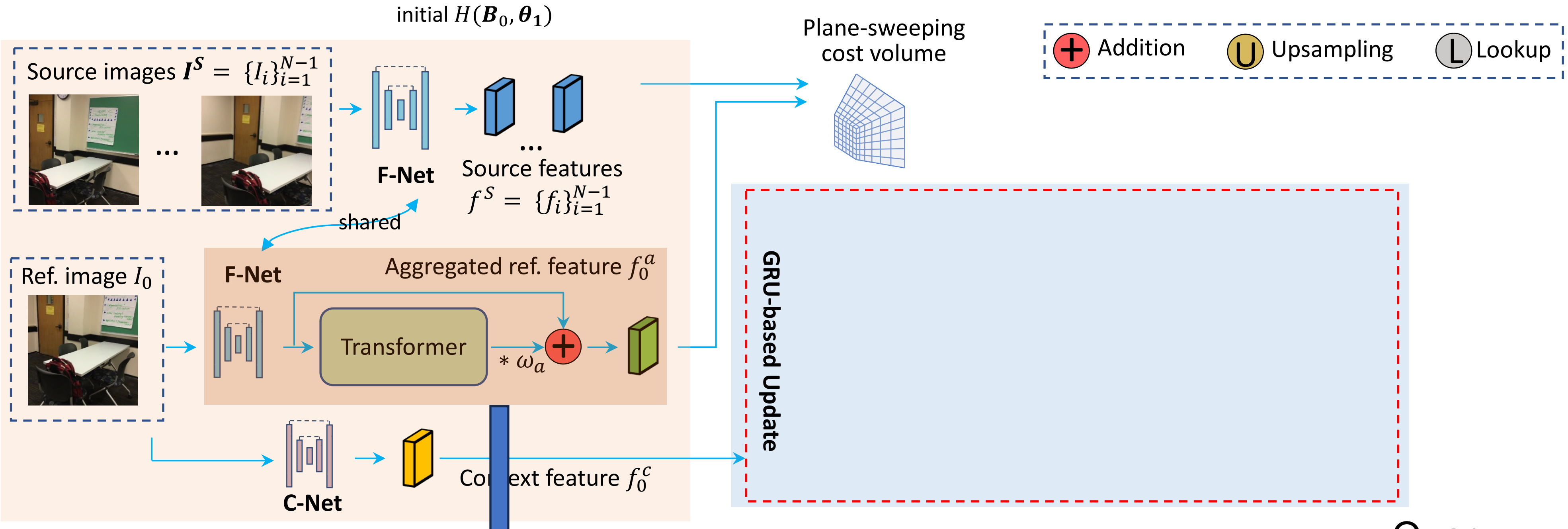
Architecture



Architecture



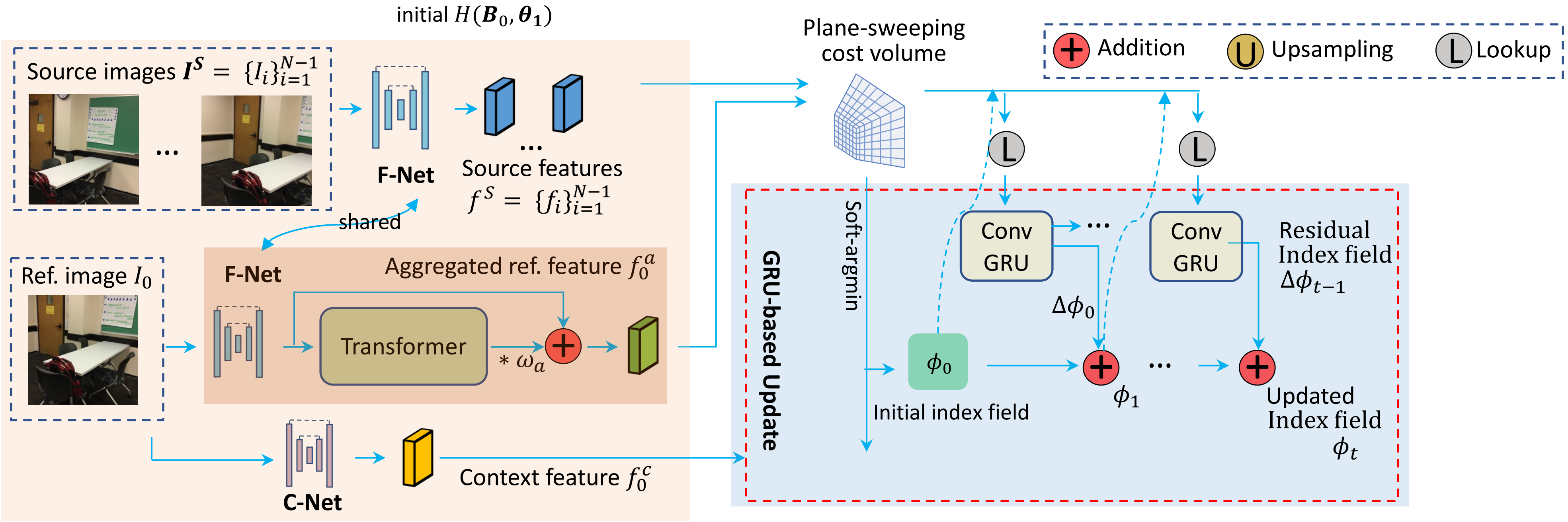
Architecture: Feature & Cost Volume



Our **asymmetric** employment of this transformer layer provides the capability to better balance the **high-frequency** (by high-pass CNNs) and **low-frequency** features (by self-attention).

$$\underbrace{f_0^a}_{\text{aggregated feature}} = \underbrace{f_0}_{\text{local feature}} + \omega_\alpha \sigma \left(\frac{\overbrace{(f_0 W^Q)}^{\text{Query}} \overbrace{(f_0 W^K)}^{\text{Key}}{}^T}{\sqrt{F_1}} \overbrace{(f_0 W^V)}^{\text{Value}} \right)$$

Architecture: GRU-based Iterative Updates



new hidden state

residual index field

index field

retrieved cost

context feature

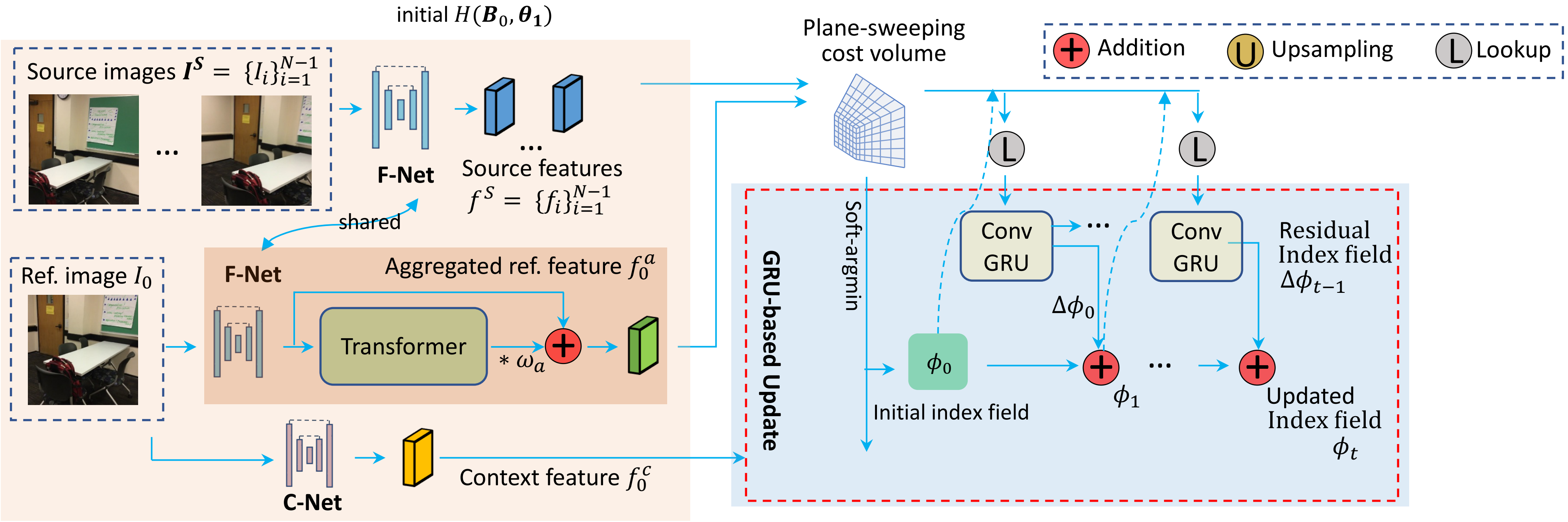
hidden state

$$\delta\phi_t, h_{t+1} \Leftarrow \mathrm{GRU}\left(<\phi_t, C_0^{\phi_t}, f_0^c>, h_t\right)$$

updated index field

$$\phi_{t+1} \Leftarrow \phi_t + \delta\phi_t$$

Architecture: GRU-based Iterative Updates



The proposed recurrent estimate of **index field** (i.e., a grid of indices to identify the depth hypotheses) enables the learning to be anchored at the cost volume domain.

new hidden state

residual index field

index field

retrieved cost

context feature

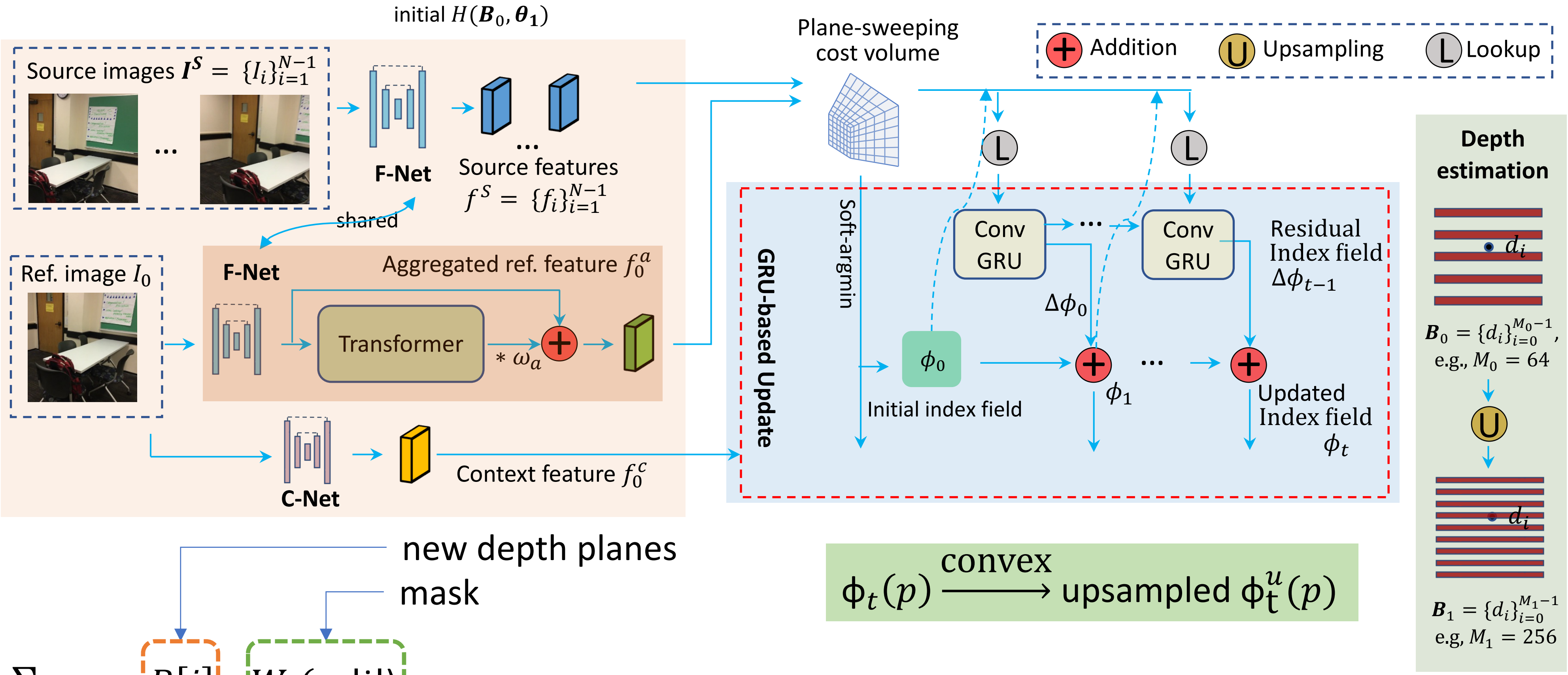
hidden state

$$\boxed{\delta\phi_t}, \boxed{h_{t+1}} \Leftarrow \mathrm{GRU}\left(<\boxed{\phi_t}, \boxed{C_0^{\phi_t}}, \boxed{f_0^c}>, \boxed{h_t}\right)$$

updated index field

$$\boxed{\phi_{t+1}} \Leftarrow \phi_t + \delta\phi_t$$

Architecture: Upsampling & Depth Estimation



$$\phi_t(p) \xrightarrow{\text{convex}} \text{upsampled } \phi_t^u(p)$$

new depth planes

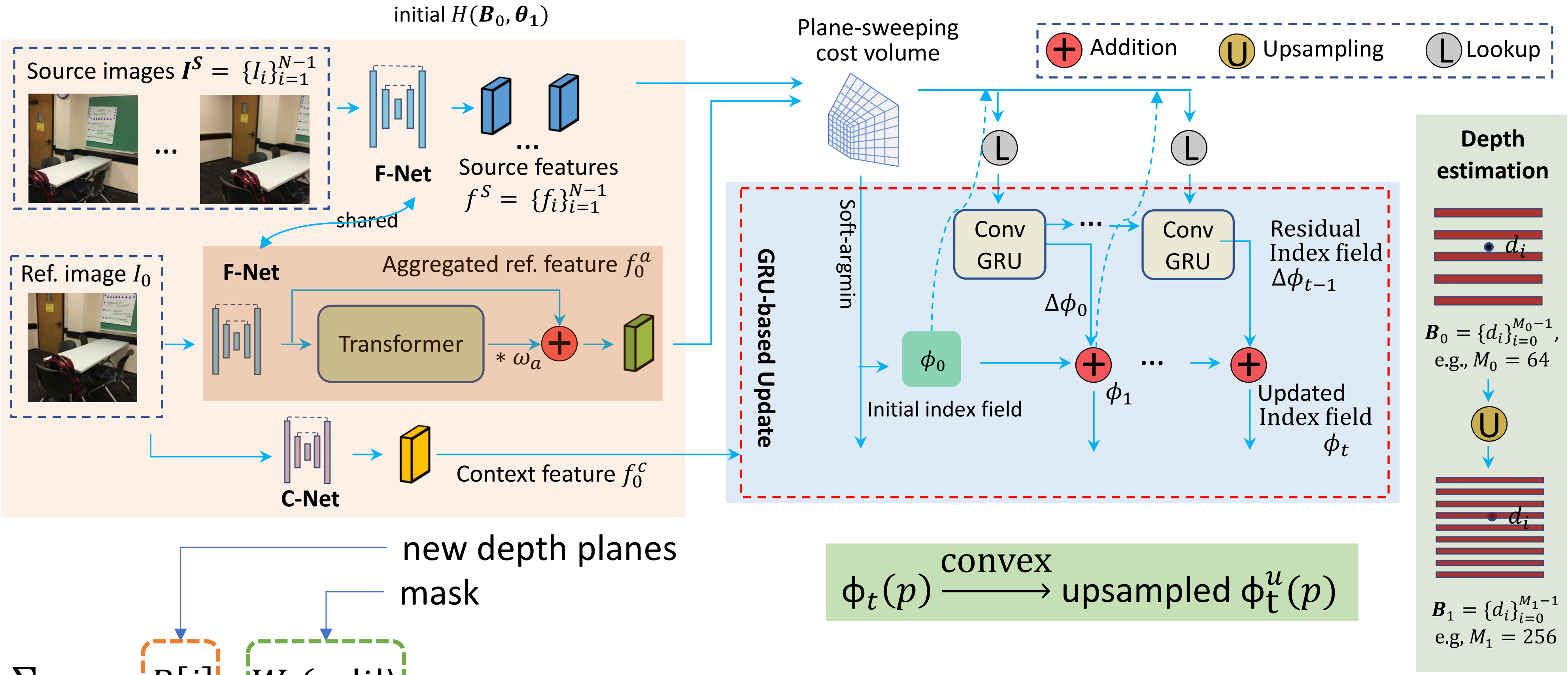
mask

Final depth

$$D_t(p) = \frac{\sum_{i \in \Omega(p)} B[i] \cdot W_1(p, \lfloor i \rfloor)}{\sum_{j \in \Omega(p)} W_1(p, \lfloor j \rfloor)}$$

Neighbor with a radius r=4 centered at upsampled index $\phi_t^u(p)$ for a given pixel p

Architecture: Upsampling & Depth Estimation



new depth planes

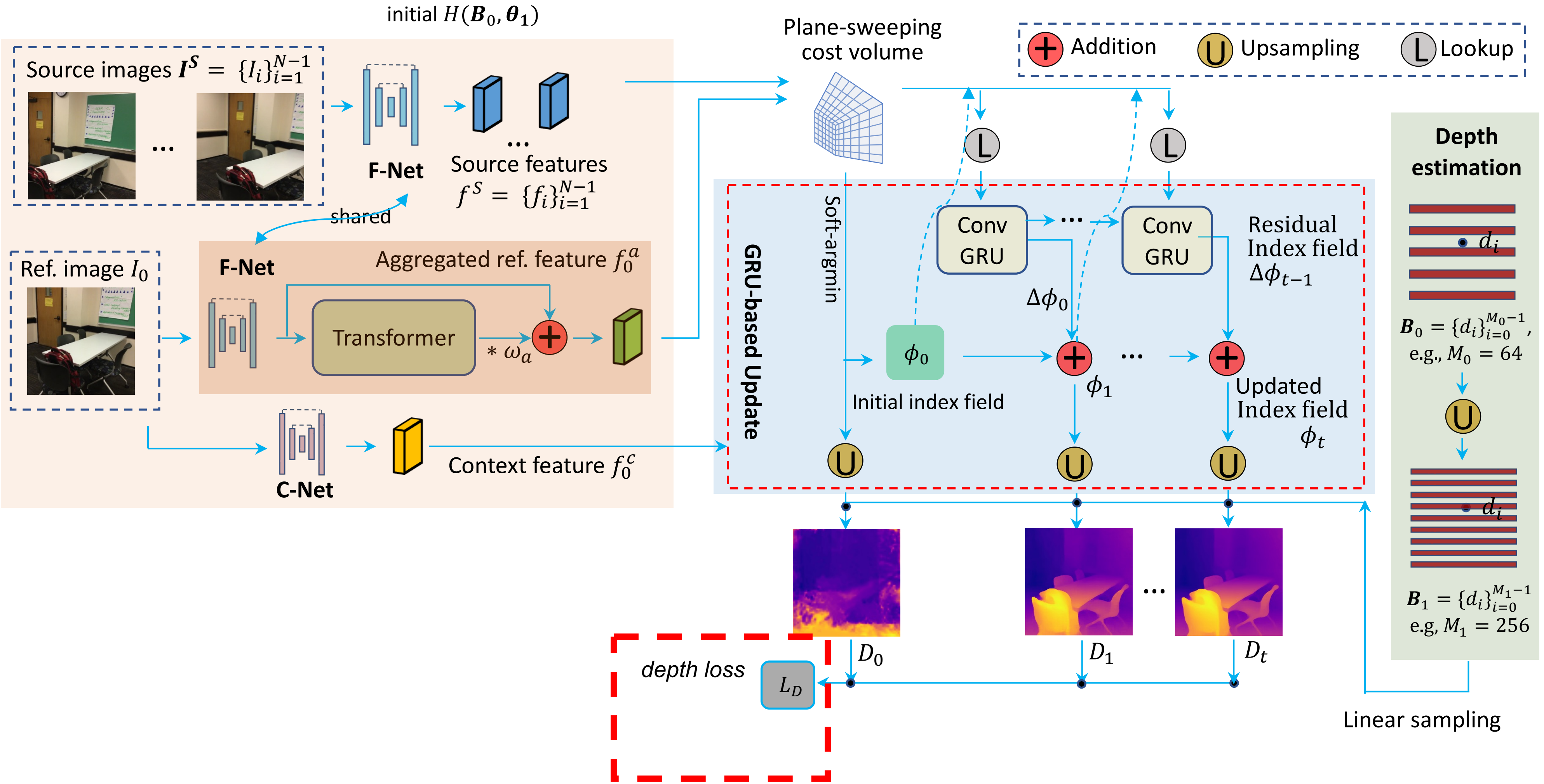
mask

Final depth

$$\boxed{D_t(p)} = \frac{\sum_{i\in \Omega(p)} \boxed{B[i]} \cdot \boxed{W_1(p, \lfloor i \rfloor)}}{\sum_{j\in \boxed{\Omega(p)}} W_1(p, \lfloor j \rfloor)}$$

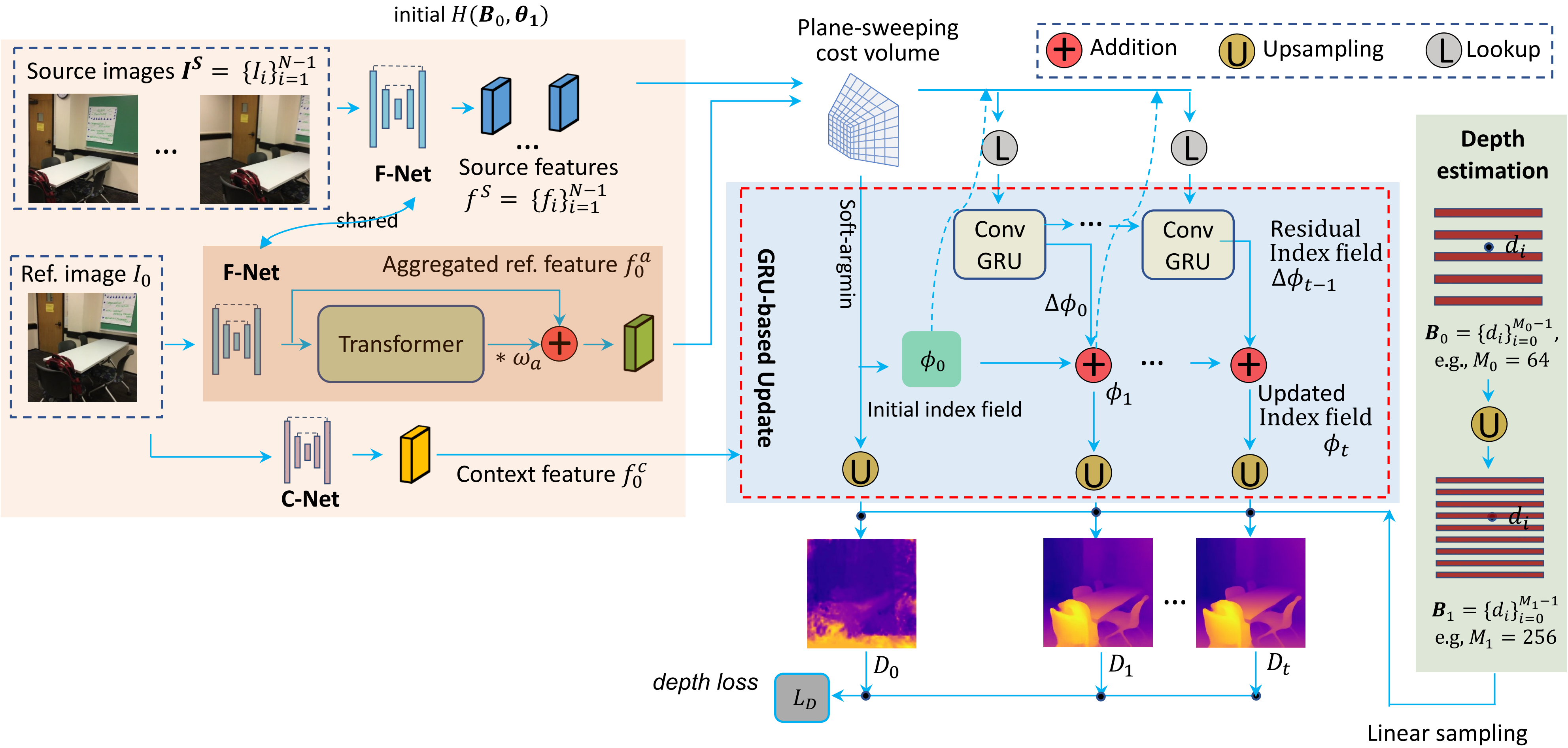
Neighbor with a radius r=4 centered at upsampled index $\phi_t^u(p)$ for a given pixel p

Architecture: Upsampling & Depth Estimation

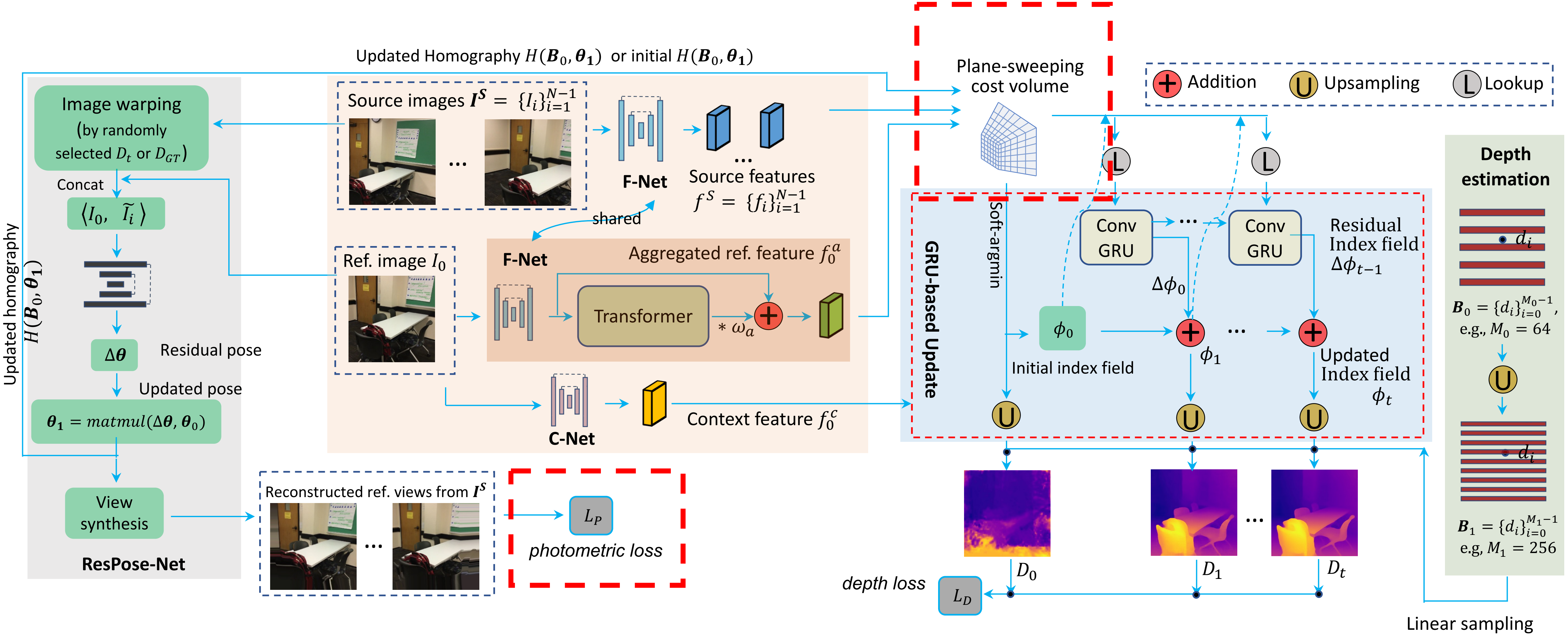


Architecture: Residual Pose Net

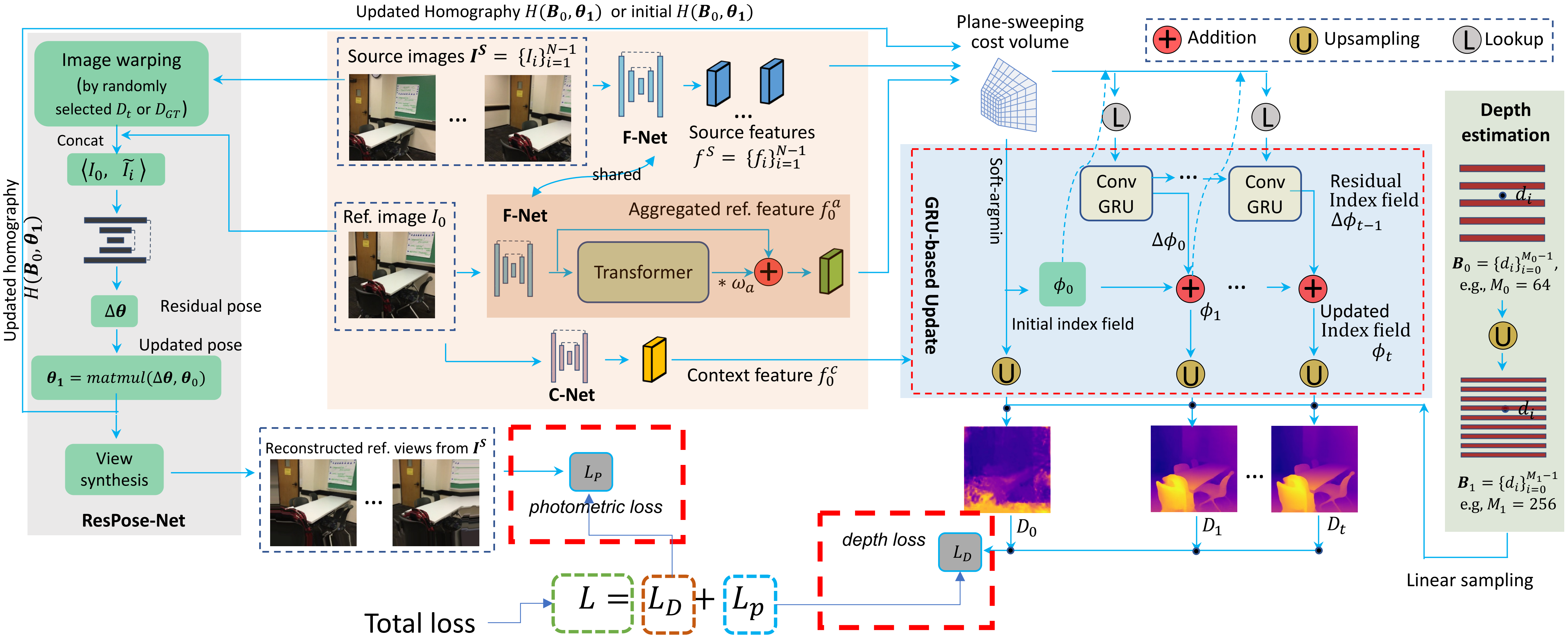
ResPose-Net



Architecture: Residual Pose Net

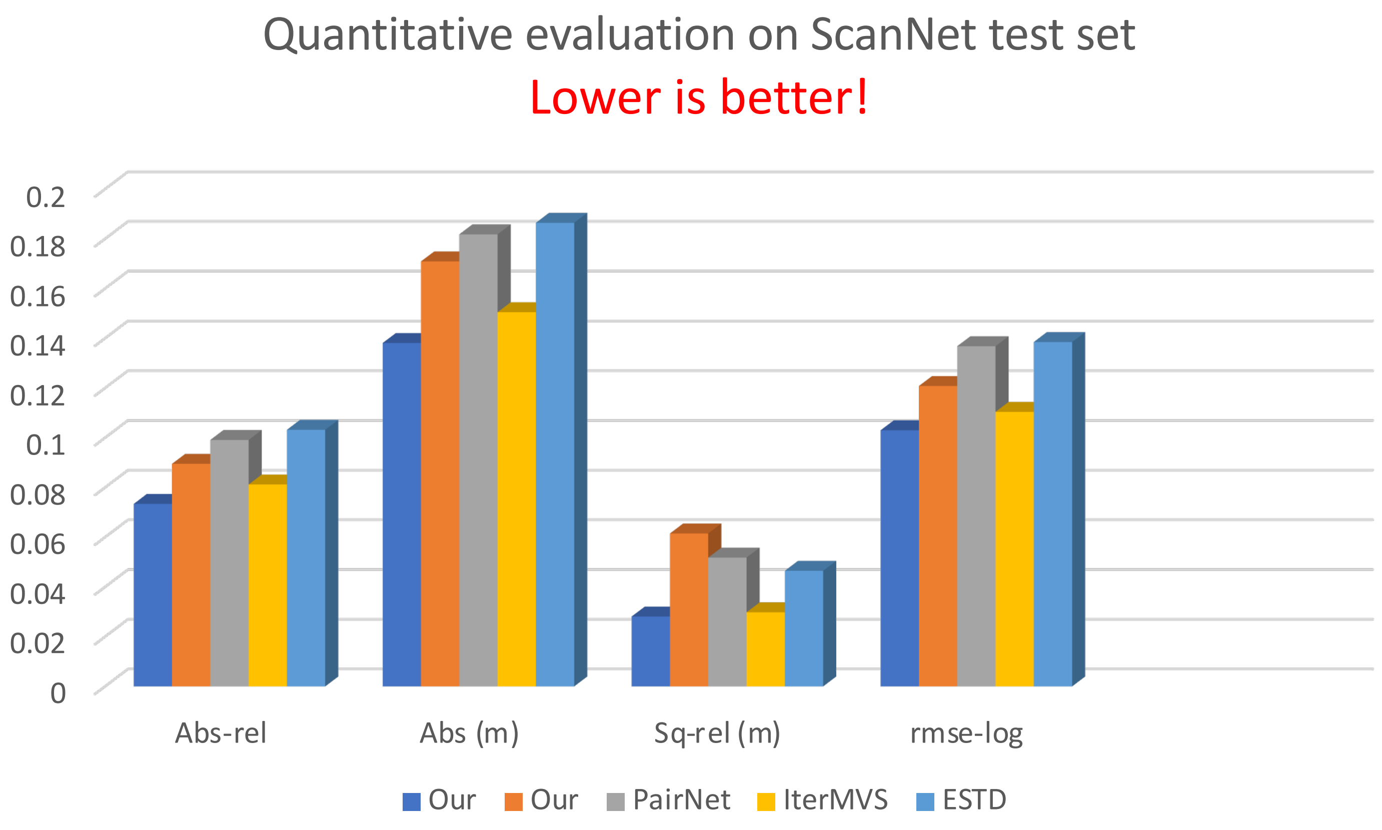


Architecture: Loss Function



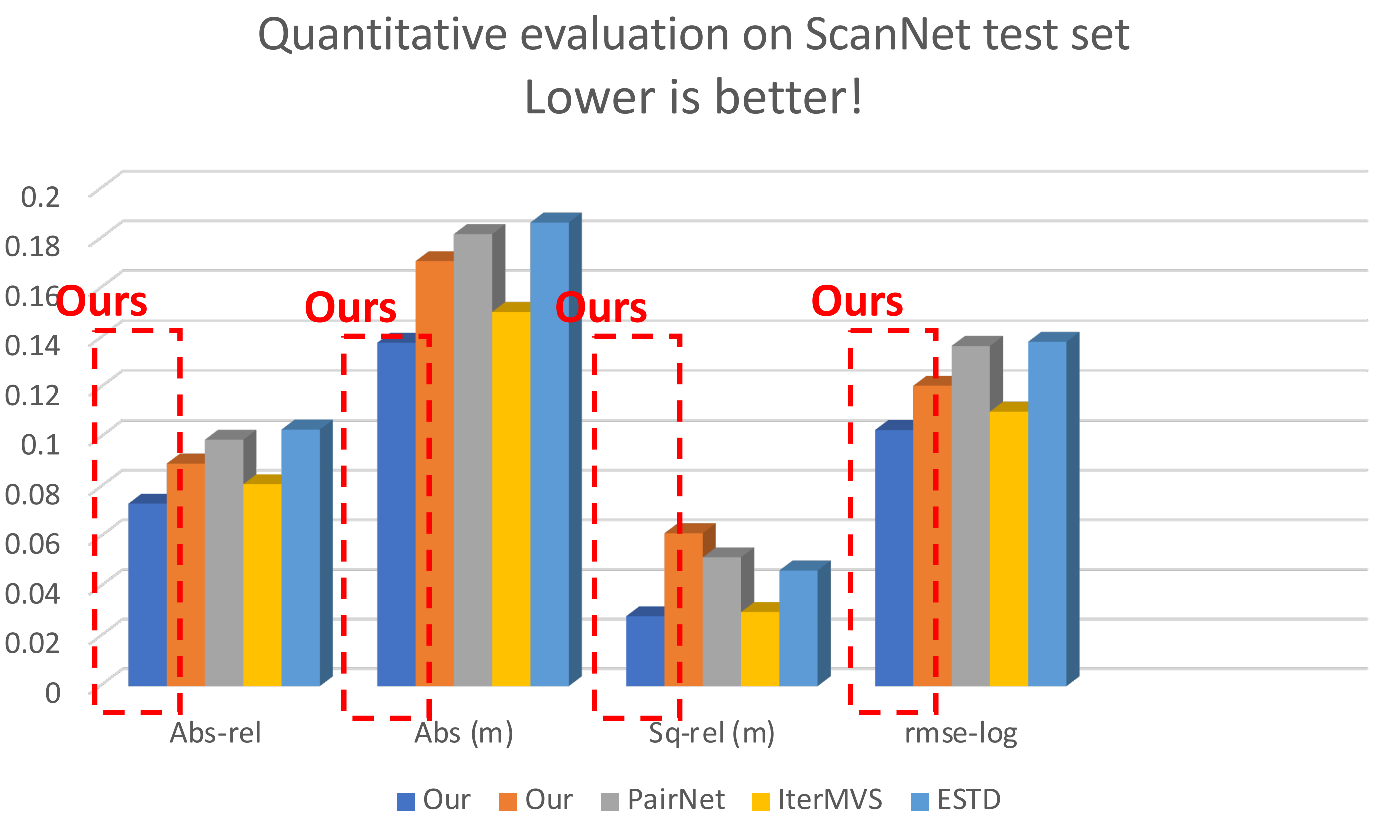
Experimental Results

- Depth map evaluation on ScanNet Testset



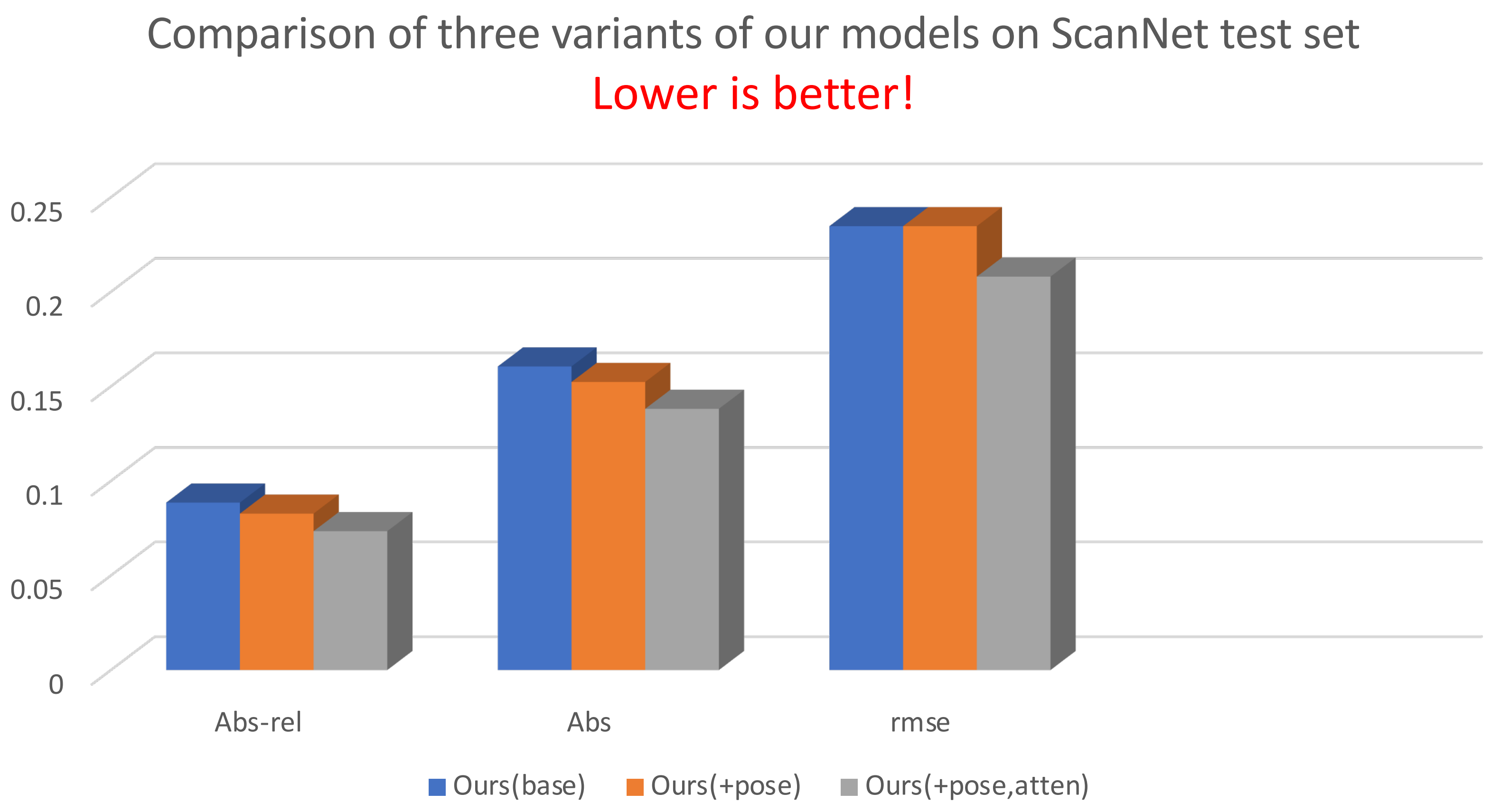
Experimental Results

- Depth map evaluation on ScanNet Testset



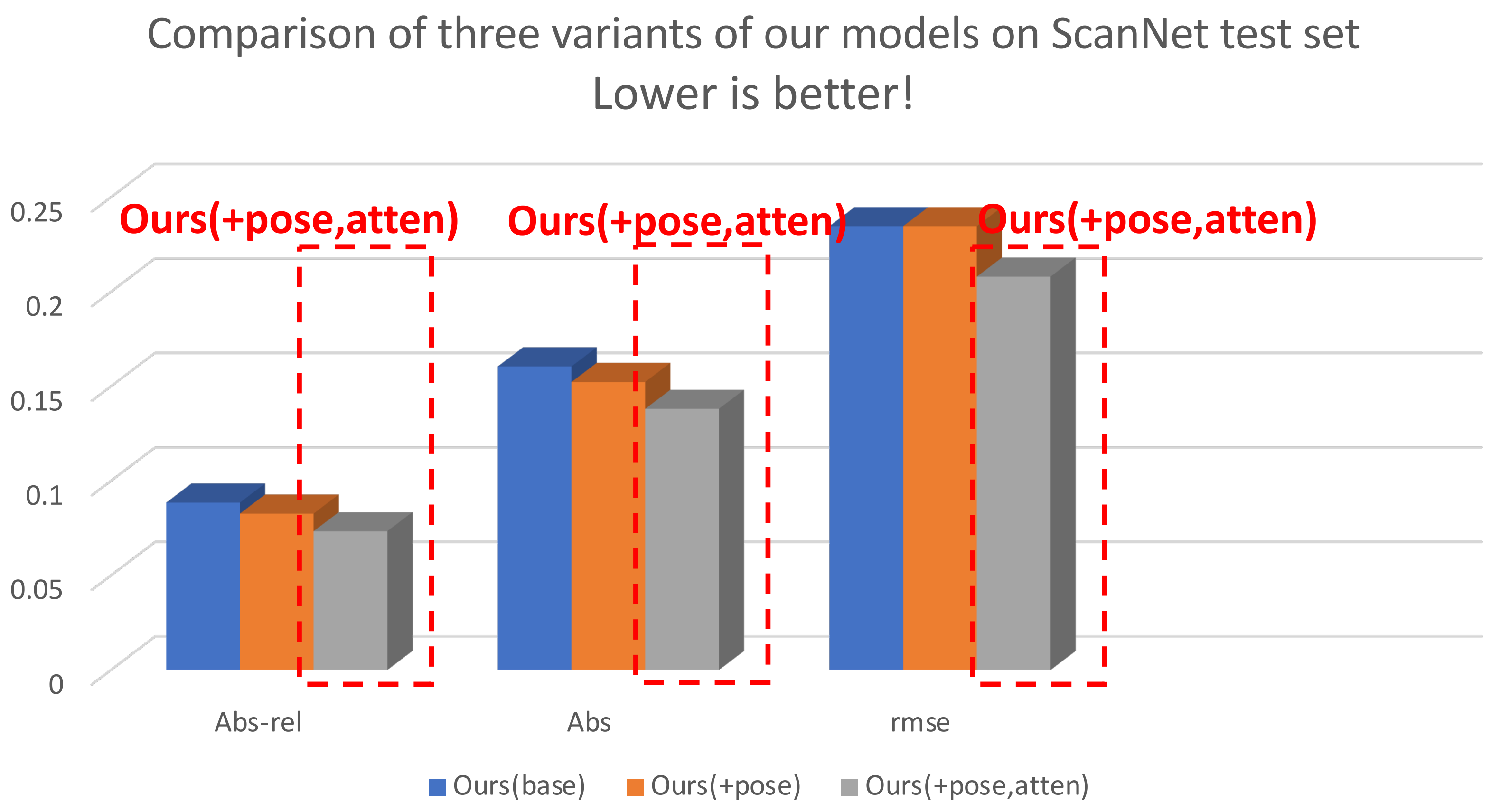
Experimental Results

- Ablation study: three variants of our method



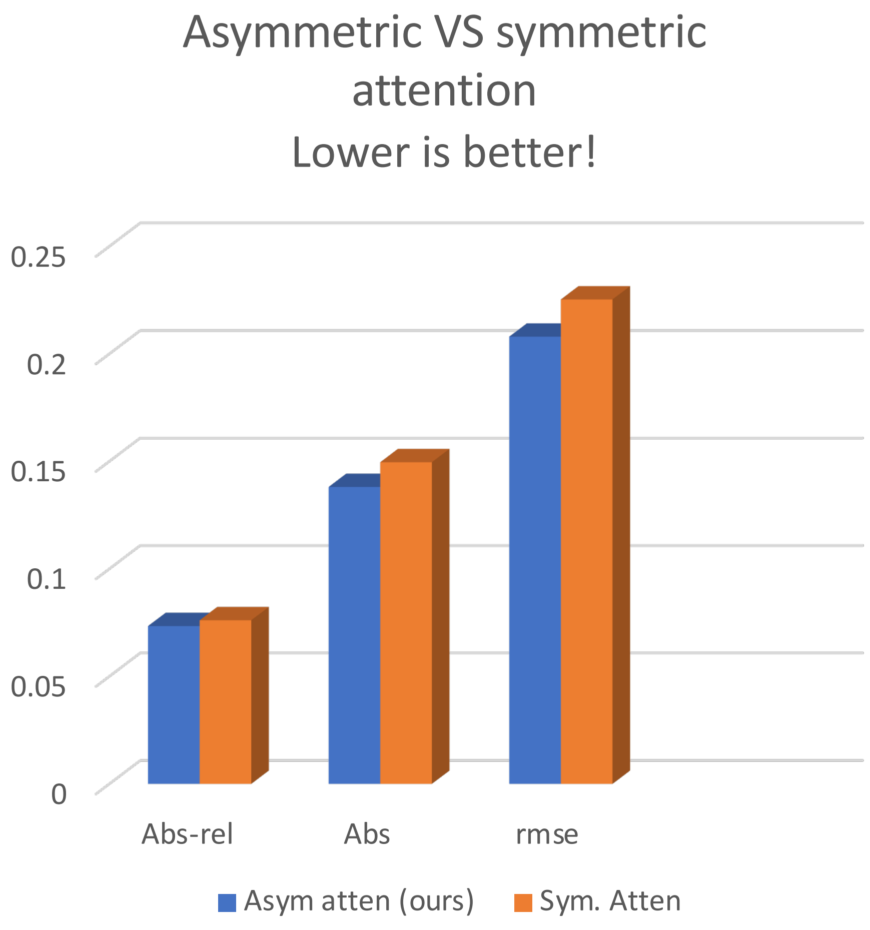
Experimental Results

- Ablation study: three variants of our method

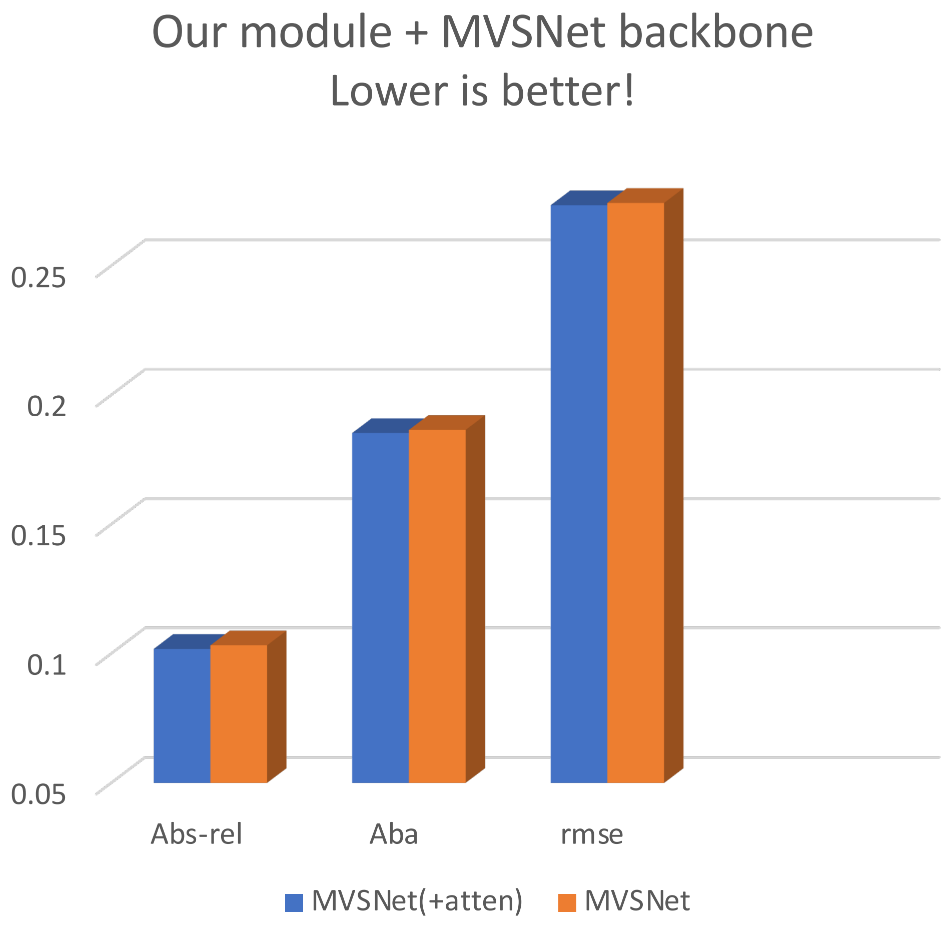


Experimental Results

- Asymmetric attention

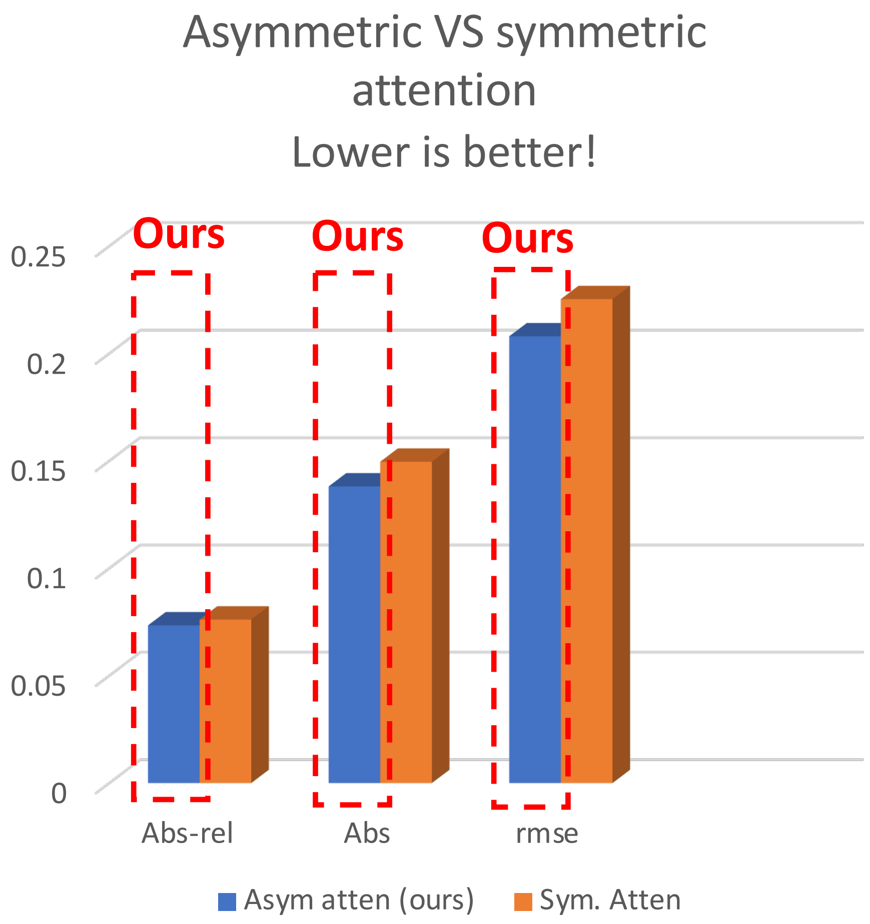


- Our attention applied to MVSNet

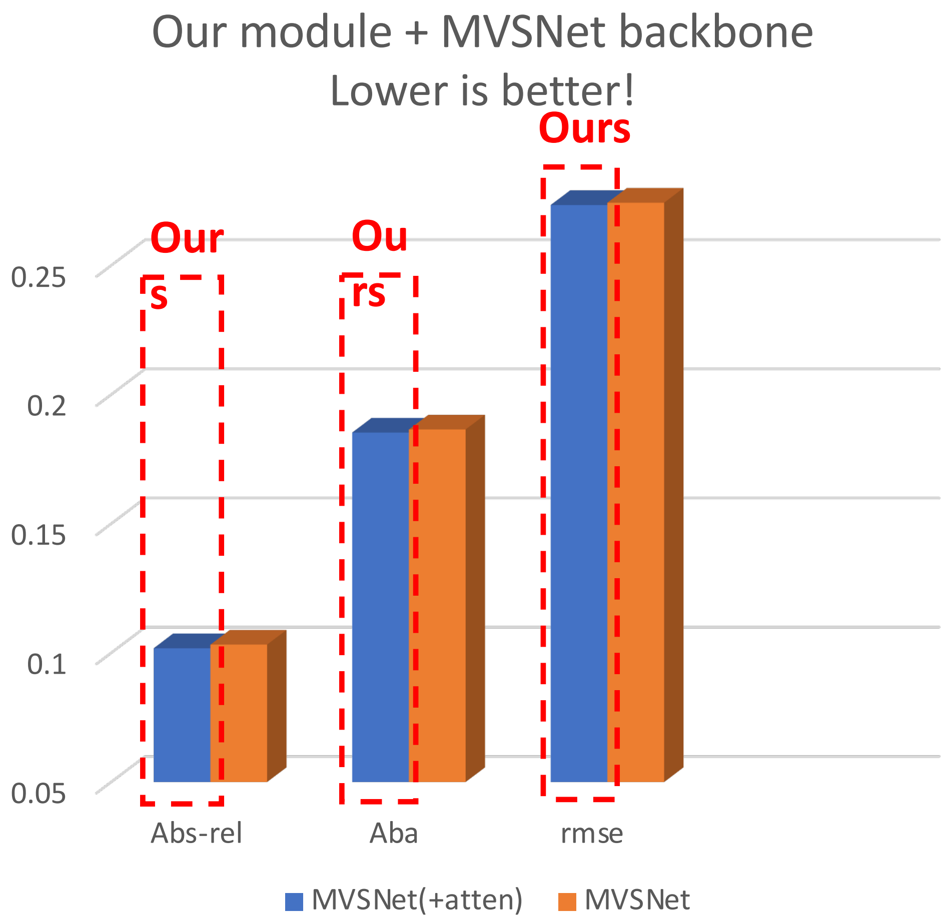


Experimental Results

- Asymmetric attention

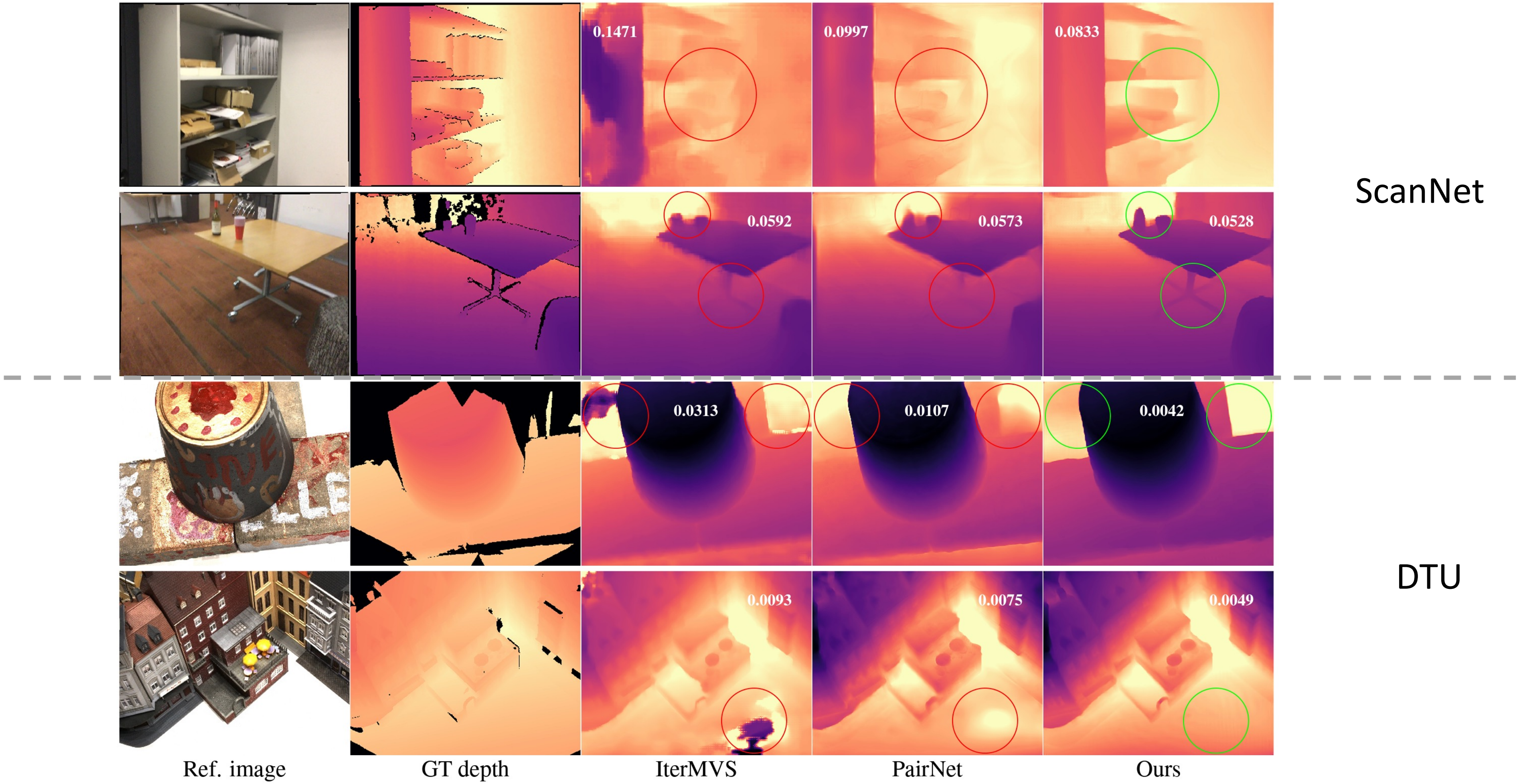


- Our attention applied to MVSNet



Experimental Results

- Qualitative results on ScanNet (top two rows) and DTU test set



Experimental Results

- More depth results and 3D point clouds on ScanNet

Conclusion

- RIAV-MVS, as a new paradigm to predict depth by learning to recurrently index cost volume via GRUs ★★★★★
- An asymmetric cost volume by a transformer block applied to the reference image ★★★☆☆
- A Residual pose network to update the relative poses to improve cost volume ★★★★☆

Conclusion

- RIAV-MVS, as a new paradigm to predict depth by learning to recurrently index cost volume via GRUs ★★★★★
- An asymmetric cost volume by a transformer block applied to the reference image ★★★☆☆
- A Residual pose network to update the relative poses to improve cost volume ★★★★☆

Conclusion

- RIAV-MVS, as a new paradigm to predict depth by learning to recurrently index cost volume via GRUs ★★★★★
- An asymmetric cost volume by a transformer block applied to the reference image ★★★☆☆
- A Residual pose network to update the relative poses to improve cost volume ★★★★☆

Thank You!

Code coming soon
https://github.com/oppo-us-research/riav-mvs