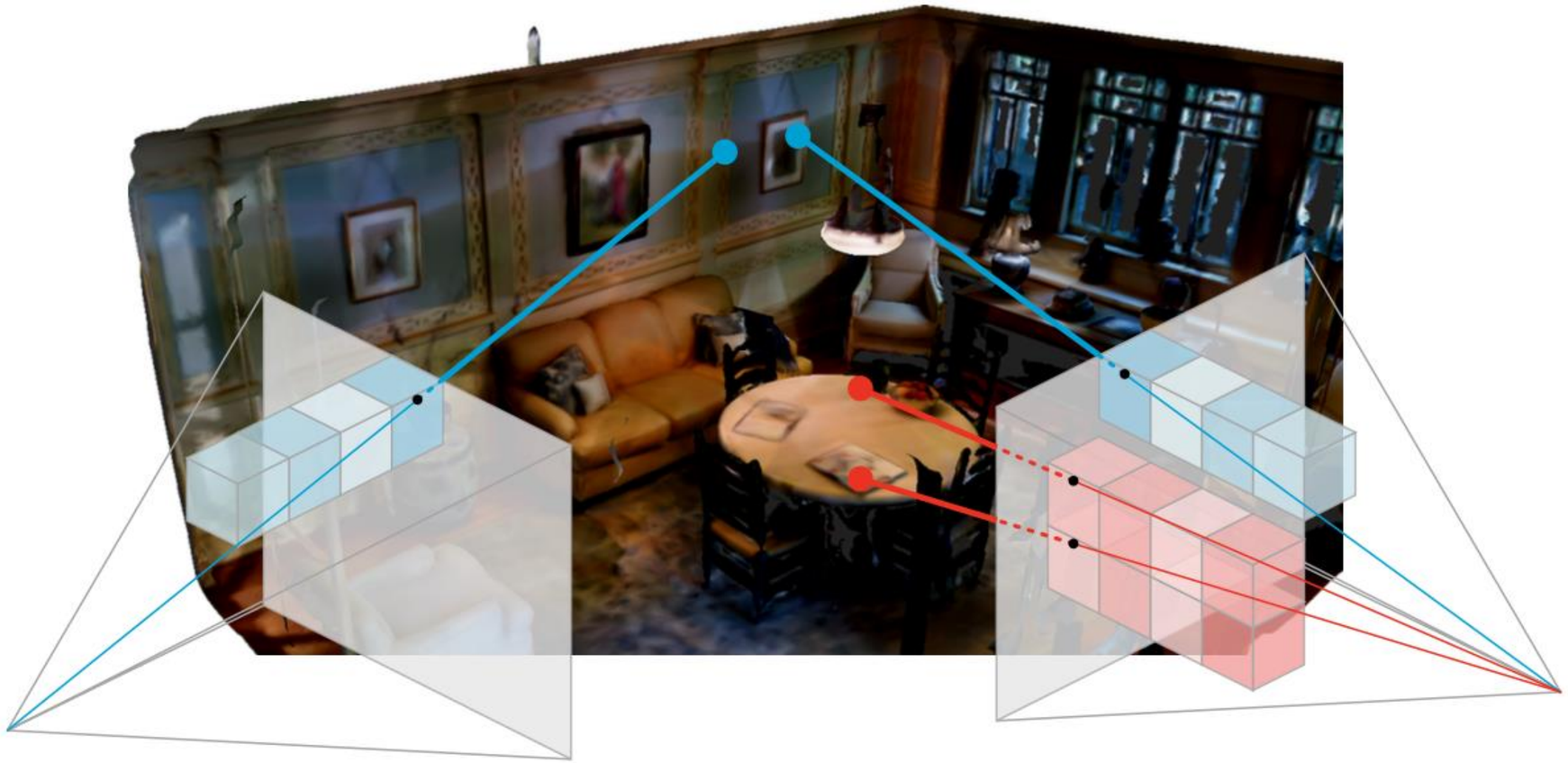
LoCo:

Learning 3D Location-Consistent Image Features with a Memory-Efficient Ranking Loss

Dominik Kloepfer (VGG), Dylan Campbell (ANU), João Henriques (VGG)

3D-Consistent Image Features

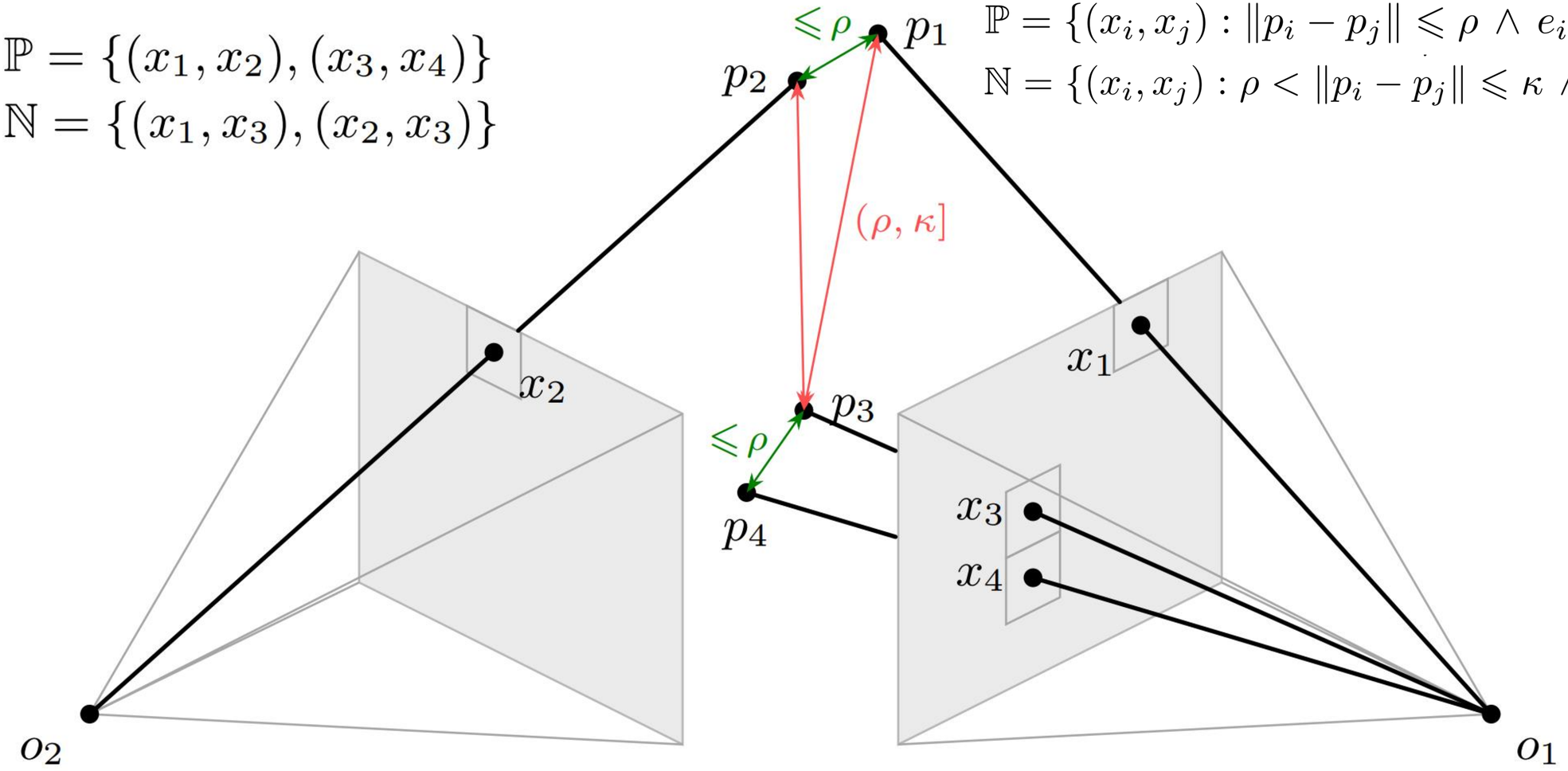


3D Consistency: same-coloured features are similar, differently coloured features are dissimilar

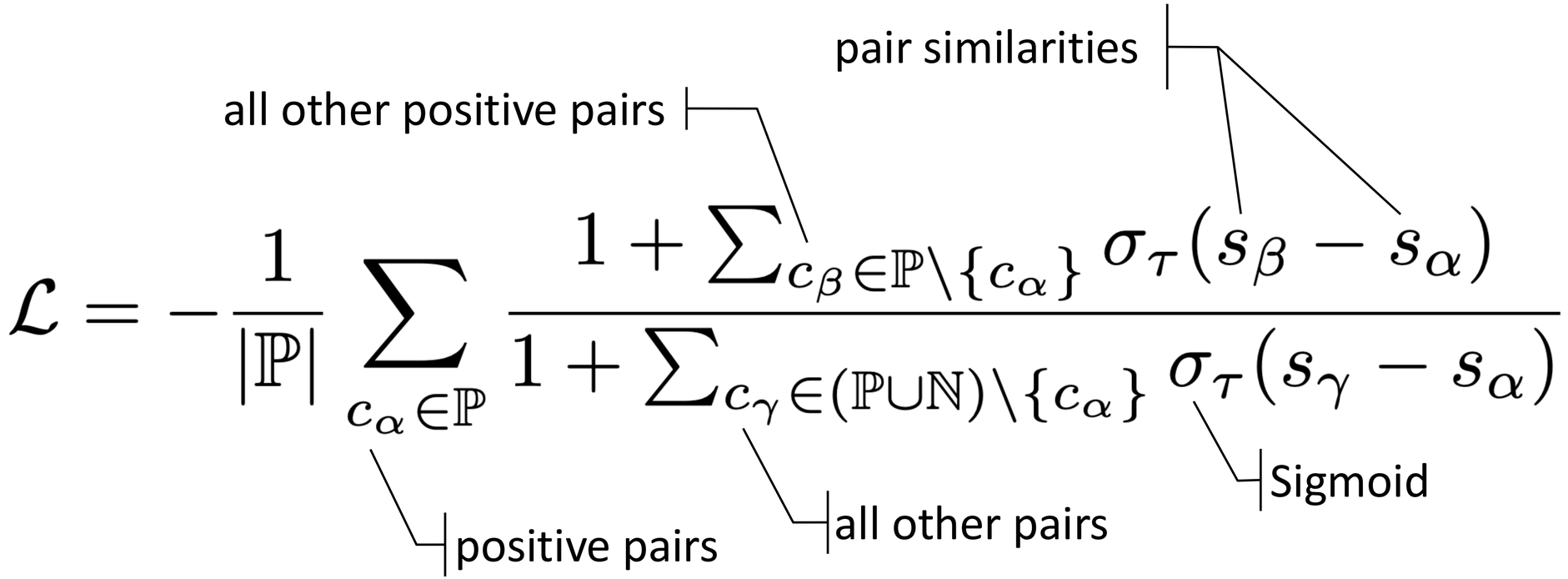
Positive & Negative Sets

$$\mathbb{P} = \{(x_1, x_2), (x_3, x_4)\}$$
$$\mathbb{N} = \{(x_1, x_3), (x_2, x_3)\}$$

$$\mathbb{P} = \{(x_i, x_j) : \|p_i - p_j\| \leqslant \rho \wedge e_i = e_j\}$$
$$\mathbb{N} = \{(x_i, x_j) : \rho < \|p_i - p_j\| \leqslant \kappa \wedge e_i = e_j\}$$



Loss Function: Smooth Vectorised AP



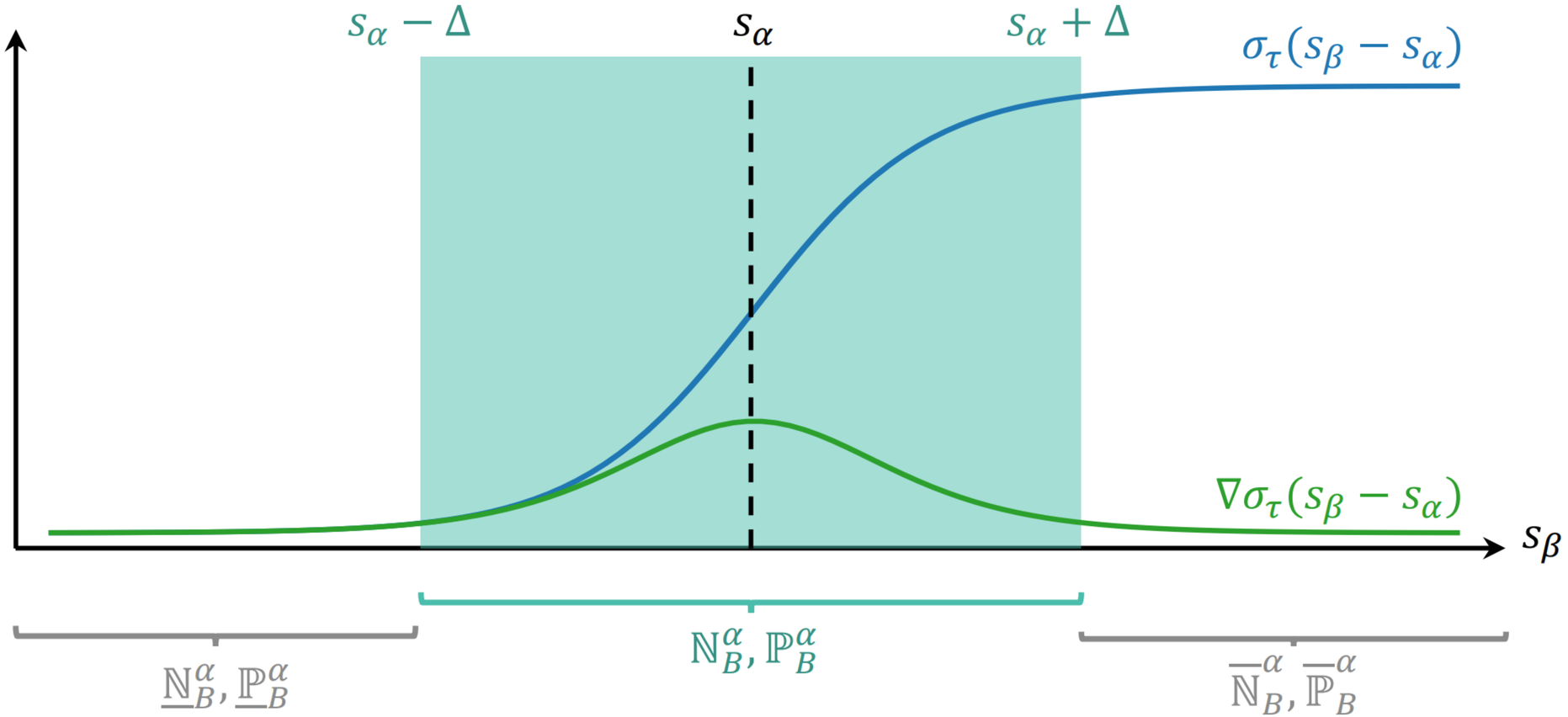
Note: Greek indices refer to ***pairs*** of patches.

Independent Positive Pair Subsets

$$\mathcal{L} = -\frac{1}{|\mathbb{P}'|} \sum_{c_\alpha \in \mathbb{P}'} \frac{1 + \sum_{c_\beta \in \mathbb{P} \setminus \{c_\alpha\}} \sigma_\tau(s_\beta - s_\alpha)}{1 + \sum_{c_\gamma \in (\mathbb{P} \cup \mathbb{N}) \setminus \{c_\alpha\}} \sigma_\tau(s_\gamma - s_\alpha)}$$

$|\mathbb{P}'| \ll |\mathbb{P}|$

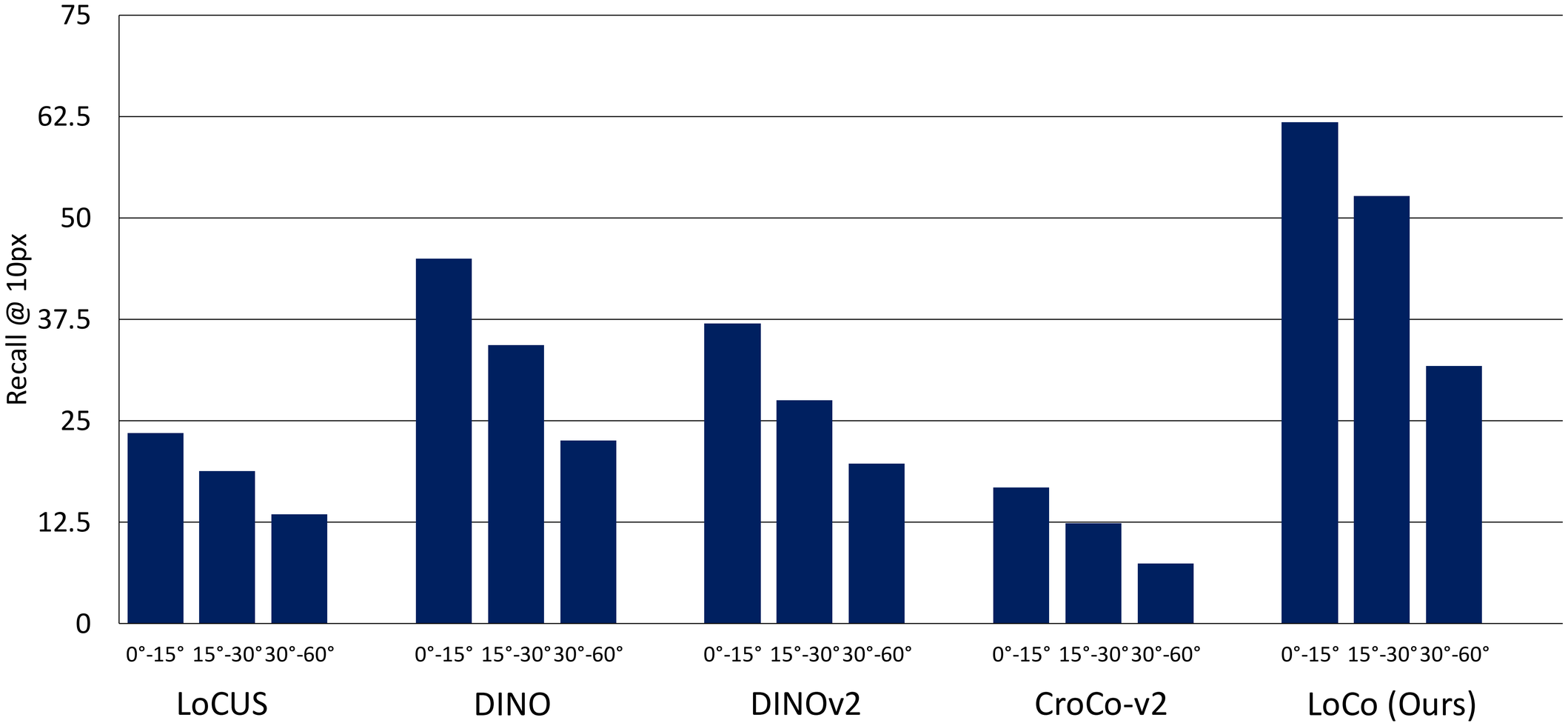
Truncate Sigmoid



Experiments

- **Model Architecture**
  - Pre-trained ViT-Base DINO model
  - Train a CNN to learn residuals
- **Training Data:**
  - 59 environments from Matterport3D

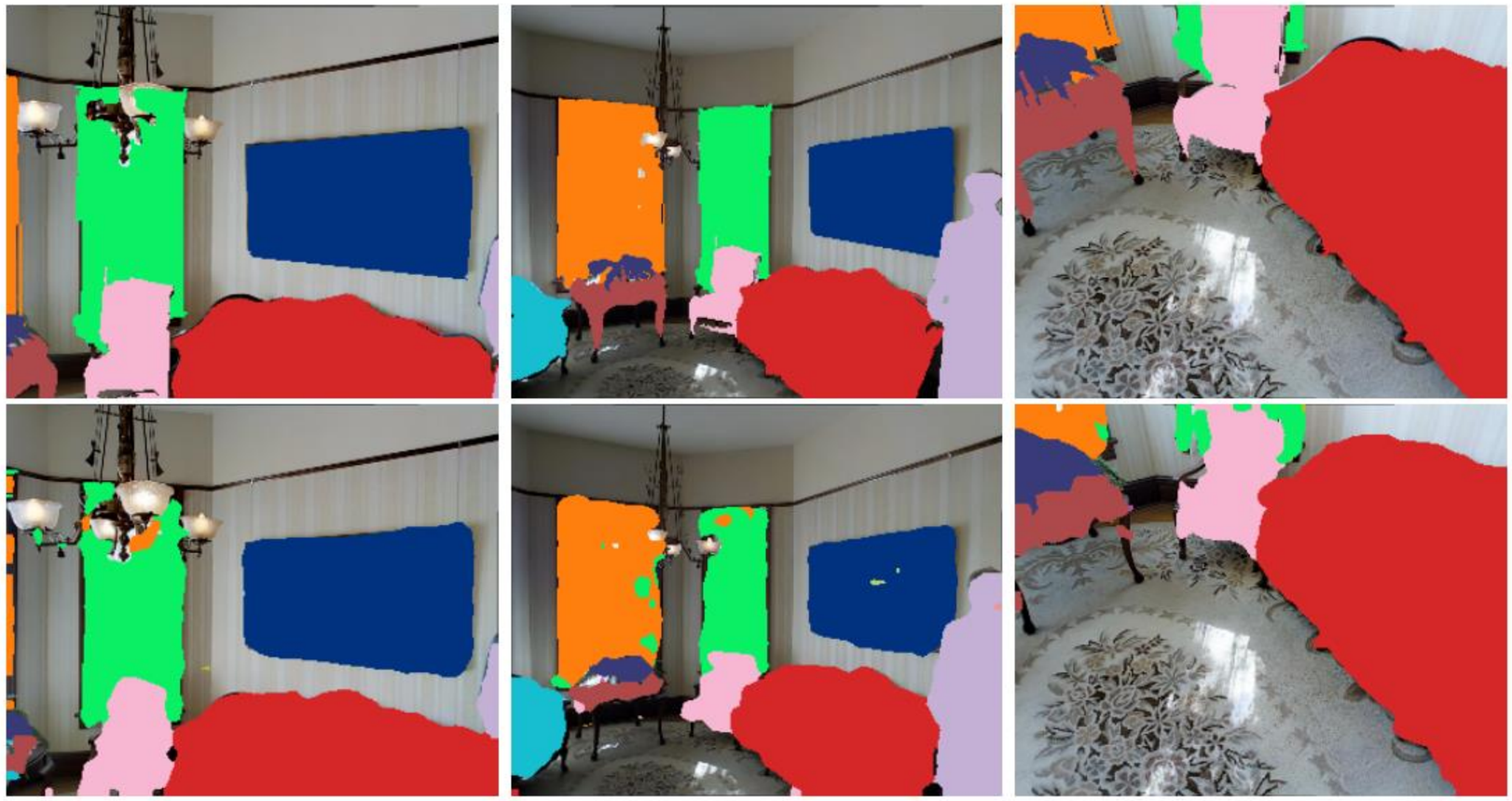
Pixel Correspondences



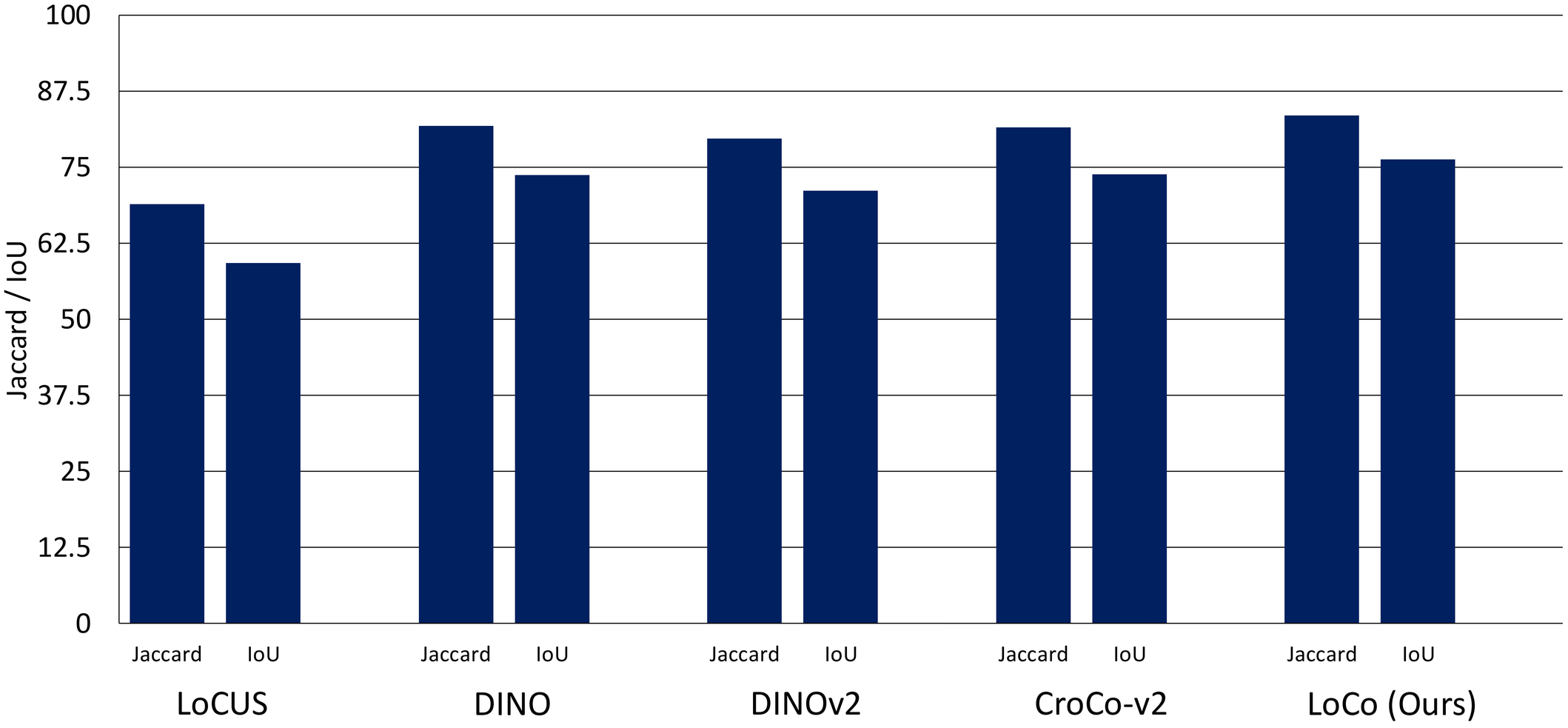
Scene-Stable Panoptic Segmentation

Ground Truth

Predictions



Scene-Stable Panoptic Segmentation



Thank you for your attention!