IMP: Iterative Matching and Pose Estimation with Adaptive Pooling

Fei Xue
Ignas Budvytis
Roberto Cipolla
Preview of IMP

Input
Two sets of keypoints

Classic pipeline
Two separate steps

Pose estimation

Output
Matches & Relative pose

Feature Matching

Ignore the geometric connections
Slow
Inaccurate
Input
Two sets of keypoints

Iterative pipeline
Matches → poses
Poses → matches

Adaptive pooling
Discard useless keypoints
- Discarded keypoints

Output
Matches & Relative pose

Retain the geometric connections
Faster
More accurate

Estimated pose
Groundtruth pose
Feature matching and pose estimation

• **Traditional approaches**
  - Two separate steps
  - Slow & inaccurate

• **Outlier filtering**
  - Promising performance
  - Accuracy limited by initial matches

• **Advanced matchers**
  - Good accuracy
  - Quadratic time cost

[1] Zhang et al., Learning two-view correspondences and geometry using order-aware network, ICCV 2019
Motivation

• Geometric connections
  • Several matches give a coarse pose
  • The pose guides the matching

• Keypoints pooling
  • Not all keypoints have matches
  • Unnecessary to update these keypoints

Progressive matching and pose estimation
More accurate matches and precise pose

• Keypoints 1024×1024
• Matches 285×285 – 27.8%
• Outliers 739×739 – 72.2%
Iterative matching & pose estimation

Input
Two sets of keypoints

Iterative pipeline
Matches $\rightarrow$ poses
Poses $\rightarrow$ matches

Adaptive pooling
Discard useless keypoints

Output
Matches & Relative pose

Discarded keypoints

Transformer-based recurrent module

Input:
- Feature Matching
- Pose estimation

Iterative pipeline:
- Matches $\rightarrow$ poses
- Poses $\rightarrow$ matches

Adaptive pooling:
- Discard useless keypoints

Output:
- Matches & Relative pose

[1] Vaswani et al., Attention is all you need, NeurIPS 2017
Transformer-based recurrent module

1. Transformer-based augmentation
   - Descriptors augmented by spatial information
   - Quadratic complexity

   \[ X(t) = X(t) + f_A(X(t), X(t)) + f_A(X(t), Y(t)) \]
   \[ Y(t) = Y(t) + f_A(Y(t), X(t)) + f_A(Y(t), Y(t)) \]

2. Cross entropy loss for matching
   - Discriminative features have high score

   \[ L_M = - \sum_{(i,j) \in M} \log(\hat{M}_{ij}) - \sum_i \log(\hat{M}_{i,n+1}) - \sum_j \log(\hat{M}_{m+1,j}) \]

3. Pose-aware loss
   - Good matches have higher score

   \[ P = f_{\text{wb}}(x_i, y_j, M_{x_i,y_j}) \]  \textit{weighted 8pt pose estimation}

   \[ L_{\text{pose}} = l_2(P, P^{\theta_t}) \]

   \[ L_{\text{geo}} = \frac{1}{n} \left( y_i^T F x_i \right)^2 \]

   \[ L_{\text{final}} = \alpha_M L_M + \alpha_{\text{pose}} L_{\text{pose}} + \alpha_{\text{geo}} L_{\text{geo}} \]

Adaptive pooling

- **Attention score tells which are inliers**
  - keypoints with high scores $\approx$ inliers

- **Our intention**
  - Keep as many inliers as possible
  - Remove as many low-contribution samples as possible

- **How to decide which one to discard**

---

**Self and cross attention scores**

**Keypoints with potential correspondences**
Adaptive pooling

- Using matching matrix as guidance

**Step 1:** samples with high matching score as seeds (inliers)

\[ X_M^{(t)}, Y_M^{(t)}, M_{X,Y} \geq \theta \]

**Step 2:** retain samples with high attention scores with guidance (keypoints with high contribution)

\[
\begin{align*}
X_A^{(t+1)} &= X_{Self}^{(t)} \cup X_{Cross}^{(t)}, S(X_{Self/Cross}) \geq md(S(X_M^{(t)})) \\
Y_A^{(t+1)} &= Y_{Self}^{(t)} \cup Y_{Cross}^{(t)}, S(Y_{Self/Cross}) \geq md(S(Y_M^{(t)}))
\end{align*}
\]

**Step 3:** merge samples with potential matches and high attention scores

\[
X^{(t+1)} = X_M^{(t)} \cup X_A^{(t+1)}, Y^{(t+1)} = Y_M^{(t)} \cup Y_A^{(t+1)}
\]

Number of keypoints: **1024 -> 496/385**
Adaptive pooling

- **Uncertainty-aware pooling**
  - Matches could be wrong due to large viewpoint changes
  - Poses reveal the quality of matches

**Step 2: retain samples with high attention scores with guidance**

\[
X_A^{(t)} = X_{Self}^{(t)} \cup X_{Cross}^{(t)}, S(X_{Self/Cross}) \geq md(S(X_M^{(t)})) * \tau
\]

\[
Y_A^{(t)} = Y_{Self}^{(t)} \cup Y_{Cross}^{(t)}, S(Y_{Self/Cross}) \geq md(S(Y_M^{(t)})) * \tau
\]

\[
\tau = \frac{|(x_i, y_i), s.t., f_{epipolar}(x_i, y_i, p_i) \leq \theta_{epipolar}|}{|(x_i, y_i) \in M^{(t)}|}
\]

Pose not accurate → matches not good → keep more samples
Pose accurate → matches good → keep fewer samples
## Quantitative results

### Training
- Megadeepth dataset from scratch without any pretraining

### Better pose accuracy
- Outdoor YFCC and Indoor Scannet datasets

<table>
<thead>
<tr>
<th>Group</th>
<th>Method</th>
<th>@5</th>
<th>@10</th>
<th>@20</th>
<th>@5</th>
<th>@10</th>
<th>@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filtering</td>
<td>NN-mutual</td>
<td>6.5</td>
<td>15.4</td>
<td>28.5</td>
<td>9.4</td>
<td>21.6</td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td>AdaLAM</td>
<td>20.8</td>
<td>36.5</td>
<td>51.9</td>
<td>6.7</td>
<td>15.8</td>
<td>27.4</td>
</tr>
<tr>
<td></td>
<td>OANet</td>
<td>19.2</td>
<td>34.5</td>
<td>50.3</td>
<td>10.0</td>
<td>25.1</td>
<td>38.0</td>
</tr>
<tr>
<td></td>
<td>CLNet</td>
<td>27.8</td>
<td>46.4</td>
<td>63.8</td>
<td>4.1</td>
<td>11.0</td>
<td>21.6</td>
</tr>
<tr>
<td>Graph-matcher</td>
<td>SuperGlue</td>
<td>37.1</td>
<td>57.2</td>
<td>73.6</td>
<td>16.2</td>
<td>32.6</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td>SGMNet</td>
<td>35.3</td>
<td>56.1</td>
<td>73.6</td>
<td>16.4</td>
<td>32.1</td>
<td>48.7</td>
</tr>
<tr>
<td>IMP</td>
<td>39.4</td>
<td>59.4</td>
<td>75.2</td>
<td>16.6</td>
<td>33.1</td>
<td>49.4</td>
<td></td>
</tr>
<tr>
<td>EIMP</td>
<td>37.9</td>
<td>57.9</td>
<td>74.0</td>
<td>15.9</td>
<td>32.4</td>
<td>48.9</td>
<td></td>
</tr>
</tbody>
</table>

Relative pose accuracy on YFCC and Scannet datasets

The **best** and **second-best** are highlighted.

### Higher speed
- IMP is faster than SuperGlue
- EIMP is close to SGMNet

---

Running time of different #keypoints
Results on Scannet dataset - case 1

Extracted keypoints

538  445
Results on Scannet dataset - case 1

Inliers/matches: 38/96, R/t error: 3.2/4.5deg
Keypoints left/right: 538/445

Inliers/matches: 34/80, R/t error: 2.6/5.3deg
Keypoints left/right: 538/445
Results on Scannet dataset - case 1

Inliers/matches: 46/93, R/t error: 1.6/1.0deg
Keypoints left/right: 538/445

Inliers/matches: 45/79, R/t error: 2.9/2.3deg
Keypoints left/right: 205/237
Results on Scannet dataset - case 1

Inliers/matches: 51/89, R/t error: 1.8/0.9deg
Keypoints left/right: 538/445

Inliers/matches: 45/80, R/t error: 2.2/0.9deg
Keypoints left/right: 205/237
Results on Scannet dataset - case 1

**IMP**

Inliers/matches: 46/93, R/t error: 1.6/1.0deg
Keypoints left/right: 538/445

**EIMP**

Inliers/matches: 45/79, R/t error: 2.9/2.3deg
Keypoints left/right: 205/237

**SuperGlue**

Inliers/matches: 8/98, R/t error: 3.2/3.5deg
Keypoints left/right: 538/445

**SGMNet**

Inliers/matches: 6/95, R/t error: 3.7/4.0deg
Keypoints left/right: 538/445

SuperGlue and SGMNet give fewer inliers, larger errors.
Results on Scannet dataset - case 2

Extracted keypoints
Results on Scannet dataset - case 2

Inliers/matches: 17/60, R/t error: 81.9/47.6deg
Keypoints left/right: 240/549

Inliers/matches: 21/46, R/t error: 3.9/3.2deg
Keypoints left/right: 240/549
Results on Scannet dataset - case 2

Inliers/matches: 17/55, R/t error: 11.9/4.1deg
Keypoints left/right: 240/549

Inliers/matches: 28/54, R/t error: 5.4/2.5deg
Keypoints left/right: 240/400
Results on Scannet dataset - case 2

Inliers/matches: 30/70, R/t error: 8.1/2.0deg
Keypoints left/right: 240/549

Inliers/matches: 27/49, R/t error: 4.9/1.8deg
Keypoints left/right: 240/381

IMP (iteration 3) EIMP (iteration 3)
Results on Scannet dataset - case 2

IMP
Inliers/matches: 30/70, R/t error: 8.1/2.0deg
Keypoints left/right: 240/549

EIMP
Inliers/matches: 27/49, R/t error: 4.9/1.8deg
Keypoints left/right: 240/381

SuperGlue
Inliers/matches: 0/1, R/t error: FAIL
Keypoints left/right: 240/549

SGMNet
Inliers/matches: 5/41, R/t error: 16.1/8.1deg
Keypoints left/right: 240/549

SuperGlue fails,
SGMNet gives much fewer inliers.
Results on YFCC100m dataset - case 1

Extracted keypoints
Results on YFCC100m dataset - case 1

Inliers/matches: 179/290, R/t error: 9.8/7.2deg
Keypoints left/right: 2000/2000

Inliers/matches: 126/235, R/t error: 7.4/5.3deg
Keypoints left/right: 2000/2000
Results on YFCC100m dataset - case 1

Inliers/matches: 266/332, R/t error: 4.8/3.5deg
Keypoints left/right: 2000/2000

Inliers/matches: 262/357, R/t error: 5.8/4.4deg
Keypoints left/right: 1167/1284
Results on YFCC100m dataset - case 1

Inliers/matches: 302/367, R/t error: 3.5/2.5deg
Keypoints left/right: 2000/2000

Inliers/matches: 274/293, R/t error: 4.2/3.1deg
Keypoints left/right: 600/677

discarded keypoints without matches
Results on YFCC100m dataset - case 1

IMP
Inliers/matches: 302/367, R/t error: 3.5/2.5deg
Keypoints left/right: 2000/2000

EIMP
Inliers/matches: 274/293, R/t error: 4.2/3.1deg
Keypoints left/right: 600/677

SuperGlue
Inliers/matches: 126/178, R/t error: 11.0/8.4deg
Keypoints left/right: 2000/2000

SGMNet
Inliers/matches: 21/73, R/t error: 11.7/8.9deg
Keypoints left/right: 2000/2000

discarded keypoints without matches

inliers spanning meaningful areas

SuperGlue and SGMNet: fewer inliers, larger errors
Conclusion and future work

• Iterative matching and pose estimation
  • Finding matches and estimating poses iteratively
  • Discarding useless keypoints dynamically

• Future work
  • Replacing traditional pose estimation with deep models