We study the task of **visually-guided audio source separation**, i.e., given an audio mixture of multiple sound sources, the task is to separate it into its constituents using the available visual information.

We leverage pseudo-3D scene geometry information encoded via scene-graphs and directionality of the object’s motion to accomplish this.

### Prior Work

- **Gao et al. (ICCV’19):** Uses visual information but neither the visual context nor motion is leveraged for this task.
- **Zhao et al. (ICCV’19):** They incorporate object motion, but the 3D nature of the scene is not exploited.
- **AVSGS (ICCV’21):** Here the visual context of the object is incorporated into the visual representation, but the 3D geometry is not.

### Audio Separation and Motion Prediction

- We present a 2.5D geometry aware scene-graph based approach for the task of **visually guided audio source separation** called **Audio Separator and Motion Predictor (ASMP)**.
- We predict the direction of motion of the sound source, aided by appropriate visual context, to derive additional supervision for training our model.

### Model Architecture and Losses

- **Orthogonality:**
  \[ L_{orth}(Y) = \sum_{i} \left(y_i - \langle y_i, y_i \rangle \right)^2 \]
- **Consistency:**
  \[ L_{cons} = \sum_{u=1}^{U} \min_{s_{u+1}} \sum_{i=1}^{N_{i+1}} \mathbb{E} \left[ \log p_{u, i} \right] \]
- **Cyclic:**
  \[ L_{cyc} = \sum_{u=1}^{U} \left| \sum_{i=1}^{N_{i+1}} M_{i}^{u} - M_{i}^{u} \right|_1 \]
- **Direction Pred:**
  \[ L_{disp} = \sum_{u=1}^{U} \min_{s_{u+1}} \sum_{i=1}^{N_{i+1}} \mathbb{E} \left[ \log q_{u, i} \right] \]

### Experimental Analysis

#### Quantitative Study

<table>
<thead>
<tr>
<th>Approach</th>
<th>ASIW</th>
<th>AVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound of Motion (SotM)</td>
<td>6.7</td>
<td>6.1</td>
</tr>
<tr>
<td>Cyclic Co-Lean [46]</td>
<td>7.8</td>
<td>7.4</td>
</tr>
<tr>
<td>Co-Separation [13]</td>
<td>6.6</td>
<td>6.2</td>
</tr>
<tr>
<td>AVSGS [8]</td>
<td>8.8</td>
<td>8.4</td>
</tr>
<tr>
<td>ASMP (only 2.5D graph)</td>
<td>9.0</td>
<td>8.5</td>
</tr>
<tr>
<td>ASMP (2.5D graph + motion)</td>
<td>9.6</td>
<td>9.4</td>
</tr>
</tbody>
</table>

#### Qualitative Study

- We explore the efficacy of geometry-aware visual representation and motion cues for the task of visually guided audio source separation.
- We propose a novel 2.5D scene-graph representation (ASMP) towards this end and train it using weakly/self-supervised loses such as predicting the direction of motion.
- We achieve state-of-the-art results on two challenging audio-visual datasets.

### Acknowledgements

MC initiated that work at UIUC and completed it at MERL. MC was partially supported, and NA was fully supported by ONR under grant N00014- 20-1-2444, and USDA National Institute of Food and Agriculture under grant 2020-67021-32799/1024178. AC was fully supported by MERL.