O-MaMa: Learning Object Mask Matching between Egocentric and Exocentric Views

Supplementary Material

Appendix A of this supplementary material explains how CMX [6] was modified and retrained for our task. On the other hand, Appendix B explains in detail the geometry baselines. We also report extra qualitative examples and attention maps for a more detailed comparison in Appendix C.

A. Implementation details of CMX

The original CMX [6] had two input images of three channels, we adapted it to input the source RGB channels concatenated with the object mask, and the destination RGB image. Since there are few examples of destination images without mask, we augment it with a probability of 10%. We use the version of [6] that incorporates Mix Transformer encoder (MiT) [5] pretrained on ImageNet [4]. In particular, we use the variant MiT-b4, changing the patch embedding input size for the source input to 4 channels. As for the decoder, we fuse the multi-level features from the backbone as done originally, and input the fused features to the XSegTx [2] decoder to generate a mask prediction. During training, we found out that freezing the pre-trained weights and training the rest worked the best, whereas fine-tuning it afterwards or training without freezing layers led to no improvement or convergence. To prevent overfitting and speeding up the epoch training time, we randomly sampled 2% of the training and validation set being different for each epoch over 25 epochs. We employ AdamW optimizer [3] with weight decay 0.01 and a starting learning rate of $1e^{-3}$, which is decreased towards 0 using a cosine scheduler.

B. Geometry Methods

Since the k-NN baseline does not take into account the mask location, objects having similar appearance can lead to false positives. For that, we decided to restrict the previous approach to fulfill the epipolar line restriction by assuming a pin-hole camera model in a self-calibration fashion. Since our epipolar constraint has been defined with a wide threshold, this assumption is good enough to detect most false positives. For this end, we used RoMa [1] to obtain the fundamental matrix F.

Given the centroid of the source mask in homogeneous coordinates $\mathbf{x}^S = (x_x^S, x_y^S, 1)$, we obtain the epipolar line in the destination image as $\mathbf{l}^D = F \cdot \mathbf{x}^S = (a, b, c)$. Then, we compute the perpendicular distance d of the most feature-similar possible masks from the centroid to its corresponding epipolar line:

$$d = \frac{|a \cdot x_x^S + b \cdot x_y^S + c|}{\sqrt{a^2 + b^2}} \tag{A.1}$$

If the distance is superior to a certain threshold, the candidate mask is discarded.

B.1. Success rate of matching methods

The success rate we use to choose the features matching method is based on an epipolar geometry criterion. With the ground truth pose and the calibration of the cameras, we verify that the matches satisfy the epipolar constraint. Even when the pose estimation is not accurate enough to obtain a precise classification of the rate of correct matches in each pair of images, the obtained result is good enough for identifying success and failure cases.

C. Extra Qualitative Results

We report extra qualitative results in Fig. A.1 and Fig. A.2, showing the best three FastSAM mask candidates predicted by our model for visualization purposes. Note that we only use the top-1 mask for reporting the official quantitative metrics. In most of the cases, the predicted mask matches with the target object, achieving a very fine-grained segmentation quality even when the source or target objects are considerably small. However, our model is also dependent on the quality of the candidate masks, showing in some cases sparse segmentations (*i.e*. the *guitar* in Fig. A.1) or inaccurate masks (the *chain* in Fig. A.2).

Finally, we visualize extra examples of the attention maps produced by our novel Ego↔Exo Cross Attention mechanism, which captures the object visual cues in the other viewpoint (Fig. A.3).

References

[1] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 1

[2] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 1

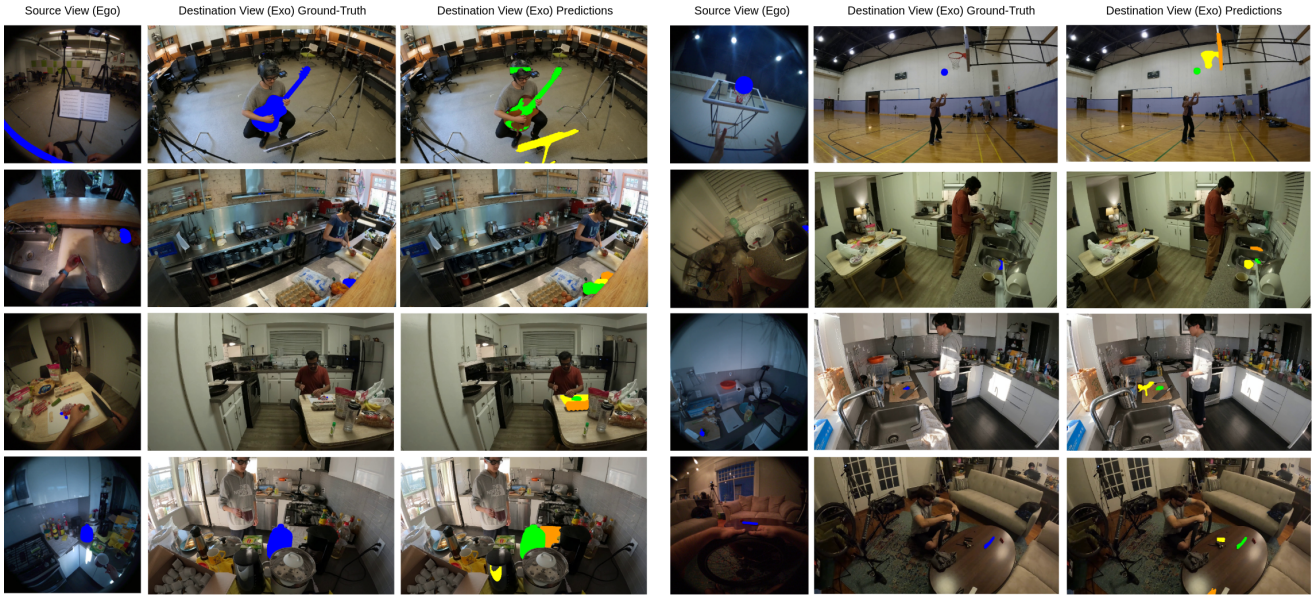

Figure A.1. **Ego2Exo Extra Qualitative Results.** For visualization purposes, we show the top 3 masks in green, yellow and orange. In blue we show the source and ground-truth masks.

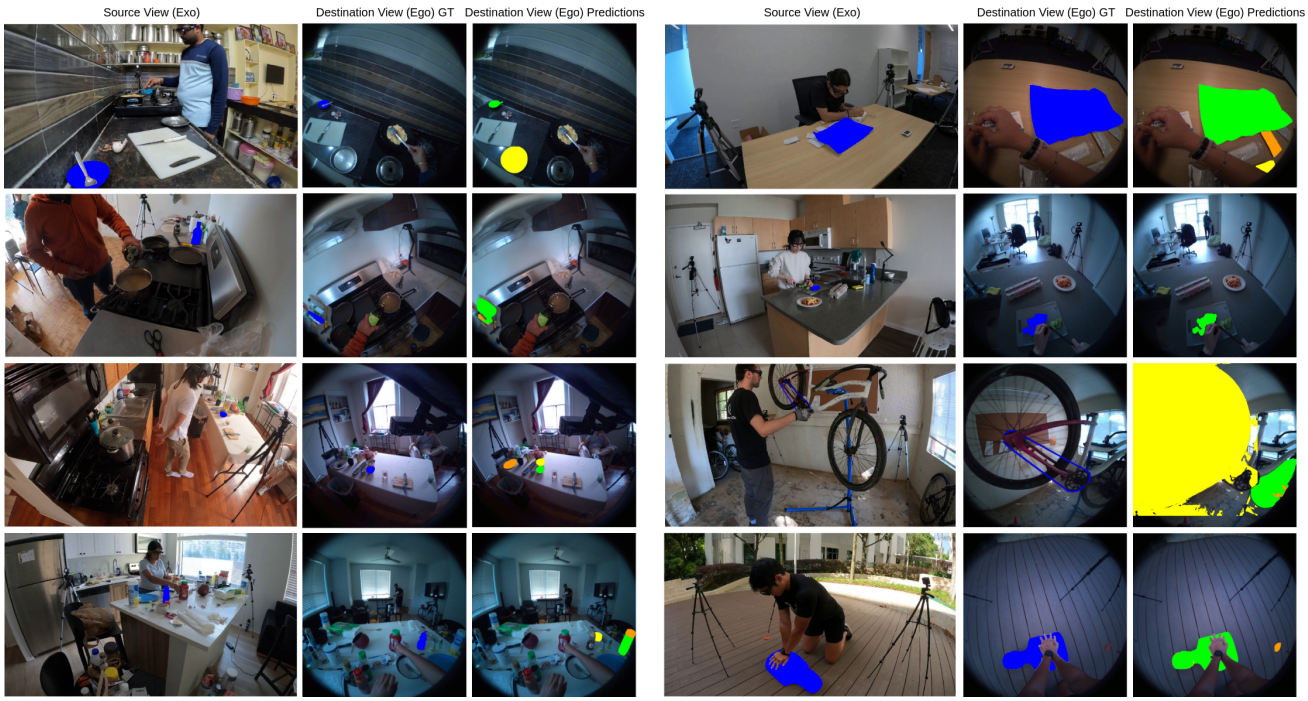

Figure A.2. **Exo2Ego Extra Qualitative Results.** For visualization purposes, we show the top 3 masks in green, yellow and orange. In blue we show the source masks.

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*, 2019. 1

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and effi-

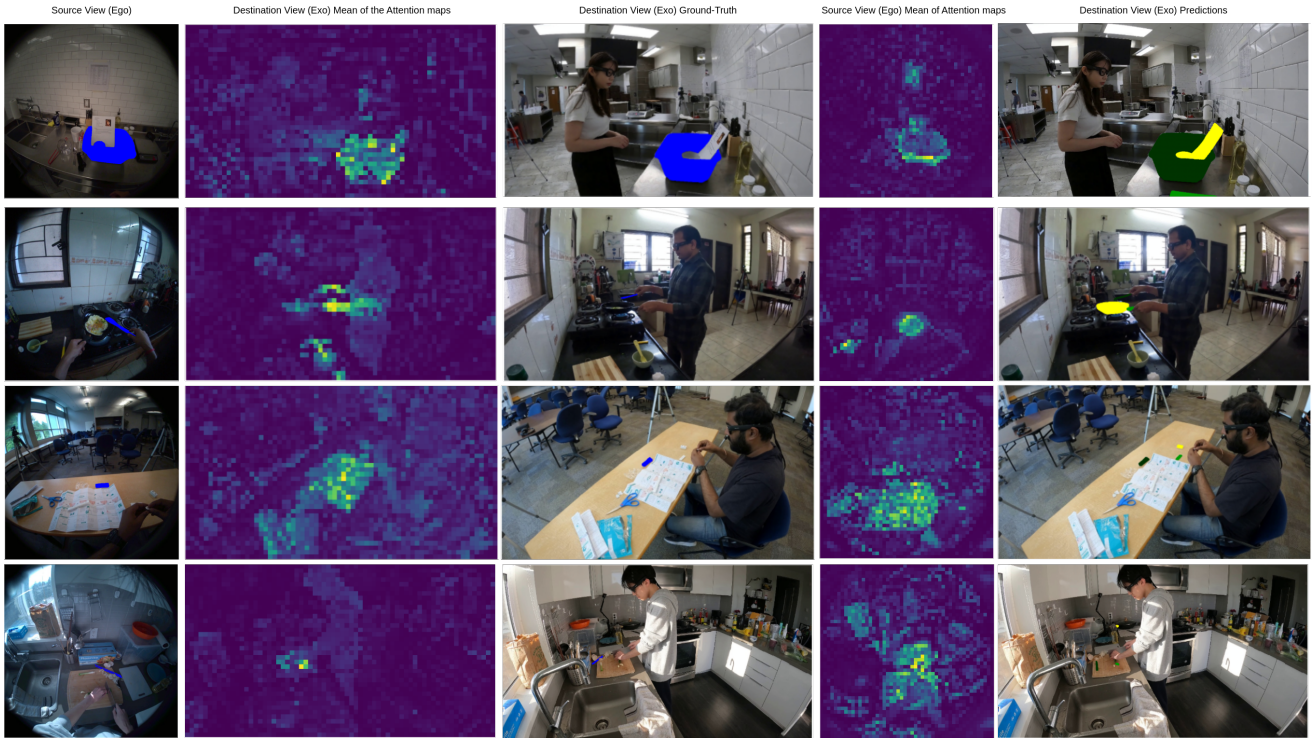

Figure A.3. **Attention maps of the Ego↔Exo Cross Attention module.** We visualize the average of the attention maps, showing how the mechanism correlates the object features from the other image perspective.

cient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 1

[6] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12): 14679–14694, 2023. 1