Layered Motion Fusion: Lifting Motion Segmentation to 3D in Egocentric Videos

Supplementary Material

The supplementary material is structured as follows. First, we show how our proposed method improves the semi-static segmentation through negative motion fusion (NMF) in Section A. Then we discuss the limitations of dynamic segmentation for egocentric videos and our attempt to address them through layered motion fusion (LMF) and test-time refinement (TR) in Section B. Next, we further investigate the precision of the motion segmentation predictions and justify their fusion in Section C and Section D respectively. Section E provides further results regarding the generalizability of our method, and Section F analyses its runtime. Section G briefly discusses broader impacts.

A. Improving semi-static segmentation

We have shown in the main paper that, besides improving the dynamic segmentation, our method also improves the semi-static segmentation as well by making use of *only dynamic* predictions of the 2D based segmentation model. We achieve this by preventing the semi-static model from predicting anything that has a dynamic (pseudo-) label assigned. We highlight the benefits in Figure 1. We see that NMF removes artefacts from the semi-static layer such as in row three. Additionally, NMF also reduces predictions of dynamic objects such as the cutting board in row one and the person in row two. As can be seen, such results cannot be achieved by training with PMF only, as this loss does not influence the semi-static layer directly.

B. Limitations of dynamic segmentation

The authors of the EPIC Fields [80] dataset observed that the performance of state-of-the-art dynamic neural rendering methods strongly depends on the type of motion. Their results show a significant gap in the reconstruction quality between the dynamic and the static parts of videos, pointing to the current limitations when handling dynamic objects. This limitation also appears in the segmentation of dynamic objects: While the 3D-based models from EPIC Fields outperform MG as a 2D baseline on the semi-static setting, they all perform worse when used for the segmentation of dynamic objects. For example, they report a mean Average Precision (mAP) score of 55.58 for NeuralDiff, while MG achieves 64.27 – a significant gap of about 15%.

We hypothesize that this difference is caused by two factors. First, we observe that the semi-static model sometimes captures dynamic objects as shown in Figure 3. The figure shows that the model reconstructs the scene well in comparison to Figure 2, but it assigns the person and the object they are holding incorrectly to *the semi-static layer*. To circumvent this problem, we make use of labels predicted from a 2D motion segmentation method, and fuse them into the 3D model. This procedure regularises both layers (semi-static and dynamic) of the model and helps specifically the dynamic layer to learn about moving objects. Second, the fusion of motion segmentation predictions can only be achieved if the geometry is captured correctly. This means that, for example, if the arms are not captured by the radiance field, it is impossible to fuse motion into them as their geometry is missing. Examples of failure cases are shown in Figure 2. The red rectangles highlight the reconstruction of dynamic objects and show a significant mismatch between the predicted RGB image and the ground truth. This in turn leads to a deterioration of the segmentation ability, since the 3D based model can only segment what it can reconstruct. We address this issue through test-time refinement.

Aside from the challenges posed by dynamic segmentation with geometry, the task also proves to be a significant hurdle for 2D-based *supervised* methods, such as EgoHOS [99], as demonstrated in Tab. 4. While this model is trained to segment hands and objects with specific supervision, it achieves a lower score compared to our method. More recent papers provide further evidence of the difficulty of segmenting dynamic objects in egocentric videos such as [2] and [63].

C. Precision of motion segmentation output

We observe that the motion segmentation predictions are unbalanced in terms of the false positive rate (FPR) and false negative rate (FNR). An increase in FPR indicates that the model is incorrectly labeling more negative instances as positive, which would decrease the precision since it is negatively impacted by an increase in FP (more false positive predictions reduce the fraction of true positive predictions among all positive predictions). While a high precision is generally desirable for the fusion of labels, the negative fusion loss benefits in particular from it, since the semi-static model can only learn to ignore dynamic objects that are actually positives. An analysis of the error rates can be found in Figure 5 and a qualitative visualization of the precision of the motion segmentation predictions can be found in Figure 4.

D. Fusion of motion segmentation predictions

The work from [101] has shown that noisy or sparse labels can be fused into 3D space through neural rendering in static scenes. They observe that the accuracy of the fusion decreases with a significant increase in noise and sparsity.

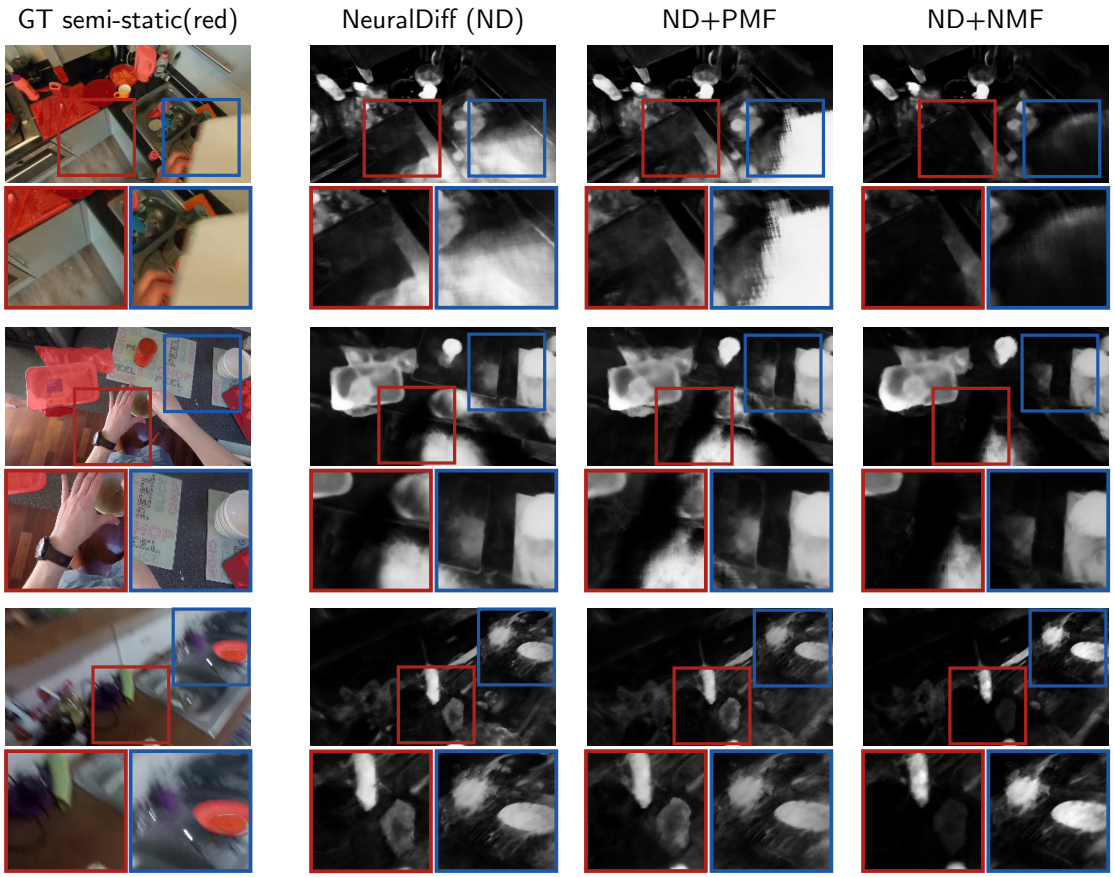

Figure 1. **Qualitative semi-static results for motion fusion**. The segmentations are produced by NeuralDiff (ND), ND + PMF, and ND + NMF. The positive motion fusion (PMF) loss does not prevent the semi-static layer from predicting dynamic objects. In comparison, the negative motion fusion (NMF) loss removes artefacts from the semi-static predictions such as in row three. It especially removes anything dynamic such as the cutting board in row one or the parts of the person and the bowl they are touching in row two.

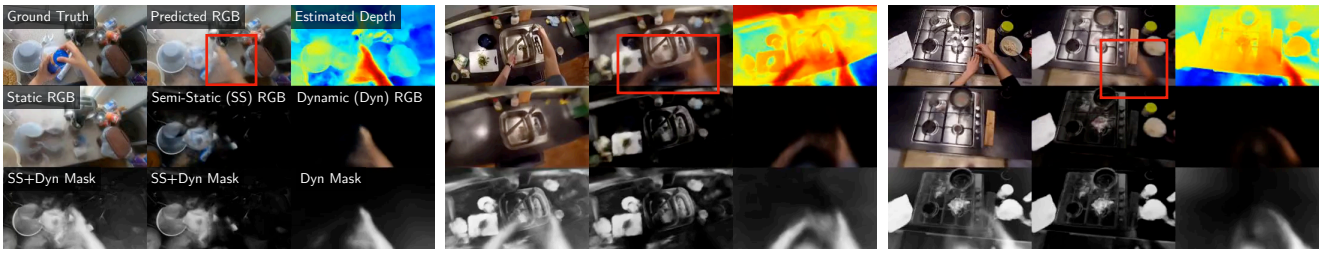

Figure 2. **Missing geometry when segmenting dynamic objects**. While the model is able to segment the semi-static components of the scene, the dynamic one is rendered with less accuracy. This lack of geometric understanding of the dynamics of the scene hurts segmentation, as the model can only segment objects whose geometry is captured.

For the case of fusing motion segmentation masks from a 2D-based model, such as MG [93], into a dynamic neural rendering representation, we therefore require labels with high precision. We analyze the FPR rate in Figure 5 and note that MG rarely predicts positives incorrectly for varying thresholds. This observation is not only important for the fusion of labels into a single layer (similar to Semantic-NeRF [101]), but even more so when fusing them into two layers as proposed in our method. While the dynamic layer simply learns to predict the dynamic labels from the motion segmentation masks, the semi-static layer is penalized for predicting anything that should belong to the dynamic layer. As the semi-static layer learns to exclude the predictions from the 2D-based segmentation model, the dynamic

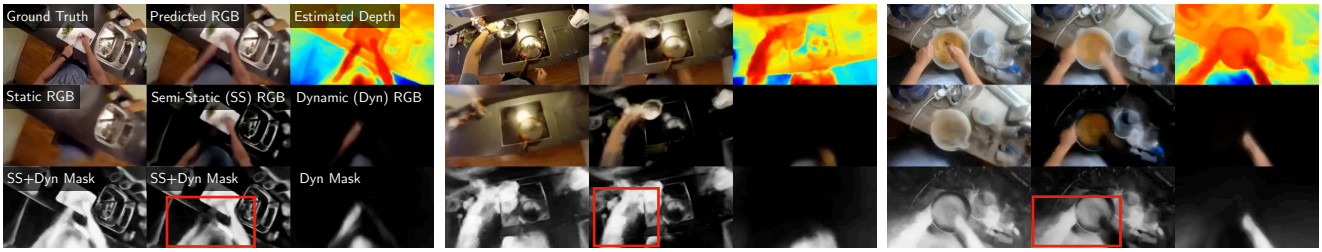

Figure 3. **The semi-static layer captures dynamic objects incorrectly**. We observe that the model segments dynamic components (the person and the objects they are holding) incorrectly into the semi-static stream. This represents a significant limitation of dynamic neural rendering methods. It can be resolved with the proposed layered motion fusion that integrates the predictions of a 2D based segmentation into its 3D representation.

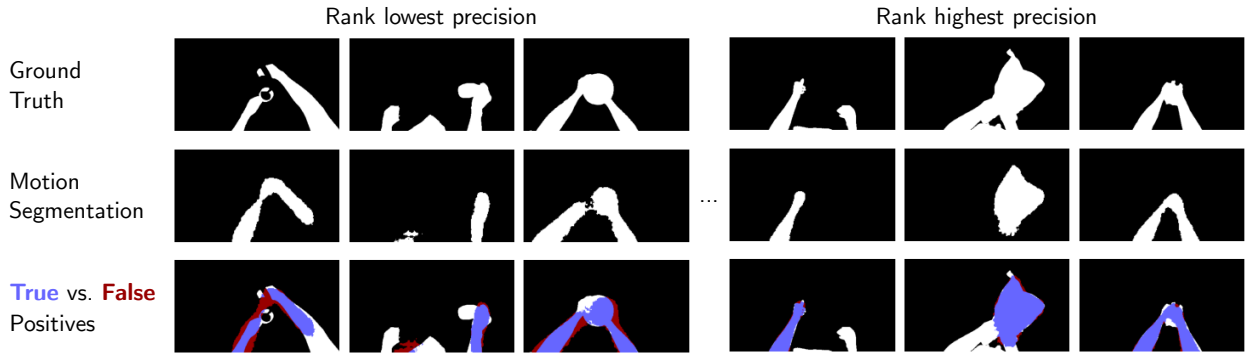

Figure 4. **Precision of motion segmentation predictions**. The motion segmentation predictions have unbalanced error probabilities for egocentric videos. We observe that the motion segmentation model is less likely to produce false positives and has a high precision as shown in the ranking of samples with respect to their precision. We rank segmentations with the lowest precision from right to left in the first three columns and those with the highest precision from left to right in the last three columns.

layer is forced indirectly to predict them instead. This results in a higher overall confidence either for the semi-static model *or* the dynamic model. In comparison, applying only one of the losses can result in predictions of the semi-static and dynamic layer that have equal confidence (probability of 0.5), as the segmentation itself depends on the rendering with multiple layers as defined in the rendering equations from [78]. Another positive side-effect of the layered motion fusion is that the semi-static model improves as well, as shown in the results, by making use of *only dynamic* labels from the 2D-based model. Qualitative examples are visualized in Figure 1.

E. Generalizability of semi-static component

The methods NeRF-W [48] and NeRF-T [21] are not designed for segmenting semi-static objects, and therefore lack a semi-static layer. This results in a lower performance, because we cannot apply NMF. We chose the architectures from EPIC Fields [80] for a fair comparison, but extend them to show the generalizability of our method. We add a semi-static layer similar to that used in NeuralDiff. The results, reported in Table 6, are extensions of the experiments from Table 2 and Table 5 in the main paper. We observe that the semi-static layer improves the performance on its own. Adding NMF improves the performance of NeRF-W and NeRF-T for all types of motion (including semi-static) even further, similar to its application in NeuralDiff.

Table 6. **Application to different 3D baselines.** We report the mean average precision (mAP) on segmenting the dynamic (Dyn) and semi-static (SS) components of the scene, and also their union (SS+Dyn). We modify NeRF-W and NeRF-T to a three-layer architecture (indicated by *), which enables us to apply LMF (*i.e.* PMF+NMF) as opposed to PMF only (Table 2 from main paper).

Method	Dyn	SS	Dyn+SS
NeRF-W [48]	28.5	20.9	45.6
+ TR + PMF	34.2 (20.0%)	19.8 (-5.3%)	47.3 (3.7%)
+ TR + PMF *	34.5 (21.1%)	21.1 (+1.0%)	48.2 (5.7%)
+ TR + PMF + NMF *	36.6 (28.4%)	21.6 (+3.3%)	49.4 (8.3%)
NeRF-T [20]	44.2	24.4	64.9
+ TR + PMF	51.1 (15.6%)	23.2 (-4.9%)	68.8 (6.0%)
+ TR + PMF *	52.9 (19.7%)	24.7 (+1.2%)	69.2 (6.6%)
+ TR + PMF + NMF *	56.1 (26.9%)	25.5 (+4.5%)	69.9 (7.7%)

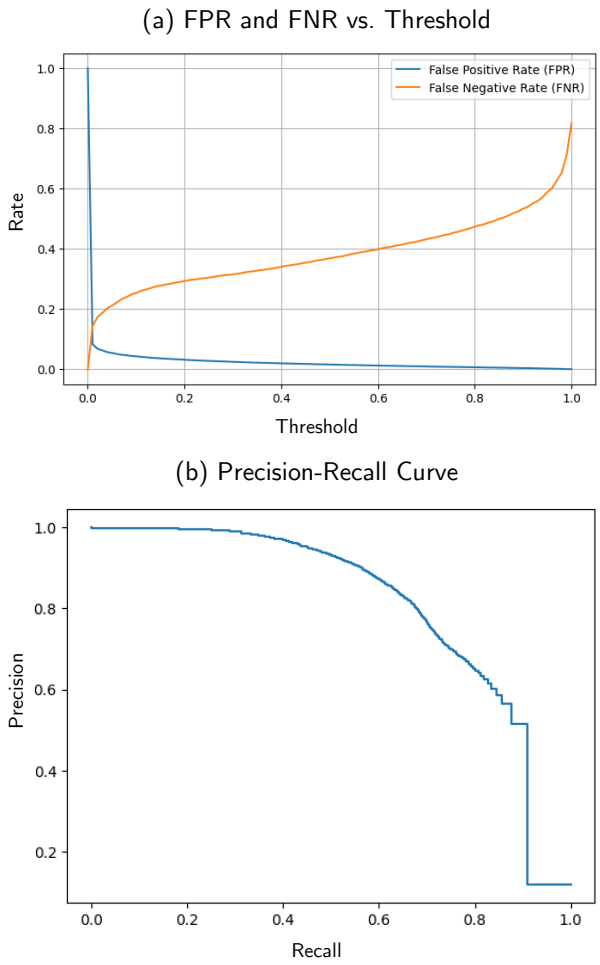

Figure 5. **Analysis of true positives from motion segmentation predictions.** (a) We observe that the probabilities from the motion segmentation method are unbalanced in terms of the false positive rate (FPR) and false negative rate (FNR). The model is rarely predicting positives incorrectly for varying thresholds. (b) The precision-recall curve inclines towards the top right and suggests that the motion segmentation method is highly effective in differentiating between the positive and negative classes.

F. Runtime analysis

The fine-tuning takes about 22 minutes for 100 frames or about 13 seconds per frame. The rendering of a frame without fine-tuning takes about 5 seconds. Therefore, the required time for rendering would increase to 18 seconds with the fine-tuning. Given that, our method improves the mAP score by up to 30%. The runtime could be significantly reduced to a fraction of its current value by adopting a more advanced architecture, such as Gaussian Splatting [29], though this would require a non-trivial adaptation.

G. Broader impact

This method, because of potential applications in augmented reality, could have some positive impact as it could be used within an AI assistant. There are also potential drawbacks to such technologies, since improved AR technologies could potentially be exploited for deceptive purposes.