2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Amodal Optical Flow

Maximilian Luz*,1, Rohit Mohan*,1, Ahmed Rida Sekkat[2],
Oliver Sawade[2], Elmar Matthes[2], Thomas Brox[1], and Abhinav Valada[1]

***Abstract*— Optical flow estimation is very challenging in situations with transparent or occluded objects. In this work, we address these challenges at the task level by introducing Amodal Optical Flow, which integrates optical flow with amodal perception. Instead of only representing the visible regions, we define amodal optical flow as a multi-layered pixel-level motion field that encompasses both visible and occluded regions of the scene. To facilitate research on this new task, we extend the AmodalSynthDrive dataset to include pixel-level labels for amodal optical flow estimation. We present several strong baselines, along with the Amodal Flow Quality metric to quantify the performance in an interpretable manner. Furthermore, we propose the novel AmodalFlowNet as an initial step toward addressing this task. AmodalFlowNet consists of a transformer-based cost-volume encoder paired with a recurrent transformer decoder which facilitates recurrent hierarchical feature propagation and amodal semantic grounding. We demonstrate the tractability of amodal optical flow in extensive experiments and show its utility for downstream tasks such as panoptic tracking. We make the dataset, code, and trained models publicly available at `http://amodal-flow.cs.uni-freiburg.de`.**

I. Introduction

Optical flow estimates the apparent motion patterns of objects between two consecutive images of a sequence. This fundamental vision problem has widespread applications, including localization [1], autonomous driving [2], and object tracking [3]. Over the years, several groundbreaking methods have been proposed [4]–[12]. One remaining challenge is still *transparency* (e.g., seeing through windows). Optical flow, by itself, cannot represent transparency accurately as it can only provide a single displacement vector per pixel. Another remaining challenge is *occlusion* and expressing the occlusion relationship of objects adeptly, which leads to incomplete and inaccurate motion patterns, particularly at occlusion boundaries. To address these challenges and enhance motion analysis, we exploit amodal perception [13]–[15], which aims to perceive objects as a whole, even when parts of them are occluded.

In this work, we introduce amodal optical flow that seamlessly incorporates the principles of optical flow with amodal perception. As illustrated in Fig. 1, amodal optical flow aims to predict a set of motion fields at the pixel level, where each pixel is distinctly associated with both the visible and occluded regions of different scene elements across consecutive frames. This pixel association is defined through the use of amodal masks of the scene elements (refer to Sec. III). Fundamentally, amodal optical flow requires the delineation of continuous object boundaries that encompass both visible and occluded regions. Thereby, it directly addresses the fragmentation problem encountered by most optical flow estimation methods, where refinement follows texture boundaries and edges rather than the true object boundaries (cf. [16]).

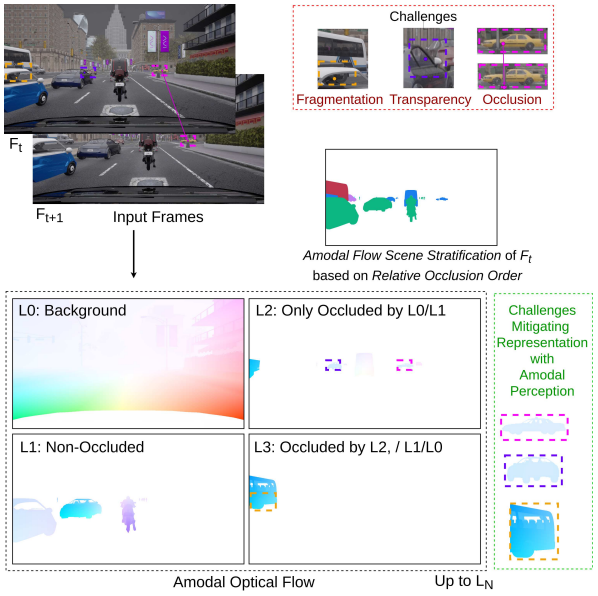

Fig. 1: Illustration of *Amodal Optical Flow*, which aims to predict a multi-layered pixel-level motion field encompassing both visible and occluded regions of the scene. This task can represent transparent and partially occluded objects while also reducing the fragmentation of object segments through amodal (visible + occluded) motion representation of scene elements.

By encouraging explicit reasoning about occlusions, amodal optical flow facilitates a more comprehensive handling thereof and lends itself more naturally to multi-frame sequence predictions. Furthermore, it moves away from the single-layer prediction that is typical in optical flow, adopting multi-layer outputs that are better suited to handle the complex interactions and overlaps between scene elements, including transparency. This is particularly relevant for scene understanding in robotics. Explicitly modeling the motion of individual scene elements is closer to the 3D nature of the scene, thus allowing for more informed, accurate decisions. Object tracking is a concrete example that can benefit from amodal optical flow.

Our main contributions are twofold. First, we formulate the amodal optical flow estimation task, and to facilitate research, we provide a labeled dataset, an appropriate evaluation metric,

*These authors contributed equally.
[1]Department of Computer Science, University of Freiburg, Germany
[2]IAV GmbH, Germany
This work was funded by the German Research Foundation Emmy Noether Program grant number 468878300 and an academic grant from NVIDIA.

and baselines. Given the inherent complexities of amodal perception and the challenges associated with acquiring real-world ground truth labels, we propose a synthetic dataset, which offers the advantage of controlled scenarios and perfect precision of annotations. We extend the recently introduced AmodalSynthDrive dataset [17] with amodal optical flow ground truth, enabling other tasks to leverage amodal flow to improve performance and vice versa. Second, we propose the novel AmodalFlowNet architecture, which predicts amodal flow in a recurrent and occlusion-ordered manner. Notably, it keeps the number of objects or occlusions in a scene flexible. It separately predicts both the visible and invisible regions of objects (i.e., amodal masks), as well as amodal semantic labels. We show the tractability and effectiveness of our approach in several experiments. Furthermore, we compare amodal optical flow and traditional optical flow in terms of their efficacy for panoptic tracking. The result demonstrates that the tracking approach based on amodal optical flow compares favorably. We will make the code and trained models publicly available at `http://amodal-flow.cs.uni-freiburg.de`.

II. Related Work

Amodal Perception: Amodal perception is an integral part of human cognitive abilities [18]. Consequently, various computer vision tasks have been proposed, often as counterparts to classical modal problems. Amodal bounding box prediction [19] and amodal instance segmentation [14] aim to infer the true extent of objects in a scene, including any occluded parts. Zhu *et al.* [13] extend the latter with semantic information; however, they still only consider salient regions of the image. Contrastively, Purkait *et al.* [20] adapt semantic segmentation to the amodal setting more directly, performing a standard modal segmentation of the background assuming no foreground objects are present and only handling the latter amodally. Amodal panoptic segmentation [15], [21] reintroduces instance information to this while keeping the separation of foreground and background classes, creating a holistic amodal segmentation approach. Due to the inherent complexity of amodal perception, however, only a few approaches go beyond the prediction of object shapes and masks, largely focusing on appearance (e.g., [22], [23]). Notably, Dhamo *et al.* [24] consider the full scene and additionally estimate per-object depth. We refer to the survey by Ao *et al.* [25] for further discussion on amodal perception.

Optical Flow Estimation: While nowadays vastly surpassed by neural networks, classical optical flow estimation approaches (e.g., [5], [26]–[28]) still contain invaluable knowledge. Particularly relevant to amodal optical flow are techniques that decompose the scene into multiple overlapping layers, each with its own flow field and mask. Although earlier methods (e.g., [29]–[32]) show limited amodal reasoning through appearance reconstruction in occluded regions, later methods (e.g., [33]–[35]) solely aim to improve their modal capabilities. Rather than constructing an amodal representation of optical flow, a major motivation for their development was the composition of complex motion via multiple simpler models. Consequently, little evaluation has been performed targeting amodal aspects, and, to our knowledge, no amodal optical flow metrics or evaluation datasets exist. The introduction of neural network-based flow estimation methods has made explicit motion decomposition as a modeling tool less important. Instead, recent methods have focused on recurrent refinement [9], motion aggregation for ambiguous or otherwise difficult matches [10], better features for correspondence estimation [11], [36], and improved matching cost representations [12].

Combining Motion and Segmentation: Even though we attempt to retain the more fundamental nature of optical flow by disentangling it from detailed object knowledge, some object reasoning is nevertheless required to address the amodal aspects. Therein, amodal optical flow is related to the detection and segmentation of independently moving objects from image sequences [37]–[39]. While amodal approaches have been proposed in this regard (e.g., [40], [41]), optical flow approaches so far only employ segmentation for direct modal flow improvements (e.g., [16], [42], [43]).

III. Amodal Optical Flow

The objective of amodal optical flow estimation is to compute motion patterns for scene components within both visible and occluded regions across consecutive frames, I_t and I_{t+1}. This involves determining a set of motion fields, $\mathscr{F}$, where each F_i in $\mathscr{F}$ corresponds to a visible scene component or a potentially occluded scene element. For each pixel p in I_t, its probable positions in I_{t+1} are given by $p+F_i(p)$, where F_i represents one of the possible displacement vectors (u,v) from the set. To accommodate the complexity of potentially diverse motion field predictions for a single pixel, considering both visible and occluded regions of scene elements, we employ the notion of relative occlusion order [21]. This facilitates the development of a scene stratification strategy for amodal flow that ensures a consistent order in the motion field set. This framework organizes scene elements hierarchically based on their visibility and occlusion potential.

Amodal Flow Scene Stratification: We introduce two fundamental categories of scene elements, inspired by panoptic segmentation:

- *Background:* This category includes amorphous regions such as roads, walls, and static objects such as traffic signs and lights.
- *Traffic participants:* This category comprises movable objects such as pedestrians and cars.

We further simplify the complexity of scenes by disregarding occlusions occurring solely within background elements. The primary focus of the task considers the interplay of occlusion between the background and traffic participants, as well as occlusions among different instances of traffic participants. This simplification ensures clarity and precision in predicting the motion fields. Based on these premises, we establish a layered stratification of the scene in frame I_t that consists of N levels, each representing a distinct occlusion degree:

- Level 0: Corresponds to the background.
- Level 1: Encompasses non-occluded traffic participants.

- Level 2 to Level $N-1$: Each subsequent level includes objects occluded by at least one object from the preceding level and any objects from earlier levels. Notably, objects within the same level do not occlude each other.

This stratified structure allows a pixel to have multiple associated motion field vectors, each tied to a specific occlusion level. It is important to highlight that this strategy remains independent of the specific scene element classes, enabling the task to prioritize motion estimation without heavy reliance on object recognition. Moreover, this simplification encourages the potential for unsupervised approaches and adaptability across diverse environments, facilitating the emergence of innovative methods beyond our approach detailed in Sec. V, which capitalizes on the semantic understanding of objects. We set N at a maximum of 8, corresponding to the highest value observed in our dataset. Fig. 1 shows a visual representation of this stratification strategy with an example scene along with the corresponding amodal optical flow ground truth for each occlusion order level.

Task Definition: For the given inputs I_t and I_{t+1}, amodal optical flow aims to predict a series of motion field maps $\{M_0, M_1, \ldots, M_{N-1}\}$, each corresponding to a specific occlusion level n within the scene. Every motion field map M_n has the same dimensions as the input, where each pixel value is represented by a triplet $\langle i, u, v \rangle$. Here, i indicates whether the pixel is associated with a scene element segment (taking a value of 0 or 1), while u and v represent the components of the displacement vector at that point. Formally, the prediction step can therefore be represented as

$$f(I_t, I_{t+1}) \rightarrow \{M_0, M_1, \ldots, M_{N-1}\}. \tag{1}$$

Evaluation Metric: To evaluate amodal optical flow estimation, we introduce a unified metric called Amodal Flow Quality (AFQ). This metric jointly assesses both the amodal motion field prediction and the class-agnostic segmentation quality of scene elements in an interpretable manner. This is achieved by combining the weighted area under the curve (WAUC) [44] metric utilized in the evaluation of optical flow with the mean intersection over union (mIoU) metric [45] which is commonly used for measuring segmentation quality. We denote the predicted amodal optical flow for occlusion level n as M_n^p and the ground truth amodal optical flow for occlusion level n as M_n^g. We then define the WAUC for each occlusion level n that integrates over distance thresholds between 0 and 5 pixels as follows:

$$\mathrm{WAUC}_n = \frac{\sum_{i=1}^{100} w_i \sum_j [e_j \leq \delta_i]}{C \cdot \sum_{i=1}^{100} w_i}, \tag{2}$$

$$e_j = \left\| \vec{v}_{M_n^p}(j) - \vec{v}_{M_n^g}(j) \right\|_2, \tag{3}$$

where e_j is the end point error at pixel j, $\vec{v}_\cdot(j)$ represent the motion vectors containing both the u and v components and $\delta_i = \frac{i}{20}$. $[.]$ is the Iverson bracket, C is the total number of pixels, and $w_i = 1 - \frac{i-1}{100}$. Similarly, we define IoU for each occlusion level n as follows:

$$\mathrm{IoU}_n = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{4}$$

where TP, FP, and FN are the number of pixels correctly segmented, falsely segmented, and missed in the segmentation, respectively. Having computed the individual WAUC and IoU values for each occlusion level, we compute AFQ as the geometric mean of the mean IoU (mIoU) and the mean WAUC (mWAUC), both averaged over all occlusion levels,

$$\mathrm{AFQ} = \sqrt{\mathrm{mWAUC} \cdot \mathrm{mIoU}}. \tag{5}$$

Hereby, mIoU and mWAUC are defined as

$$\mathrm{mIoU} = \frac{1}{\sum_{n=1}^{N-1} w_n} \sum_{n=1}^{N-1} w_n \mathrm{IoU}_n, \tag{6}$$

$$\mathrm{mWAUC} = \frac{1}{\sum_{n=0}^{N-1} w_n} \sum_{n=0}^{N-1} w_n \mathrm{WAUC}_n, \tag{7}$$

with exponentially decaying weights

$$w_n = \exp\left(-\max\left(-\frac{n-k}{N-1-k} \log(w_{N-1}),\ 0 \right) \right), \tag{8}$$

where N is the total number of occlusion levels, $k = 3$ the number of amodal levels with equal weighting, and $w_{N-1} = 0.25$ the weight of the last layer. The choice of k is determined by the dataset's relative occlusion order distribution.

IV. Dataset

AmodalSynthDrive [17] is the first comprehensive dataset for multi-task multi-modal amodal perception tailored to the automotive domain. It consists of 60,000 multi-view camera images, 3D bounding boxes, LiDAR data, odometry, and amodal semantic/instance/panoptic segmentation annotations, distributed across 150 distinct driving sequences, with over 1 million object annotations. It covers various scenarios involving diverse traffic, weather conditions, and lighting conditions. In this work, we extend this dataset with amodal optical flow ground truth for images of the original training, validation, and test splits, consisting of 105 video sequences with 42,000 images, 15 video sequences with 6000 images, and 30 video sequences with 12,000 images, respectively.

We compute the amodal optical flow by individually rendering all visible objects, excluding the background, frame by frame. In the rendering procedure, we individually extract ground truth labels for each object. This ground truth encompassed the data required to compute the optical flow for each individual object separately, namely depth maps, instance segmentation, and precise positions and rotations. For every object in the scene, we compute both its displacement between two consecutive frames and the corresponding camera displacement representing the displacement transformations. We also compute the amodal point cloud by utilizing the corresponding amodal depth maps. By applying the displacement transformations, we generate the amodal point cloud that underwent the aforementioned movements. This allows us to derive 3D movement vectors for each pixel representing the object in the amodal instance segmentation. These 3D vectors are then reprojected into the different image planes to obtain modal and amodal optical flow. To gain a more comprehensive understanding of the optical flow ground

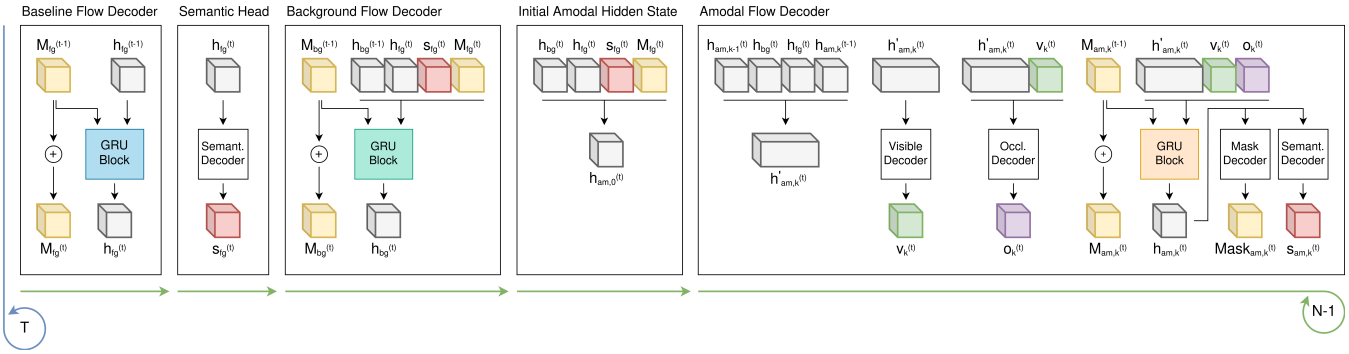

Fig. 2: AmodalFlowNet architecture. Flow and corresponding amodal masks (yellow blocks) are estimated recurrently over both refinement steps (outer, blue arrow) and amodal layers (inner, green arrows). The decoder structure for the standard optical flow is retained from the baseline model. Additional semantic and mask predictions (red, green, and purple blocks) guide the network.

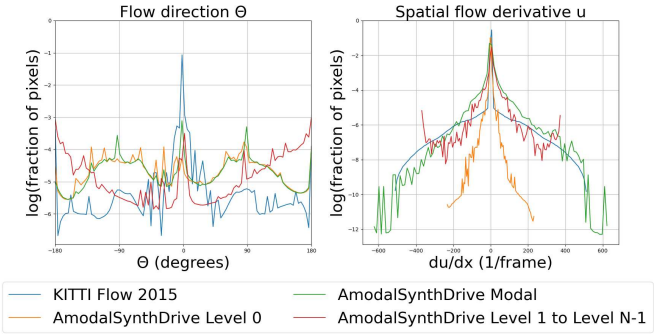

Fig. 3: Log histograms of the optical flow direction and the spatial derivative of the horizontal optical flow velocity u. The modal statistics of AmodalSynthDrive (green) demonstrate similarities to KITTI (blue), indicating similar motion patterns, albeit the existence of spatial irregularity. This distinction can be attributed to the fact that the optical flow ground truth in our dataset is characterized by its high level of detail and precision.

truth in our dataset, we perform statistical analyses as depicted in Fig. 3 on both our dataset and the KITTI dataset [46]. Our motivation for this analysis is grounded in the idea that when the statistical characteristics of the computed flow align, it hints at a likely similarity in the depicted scenes and motions.

V. Methodology

In order to demonstrate the feasibility of amodal optical flow estimation, we present two baselines. To this end, we adapt three pre-trained standard optical flow estimation methods to the amodal setting: GMA [10], GMFlow+ [11], [36], and FlowFormer++ [12]. We chose these approaches specifically due to their respective contributions: GMA represents the natural successor to the transformational RAFT [9] approach, extending it with a motion aggregation module to aid refinement. GMFlow+ provides convincing results for the standard optical flow task, resulting from its strong transformer-based feature extractor. FlowFormer++ similarly achieves state-of-the-art performance but uses a more expressive matching cost representation by applying a transformer encoder-decoder network to the cost volume instead. We presume the latter to be particularly valuable for the amodal optical flow task. Our proposed AmodalFlowNet module integrates well with all three approaches.

AmodalFlowNet Architecture: Our core amodal optical flow prediction is structured in a recurrent hierarchical manner using three flow decoder modules as shown in Fig. 2. In summary, we first estimate the standard optical flow using the respective baseline approach, representing the first decoder. From this, the second decoder predicts the flow field for the background (M_0). Lastly, the third decoder recurrently predicts amodal flow and masks for all amodal object layers, ordered from front to back (i.e., $M_1, M_2, \ldots, M_{N-1}$), again employing the previously predicted standard flow field as an additional input. This recurrent hierarchical structure allows the third decoder to carry over information from the previous layers, enabling occlusion-aware prediction.

Notably, both baselines employ a RAFT-like recurrent refinement framework [9]. This signifies that flow is estimated over the course of multiple refinement iterations, each predicting a residual flow update. The full flow output is then computed as the sum over all residual updates. In this context, the aforementioned amodal estimation process takes place during each (outer) refinement iteration. Furthermore, we exploit the hidden state of the recurrent update module proposed by RAFT to exchange information between the different decoders. In particular, we maintain $N+1$ hidden states: one for the standard flow (h^t_{fg}), one for the background (h^t_{bg}), and one for each amodal object layer ($h^t_{\text{am},k}$, $k \in \{1,\ldots,N-1\}$). While h^t_{fg} is kept as-is from the baseline, the hidden state of the background decoder h^t_{bg} is fused with the hidden state of the standard flow decoder h^t_{fg} before being passed into the background update module in each outer refinement iteration. Similarly, the hidden state of each amodal layer $h^t_{\text{am},k}$ is fused with the hidden states corresponding to the standard flow h^t_{fg}, background flow h^t_{bg}, and previous layer $h^t_{\text{am},k-1}$. We construct $h^t_{\text{am},0}$ as a fusion of hidden states from the standard (h^t_{fg}) and background (h^t_{bg}) decoders. This therefore creates a network that is recurrent over both time steps and amodal layers.

To provide additional guidance to our architecture, we predict several auxiliary targets and consider them in our training loss. For the object layers (1 to N), we decompose the amodal masks into visible and occluded regions akin to [15], [47], predicting both separately from the fused hidden state of the layer. The decomposed masks are then fed back to the update module to aid in the estimation of the complete amodal mask and flow. Additionally, we predict semantic

labels for both the full frame and the amodal object layers. While the semantic labels of the amodal layers are only used as an additional training target, the semantic labels of the full frame are fed to the update module of the background flow and the initial object layer hidden state $h^t_{a,0}$.

Training Loss: We adapt the sequence loss introduced by RAFT [9] by adding amodal and background flow, semantic, and mask terms to the per-iteration loss. In particular, we employ a L_1 end-point loss term for each flow output per iteration, forcing amodal flow predictions to zero outside the area of amodal objects. The semantic labels are trained with a cross-entropy loss, and the visible and occlusion masks with a binary cross-entropy loss. Therefore, the total loss can be written as follows:

$$\mathscr{L} = \sum_{t=0}^{T} \gamma^{T-t} \left(F_{t,\text{fg}} + \mathscr{L}_{\text{am},t} \right), \tag{9}$$

$$\mathscr{L}_{\text{am},t} = F_{t,\text{bg}} + S_{t,\text{fg}} + \sum_{k=1}^{N-1} F_{t,k} + S_{t,k} + M_{t,k,\text{vis}} + M_{t,k,\text{occ}}, \tag{10}$$

where $S_{t,\text{fg}}$ and $S_{t,k}$ are the cross-entropy losses for the full and k-th layer amodal segmentation labels, respectively; $M_{t,k,\text{vis}}$ and $M_{t,k,\text{occ}}$ the binary cross-entropy losses for the amodal visible and occlusion masks; and $F_{t,\text{fg}}$, $F_{t,\text{bg}}$, and $F_{t,k}$ the L_1 end-point error losses for the standard, background, and zero-extended amodal optical flow.

VI. Experimental Evaluation

In this section, we first present the benchmarking results on the AmodalSynthDrive dataset [17] in Sec. VI-A, followed by an ablation study on the different architectural design choices of AmodalFlowNet in Sec. VI-B. We then evaluate the utility of amodal optical flow for the downstream task of panoptic tracking [48]. Finally, we present qualitative comparisons in Sec. VI-D.

We trained all models for 120,000 iterations with a batch size of six and four recurrent refinement steps. We initialize the modal base architecture of the baselines with pre-trained Sintel [49] checkpoints, using Xavier initialization for the rest of the network. For the remaining hyperparameters, we follow the Sintel fine-tuning strategy of FlowFormer++ [12].

A. Benchmarking Results

We present results from evaluations of the baselines on the validation and test sets of AmodalSynthDrive in Tab. I, again using 4 recurrent iterations. More specifically, we compare our FlowFormer++-based AmodalFlowNet architecture (as described in Sec. V) against GMA- and GMFlow+-based learned baselines without mask and semantic guidance (cf. M1 in Sec. VI-B and Tab. II). Additionally, we evaluate two strategies for performing simple non-learned flow infilling of occluded areas based on the amodal mask predictions from our network, one extending the border between visible and occluded regions to the occluded region, whereas the other uses the mean of the visible area.

Comparing learned and non-learned baselines, we observe that there is clearly a need for techniques with a more comprehensive understanding of objects and their motion to effectively address the amodal optical flow task. In particular, AmodalGMFlow+ outperforms both non-learned baselines for pure optical flow (mWAUC) by over 64 % and 46 % (16.2 pp and 11.0 pp) on the validation and test split, respectively, leading to clear gains in the AFQ score. Notably, however, it performs worse than our GMA-based baseline, indicating that global reasoning (in this case through motion aggregation) is better suited to the amodal task than pure improvements in the feature representation. Our complete AmodalFlowNet achieves an additional 20 % (6.7 pp) improvement in AFQ over AmodalGMA on the test split.

TABLE I: Comparison of amodal optical flow performance on the AmodalSynthDrive dataset. All scores are in [%].

Method	Val Set			Test Set		
	AFQ	mWAUC	mIoU	AFQ	mWAUC	mIoU
Near boundary	21.7	20.1	23.5	20.1	18.8	21.5
Mean	24.3	25.3	23.5	22.5	23.6	21.5
AmodalGMFlow+	31.2	41.5	23.5	27.3	34.6	21.5
AmodalGMA	41.6	43.7	39.6	32.9	34.8	32.8
AmodalFlowNet (ours)	**45.8**	**49.4**	**42.4**	**39.6**	**42.0**	**37.3**

TABLE II: Ablation study on the various architectural components of our proposed AmodalFlowNet on the validation set of the AmodalSynthDrive dataset. All scores are in [%].

Model	Guidance	AFQ	mWAUC	mIoU
M1	—	44.6	48.9	40.7
M2	masks	43.8	46.8	41.0
M3	masks + semantics	**45.8**	**49.4**	**42.4**

B. Ablation Study

We ablate the architecture of our proposed AmodalFlowNet on the AmodalSynthDrive validation split to study the impact of our semantics- and mask-based guidance strategy. Results from this experiment are shown in Tab. II. We compare our model with full guidance, as described in Sec. V (M3), against a simplified version (M2) without semantic training. Therein, the amodal semantic decoder is dropped completely, whereas the semantic decoder for the full scene is replaced by a decoder predicting a motion mask, representing the union over all amodal masks of any moving objects in the scene. Lastly, we also evaluate a model (M1) without any guidance, removing the prediction of semantics/motion masks as well as the decomposed amodal visible and occlusion masks entirely.

The results show that the prediction of decomposed amodal masks in combination with guidance through semantic information leads to significant performance improvements. Notable hereby is that while predicting masks without semantics improves the IoU of the amodal masks, it has a detrimental effect on the optical flow quality as measured by the mWAUC score. Including semantics, however, improves both mask and flow predictions. We argue that providing semantic information likely helps our network to better localize and refine flow in amodal occluded regions. Per-layer statistics confirm that improvements are largely due to better performance on the amodal object layers. This demonstrates that our proposed AmodalFlowNet can successfully associate

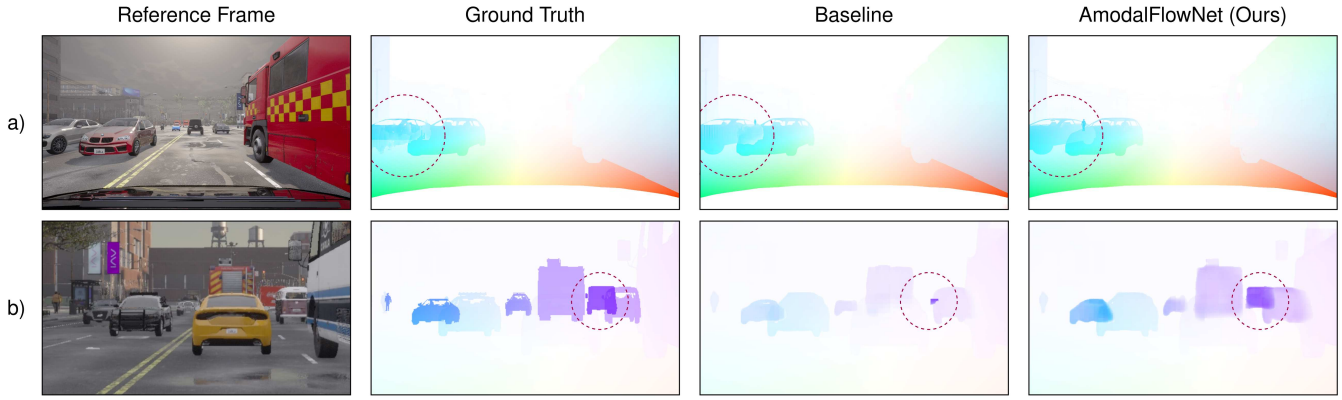

Fig. 4: Qualitative comparison of amodal optical flow prediction from our proposed AmodalFlowNet with the baseline M1 on the AmodalSynthDrive dataset. For visualization, we sequentially superimpose the multi-layer amodal optical flow predictions, in order of $M_0, M_1, \ldots, M_{N-1}$

shape and motion information and exploit them for both amodal flow and mask refinement.

C. Exploiting Amodal Flow for Panoptic Tracking

In this section, we present experimental results, showing the benefits of amodal optical flow for the complex downstream task of panoptic tracking by building upon a mask propagation framework [50] that leverages the output of a panoptic segmentation network to extract object masks. This framework uses optical flow to warp each predicted object between consecutive frames, followed by IoU matching with the Hungarian algorithm to assign consistent tracking IDs.

To study the effectiveness of amodal optical flow for panoptic tracking, we developed two distinct tracking models: Modal-MaskProp (MMP) and Amodal-MaskProp (AMP). We employ APSNet [15] for generating panoptic segmentation predictions for both models to ensure a fair comparison. MMP uses FlowFormer++ for optical flow estimation. Conversely, AMP employs our proposed AmodalFlowNet and adapts the mask propagation framework to accommodate amodal optical flow and amodal object masks. This model takes both amodal and modal object masks from APSNet as input in conjunction with amodal optical flow predictions. To identify the correct motion field map for a given object from the N amodal optical flow layers, we compute the overlap between the amodal mask of the object and the masks of each layer in the amodal optical flow, selecting the layer with the highest overlap. We then use this layer to warp the amodal object mask and perform IoU matching to assign consistent tracking IDs. We evaluate the performance using the standard Panoptic Tracking (PAT) metric [51].

Results are shown in Tab. III. We observe that AMP outperforms MMP by margins of 4.7 % and 4.4 % on the validation and test sets, respectively. Given that both models use the same panoptic segmentation architecture, the Panoptic Quality (PQ) component of the PAT metric remains the same. The Tracking Quality (TQ) component of the PAT metric, which assesses object association performance across frames, demonstrates improvements of 8.8 % for both validation and test sets. These results substantiate the utility of amodal optical flow for enhancing tracking quality when paired with an appropriate amodal segmentation approach.

TABLE III: Comparison of panoptic tracking performance on the AmodalSynthDrive dataset. All scores are in [%].

Method	Val Set			Test Set		
	PAT	TQ	PQ	PAT	TQ	PQ
Modal-MaskProp	50.6	46.2	56.1	49.0	44.3	54.8
Amodal-MaskProp	**53.0**	**50.3**	**56.1**	**51.2**	**48.2**	**54.8**

D. Qualitative Evaluation

Fig. 4 presents qualitative comparisons of the amodal optical flow prediction from the proposed AmodalFlowNet with the M1 baseline (see Sec. VI-B). We observe that AmodalFlowNet distinguishes between visible and occluded scene elements, as highlighted by the red circle, owing to its amodal semantic grounding coupled with the recurrent hierarchical feature propagation from background to amodal flow layer M_{N-1}. However, it falls short of providing a complete amodal flow profile for the occluded objects, as seen in Fig. 4(b). Despite its limitations, we believe that it poses a promising baseline for future research on amodal optical flow.

VII. Conclusion

In this work, we introduced the novel amodal optical flow estimation task, bringing optical flow to the invisible and occluded regions of scenes and therewith extending it to the amodal setting. To enable quantitative analysis and evaluation, we formulated the amodal flow quality metric. We demonstrated the feasibility of the task by extending the AmodalSynthDrive dataset with ground truth labels for amodal optical flow and by providing the AmodalFlowNet architecture and comparing it with other learned and non-learned baselines. Ablations validate the choices made in our architecture, in particular, that guidance through semantic information and decomposed amodal masks provides valuable information and structure to the network. Lastly, we show that the amodal optical flow estimated from our AmodalFlowNet method is indeed beneficial to panoptic tracking as a downstream application. We therefore believe that amodal optical flow shows great potential for the robotics community and will prove valuable for aiding dynamic scene understanding.

REFERENCES

[1] D. Cattaneo, D. G. Sorrenti, and A. Valada, "Cmrnet++: Map and camera agnostic monocular visual localization in lidar maps," *arXiv preprint arXiv:2004.13795*, 2020.

[2] N. Gosala, K. Petek, P. L. Drews-Jr, W. Burgard, and A. Valada, "Skyeye: Self-supervised bird's-eye-view semantic mapping using monocular frontal view images," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 14 901–14 910.

[3] M. Büchner and A. Valada, "3d multi-object tracking using graph neural networks with cross-edge modality attention," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9707–9714, 2022.

[4] B. K. P. Horn and B. G. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[5] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High Accuracy Optical Flow Estimation Based on a Theory for Warping," in *Europ. Conf. on Computer Vision*, vol. 3024, 2004, pp. 25–36.

[6] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning Optical Flow with Convolutional Networks," in *Int. Conf. on Computer Vision*, 2015, pp. 2758–2766.

[7] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1655.

[8] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[9] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Europ. Conf. on Computer Vision*, vol. 12347, 2020, pp. 402–419.

[10] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning To Estimate Hidden Motions With Global Motion Aggregation," in *Int. Conf. on Computer Vision*, 2021, pp. 9772–9781.

[11] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "GMFlow: Learning Optical Flow via Global Matching," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2022.

[12] X. Shi, Z. Huang, D. Li, M. Zhang, K. C. Cheung, S. See, *et al.*, "FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 1599–1610.

[13] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollar, "Semantic amodal segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 1464–1472.

[14] K. Li and J. Malik, "Amodal instance segmentation," in *Europ. Conf. on Computer Vision*, 2016, pp. 677–693.

[15] R. Mohan and A. Valada, "Amodal panoptic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 21 023–21 032.

[16] S. Zhou, R. He, W. Tan, and B. Yan, "SAMFlow: Eliminating Any Fragmentation in Optical Flow with Segment Anything Model," *arXiv preprint arXiv:2307.16586*, 2023.

[17] A. R. Sekkat, R. Mohan, O. Sawade, E. Matthes, and A. Valada, "Amodalsynthdrive: A synthetic amodal perception dataset for autonomous driving," *arXiv preprint arXiv:2309.06547*, 2023.

[18] E. E. Smith and S. M. Kosslyn, *Cognitive Psychology: Mind and Brain: Pearson New International Edition*. Pearson Higher Ed, 2013.

[19] Z. Deng and L. Jan Latecki, "Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in RGB-depth images," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 5762–5770.

[20] P. Purkait, C. Zach, and I. Reid, "Seeing behind things: Extending semantic segmentation to occluded regions," in *Int. Conf. on Intelligent Robots and Systems*, 2019, pp. 1998–2005.

[21] R. Mohan and A. Valada, "Perceiving the invisible: Proposal-free amodal panoptic segmentation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9302–9309, 2022.

[22] K. Ehsani, R. Mottaghi, and A. Farhadi, "SeGAN: Segmenting and generating the invisible," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6144–6153.

[23] X. Yan, F. Wang, W. Liu, Y. Yu, S. He, and J. Pan, "Visualizing the Invisible: Occluded Vehicle Segmentation and Recovery," in *Int. Conf. on Computer Vision*, 2019, pp. 7618–7627.

[24] H. Dhamo, K. Tateno, I. Laina, N. Navab, and F. Tombari, "Peeking behind objects: Layered depth prediction from a single image," *Pattern Recognition Letters*, vol. 125, pp. 333–340, 2019.

[25] J. Ao, Q. Ke, and K. A. Ehinger, "Image amodal completion: A survey," *Computer Vision and Image Understanding*, vol. 229, p. 103661, 2023.

[26] M. Black and P. Anandan, "Robust dynamic motion estimation over time," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1991, pp. 296–302.

[27] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 2432–2439.

[28] ——, "A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles Behind Them," *Int. Journal of Computer Vision*, vol. 106, no. 2, pp. 115–137, 2014.

[29] J. Wang and E. Adelson, "Representing moving images with layers," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625–638, 1994.

[30] N. Jojic and B. Frey, "Learning flexible sprites in video layers," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[31] A. Kannan, B. Frey, and N. Jojic, "A generative model of dense optical flow in layers," in *Spatial Coherence for Visual Motion Analysis*, 2006, pp. 104–114.

[32] D. Sun, E. Sudderth, and M. Black, "Layered image motion with explicit occlusions, temporal consistency, and depth ordering," in *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[33] D. Sun, E. B. Sudderth, and M. J. Black, "Layered segmentation and optical flow estimation over time," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 1768–1775.

[34] D. Sun, J. Wulff, E. B. Sudderth, H. Pfister, and M. J. Black, "A Fully-Connected Layered Model of Foreground and Background Flow," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 2451–2458.

[35] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical flow with semantic segmentation and localized layers," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 3889–3898.

[36] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying Flow, Stereo and Depth Estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2023.

[37] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Europ. Conf. on Computer Vision*, ser. Lecture Notes in Computer Science, 2010, pp. 282–295.

[38] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1187–1200, 2014.

[39] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 4083–4090.

[40] H. Lamdouar, W. Xie, and A. Zisserman, "Segmenting invisible moving objects," in *Proceedings of the The 32nd British Machine Vision Conference*, 2021.

[41] J. Xie, W. Xie, and A. Zisserman, "Segmenting moving objects via an object-centric layered representation," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 28 023–28 036.

[42] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 3899–3908.

[43] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "SegFlow: Joint Learning for Video Object Segmentation and Optical Flow," in *Int. Conf. on Computer Vision*, 2017, pp. 686–695.

[44] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for benchmarks," in *Int. Conf. on Computer Vision*, 2017, pp. 2213–2222.

[45] J. V. Hurtado and A. Valada, "Semantic scene segmentation for robotics," in *Deep Learning for Robot Perception and Cognition*, 2022, pp. 279–311.

[46] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 3061–3070.

[47] M. Tran, K. Vo, K. Yamazaki, A. Fernandes, M. Kidd, and N. Le, "Aisformer: Amodal instance segmentation with transformer," *arXiv preprint arXiv:2210.06323*, 2022.

[48] J. V. Hurtado, R. Mohan, W. Burgard, and A. Valada, "Mopt: Multi-object panoptic tracking," *arXiv preprint arXiv:2004.08189*, 2020.

[49] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Europ. Conf. on Computer Vision*, 2012, pp. 611–625.

[50] M. Weber, J. Xie, M. Collins, Y. Zhu, P. Voigtlaender, H. Adam, B. Green, A. Geiger, B. Leibe, D. Cremers, *et al.*, "Step: Segmenting and tracking every pixel," *arXiv preprint arXiv:2102.11859*, 2021.

[51] W. K. Fong, R. Mohan, J. V. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada, "Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3795–3802, 2022.