Can we predict **where** and **what** a person is looking at?

**Where**: predict the image region on which the person is looking.

**What**: if a person is looking at an object, predict box and class of it.

- ✅ Single end-to-end method for person and gazed-object detection.
- ✅ Detect the gaze of all people in a single forward pass.
- ✅ Detect heads and objects with a single object backbone
- ❏ Predict object gaze scores for each person's gaze.
- ❏ Estimate a person's gaze in absence of objects.

---

**OUR PROPOSAL - GAZE OBJECT TARGET DETECTOR**

1. Detect and classify **objects/heads** in the image.
2. Predict the 2D/3D gaze cone (field-of-view) for each head.
3. Calculate the **probability** that an object is gazed by a person based on the gaze cone scores.
4. Model the relationships for each head-object pair.
5. Predict gaze heatmap, **object box and class**.
6. If no object is gazed, we predict a gaze heatmap from head features only.

---

**QUALITATIVE RESULTS**

**QUANTITATIVE RESULTS & THE EFFECTS OF VARIANCE IN ANNOTATIONS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Modality</th>
<th>Multiperson Gaze</th>
<th>GazeFollow AUC</th>
<th>VideoAttentionTarget AUC</th>
<th>Distance Avg.</th>
<th>Distance Min.</th>
<th>Distance Dist.</th>
<th>Distance Out.</th>
<th>Attention Target Avg.</th>
<th>Attention Target Min.</th>
<th>Attention Target Dist.</th>
<th>Attention Target Out.</th>
<th>AP</th>
<th>AP Dist.</th>
<th>AP Out.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recasens et al.</td>
<td>R</td>
<td>✗</td>
<td>0.804</td>
<td>0.812</td>
<td>0.146</td>
<td>0.124</td>
<td></td>
<td></td>
<td>0.849</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chong et al.</td>
<td>R+T</td>
<td>✗</td>
<td>0.902</td>
<td>0.812</td>
<td>0.146</td>
<td>0.124</td>
<td></td>
<td></td>
<td>0.849</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tonini et al.</td>
<td>R+D</td>
<td>✓</td>
<td>0.894</td>
<td>0.894</td>
<td>0.182</td>
<td>0.165</td>
<td></td>
<td></td>
<td>0.854</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tu et al.</td>
<td>R</td>
<td>✓</td>
<td>0.917</td>
<td>0.904</td>
<td>0.126</td>
<td>0.133</td>
<td>0.069</td>
<td></td>
<td>0.854</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>R</td>
<td>✓</td>
<td>0.922</td>
<td>0.922</td>
<td>0.102</td>
<td>0.072</td>
<td>0.033</td>
<td></td>
<td>0.944</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>R+D</td>
<td>✓</td>
<td>0.922</td>
<td>0.933</td>
<td>0.104</td>
<td>0.069</td>
<td>0.029</td>
<td></td>
<td>0.934</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Due to the low consensus across annotators, we evaluate our method under different levels of variance across individual gaze annotation.