DualRefine: Self-Supervised Depth and Pose Estimation
Through Iterative Epipolar Sampling and Refinement Toward Equilibrium
Supplementary Material

Antyanta Bangunharcana\textsuperscript{1}, Ahmed Magd\textsuperscript{2}, Kyung-Soo Kim\textsuperscript{1}
\textsuperscript{1}Mechatronics, Systems, and Control Laboratory, \textsuperscript{2}Vehicular Systems Design and Control Lab
Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
\{antabangun, a.magd, kyungsoo\}@kaist.ac.kr

1. DEQ Framework

We adhere to the general framework of DEQ \cite{2, 3} and employ a quasi-Newton solver to accelerate convergence. In our experiments, we utilize the Anderson solver \cite{1}. A DEQ model computes \( A = I - \frac{\partial U}{\partial z} \) at the fixed point \( z^* \) to obtain the gradient. This is typically achieved by performing another fixed-point iteration. However, in line with \cite{2, 5, 8}, we approximate \( A = I \) and utilize the inexact gradient for training.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{kitti_depth_error.png}
\caption{The progression of Abs Rel errors in each DualRefine iteration for KITTI depth.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{kitti_improved_depth_error.png}
\caption{The progression of Abs Rel errors in each DualRefine iteration for KITTI improved depth.}
\end{figure}

2. Training Loss Combinations

Determining the optimal pairings to calculate the self-supervision losses at the refined fixed point is not straightforward. For each refined estimate (\( D^* \) and \( T^* \)), we can calculate the self-supervision loss using either the detached initial estimates (\([D^* \leftrightarrow \text{detached } T_0]\) pair and \([T^* \leftrightarrow \text{detached } D_0]\) pair) or the corresponding refined estimate (\([D^* \leftrightarrow T^*]\) pair and \([D^* \leftrightarrow T^*]\)).

We observe a worse accuracy when both final estimates are paired with the corresponding initial estimates. We infer that, by pairing the final estimates with the initial ones, we impose a strong constraint on the model, limiting the scope of the output. We observe the best results when at least one of the final estimates is paired with the corresponding initial estimate. One example is when the depth loss is computed using the \([D^* \leftrightarrow \text{detached } T_0]\) pair, while the pose loss is computed using the \([T^* \leftrightarrow D^*]\) pair. From this experiment, pairing the refined estimates with each other seems to display the best accuracy. However, to ensure scale consistency with the teacher networks, we follow the third loss pairing in the table.

3. Additional results on KITTI Depth

3.1. KITTI improved depth

In Tab. 1 we present evaluation results on the improved dense ground truth \cite{19} of the KITTI \cite{7} eigen split \cite{4}. We perform garg cropping \cite{6} and report the error for distances up to 80m. Our refinement module improves the initial estimates and outperforms most previous models while still being competitive with the Transformer \cite{20}-based DepthFormer \cite{12} model.

3.2. DEQ results

In Tab. 2 we present the error for the output of our model in each DEQ iteration. Iteration 0 corresponds to the depth...

<table>
<thead>
<tr>
<th>Loss pairs</th>
<th>Test frames</th>
<th>W x H</th>
<th>Abs Rel ↓</th>
<th>Sq Rel ↓</th>
<th>RMSE ↓</th>
<th>RMSE log ↓</th>
<th>δ↓</th>
<th>δ↑</th>
<th>δ↑†</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $T_0$ $D_0$</td>
<td>1 832 x 256</td>
<td>0.123</td>
<td>0.881</td>
<td>4.834</td>
<td>0.181</td>
<td>0.860</td>
<td>0.959</td>
<td>0.985</td>
<td></td>
</tr>
<tr>
<td>2 $T^*$ $D_0$</td>
<td>1 832 x 256</td>
<td>0.120</td>
<td>0.789</td>
<td>4.755</td>
<td>0.177</td>
<td>0.856</td>
<td>0.961</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>3 $T_0$ $D^*$</td>
<td>1 832 x 256</td>
<td>0.081</td>
<td>0.484</td>
<td>3.716</td>
<td>0.126</td>
<td>0.927</td>
<td>0.985</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td>4 $T^<em>$ $D^</em>$</td>
<td>1 832 x 256</td>
<td>0.092</td>
<td>0.657</td>
<td>4.342</td>
<td>0.137</td>
<td>0.914</td>
<td>0.983</td>
<td>0.995</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The progression of the errors on the KITTI [7] Eigen split in each DualRefine iteration.

<table>
<thead>
<tr>
<th># iters</th>
<th>Abs Rel ↓</th>
<th>Sq Rel ↓</th>
<th>RMSE ↓</th>
<th>RMSE log ↓</th>
<th>δ↓</th>
<th>δ↑</th>
<th>δ↑†</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.103</td>
<td>0.726</td>
<td>4.497</td>
<td>0.181</td>
<td>0.893</td>
<td>0.965</td>
<td>0.983</td>
</tr>
<tr>
<td>1</td>
<td>0.103</td>
<td>0.702</td>
<td>4.360</td>
<td>0.179</td>
<td>0.900</td>
<td>0.967</td>
<td>0.984</td>
</tr>
<tr>
<td>2</td>
<td>0.099</td>
<td>0.700</td>
<td>4.321</td>
<td>0.174</td>
<td>0.906</td>
<td>0.968</td>
<td>0.984</td>
</tr>
<tr>
<td>3</td>
<td>0.098</td>
<td>0.695</td>
<td>4.318</td>
<td>0.175</td>
<td>0.905</td>
<td>0.968</td>
<td>0.984</td>
</tr>
<tr>
<td>4</td>
<td>0.095</td>
<td>0.690</td>
<td>4.308</td>
<td>0.174</td>
<td>0.908</td>
<td>0.967</td>
<td>0.984</td>
</tr>
<tr>
<td>5</td>
<td>0.092</td>
<td>0.673</td>
<td>4.264</td>
<td>0.172</td>
<td>0.911</td>
<td>0.968</td>
<td>0.984</td>
</tr>
<tr>
<td>6</td>
<td>0.090</td>
<td>0.658</td>
<td>4.237</td>
<td>0.171</td>
<td>0.912</td>
<td>0.968</td>
<td>0.984</td>
</tr>
<tr>
<td>7</td>
<td>0.089</td>
<td>0.653</td>
<td>4.234</td>
<td>0.172</td>
<td>0.912</td>
<td>0.967</td>
<td>0.983</td>
</tr>
<tr>
<td>8</td>
<td>0.090</td>
<td>0.653</td>
<td>4.234</td>
<td>0.173</td>
<td>0.910</td>
<td>0.967</td>
<td>0.983</td>
</tr>
<tr>
<td>9</td>
<td>0.091</td>
<td>0.655</td>
<td>4.251</td>
<td>0.174</td>
<td>0.909</td>
<td>0.966</td>
<td>0.983</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Loss pairs</th>
<th>Test frames</th>
<th>W x H</th>
<th>Abs Rel ↓</th>
<th>Sq Rel ↓</th>
<th>RMSE ↓</th>
<th>RMSE log ↓</th>
<th>δ↓</th>
<th>δ↑</th>
<th>δ↑†</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $T_0$ $D_0$</td>
<td>1 832 x 256</td>
<td>0.099</td>
<td>0.765</td>
<td>4.449</td>
<td>0.898</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 $T^*$ $D_0$</td>
<td>1 832 x 256</td>
<td>0.093</td>
<td>0.698</td>
<td>4.342</td>
<td>0.907</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 $T_0$ $D^*$</td>
<td>1 832 x 256</td>
<td>0.092</td>
<td>0.657</td>
<td>4.34</td>
<td>0.908</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 $T^<em>$ $D^</em>$</td>
<td>1 832 x 256</td>
<td>0.089</td>
<td>0.632</td>
<td>4.305</td>
<td>0.907</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3. KITTI improved depth DEQ results

We also present detailed DEQ errors in Tab. 3 and plot the Abs Rel error in each iteration on Fig. 2 for the KITTI improved depth ground truth. Similarly, our model converges around the 6th iteration.

3.4. Additional qualitative results

We illustrate through Figs. 3 and 4 additional results in the KITTI dataset. An interesting observation is how the model learns to give low confidence to vehicles and texture-less image regions. We also show in Fig. 4 how the epipolar geometry differs between the initial estimates and the refined estimates, which may cause inaccurate photometric losses.

estimates produced by the initial depth estimator. We can see that our model converges around the 6th iteration. We also plot the Abs Rel error on Fig. 1.
Table 5. Additional results on KITTI odometry test split (Seq. 11 ~ 21) using ORB-SLAM2 stereo as pseudo-GT. We provide a comparison with representative state-of-the-art self-supervised depth and odometry methods. ORB-SLAM2 is included as a representative non-learning based method.

### 4. Additional results on KITTI odometry

We perform an additional evaluation on Seq. 11-21 of the KITTI odometry dataset, using the stereo version of ORB-SLAM2 as a pseudo-GT following Zou et al. [23]. We present the average results in Tab. 5. The refinement greatly improves over the initial predictions and also displays better ATE even in comparison to ORB-SLAM2 with loop closure.

### 5. Conv-GRU Update Implementation

In our approach, we use the standard Conv-GRU block [18] to compute the updates as follows:

\[
\begin{align*}
    z_{k+1} & = \sigma(\text{CNN}_z([h_k, x_k])) \\
    r_{k+1} & = \sigma(\text{CNN}_r([h_k, x_k])) \\
    \tilde{h}_{k+1} & = \tanh(\text{CNN}_h([r_{k+1} \odot h_k, x_k])) \\
    h_{k+1} & = (1 - z_{k+1}) \odot h_k + z_{k+1} \odot \tilde{h}_{k+1}
\end{align*}
\]

(1)

where $\sigma$ represents the sigmoid activation function. Exploring other variants of the Conv-GRU block will be considered in the future.
Figure 3. Qualitative results in the KITTI [7] dataset. Top left: input image, top center: initial disparity, top right: refined disparity, middle center: initial error map, middle right: refined error map, bottom center: fixed confidence weights, bottom right: 6th iter confidence weights.
Figure 4. The epipolar line in the source image, calculated from yellow points in the target image, for the PoseNet [14] initial pose (red) and our refined pose (green). The yellow point in the source image is calculated based on our final depth and pose estimates.
References


