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

Vadim Tschernezki[1,2] Diane Larlus[2] Iro Laina[1] Andrea Vedaldi[1]

[1]Visual Geometry Group
University of Oxford
{vadim,iro,vedaldi}@robots.ox.ac.uk

[2]NAVER LABS Europe
diane.larlus@naverlabs.com

Abstract

Computer vision is largely based on 2D techniques, with 3D vision still relegated to a relatively narrow subset of applications. However, by building on recent advances in 3D models such as neural radiance fields, some authors have shown that 3D techniques can at last improve outputs extracted from independent 2D views, by fusing them into 3D and denoising them. This is particularly helpful in egocentric videos, where the camera motion is significant, but only under the assumption that the scene itself is static. In fact, as shown in the recent analysis conducted by EPIC Fields, 3D techniques are ineffective when it comes to studying dynamic phenomena, and, in particular, when segmenting moving objects. In this paper, we look into this issue in more detail. First, we propose to improve dynamic segmentation in 3D by fusing motion segmentation predictions from a 2D-based model into layered radiance fields (Layered Motion Fusion). However, the high complexity of long, dynamic videos makes it challenging to capture the underlying geometric structure, and, as a result, hinders the fusion of motion cues into the (incomplete) scene geometry. We address this issue through test-time refinement, which helps the model to focus on specific frames, thereby reducing the data complexity. This results in a synergy between motion fusion and the refinement, and in turn leads to segmentation predictions of the 3D model that surpass the 2D baseline by a large margin. This demonstrates that 3D techniques can enhance 2D analysis even for dynamic phenomena in a challenging and realistic setting.

1. Introduction

Forty years ago, pioneers like Marr [47] argued that 3D representation should be a foundation of computer vision. However, the field has since evolved differently, reducing image and video understanding to 2D pattern recognition, first with the introduction of visual representations like bags of visual words [11, 45, 62, 70] and then learned ones [13, 22, 24, 33, 69, 81].

Recently, several authors have shown that 3D representations can, if not replace, at least enhance 2D techniques. An approach is to *fuse* 2D information extracted from individual views of a scene into a coherent 3D reconstruction. Examples include methods like Semantic NeRF [101], N3F [79], DFF [32], LERF [30], GARField [31]. These techniques take various types of 2D outputs (*e.g.*, semantic or instance segmentation, image or text features) and project them into a coherent 3D reconstruction while also removing noise and compressing information.

While these methods show that 3D *can* improve 2D processing, so far this has been demonstrated only in restricted settings; most of these methods require multi-view images of a *static* scene to work. However, in real applications, the scene itself is often dynamic, and, arguably, the dynamic part is often the most interesting one. Thus, 3D techniques need to be able to handle dynamic content in order to be useful in practice. As 3D methods mature, however, we can expect them to become useful in these more challenging scenarios as well, particularly when the scene content is dynamic *and* the camera moves significantly. A great example of such a scenario is egocentric videos, in which a camera is attached to a person as they navigate and interact with their environment, resulting in camera movement that mirrors their actions within the scene.

A few authors have explored this challenging scenario in the broader context of monocular video understanding [43, 78, 88]. Out of those, NeuralDiff [78] focuses on monocular egocentric videos. It uses an architecture based on neural radiance fields (NeRF) [53] to decompose these videos into three layers: a static background and two foreground layers, respectively modeling the semi-static part (objects that are currently stationary, but move at some point in the video) and the dynamic part (objects that presently move) of the scene. More recently, EPIC Fields [80] conducted a systematic study of 3D techniques for long egocentric videos, introducing a large annotated dataset and performing comprehensive empirical evaluations of existing techniques for novel view synthesis and scene decomposition into static, semi-static and dynamic objects. One of the main findings of that study is that *the performance of*

3D methods lags behind that of the 2D baseline in terms of segmenting dynamic components.

Inspired by recent developments in the distillation of features [30, 32, 79] and the fusion of labels [35, 68, 101] into 3D representations, we aim to improve the 3D segmentation by fusing the motion of a 2D-based method into a 3D representation. This leads us to our key question: *can a 3D representation improve the performance of a 2D neural network in understanding dynamic phenomena?* To answer it, we consider a strong unsupervised 2D-based motion segmentation method, Motion Grouping (MG) [93], and suggest to fuse its predictions into our *dynamic* 3D model. We observe that while such models capture only parts of the dynamic objects (making them incomplete), they are precise and therefore suitable for being fused into a 3D representation as they are similar to "sparse" (incomplete, but precise) labels as in Semantic-NeRF [101]. To do so, we develop a new *motion fusion* technique based on a layered representation of dynamic NeRFs. We first show that the fusion of the motion segmentation predictions into the dynamic layer already results in significant improvements of the segmentation capability. Additionally, we find that we can further improve the 3D model by regularizing the semi-static layer with the *same* segmentation predictions from the 2D-based model. We achieve this by penalizing the semi-static layer from predicting anything that the motion segmentation model thinks is dynamic. These constraints work in synergy to enhance the overall performance of the model. Because we fuse motion into both the semi-static and dynamic layers, we refer to this method as Layered Motion Fusion (LMF).

While the fusion of motion improves segmentation significantly, we find that this process is still limited by the high data complexity of egocentric videos. Specifically, we observe that 3D-based models can only fuse motion segmentation predictions into the geometry as they have learned it from the scene. If the scene's geometry is too complex — such as in long, complex egocentric videos — the model fails to capture it and, in turn, cannot fuse motion into the (missing) geometry. To address this issue, we suggest considering a setting where test-time adaptation [42] and test-time refinement [26] can be applied. We show that fine-tuning the model to the subset of frames we wish to analyze enhances the model's ability to capture the scene's geometry, thereby allowing it to fuse motion more accurately.

In summary, our contributions are as follows: (1) We propose a new motion fusion technique for layered radiance fields that boosts the segmentation of dynamic objects in egocentric videos by a large margin. Additionally, we observe improvements in the segmentation of semi-static objects, highlighting the benefits of a layered fusion. To our knowledge, this represents the first attempt at fusing motion segmentation into dynamic radiance fields. (2) We propose test time refinement to further boost segmentation performance by focusing the optimization on selected frames to reduce scene complexity. (3) We solve the issue observed in [80] of the inferior results of 3D models compared to 2D ones for unsupervised dynamic object segmentation, and show that our proposed method outperforms all results reported in [80]. This underscores the potential of 3D computer vision techniques to *enhance* the performance of 2D video understanding methods, even in challenging, highly dynamic scenarios. We hope that this finding will encourage others to explore 3D vision for understanding this data, where we believe it has considerable potential.

2. Related work

NeRF and dynamics. Neural Radiance Fields (NeRFs) were introduced in [53] as a way to synthesize novel views in 3D scenes. Initially restricted to static scenes, several methods have extended NeRF to dynamic scenes. There are two main approaches. The first one adds time as an additional dimension to radiance fields [7, 17, 19, 38, 48, 78, 86, 89]. The other one explicitly learns a 3D flow and reduces the reconstruction to a canonical (static) one [14, 16, 23, 39, 40, 44, 61, 64, 71, 77, 85, 95, 97]. As noted in [80], dynamic neural rendering has been mostly applied to synthetic or simple environments (small camera baseline and sequence lengths up to 60 seconds). To encourage the use of dynamic neural rendering in more complex and realistic environments, the EPIC Fields benchmark [80] was proposed. It consists of long and complex egocentric video sequences and is associated with difficult scene understanding tasks. This benchmark highlights the unresolved issues of recent NeRF-like methods in rendering dynamic parts of scenes in long videos. Our work explicitly addresses these challenges.

NeRF and semantics. Other research has focused on the potential of neural rendering to enhance the semantic understanding of the 3D model of a scene. For example, Semantic NeRF [83, 101] fuses semantic labels with static scenes. Others explore panoptic segmentation [3, 18, 35, 68, 84], which extends semantic segmentation with the ability to differentiate between instances of the same class. Besides integrating labels into a 3D scene representation, a related line of work [30, 32, 43, 79] has proposed to enhance NeRFs with a separate prediction head that captures semantic *features* from pre-trained ViTs [8, 37, 76] or vision-language features from CLIP [66]. This merges the open-world knowledge from 2D models into 3D scene representations and extends them to applications such as the retrieval or editing of objects inside of scenes [32, 79]. The authors of [93, 100] even capture the semantics of a scene through the decomposition of objects of this scene with individual radiance fields. Other related methods

[15, 54, 59, 67, 78, 88, 91, 96] rely on neural rendering techniques to distinguish between objects and backgrounds with no or minimal supervisory signals. We specifically focus on the segmentation of dynamic objects, which has only been explored by few works [43, 78, 88].

Distillation of 2D models into 3D. While most work related to 'NeRF and semantics' directly use labels, another option is to integrate semantic knowledge through the distillation of models designed to receive 2D images as input into 3D representations. This idea has been explored already before NeRF and is known as multi-view semantic fusion methods [25, 46, 49, 52, 73, 82]. Similar to approaches such as Semantic NeRF [101] and Panoptic Lifting [68], these methods combine multiple, potentially noisy or partial, 2D semantic observations and re-render them to obtain clean and multi-view consistent labels. Other research incrementally builds semantic maps using SLAM [34, 56, 74]. The advent of NeRF has increased the application of distillation for 3D representations as shown in DFF [32], N3F [79], and LERF [30]. These methods apply 3D fusion directly to supervised and unsupervised dense features in order to transfer semantics into the 3D space. This benefits applications such as scene editing, object retrieval and zero-shot 3D segmentation. More recent approaches such as [1, 65, 102, 103] extend these ideas to the more efficient Gaussian Splatting [29] rendering technique that significantly speeds up training and rendering, which in turn makes the fusion of labels/features more easily applicable. Similarly, we distill 2D knowledge from a model, but with the difference that our 2D model is specialized in motion segmentation.

Motion and object segmentation. As shown in [78], objects can be segmented in dynamic videos without supervision by combining motion cues with a NeRF-like architecture. This can also be done using standard 2D approaches such as background subtraction [5] or motion segmentation [50, 58, 98]. The latter typically requires optical flow, which is subject to ambiguities [55] and only reasons locally. Such approaches are also prone to errors in the presence of occlusions or if dynamic objects temporarily remain static [50, 94, 98]. The mechanism behind motion segmentation has been extended to the segmentation of specific objects [4, 27, 60, 75, 90]. In [4] a probabilistic model acts upon optical flow to segment moving objects from the background. In [6, 72] pixel trajectories and spectral clustering are combined to produce motion segments, while [51] reconstructs urban scenes and discovers their dynamic elements such as billboards or street art by clustering 3D points in space and time. More recent work [10, 36, 41, 87, 90, 92] such as Motion Grouping (MG) [93] combines classical motion segmentation with deep learning. Other examples include [41], which learns an image segmenter in the loop of approximating optical flow with constant segment flow and then refines it for more coherent appearance and statistical figure-ground relevance, and [92], which segments moving objects via an object-centric layered representation. Our method can be combined with these techniques and refine them with dynamic neural rendering, resulting in cleaner and more accurate motion segmentation masks.

Test-time refinement. The complexity of egocentric videos can be approached by enabling the model to focus on specific parts of the video at test time. Concretely, we are interested in improving the segmentation of *specific frames* that the model receives as input. This is closely related to test-time adaptation [42], which aims at adjusting a pre-trained model as test data becomes available. Predictions are made using the adjusted model. Similarly, we can adapt 3D models to user selected frames and use their corresponding predicted motion segmentation as pseudo-labels. In the context of 3D vision, this paradigm is more specifically known as test-time refinement [26]. For example, SfM-TTR [26] boosts the performance of single-view depth networks at test time using SfM multi-view cues, while [9] learns depth, optical flow, camera pose and intrinsic parameters on a test sample in an online refinement setting.

3. Method

Our method is given an egocentric video sequence $\mathbf{V} = \{I_1, I_2, \dots, I_T\}$ with corresponding camera geometry, where each I_t is a frame at time t out of a total number of T frames, and a 2D motion segmentation model $\mathcal{M}$ that outputs a sequence of motion segmentation masks $\mathbf{M} = \{M_1, M_2, \dots, M_T\}$ corresponding to the frames of the video. It then integrates the motion segmentation masks $\mathbf{M}$ into a dynamic 3D model. The desired output is a set of enhanced 3D segmentation masks $\hat{\mathbf{M}} = \{\hat{M}_1, \hat{M}_2, \dots, \hat{M}_T\}$ that accurately represent the segmented dynamic and semi-static objects in the 3D space across the video sequence.
The following subsections first provide background on Neural Radiance Fields (NeRFs) in Sec. 3.1, and then describe the integration of motion segmentation into layered NeRFs in Sec. 3.2 and how to boost their segmentation capability even further with test-time refinement in Sec. 3.3.

3.1. Background on Neural Radiance Fields

Let $I : \Omega \to \mathbb{R}^3$ be an image, where $\Omega \subset \mathbb{R}^2$ is the image domain (generally a rectangle). Let $\boldsymbol{u} = (u_x, u_y, 1)$ be the homogeneous coordinate of a pixel, where $(u_x, u_y) \in \Omega$. Let π be a camera that images the 3D scene. We define a ray as the parametric curve $\boldsymbol{x}_\tau = \boldsymbol{x}_0 - \boldsymbol{\nu}\tau$, where $\boldsymbol{x}_0$ is the center of projection (camera center), $\tau \in [0, \infty)$ is the distance traveled, and $\boldsymbol{\nu}$ is the direction of the ray, intersect-

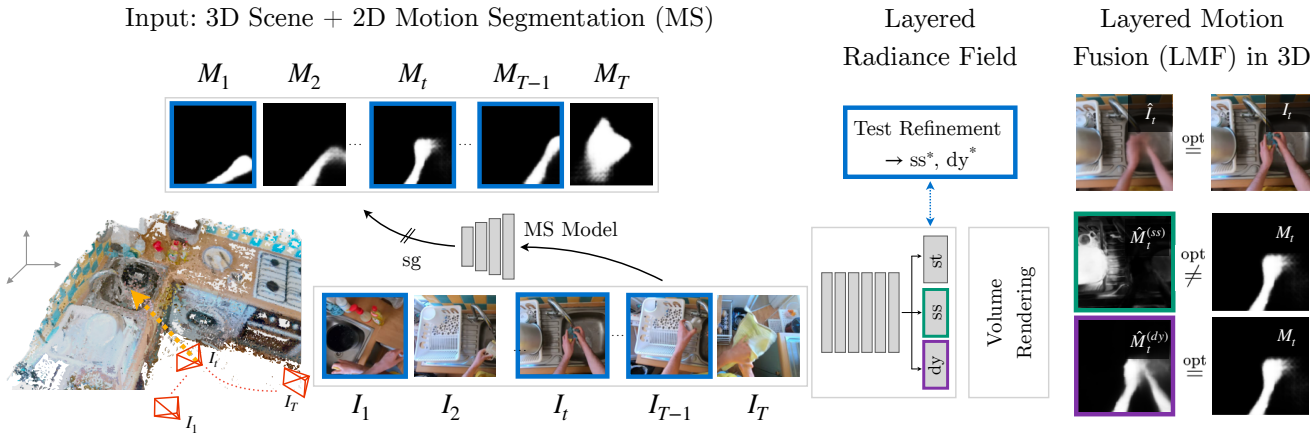

Figure 1. **Overview of our method**. Given a layered radiance field with static, semi-static and dynamic layers, our method fuses pseudo-labels M_t from a 2D segmentation method into its 3D representation. The static layer (st) is not updated as it does not learn any dynamics. The semi-static (ss) and dynamic (dy) layer produce segmentation masks $\hat{M}_t^{(ss)}, \hat{M}_t^{(dy)}$, which are improved through Layered Motion Fusion (LMF) that consists of the RGB loss, the positive motion fusion and the negative motion fusion loss. The semi-static and dynamic models are updated to ss* and dy* through test-time refinement by focusing on frames that are selected for analysis (highlighted in blue).

ing pixel $\boldsymbol{u}$, where all the quantities are defined in the world reference frame system (rather than the camera's one).

A *radiance field* is a pair of functions (σ, c). The opacity $\sigma : \mathbb{R}^3 \to \mathbb{R}_+$ maps 3D points to opacity values and the second function assigns a directional color $c : \mathbb{R}^3 \times \mathbb{S}^2 \to \mathbb{R}^3$ to each 3D point $\boldsymbol{x}$ and emission direction $\boldsymbol{\nu}$. The color $I(\boldsymbol{u})$ of the image at pixel $\boldsymbol{u}$ is given by the *emission-absorption equation*

$$I(\boldsymbol{u}) = \int_0^\infty c(\boldsymbol{x}_\tau, \boldsymbol{\nu}) \sigma(\boldsymbol{x}_\tau) e^{-\int_0^\tau \sigma(\boldsymbol{x}_\mu)\, d\mu}\, d\tau. \quad (1)$$

A *dynamic* radiance field just adds an additional time variable $t \in [0, T]$ to these functions, so that $I : \Omega \times [0, T] \to \mathbb{R}^3$, $\sigma : \mathbb{R}^3 \times [0, T] \to \mathbb{R}_+$ and $c : \mathbb{R}^3 \times \mathbb{S}^2 \times [0, T] \to \mathbb{R}^3$.

3.2. Layered motion fusion (LMF)

Layered neural radiance fields. Radiance fields have been originally designed for static scenes. In order to model dynamic videos, a natural extension assigns an individual radiance field per scene component as done in [48, 78, 88]. We refer to such models as *layered neural radiance fields*. Since we are interested in the segmentation of dynamic objects in egocentric videos, we describe our method with respect to NeuralDiff [78], that was explicitly designed for such videos, but the fusion itself can easily be applied to other layered architectures as shown in the experiments (Section 4.3). Assuming that we follow NeuralDiff, we decompose the scene into three layers: a static background layer, a semi-static layer, and a dynamic layer. This corresponds to the functions $\sigma_{\text{st}}(\boldsymbol{x}), c_{\text{st}}(\boldsymbol{\nu}, \boldsymbol{x})$ for the static layer, $\sigma_{\text{ss}}(\boldsymbol{x}, t), c_{\text{ss}}(\boldsymbol{\nu}, \boldsymbol{x}, t)$ for the semi-static one and $\sigma_{\text{dy}}(\bar{\boldsymbol{x}}, t), c_{\text{dy}}(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{x}}, t)$ for the dynamic one, where only the last two layers are time-dependent. The dynamic layer is defined w.r.t. camera coordinates $\bar{\boldsymbol{x}} = \pi(\boldsymbol{x})$ instead of world coordinates, where $\pi : \mathbb{R}^3 \to \mathbb{R}^3$ is the world to camera coordinate transformation. This is because in egocentric videos such as [80], the dynamic part is caused by the observer interacting with the world and so it is easier to model this part of the scene from their perspective.

The three fields (*i.e.* layers) are combined into a single one via the equations

$$\sigma(\boldsymbol{x}, t) = \sigma_{\text{st}}(\boldsymbol{x}) + \sigma_{\text{ss}}(\boldsymbol{x}, t) + \sigma_{\text{dy}}(\pi(\boldsymbol{x}), t) \quad (2)$$

$$c(\boldsymbol{x}, \boldsymbol{\nu}) = \frac{c_{\text{st}}(\boldsymbol{\nu}, \boldsymbol{x})\sigma_{\text{st}}(\boldsymbol{x}) + c_{\text{ss}}(\boldsymbol{\nu}, \boldsymbol{x}, t)\sigma_{\text{ss}}(\boldsymbol{x}, t)}{\sigma(\boldsymbol{x}, t)} + \frac{c_{\text{dy}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\nu}}, t)\sigma_{\text{dy}}(\bar{\boldsymbol{x}}, t)}{\sigma(\boldsymbol{x}, t)}. \quad (3)$$

Note that colors are mixed and weighed by the opacity [78]. These equations can be used to render image $\hat{I}(\boldsymbol{u}, t)$ as a function of time using Equation (1).

On top of colors and opacities, we also predict an uncertainty value for each layer: $\beta_{\text{st}}(\boldsymbol{x}), \beta_{\text{ss}}(\boldsymbol{x}, t)$, and $\beta_{\text{dy}}(\bar{\boldsymbol{x}}, t) \in \mathbb{R}_+$. These are then projected to the image domain using Equations (1) and (2) where c is replaced by β to render an uncertainty image $B(\boldsymbol{u}, t)$. This is used in a self-calibrated robust loss [28, 57]:

$$\mathcal{L}_{\text{rgb}}(\hat{I}, I, B, t) = \frac{1}{|\Omega|} \sum_{\boldsymbol{u} \in \Omega} \frac{\|\hat{I}(\boldsymbol{u}, t) - I(\boldsymbol{u}, t)\|^2}{2B(\boldsymbol{u}, t)^2} + \log B(\boldsymbol{u}, t)^2 \quad (4)$$

Network architecture and information sharing. The other important characteristic of layered radiance fields is how the different functions are implemented by a neural network and how the parameters are shared between the layers. This is achieved by composing networks as follows:

$$(\sigma_{\text{st}}(\boldsymbol{x}), c_{\text{st}}(\boldsymbol{x}, \boldsymbol{\nu})) = \Phi_{\text{st}}(\Phi_0(\boldsymbol{x}), \boldsymbol{\nu}) \quad (5)$$

$$(\sigma_{\text{ss}}(\boldsymbol{x}, t), c_{\text{ss}}(\boldsymbol{x}, \boldsymbol{\nu}, t)) = \Phi_{\text{ss}}(\Phi_0(\boldsymbol{x}), \boldsymbol{\nu}, \boldsymbol{z}_t) \quad (6)$$

$$(\sigma_{\text{dy}}(\bar{\boldsymbol{x}}, t), c_{\text{dy}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\nu}}, t)) = \Phi_{\text{dy}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\nu}}, \boldsymbol{z}_t). \quad (7)$$

Thus, the static and semi-static layers share the same spatial features Φ_0. The semi-static layer also takes as input the time t encoded a time-dependent feature vector $\boldsymbol{z}_t \in \mathbb{R}^D$. The dynamic layer does not share features since it is defined in a different reference frame, but does use the time encoding $\boldsymbol{z}_t$. Taken together, the vectors $\boldsymbol{z}$ thus form a matrix $Z \in \mathbb{R}^{T \times D}$. In order to ensure smoothness in the motion, the matrix Z is decomposed as the product $Z = \tilde{Z}F$ where $F \in \mathbb{R}^{P \times D}$ is a fixed Fourier-like basis with $P \ll T$ and $\tilde{Z} \in \mathbb{R}^{T \times P}$ are learned coefficients.

Fusing motion into the semi-static and dynamic layer. Recall that our goal is to obtain a segmentation that separates dynamic and semi-static objects, which naturally emerges from the decomposition offered by layered neural fields. However, experiments conducted in [80] show that, while layered radiance fields improve the segmentation of semi-static components over off-the-shelf 2D motion segmentation methods [93], they struggle with dynamic ones. Our idea is thus to use 3D reconstruction as a *fusion* network, in line with [3, 32, 79, 101]. The key difference is that our scene is in motion instead of being static. Additionally, fusion involves two different layers, representing the semi-static and dynamic motion.

We consider a motion segmentation algorithm that takes as input a video and outputs a segmentation mask $M(\boldsymbol{u}, t) \in [0, 1]$ for each frame, where 0 means background and 1 means foreground. We use these sparse, noisy and partial labels and fuse them into a joint implicit 3D space. This in turn enables the method to render *denoised* labels back to frames through its learned representation. We render the fused labels through masks obtained from the dynamic and semi-static layers as follows. We associate pseudo-colors to both layers with $\boldsymbol{p}_{\text{ss}} = (0, 1, 0)$ and $\boldsymbol{p}_{\text{dy}} = (0, 0, 1)$ in order to calculate the pixel-based (rendered) mask values $\hat{M}_{\text{ss}}(\boldsymbol{u}, t)$ and $\hat{M}_{\text{dy}}(\boldsymbol{u}, t)$, and point-based mask values $m_{\text{ss}}(\boldsymbol{x}_\tau, t)$ and $m_{\text{dy}}(\boldsymbol{x}_\tau, t)$ respectively. This results in the following volume rendering equation for *negative* motion fusion as the semi-static layer is supposed to be dissimilar to purely dynamic motion (push it away):

$$\hat{M}_{\text{ss}}(\boldsymbol{u}, t) = \int_0^\infty m_{\text{ss}}(\boldsymbol{x}_\tau, t)\sigma(\boldsymbol{x}_\tau, t)e^{-\int_0^\tau \sigma(\boldsymbol{x}_\mu, t)\, d\mu}\, d\tau\,, \quad (8)$$

with

$$m_{\text{ss}}(\boldsymbol{x}, t) = \frac{\boldsymbol{p}_{\text{ss},1}\sigma_{\text{st}}(\boldsymbol{x}) + \boldsymbol{p}_{\text{ss},2}\sigma_{\text{ss}}(\boldsymbol{x}, t)}{\sigma(\boldsymbol{x}, t)} + \frac{\boldsymbol{p}_{\text{ss},3}\sigma_{\text{dy}}(\bar{\boldsymbol{x}}, t)}{\sigma(\boldsymbol{x}, t)} = \frac{0 + \sigma_{\text{ss}}(\boldsymbol{x}, t) + 0}{\sigma(\boldsymbol{x}, t)}\,. \quad (9)$$

For positive motion fusion, we similarly calculate $\hat{M}_{\text{dy}}$ with Equation (8) where $\boldsymbol{p}_{\text{dy}}$ is used for m_{dy} with

$$m_{\text{dy}}(\boldsymbol{x}, t) = \frac{\sigma_{\text{dy}}(\bar{\boldsymbol{x}}, t)}{\sigma(\boldsymbol{x}, t)}. \quad (10)$$

With the rendered masks from the semi-static and dynamic layers, we calculate the positive motion fusion (PMF) and the negative motion fusion (NMF) losses respectively. Compared to NMF, the PMF loss pulls dynamic motion of the 2D model to the dynamic layer and is defined as

$$\mathcal{L}_{\text{PMF}}(\hat{M}_{\text{dy}}, M, t) = \lambda_{\text{PMF}} \frac{1}{|\Omega|} \sum_{\boldsymbol{u} \in \Omega} \|\hat{M}_{\text{dy}}(\boldsymbol{u}, t) - M(\boldsymbol{u}, t)\|^2. \quad (11)$$

The NMF loss penalizes the semi-static (ss) model when the predicted mask $\hat{M}_{\text{ss}}(\mathbf{u}, t)$ deviates from the target value of 0 for the set of pixels $\mathbf{u} \in \Omega$. This function is weighted by a factor λ_{NMF}, adjusting the emphasis on the penalty for incorrect predictions over these *negative* samples. Furthermore, we binarize the mask from the motion segmentation model to $\bar{M} \in \{0, 1\}$ and use it to select the pixels that are dynamic with $\bar{\Omega} = \{u \in \Omega \,|\bar{M}(u) = 1\}$, resulting in the loss:

$$\mathcal{L}_{\text{NMF}}(\hat{M}_{\text{ss}}, M, t) = \lambda_{\text{NMF}} \frac{1}{|\bar{\Omega}|} \sum_{\boldsymbol{u} \in \bar{\Omega}} \|\hat{M}_{\text{ss}}(\boldsymbol{u}, t)\|^2. \quad (12)$$

The final loss is then $\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{PMF}} + \mathcal{L}_{\text{NMF}}$, as defined in Equations (4), (11) and (12).

3.3. Test-time refinement

We follow [9, 26] and optimize the model at test time over a set of specified frames $\mathcal{T}$ *before* rendering the masks $\hat{M}_{\text{ss}}$ and $\hat{M}_{\text{dy}}$. We refer to this procedure throughout the paper as TR (test refinement). In comparison to the more typical refinement setting, we only refine the semi-static and dynamic model Φ_{st} and Φ_{dy}. The rationale for that is the independence of the static model from the task of motion segmentation, *i.e.* it does not capture any motion by design. Let W_{st}, W_{ss} and W_{dy} be the sets of parameters of the static, semi-static and dynamic models respectively. We obtain the parameters W^*_{ss} and W^*_{dy} of the refined models Φ^*_{ss} and Φ^*_{dy} with

$$(W^*_{\text{ss}}, W^*_{\text{dy}}) = \arg\min_{W_{\text{ss}}, W_{\text{dy}}} \sum_{t \in \mathcal{T}} \mathcal{L}(W_{\text{st}}, W_{\text{ss}}, W_{\text{dy}}; I_t, M_t, t). \quad (13)$$

Table 1. **Comparison with the state of the art**. Mean average precision (mAP) on segmenting the dynamic (Dyn) and semi-static (SS) components and their union (SS+Dyn) for the UDOS task from EPIC Fields [80]. We report the results for our method (TR + LMF) combined with NeuralDiff (improvements w.r.t. ND shown in brackets). The original 3D baselines do not use any 2D fusion. Our approach enhances the segmentation of semi-static objects as a secondary benefit.

Method	3D	2D	Dyn	SS	Dyn+SS
MG [93]	✗	✓	64.27	12.78	55.53
NeRF-W [48]	✓	✗	28.52	20.97	45.62
NeRF-T [20]	✓	✗	44.27	24.48	64.91
NeuralDiff (ND) [78]	✓	✗	55.58	25.55	69.74
ND + TR + LMF (ours)	✓	✓	**72.51**	**27.70**	**74.21**
			+30.5%	+8.4%	+6.4%

The masks from Equations (9) and (10) are rendered as described in Equation (8) with the densities and colors of the static model from Equation (5), and outputs from the refined semi-static and dynamic model with

$$(\sigma_{\text{ss}}(\boldsymbol{x}, t), c_{\text{ss}}(\boldsymbol{x}, \boldsymbol{\nu}, t)) = \Phi^*_{\text{ss}}(\Phi_0(\boldsymbol{x}), \boldsymbol{\nu}, \boldsymbol{z}_t) \quad (14)$$

$$(\sigma_{\text{dy}}(\bar{\boldsymbol{x}}, t), c_{\text{dy}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\nu}}, t)) = \Phi^*_{\text{dy}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\nu}}, \boldsymbol{z}_t). \quad (15)$$

Besides selecting only reference frames, we also explore sampling additional frames within the vicinity of reference frames, to see if the added temporal context further facilitates the refinement. Formally, given a set of frames $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$ with $t_i \in [0, T], \forall i \in \{1, 2, \ldots, M\}$, we define a set of neighboring frames for each test frame t_i within a window N. The combined set of test frames and their neighboring frames is denoted as $\mathcal{T}_N$. For each test frame $t_i \in \mathcal{T}$, the neighboring frames within a window N are defined as:

$$\mathcal{N}(t_i, N) = \{t_{i-N}, \ldots, t_{i-1}, t_i, t_{i+1}, \ldots, t_{i+N}\}. \quad (16)$$

The set of frames used for refinement, including both the test frames and their neighbors, is given by:

$$\mathcal{T}_N = \bigcup_{t_i \in \mathcal{T}} \mathcal{N}(t_i, N). \quad (17)$$

4. Experiments

In the following, we will first describe the experimental details, and then compare our proposed method with the state of the art. The last sections contain an analysis of our method such as its application to other 3D methods and an ablation study.

4.1. Experimental details

Our experiments follow the *Unsupervised Dynamic Object Segmentation* (UDOS) benchmark from the EPIC Fields dataset [80]. EPIC Fields augments the EPIC-KITCHENS dataset [12] with 3D camera information and use the provided segmentation masks of dynamic and semi-static objects for evaluation. We use the provided evaluation script and the pre-trained models (NeuralDiff, NeRF-W and T-NeRF – referred to NeRF-T in our paper) for our experiments. The pre-trained models are trained for 20 epochs with a learning rate of 5×10^4 and cosine annealing schedule with an NVIDIA RTX A4000 per experiment. We set the parameters λ_{PMF} and λ_{NMF} of the LMF loss to 1.1 and 1.0 respectively. The fine-tuning takes about 22 minutes for 100 frames (about 13 seconds per frame). The rendering of a frame without fine-tuning takes about 5 seconds. For a fair comparison with the results from EPIC Fields, we use exactly the same frames as used to train their models. For motion fusion, we extract labels from the 2D motion segmentation model *Motion Grouping* (MG) [93] that is used as 2D baseline in EPIC Fields. Further details such as network architecture or the training setup of MG can be found in the supplementary material of EPIC Fields [80].

4.2. Comparison with the state of the art

We evaluate our method on the *Unsupervised Dynamic Object Segmentation* (UDOS) task from EPIC Fields [80] and compare to the 3D baselines NeuralDiff [78], NeRF-W [48] and NeRF-T [20]. The results are shown in Table 1. We observe improvements of up to 30%. A positive byproduct of our method is the improvement of the semi-static segmentation, by up to 8% in comparison to NeuralDiff. The joint segmentation of dynamic and semi-static objects improves by up to 6.4%. The most important result is the improvement of NeuralDiff as the 3D baseline over MG as the 2D baseline, which was posed as an open problem and question in [80]. Our method outperforms MG by up to 8.2 mAP, while MG outperformed NeuralDiff previously by about 8.7 mAP. Qualitative results comparing improvements over NeuralDiff and MG are shown in Figure 2. Results that highlight the improvements of the semi-static segmentation are shown in the supplementary material in Figure 1.

Additionally, we compare our approach to a video object segmentation method specifically tailored for egocentric videos, focusing on the task of fine-grained hand-object segmentation. This comparison highlights the potential of 3D computer vision techniques to improve the performance of 2D video understanding methods. For this purpose, we select the state-of-the-art method EgoHOS [99] and report the results in Table 4. We observe that applying our method to MG [93] results in a performance that is slightly better than EgoHOS. This result is significant, since EgoHOS requires supervision, while applying our method to MG works without any supervision. In addition we show that we can boost the performance of EgoHOS even further by combining it with our method. These results highlight the

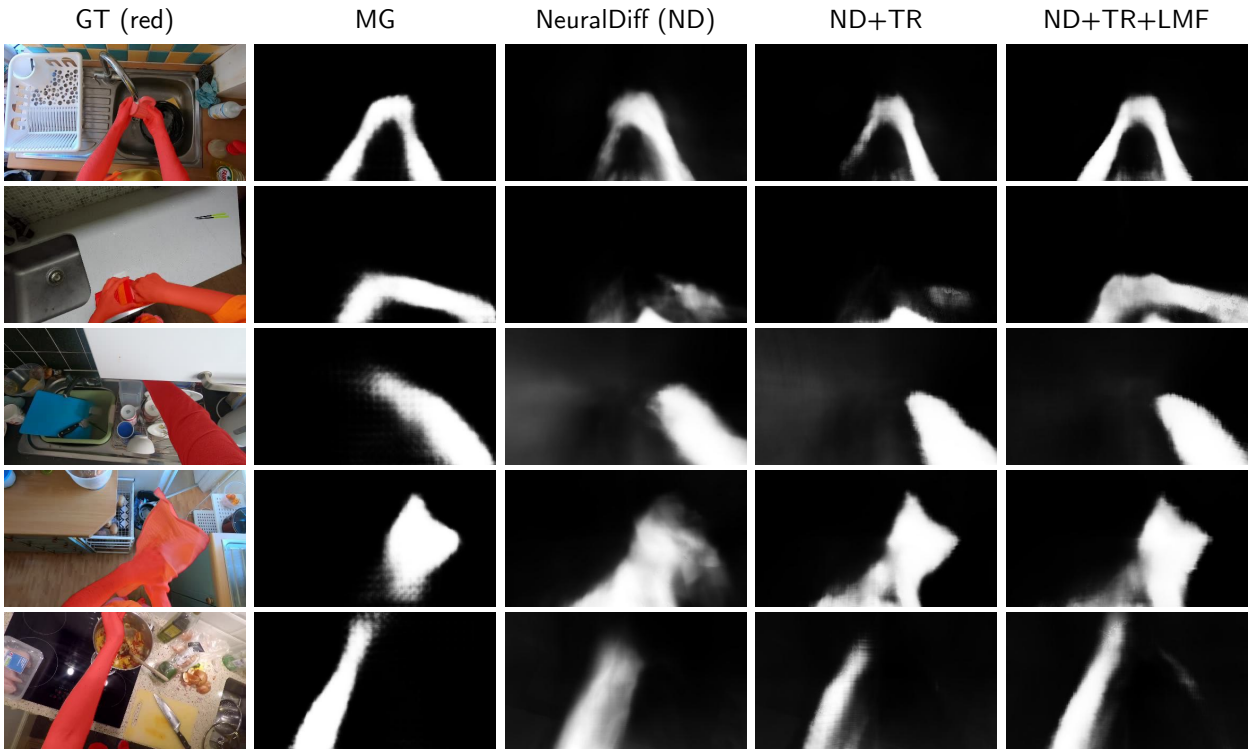

Figure 2. **Qualitative results.** The segmentations are produced by Motion Grouping (MG), NeuralDiff (ND), ND + Test-Time Refinement (TR), and ND + TR + Layered Motion Fusion (LMF). This shows the clear benefits of applying TR to ND, resulting in sharper segments such as in row 3, 4 and 5. The segmentation can be improved even further through LMF as shown in row 1 and 2.

Table 2. **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) for the UDOS task from EPIC Fields [80]. We apply a variant of our method to NeRF-W, NeRF-T and NeuralDiff. We observe consistent gains across all architectures. The relative improvements are shown in brackets.

Method	3D	2D	Dyn	SS	Dyn+SS
NeRF-W [48]	✓	✗	28.52	20.97	45.62
NeRF-W + TR + PMF	✓	✓	34.20 (19.9%)	19.88 (-5.2%)	47.37 (3.8%)
NeRF-T [20]	✓	✗	44.27	24.48	64.91
NeRF-T + TR + PMF	✓	✓	51.11 (15.4%)	23.24 (-5.1%)	68.87 (6.1%)
NeuralDiff [78]	✓	✗	55.58	25.55	69.74
NeuralDiff + TR + PMF	✓	✓	**67.23** (20.9%)	**26.61** (4.1%)	**72.53** (4.0%)

effectiveness and versatility of the proposed approach.

4.3. Analysis of model components

Application to other architectures. To show that our contributions can be applied beyond NeuralDiff, we also experiment with the architectures NeRF-T and NeRF-W (NeRF plus time [20] and NeRF in the wild [48]). While NeRF-W contains already a dynamic layer, we extend NeRF-T with the same dynamic layer as we use for our method as described in Section 3. Furthermore, in NeRF-T, time is encoded using a positional encoding, whereas in NeRF-W, time is encoded by specifying and learning a separate latent vector z_t for each frame (similar to setting $P = D$ in the representation of Z).

We analyze the generalization of our method across these architectures in Table 2. Since NeRF-T and NeRF-W contain static and dynamic layers only, we can apply PMF, but omit NMF as it depends on a semi-static layer. We observe that a combination of TR and PMF boosts the performance of all architectures in terms of the segmentation

Table 3. **Effect of the temporal context on test refinement.** Mean average precision (mAP) on segmenting the dynamic (Dyn) and semi-static (SS) components and their union (SS+Dyn) for the unsupervised dynamic object segmentation (UDOS) task from EPIC Fields [80], on a subset of 5 scenes and for 4 different numbers of neighbouring frames N as defined in Eq. (16). We set $N = 0$ for the default setting without any neighboring frames, and compare to sampling 2, 5 and 20 frames. The relative improvements are shown in brackets.

Method	Dyn	SS	Dyn+SS
NeuralDiff (ND) [78]	58.12	25.84	70.49
ND + TR	**64.29** (10.6%)	**26.26** (1.6%)	**72.07** (2.2%)
ND + TR-2	64.23 (10.5%)	26.23 (1.5%)	72.02 (2.2%)
ND + TR-5	63.14 (8.6%)	26.21 (1.4%)	71.98 (2.1%)
ND + TR-20	61.64 (6.1%)	25.95 (0.4%)	71.45 (1.4%)

Table 4. **Hand-object segmentation in the dynamic setting.** We combine our method with the state-of-the-art hand-object segmentation method EgoHOS [99]. Our method with MG [93] requires no supervision and is slightly better than EgoHOS, which is trained with supervision. Applying our method to EgoHOS improves it even further.

Method	3D	2D	Supervision	mAP
MG [93]	✗	✓	✗	64.27
MG + Ours	✓	✓	✗	72.51
EgoHOS [99]	✗	✓	✓	71.20
EgoHOS + Ours	✓	✓	✓	**77.31**

of dynamic objects. The decrease in performance of the semi-static prediction of NeRF-T and NeRF-W is expected as both architectures use only one layer to predict both dynamic and semi-static objects. Adding a semi-static layer to both models as shown in Table 6 of the supplementary results in improved semi-static performance.

Effect of temporal context on test refinement. We defined the test refinement with respect to a number of neighboring frames in Equation (17). We analyze in Table 3 the influence of the temporal context on test-time refinement to find out if additional frames around test frames provide additional guidance to the model. We observe no benefit from adding frames. This highlights the importance of enhancing the focus of the 3D model to achieve better segmentation.

Ablation study. Table 5 compares different combinations of components with NeuralDiff. We observe that motion fusion results overall in the highest relative gains for segmenting dynamic objects. The best results are obtained with test time refinement and layered motion fusion (LMF = PMF + NMF), resulting in an improvement of 30.4%. The semi-static segmentation also improves significantly in this case, by 8.4%. We observe as well that a joint combination of PMF and NMF is better than any of those components on their own – with and without TR.

Table 5. **Ablation**. Mean average precision (mAP) on the segmentation of dynamic (Dyn) and semi-static (SS) components and their union (SS+Dyn), for the UDOS task from EPIC Fields [80]. The relative improvements are w.r.t. NeuralDiff. The best results are obtained with layered motion fusion (LMF = PMF + NMF) and test time refinement (TR), and show that the different components improve the segmentation additively.

Method	3D	2D	Dyn	SS	Dyn+SS
NeuralDiff (ND) [78]	✓	✗	55.58	25.55	69.74
ND + NMF	✓	✓	63.73 +14.7%	27.14 +6.2%	72.69 +4.2%
ND + PMF	✓	✓	63.38 +14.0%	26.33 +3.1%	72.10 +3.4%
ND + PMF + NMF	✓	✓	66.14 +19.0%	26.44 +3.5%	72.29 +3.7%
ND + TR	✓	✗	62.29 +12.0%	26.02 +1.8%	71.66 +2.8%
ND + TR + NMF	✓	✓	69.79 25.5%	**27.94** 9.3%	**75.57** 8.3%
ND + TR + PMF	✓	✓	67.23 +20.9%	26.61 +4.1%	72.53 +4.0%
ND + TR + PMF + NMF	✓	✓	**72.51** +30.4%	27.70 +8.4%	74.21 +6.4%

5. Conclusion

In this paper, we analyzed the limitations of current 3D methods when applied to the task of unsupervised dynamic object segmentation of long egocentric videos, as observed in the EPIC Fields benchmark [80]. We address these limitations through two contributions. First, we introduce *layered motion fusion*, which lifts motion segmentation predictions from a 2D-based model into layered radiance fields. This approach includes positive motion fusion, which pulls the predictions of the dynamic layer closer to the segmentation of the 2D model, and negative motion fusion, which pushes the predictions of the semi-static layer away from the segmentation of the 2D model. Both losses work in synergy, resulting in significant performance gains. We demonstrate that these effects can be further leveraged through test-time refinement, enabling the final model to outperform the 2D model used for training by a large margin. The proposed method is straightforward to implement, effective, adaptable to various 3D architectures and can be combined with a state-of-the-art hand-object segmentation approach to boost its performance. We believe that these results will inspire further research on the use of 3D geometry for egocentric scene and video understanding.

References

[1] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 197–214. Springer, 2024. 3

[2] Yash Bhalgat, Vadim Tschernezki, Iro Laina, Joao F. Henriques, Andrea Vedaldi, and Andrew Zisserman. 3d-aware instance segmentation and tracking in egocentric videos. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. IEEE, 2024. 13

[3] Yash Sanjay Bhalgat, Iro Laina, Joao F. Henriques, Andrea Vedaldi, and Andrew Zisserman. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 5

[4] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3

[5] Thierry Bouwmans. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer science review*, 11:31–66, 2014. 3

[6] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 3

[7] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 2021. 2

[9] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 7063–7072, 2019. 3, 5

[10] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating Motion. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 3

[11] G. Csurka, C. R. Dance, L. Dan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. ECCV Workshop on Stat. Learn. in Comp. Vision*, 2004. 1

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2022. 6

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021. 1

[14] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proc. ICCV*, 2021. 2

[15] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. NeRF-SOS: Any-view self-supervised object segmentation on complex scenes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3

[16] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *Proc. SIGGRAPH Asia*, 2022. 2

[17] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[18] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic NeRF: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv.cs*, abs/2203.15224, 2022. 2

[19] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. *arXiv.cs*, abs/2105.06468, 2021. 2

[20] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6, 7, 15

[21] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *arXiv.cs*, abs/2210.13445, 2022. 15

[22] Ross B. Girshick. Fast R-CNN. In *Proc. ICCV*, 2015. 1

[23] Xiang Guo, Guanying Chen, Yuchao Dai, Xiaoqing Ye, Jiadai Sun, Xiao Tan, and Errui Ding. Neural deformable voxel grid for fast optimization of dynamic view synthesis. In *Proc. ACCV*, 2022. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1

[25] Alexander Hermans, Georgios Floros, and Bastian Leibe. Dense 3d semantic mapping of indoor scenes from rgb-d images. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2631–2638. IEEE, 2014. 3

[26] Sergio Izquierdo and Javier Civera. Sfm-ttr: Using structure from motion for test-time refinement of single-view depth networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21466–21476, 2023. 2, 3, 5

[27] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[28] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Proc. NeurIPS*, 2017. 4

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 3, 16

[30] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3

[31] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. *arXiv.cs*, abs/2401.09419, 2024. 1

[32] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing neRF for editing via feature field distillation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3, 5

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NeurIPS*, 2012. 1

[34] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 703–718. Springer, 2014. 3

[35] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12871–12881, 2022. 2

[36] Dong Lao, Zhengyang Hu, Francesco Locatello, Yanchao Yang, and Stefano Soatto. Divided attention: Unsupervised multiple-object discovery and segmentation with interpretable contextually separated slots, 2024. 3

[37] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2

[38] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. Neural 3D video synthesis from multi-view video. In *Proc. CVPR*, 2022. 2

[39] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proc. CVPR*, 2021. 2

[40] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[41] Long Lian, Zhirong Wu, and Stella X. Yu. Bootstrapping objectness from videos by relaxed common fate and visual grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14582–14591, 2023. 3

[42] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts, 2023. 2, 3

[43] Yiqing Liang, Eliot Laidlaw, Alexander Meyerowitz, Srinath Sridhar, and James Tompkin. Semantic attention flow fields for monocular dynamic scene decomposition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 21797–21806, 2023. 1, 2, 3

[44] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 35: 36762–36775, 2022. 2

[45] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, 1999. 1

[46] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with RGB-D cameras. In *Proc.IROS*, 2017. 3

[47] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. New Yorl: WH Freeman, 1982. 1

[48] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4, 6, 7, 15

[49] Ruben Mascaro, Lucas Teixeira, and Margarita Chli. Diffuser: Multi-view 2D-to-3D label diffusion for semantic scene segmentation. In *Proc. ICRA*, 2021. 3

[50] Jana Mattheus, Hans Grobler, and Adnan M Abu-Mahfouz. A review of motion segmentation: Approaches and major challenges. In *2020 2nd International multidisciplinary information technology and engineering conference (IMITEC)*, pages 1–8. IEEE, 2020. 3

[51] Kevin Matzen and Noah Snavely. Scene chronology. In *Proc. ECCV*, 2014. 3

[52] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017. 3

[53] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020. 1, 2

[54] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[55] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning visual motion segmentation

using event surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[56] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212. IEEE, 2019. 3

[57] David Novotný, Diane Larlus, and Andrea Vedaldi. Capturing the geometry of object categories from video supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 4

[58] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *PAMI*, 36(6):1187 – 1200, 2014. Preprint. 3

[59] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[60] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 3

[61] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[62] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorizaton. In *Proc. CVPR*, 2006. 1

[63] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *2025 International Conference on 3D Vision (3DV)*, 2025. 13

[64] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[65] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20051–20060, 2024. 3

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763, 2021. 2

[67] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Andrei Ambrus, Adrien Gaidon, William T. Freeman, Fredo Durand, Joshua B. Tenenbaum, and Vincent Sitzmann. Neural groundplans: Persistent neural scene representations from a single image. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3

[68] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[69] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 1

[70] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*. Springer, 2006. 1

[71] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. NeRFPlayer: A streamable dynamic scene representation with decomposed neural radiance fields. *arXiv.cs*, abs/2210.15947, 2022. 2

[72] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 3

[73] Niko Sünderhauf, Trung T Pham, Yasir Latif, Michael Milford, and Ian Reid. Meaningful maps with object-oriented semantic mapping. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5079–5085. IEEE, 2017. 3

[74] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6243–6252, 2017. 3

[75] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[76] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. ICML*, 2021. 2

[77] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proc. ICCV*, 2021. 2

[78] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2021. 1, 2, 3, 4, 6, 7, 8, 15

[79] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022. 1, 2, 3, 5

[80] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea

Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 4, 5, 6, 7, 8, 13, 15

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 1

[82] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Nießner, Stuart Golodetz, Victor A Prisacariu, Olaf Kähler, David W Murray, Shahram Izadi, Patrick Pérez, et al. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 75–82. IEEE, 2015. 3

[83] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi S. M. Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. NeSF: Neural semantic fields for generalizable semantic segmentation of 3D scenes. *arXiv.cs*, abs/2111.13260, 2021. 2

[84] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022. 2

[85] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv.cs*, abs/2105.05994, 2021. 2

[86] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier PlenOctrees for dynamic radiance field rendering in real-time. In *Proc. CVPR*, 2022. 2

[87] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 3

[88] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 35:32653–32666, 2022. 1, 3, 4

[89] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proc. CVPR*, 2021. 2

[90] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[91] Christopher Xie, Keunhong Park, Ricardo Martin-Brualla, and Matthew Brown. Fig-nerf: Figure-ground neural radiance fields for 3d object category modelling. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 962–971. IEEE, 2021. 3

[92] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[93] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 5, 6, 8, 14

[94] Gengshan Yang and Deva Ramanan. Learning to segment rigid motions from two frames. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[95] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proc. CVPR*, 2020. 2

[96] Hong-Xing Yu, Leonidas Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 3

[97] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[98] Luca Zappella, Xavier Lladó, and Joaquim Salvi. Motion segmentation: A review. *Artificial Intelligence Research and Development*, pages 398–407, 2008. 3

[99] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 127–145. Springer, 2022. 6, 8, 13

[100] Xiaoshuai Zhang, Abhijit Kundu, Thomas Funkhouser, Leonidas Guibas, Hao Su, and Kyle Genova. Nerflets: Local radiance fields for efficient structure-aware 3d scene representation from 2d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[101] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *Proc. ICCV*, 2021. 1, 2, 3, 5, 13, 14

[102] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21676–21685, 2024. 3

[103] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding, 2024. 3