Semantic-guided RGB-Thermal Crowd Counting with Segment Anything Model

Yaqun Fang, Yi Shi, Jia Bei*, Tongwei Ren
State Key Laboratory for Novel Software Technology, Nanjing University
Introduction & Related Works

- Problem
- Related Work
- Contribution
Introduction

Problem

RGB Image

Thermal Image

estimate: 342

RGB-Thermal Crowd Counting

RGB Images

Thermal Images

Ground Truth

Density Map

(a) Darkness
(b) Thermal blurring
(c) High Quality
• Need segmentation labels
• Face difficulty in transferring between datasets with different crowd sizes

As the level of congestion increases, the incidence of missed detections becomes more pronounced.
Contribution

In this paper, we propose a novel method which utilize SAM to generate semantic map, and guide the interaction between modalities using semantic features.

(1) Our research is the inaugural effort to integrate SAM into RGB-T crowd counting. Leveraging SAM, we innovatively generate semantic maps in both RGB and thermal modalities.

(2) We employ semantic features to guide and enhance the representation of modal features within both RGB and thermal modalities. This approach significantly boosts the effectiveness of cross-modal feature fusion, leading to enhanced performance in crowd counting tasks.
Method & Experiment

- Method
- Experiment
Overview

(a) Multi-modal Feature Extraction

- RGB Image
- Thermal Image

Semantic Prompt

Grounded-SAM

(b) Semantic-guide Feature Fusion

- Feature Extractor
- Count Token

MHS

MLP

Multi-level Decoder

(c) Multi-level Decoder

- GT Count
- GT Point
- GT Density Map

Linear & Reshape

3x3 Conv

Predicted Count

Element-wise Add

Concatenation

Multi-Head Self-Attention

RGB Flow

Thermal Flow

Multilayer Perceptron
Multi-modal Feature Extraction

- RGB/T images
- Semantic maps
- Feature maps

G-SAM → Feature maps
PVT → Feature maps
**Semantic-guide Feature Fusion**

- Add semantic features to modal features
- Concat with count token
- Use Multi-head self-attention to enhance features
- Get the fused features
Method

Multi-level Decoder

- Split fused feature into RGB feature, thermal feature and count token
- Concat thermal feature with count token
- Use Multi-level Deformable Attention to integrate low-level feature
- Get the output feature
Method

Loss Function

- Output feature from decoder is split to count token and density map.
- The overall loss is composed of two parts: the loss of the density map and the loss of counting.

\[ \mathcal{L}_{total} = \mathcal{L}_{map}(D, \hat{D}) + \mathcal{L}_{count}(C, \hat{C}), \]
Dataset and Metrics

- Dataset
  - RGBT-CC Dataset (2030 pairs)
- Metrics
  - Grid Average Mean Absolute Error (GAME)
  - Root Mean Square Error (RMSE)

\[
GAME(l) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{4} |\hat{p}_i^j - P_i^j|, \quad RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{P}_i - P_i)^2},
\]
## Comparison with State-of-the-Arts

<table>
<thead>
<tr>
<th>Methods</th>
<th>Publisher</th>
<th>Year</th>
<th>GAME(0) ↓</th>
<th>GAME(1) ↓</th>
<th>GAME(2) ↓</th>
<th>GAME(3) ↓</th>
<th>RMSE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAT [28]</td>
<td>ICME</td>
<td>2022</td>
<td>12.35</td>
<td>16.29</td>
<td>20.81</td>
<td>29.09</td>
<td>22.53</td>
</tr>
<tr>
<td>DEFNet [35]</td>
<td>TITS</td>
<td>2022</td>
<td>11.90</td>
<td>16.08</td>
<td>20.19</td>
<td>27.27</td>
<td>21.09</td>
</tr>
<tr>
<td>CSCA [32]</td>
<td>ACCV</td>
<td>2022</td>
<td>14.32</td>
<td>18.91</td>
<td>23.81</td>
<td>32.47</td>
<td>26.01</td>
</tr>
<tr>
<td>CGINet [21]</td>
<td>EAAI</td>
<td>2023</td>
<td>12.07</td>
<td>15.98</td>
<td>20.06</td>
<td>27.73</td>
<td>20.54</td>
</tr>
</tbody>
</table>

Ours: **10.51** 14.52 18.92 26.28 17.71
Experiment

Visualization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GT:73</td>
<td>Estimate:8.4</td>
<td>Difference:64.6</td>
<td>Estimate:25.2</td>
<td>Difference:47.8</td>
<td>Estimate:57.3</td>
<td>Difference:15.7</td>
<td>Estimate:71.7</td>
<td>Difference:1.3</td>
</tr>
<tr>
<td>GT:37</td>
<td>Estimate:61.7</td>
<td>Difference:24.7</td>
<td>Estimate:22.8</td>
<td>Difference:14.2</td>
<td>Estimate:33.2</td>
<td>Difference:3.8</td>
<td>Estimate:35.2</td>
<td>Difference:1.8</td>
</tr>
</tbody>
</table>

(a) RGB (b) Thermal (c) GT (d) CMCRL (e) CSCA (f) LIURGBT (g) Ours
## Ablation Studies

<table>
<thead>
<tr>
<th>$S_{rgb}$</th>
<th>$S_t$</th>
<th>GAME(0) ↓</th>
<th>GAME(1) ↓</th>
<th>GAME(2) ↓</th>
<th>GAME(3) ↓</th>
<th>RMSE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
<td>11.44</td>
<td>15.43</td>
<td>19.67</td>
<td>26.70</td>
<td>20.44</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td>10.85</td>
<td>15.17</td>
<td>19.56</td>
<td>26.78</td>
<td>19.08</td>
</tr>
<tr>
<td>✗</td>
<td>✓</td>
<td>10.84</td>
<td>15.14</td>
<td>19.52</td>
<td>26.59</td>
<td>18.53</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td><strong>10.51</strong></td>
<td><strong>14.52</strong></td>
<td><strong>18.92</strong></td>
<td><strong>26.28</strong></td>
<td><strong>17.71</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy</th>
<th>GAME(0) ↓</th>
<th>GAME(1) ↓</th>
<th>GAME(2) ↓</th>
<th>GAME(3) ↓</th>
<th>RMSE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiply</td>
<td>10.92</td>
<td>14.96</td>
<td>19.51</td>
<td>26.62</td>
<td>19.70</td>
</tr>
<tr>
<td>Concat</td>
<td>10.95</td>
<td>15.30</td>
<td>19.68</td>
<td>27.10</td>
<td>18.49</td>
</tr>
<tr>
<td>Avg</td>
<td>11.48</td>
<td>15.55</td>
<td>19.96</td>
<td>27.94</td>
<td>19.30</td>
</tr>
<tr>
<td>Avg+Concat</td>
<td>11.00</td>
<td>14.76</td>
<td>19.02</td>
<td><strong>26.24</strong></td>
<td>19.33</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>10.51</strong></td>
<td><strong>14.52</strong></td>
<td><strong>18.92</strong></td>
<td><strong>26.28</strong></td>
<td><strong>17.71</strong></td>
</tr>
</tbody>
</table>
Limitation &
Conclusion

- Limitation
- Conclusion
Limitation

Failure cases

- Challenge in scenarios contain excessive crowd
- Constrain by the quality of the original image
Conclusion

In this paper, we proposed a novel semantic-guided RGB-T crowd counting method, which generates semantic maps of crowd on both RGB and thermal modalities by leveraging SAM. Our method explored the utilization of semantic features to guide and enhance the representation of modal features through the semantic-guided fusion module. With semantic information, the false-positive counting in background is reduced, while the counting accuracy in crowd regions is improved. The experiments on the RGBT-CC dataset demonstrate that our proposed method outperforms the state-of-the-art methods.
Thank You!

ICMR 2024
The 14th International Conference on Multimedia Retrieval
June 10-14, 2024
Dusit Thani Laguna Phuket, Thailand