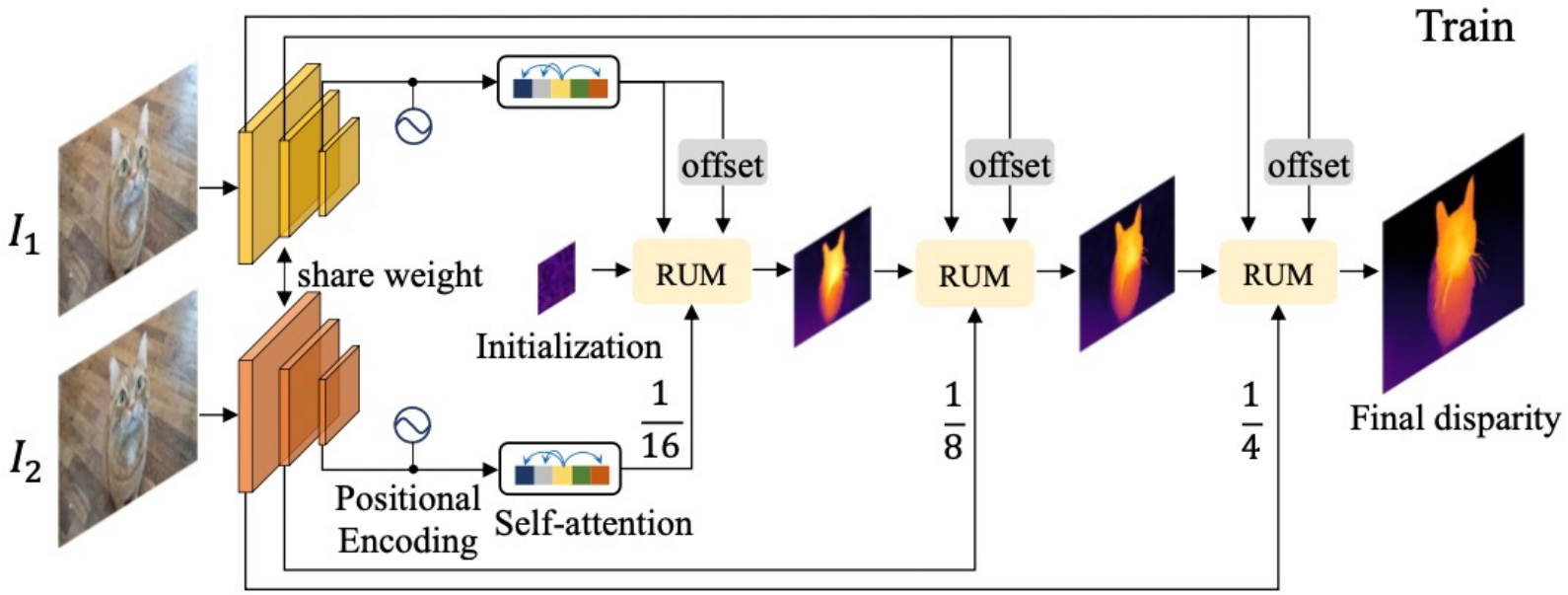
Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation

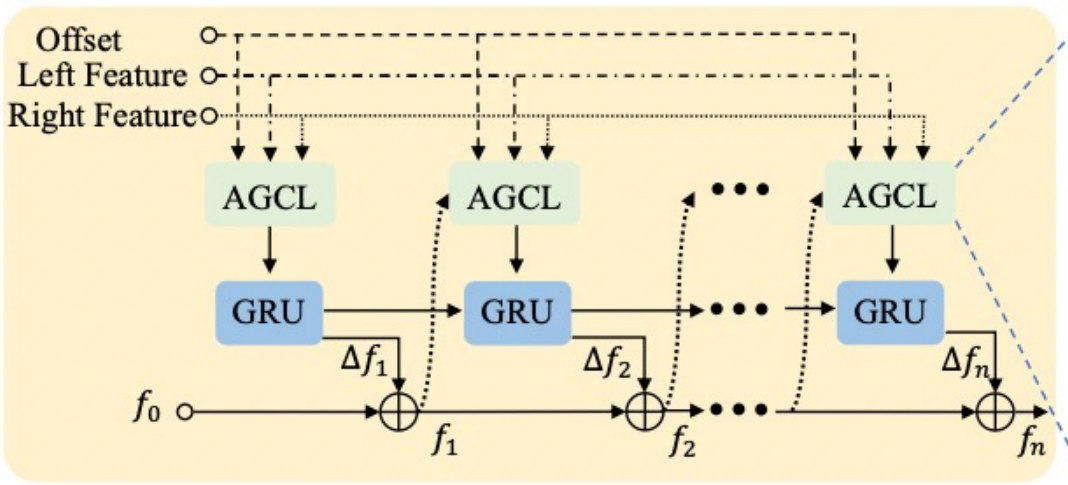
Youngju Yoo

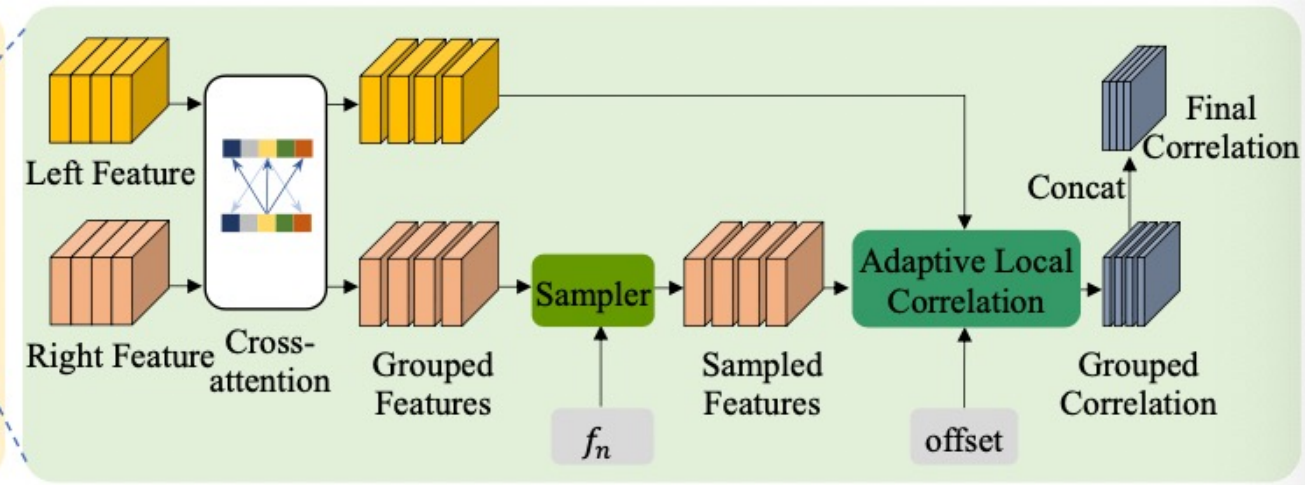
Introduction

- **CREStereo** : **C**ascaded **RE**current **S**tereo matching network
- Complicated issue for most existing algorithms to precisely recover the disparity of fine image **details, or thin structure**s
- Perfect rectification is hard to obtain for real-world stereo image pairs
- It is difficult to estimate disparity in typical hard cases, like non-texture or repetitive-texture regions.

Architecture

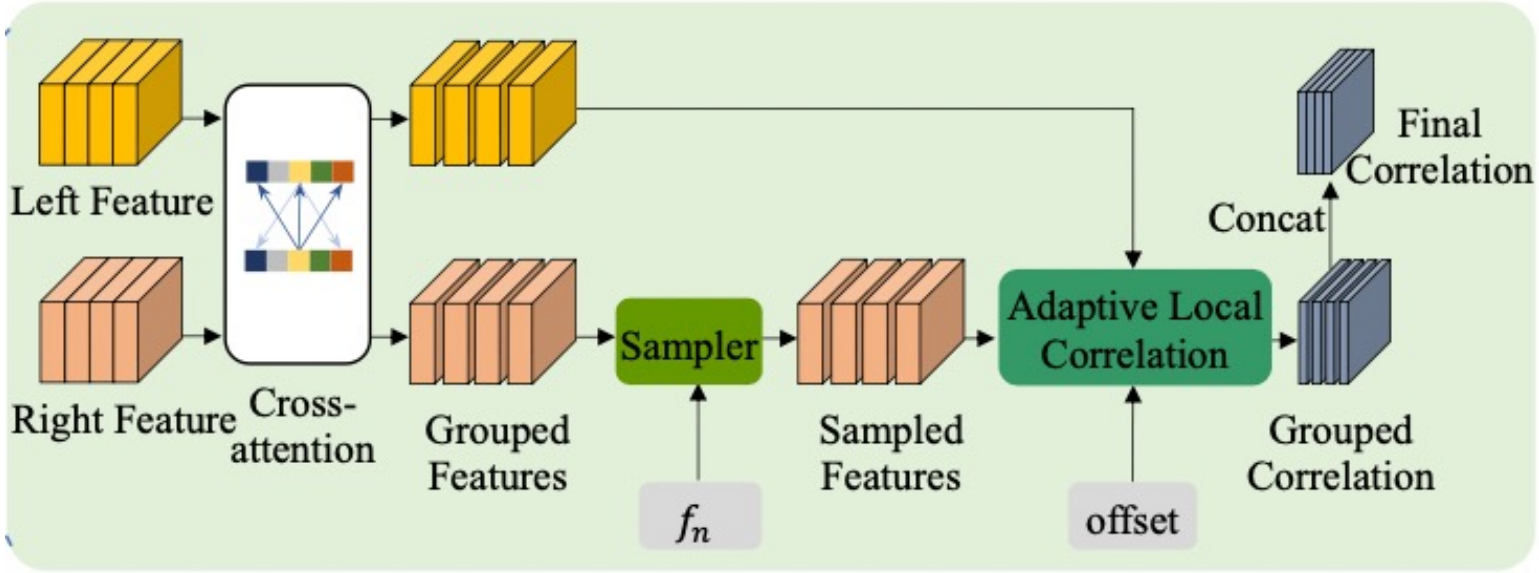






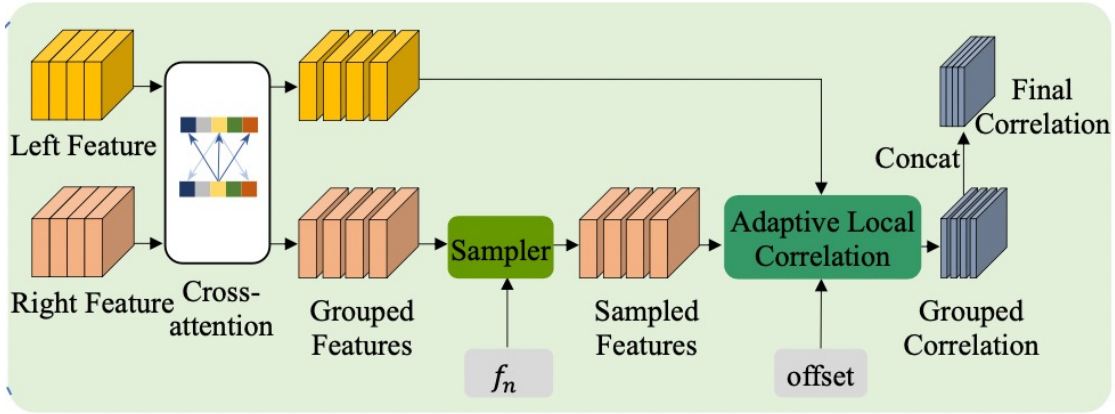
AGCL : Adaptive Group Correlation Layer

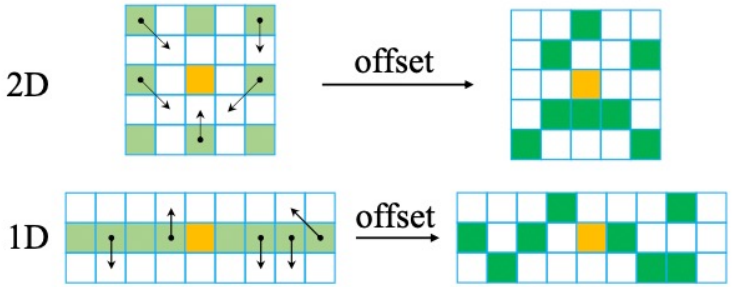
- To reduce matching ambiguity in the situation : the corresponding points may not locate on the horizontal epipolar line
- Achieving better performance compared to all pairs matching while only local correlation is computed.



AGCL : Adaptive Group Correlation Layer

- **Local Feature Attention** :
  - Only match points in a local window to avoid computation cost
  - add an attention module to aggregate global context information.
- **2D-1D Alternate Local Search** :
  - To deal with non-ideal stereo rectification cases
- **Deformable search window :**
  - To improve accuracy in occlusion and textureless areas
  - Additional learnable offset dx, dy
- **Group-wise correlation**

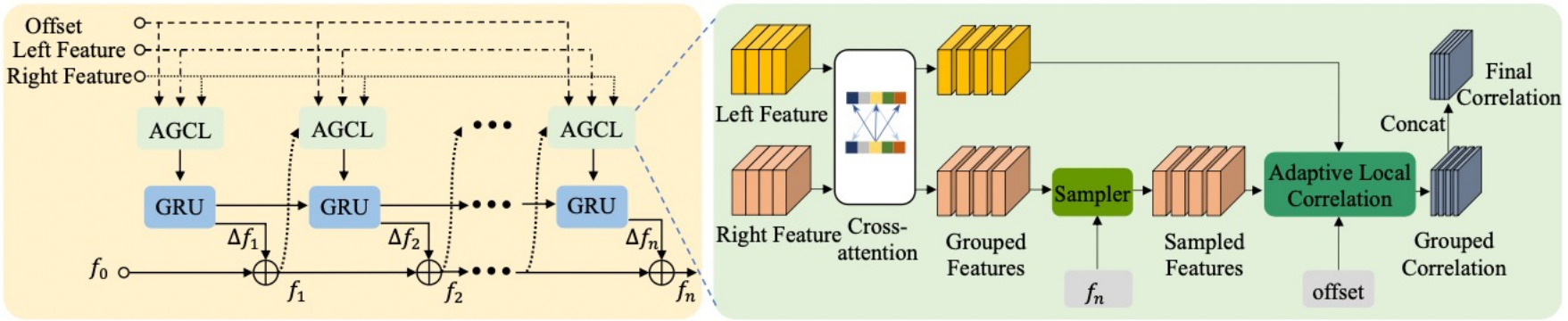




$$\mathrm{Corr}(x, y, d) = \frac{1}{C}\sum_{i=1}^{C}\mathbf{F}_1(i, x, y)\mathbf{F}_2(i, x'', y'') \text{ where } x'' = x + f(d) + \mathrm{d}x,\ y'' = y + g(d) + \mathrm{d}y.$$

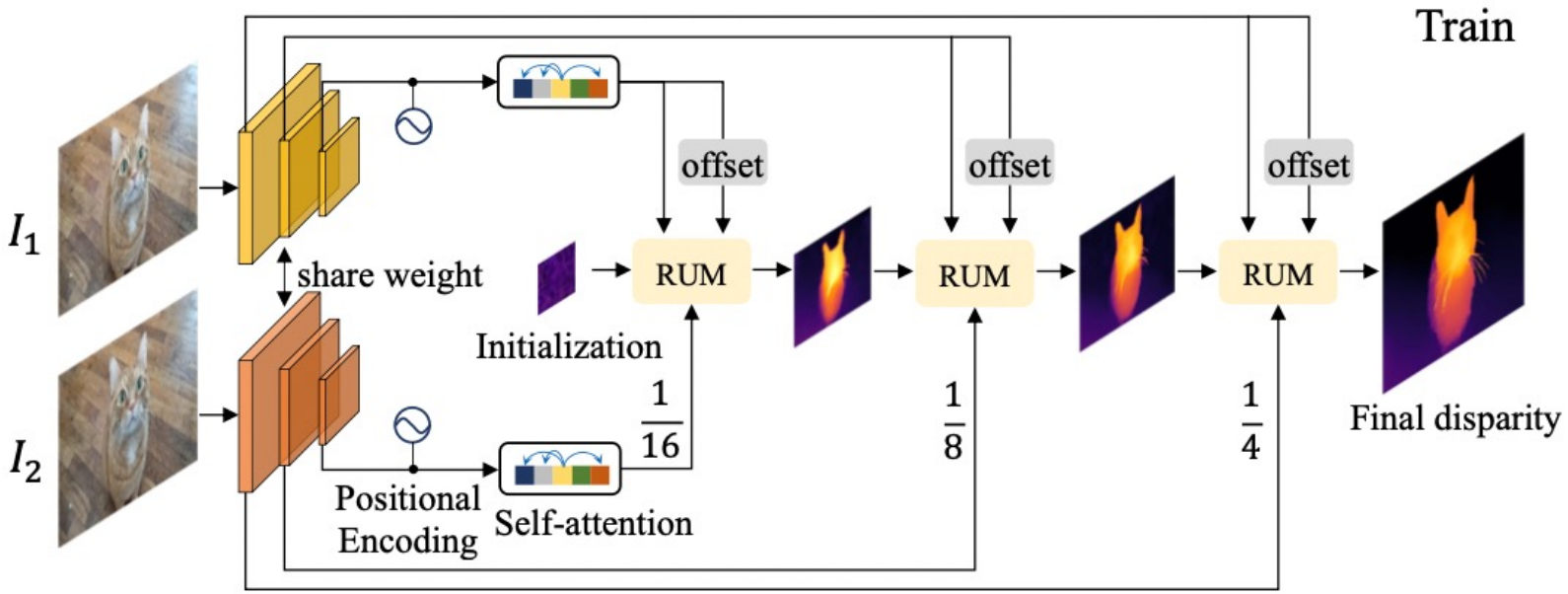
Cascaded Recurrent Network

- To maintain robustness in non-texture and repetitive-texture areas and preserve the details in high resolution input
- **Recurrent Update Module(RUM)** : based on GRU blocks and AGCL
  - Compute correlations for every feature map respectively in different cascade levels
  - Refine the disparities for several iterations independently.
  - **sampler** $\{f_1, \ldots, f_n\}$: intermediate predictions of n iterations



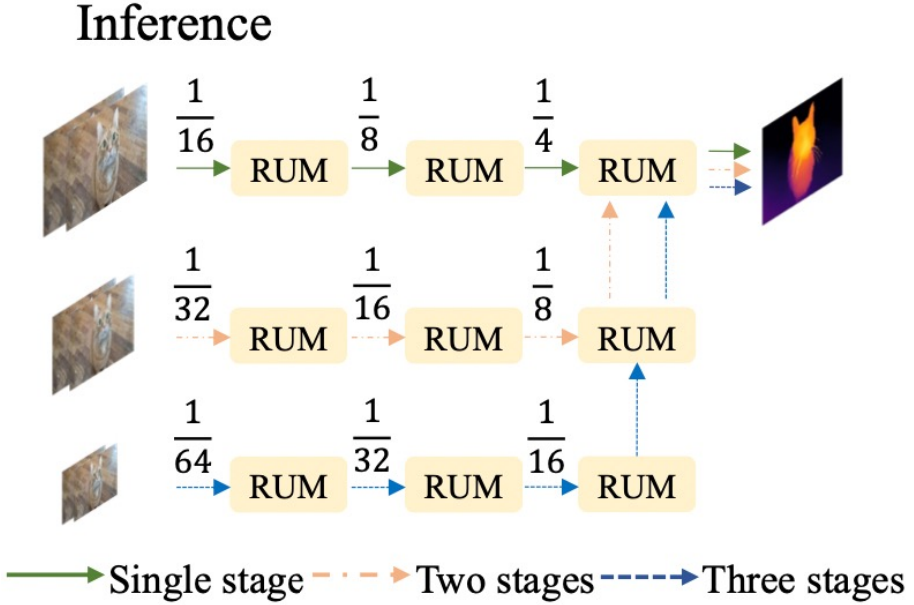
Cascaded Recurrent Network

- **Cascaded Refinement** :
  - Except for the first level of cascades, other levels take the upsampled version of prediction from previous level.
  - Share the same weights.

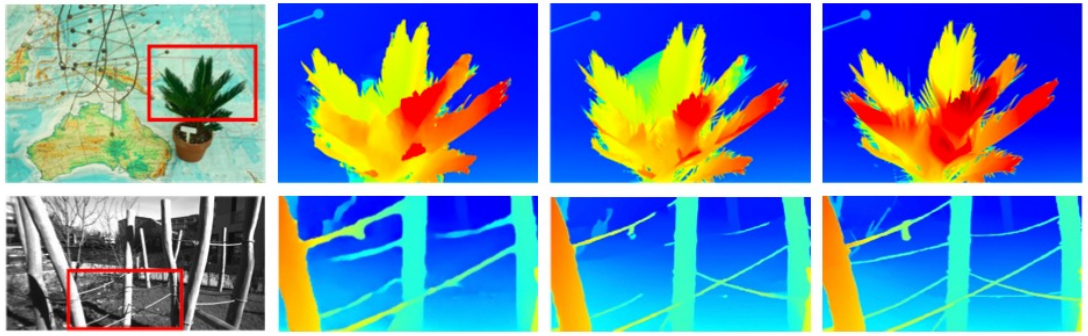


Stacked Cascades for Inference

- For small objects in high resolution images, features in these regions may suffer from deterioration with direct downsampling.
- Takes an image pyramid as input and feed them into the same trained feature extraction network to take advantage of multi-level context..

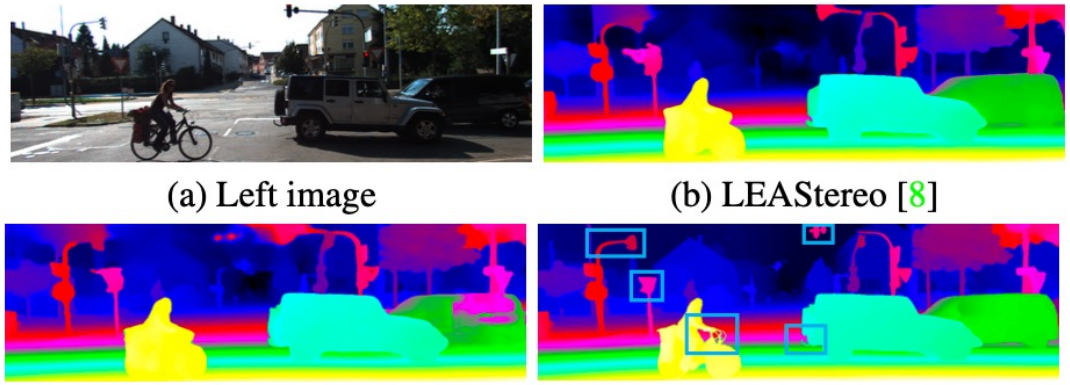


Result



(a) Left image (b) HITNet (c) RAFT-Stereo (d) Ours

Figure 7. Visual comparisons on Middelbury and ETH3D with HITNet [44] and RAFT-Stereo [23].



(a) Left image (b) LEAStereo [8]

(c) AANet [49] (d) Ours

Figure 8. Visual comparisons with other methods on one case of KITTI 2015 test set. Our method preserves more details.

RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching

Youngju Yoo

Introduction

- **RAFT-Stereo** : A new deep architecture for rectified stereo based on the optical flow network RAFT
- RAFT-Stereo uses only 2D convolutions and a lightweight cost volume constructed using a single matrix multiplication.
- Avoiding the high computation and memory cost of 3D convolutions.

Architecture

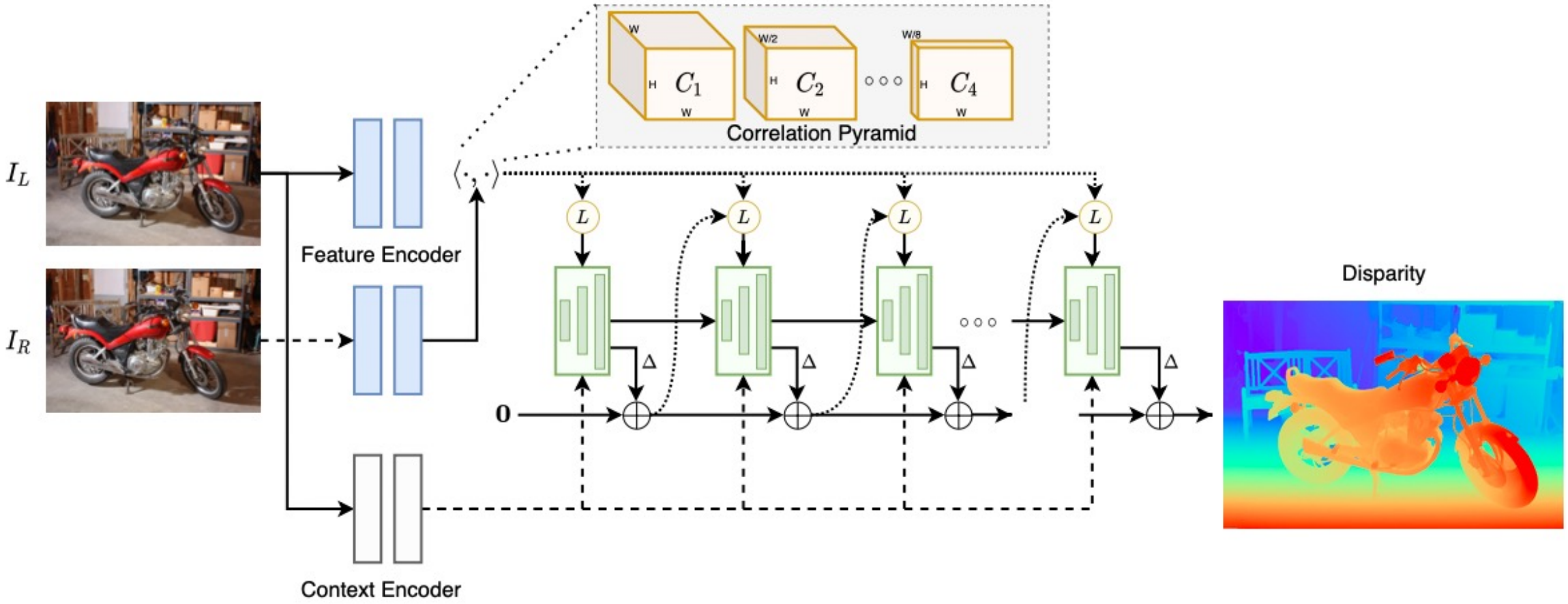


Figure 1. Correlation features (blue) are extracted from each of the images and are used to construct the correlation pyramid. "Context" image features (white) and an initial hidden state are also extracted from the context encoder. The disparity field is initialized to zero. Every iteration, the GRU(s) (green) use the current disparity estimate to sample from the correlation pyramid. The resulting correlation features, initial image features and current hidden state(s) are used by the GRU(s) to produce a new hidden state and an update to the disparity.

Correlation Pyramid

- **Correlation Volume** : $\mathbf{C}_{ijk} = \sum_h \mathbf{f}_{ijh} \cdot \mathbf{g}_{ikh}, \quad \mathbf{C} \in \mathbb{R}^{H \times W \times W}$
- **Correlation Pyramid** :
  - construct a 4 level pyramid of correlation volumes through repeated average pooling of the last dimension.
  - Each level of the pyramid has an increased receptive field
- **Correlation Lookup** :

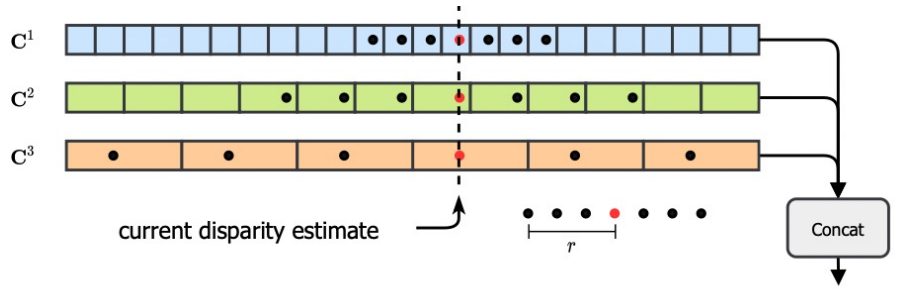


Figure 2. Lookup from the correlation pyramid. We use the current estimate of disparity to retrieve values from the each level of the correlation pyramid. We index from each level in the pyramid by linear interpolating at the current disparity estimate and at integer offsets, whose size depends on the correlation pyramid level.

Multi-Level GRU

- The correlation, disparity, and context features and then concatenated and injected into the GRU.

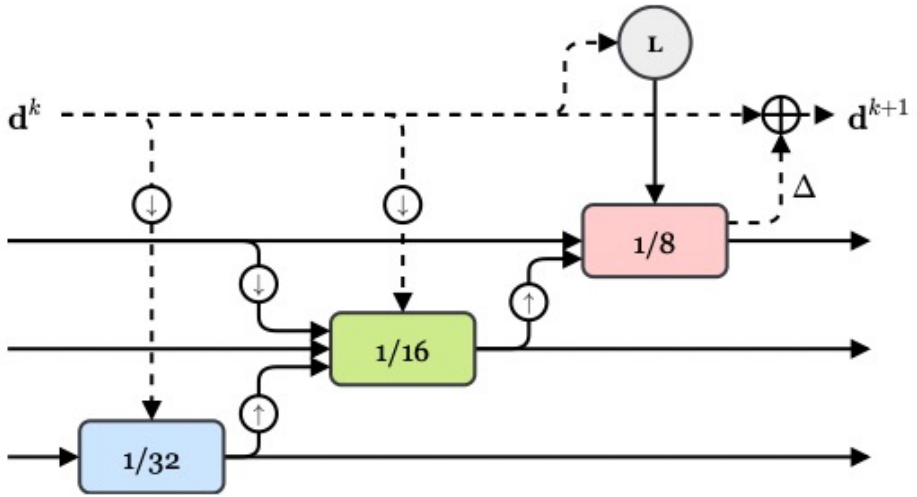


Figure 3. Multilevel GRU. We use a 3-level convolutional GRU which acts on feature maps at 1/32, 1/16, and 1/8 the input image resolution. Information is passed between GRUs at adjacent resolutions using upsampling and downsampling operations. The GRU at the highest resolution (red) performs lookups from the correlation pyramid and updates the disparity estimate.

Result

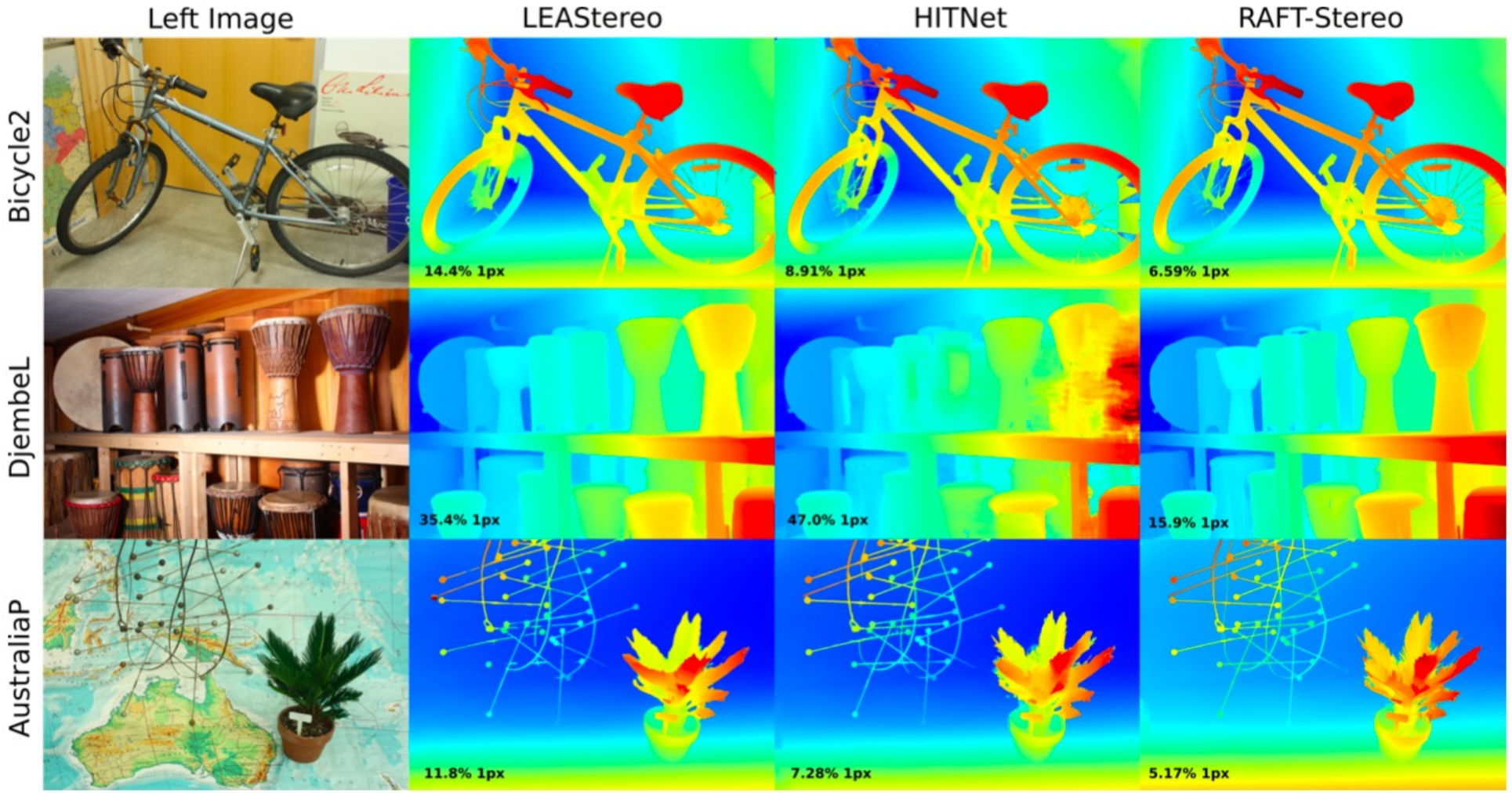


Figure 5. Results on the Middlebury [28] test set compared to the top end-to-end deep learning approaches. We also report the 1px error of each output in the corner. RAFT-Stereo is able to recover extremely fine details that other approaches cannot, such as the spokes of the bike wheel, the individual leaves of the plant, and sharp object boundaries.