27^{th} Computer Vision Winter Workshop
Terme Olimia, Slovenia, February 14–16, 2024

Dense Matchers for Dense Tracking

Tomáš Jelínek, Jonáš Šerých, Jiří Matas
CMP Visual Recognition Group, Faculty of Electrical Engineering,
Czech Technical University in Prague
{tomas.jelinek,serycjon,matas}@fel.cvut.cz

Abstract. *Optical flow is a useful input for various applications, including 3D reconstruction, pose estimation, tracking, and structure-from-motion. Despite its utility, the problem of dense long-term tracking, especially over wide baselines, has not been extensively explored. This paper extends the concept of combining multiple optical flows over logarithmically spaced intervals as proposed by MFT. We demonstrate the compatibility of MFT with two dense matchers, DKM and RoMa. Their incorporation into the MFT framework optical flow networks yields results that surpass their individual performance. Moreover, we present simple yet effective ensembling strategies that prove to be competitive with more sophisticated, non-causal methods in terms of position prediction accuracy, highlighting the potential of MFT in long-term tracking applications.*

1. Introduction

Obtaining point-to-point correspondences is a classical task in computer vision, useful for a wide range of applications including tracking, structure-from-motion, and localization. Despite the extensive research in wide baseline stereo methods, including those with a time baseline, the domain of dense point correspondences in videos has not been explored until recently [38, 53]. The emergence of the TAP-Vid dataset [11] has further fueled interest in long-term point-tracking methods.

Point-trackers usually [12, 27, 45, 11] track sparse sets of points. However, dense correspondences are useful in various applications, such as video editing, object tracking, and 3D reconstruction. While optical flow techniques provide dense correspondences, they are typically limited to pairs of consecutive frames.

Long-term dense tracking has been recently addressed by Neoral *et al*. [38] MFT tracker, which computes optical flow not only for consecutive frames but also for pairs of more temporally distant frames, including flow computation between the reference and every other frame of the video. At every frame, optical flow is computed w.r.t. the previous, first, and a constant number of logarithmically spaced frames. Such approach is linear in the number of frames and thus not computationally prohibitive.

In the original MFT[38], all optic flow computations are based on RAFT [50], which has performed well in both standard benchmarks [2, 36] and in applications. However, the RAFT optical flow network was trained on pairs of consecutive frames, which is likely sub-optimal for large baselines.

Recently, dense matchers such as DKM [14] and RoMa [15] have been published. This development opens the possibility to apply the MFT framework with different dense matchers, or to use RAFT for pairs of frames with short temporal, and thus probably spatial, baseline. The only requirement of the MFT "meta optic flow algorithm" is that the basis dense two view optic flow or matcher provides confidence in its predictions.

In this paper, we evaluate the MFT approach with the DKM and RoMa matchers instead of RAFT. We show that both of these matchers provide accurate matches, but inaccurate occlusion predictions. Addressing the strengths and weaknesses of optical-flow-based and dense-matching-based methods, we propose a combined tracker, that outperforms the original MFT design.

In summary, our contributions are: (1) We show how to adapt dense matchers DKM and RoMa for use in the MFT framework, and experimentally evaluate their performance. (2) We show that the MFT algorithm outperforms both direct flow between the first and the current frame, and the chaining of optical flows computed on consecutive frames for RAFT,

DKM, and RoMa. (3) Based on better results of RoMa over DKM in our experiments, we propose a dense long-term tracker that combines the strengths of RAFT-based MFT and RoMa-based MFT.

2. Related Work

Tracking, 3D Reconstruction, and SLAM Object tracking algorithms [1, 26, 10] traditionally outputted the track of an object specified in the first frame in the form of bounding boxes. Later, the focus shifted towards segmentation-based tracking [29, 41, 34].

Modern model-free trackers based on differentiable rendering [54, 43], that can simultaneously track and reconstruct any object specified in the first frame are naturally able to provide point-to-point correspondences for the tracked object; however, to the best of our knowledge, they can track a single object only or require multi-camera input [33]. Additionally, recent methods [56, 57, 55], involving differentiable rendering of neural radiance fields (NeRFs) [37], show potential in creating deformable 3D models for point tracking. Nonetheless, the extensive computational demands of these methods limit their practical applicability in real-world scenarios.

The traditional SLAM methods [46] produced sparse point clouds. Later on, semi-dense [16, 51] SLAM methods appeared. Some SLAM-based trackers, [17] can densely estimate point positions in static scenes, and recent advances in differentiable rendering opened the avenue for differentiable-rendering-based monocular SLAMs [42] but their application remains constrained to static scenes.

Optical Flow estimation is a classical problem in computer vision, with the early works [32, 20] relying on the brightness-constancy assumption. With the advent of deep neural networks, the focus shifted towards learning-based approaches [13, 49, 23, 50, 21] trained on synthetic data.

Optical flow estimation in state-of-the-art methods, exemplified by RAFT [50] and FlowFormer [21], is achieved through the analysis of a 4D correlation cost volume, considering features of all pixel-pairs. These techniques excel in densely estimating flow between consecutive frames, yet they encounter challenges in accurately determining flow across distant frames, particularly in scenarios with large displacements or significant object deformation.

Multi-step-flow algorithms [7, 6, 8] address the limitations of concatenation-based approaches for long-term dense point tracking. These algorithms create extended dense point tracks by merging optical flow estimates across variable time steps, effectively managing temporarily occluded points by bypassing them until their re-emergence. However, their dependence on the brightness constancy assumption renders them less effective over distant frames. Subsequent works in multi-step-flow, such as the multi-step integration and statistical selection (MISS) approach by Conze et al. [4, 5], further refine this process. This approach relies on generating a multitude of candidate motion paths from random reference frames, with the best path selected through a global spatial smoothness optimization process. However, this strategy makes these methods computationally demanding. Although certain optical flow techniques [24, 39, 22, 59, 31, 58] address occlusions and flow uncertainty, most leading optical flow methods, influenced by standard benchmarks like those in Butler *et al*. [3] and Menze *et al*. [35], do not detect occlusions. Jiang *et al*. [25], building on RAFT [50], has taken a different approach in which they handle occlusion implicitly by computing hidden motions of the occluded objects. However, the method still falls short in the context of tracking dynamic, complex motions.

We now describe in greater depth three methods that are most relevant to our paper: RAFT [38], DKM [14], and RoMa [15]. While the latter two are in fact dense matchers, we will use the term interchangeably with long-ranged optical flow estimation with occlusion prediction.

MFT extends optical flow into dense long-term trajectories by constructing multiple chains of optical flows and selecting the most reliable one [38]. The flow chains consist of optical flow computed both between consecutive frames, and between more distant frames, which allows for re-detecting points after occlusions. The intervals between distant frames are chosen to be logarithmically spaced.

MFT extends the RAFT optical flow method with two heads, estimating occlusion and uncertainty for each flow vector. Like the optical flow, the uncertainty and the occlusion are accumulated over each chain, and the non-occluded flow chain with the least overall uncertainty is selected as the most reliable candidate. The long-term tracks of different points thus chain possibly different sequences of optical flows. This strategy on one hand takes into account

that changes in appearance and viewpoint gradually accumulate over time, which makes it more reliable to chain flows on easier-to-match frames rather than estimating matches directly between the template and the current frame. On the other hand, short chains containing longer jumps with low uncertainty result in less error accumulation.

DKM proposed by Edstedt *et al*. [14], a dense point-matching method, employing a ResNet [19]-based encoder pre-trained on ImageNet-1K [44] for generating both fine and coarse features. The coarse features undergo sparse global matching, modeled as Gaussian process regression, to determine embedded target coordinates and certainty estimates. Fine features are refined using CNN refiners, following a methodology similar to Truong *et al*. [52] and Shen *et al*. [47]. DKM's match certainty estimation relies on depth consistency, necessitating 3D supervision. The process concludes by filtering matches below a certainty threshold of 0.05 weighted sampling for match selection. Edstedt *et al*. [14] released outdoor and indoor models trained on MegaDepth [30]) and ScanNet [9] respectively.

RoMa similarly to DKM, RoMa [15] is a dense matching method that provides pixel displacement vectors along with their estimated certainty, building upon the foundation set by DKM [14]. RoMa differentiates itself by employing a two-pronged approach for feature extraction: using frozen DINOv2 [40] for sparse features and a specialized ConvNet with a VGG19 backbone [48] for finer details. Unique to RoMa is their transformer-based match decoder, which matches features through a regression-by-classification approach, better handling the multi-modal nature of coarse feature matching. In contrast to DKM, RoMa's pipeline omits the use of dense depth maps for match certainty supervision, relying instead on pixel displacements for match supervision. Their model is trained on datasets like MegaDepth [30] and ScanNet [9], similar to DKM.

Long-Term Point Tracking aiming to track a set of physical points in a video has emerged significantly since the release of TAP-Vid [11]. The dataset's baseline method TAP-Net [11] computes a cost volume for each frame, employing a technique akin to RAFT's approach [50]. It focuses on tracking individual query points. PIPs [18] takes this approach to an extreme by completely trading off spatial awareness about other points for temporal awareness within fixed-sized temporal windows, making it unable to re-detect the target after longer occlusions. TAPIR [12] combines TAP-Net's track initialization with PIPs' refinement while removing the PIPs' temporal chunking, using a time-wise convolution instead. CoTracker [27] models the temporal correlation of different points via a sliding-window transformer, modeling multiple tracks' interactions. While these methods are designed for sparse tracking, they can provide dense tracks by querying all points in the first frame.

Notably, differentiable rendering has been leveraged in recent approaches, with OmniMotion representing 3D points' motion implicitly using learned bijections [53] enabling it to provide dense tracks. Alternative methods like [33] which models the scene as temporally-parametrized Gaussians[28]. However, these methods have their limitations, such as OmniMotion's quadratic complexity and the multi-camera requirement of [33].

3. Method

For a stream $\{\mathcal{I}_1, ..., \mathcal{I}_N\}$ of N video frames defined on a common image domain Ω, we denote the optical flow between frames i and j as $\mathcal{F}^{(i,j)}$. Moreover, we use $\sigma^{(i,j)} \in \mathbb{R}_+^{\Omega}$ to denote the estimated flow variance, and $\rho^{(i,j)} \in [0,\, 1]^{\Omega}$ to represent the estimated certainty of $\mathcal{F}^{(i,j)}$. Finally, occlusion score $o^{(i,j)} \in [0,\, 1]^{\Omega}$ denotes the estimated probability of pixels appearing in frame i being occluded in frame j. To simplify notation, although $\mathcal{F}^{(i,j)}$, $\rho^{(i,j)}$, and $\sigma^{(i,j)}$ are 2D or 3D tensors, we will use these symbols to denote their values at a specific point $\boldsymbol{p} = (x,\, y)$ in the image. Moreover, for every point $\boldsymbol{p}_i$ in frame i, its predicted position $\boldsymbol{p}_j$ in frame j relates to the optical flow $\mathcal{F}^{(i,j)}$ as follows:

$$\boldsymbol{p}_j = \boldsymbol{p}_i + \mathcal{F}^{(i,j)}(\boldsymbol{p}_i). \tag{1}$$

Let us denote by ϕ_{RAFT}, ϕ_{DKM}, ϕ_{RoMa}, ϕ_{MFT} the functions computed by RAFT, DKM, RoMa, and MFT respectively. By RAFT we mean the MFT's adaptation of RAFT with additional uncertainty and occlusion heads [38]. The output vectors of these methods are as follows:

$$\phi_{\text{RAFT}} = (\mathcal{F}^{(i,j)},\, \sigma^{(i,j)},\, \rho^{(i,j)}) \tag{2}$$

$$\phi_{W} = (\mathcal{F}^{(i,j)},\, \rho^{(i,j)}) \tag{3}$$

$$\phi_{\text{MFT}} = (\mathcal{F}^{(i,j)},\, \sigma^{(i,j)},\, o^{(i,j)}), \tag{4}$$

where W is one of the wide-baseline methods, either DKM or RoMa.

3.1. MFT Flow Chaining

MFT [38] achieves long-term optical flow estimation by combining multiple optical flows. These flows are obtained from ϕ_{RAFT} over logarithmically spaced distances. When estimating the flow $\mathcal{F}^{(1,j)}$, MFT utilizes a sequence of intermediate flows. This sequence, denoted as $\mathcal{S}$, comprises flows $\mathcal{F}^{(j-\Delta_1,j)}$, ..., $\mathcal{F}^{(j-\Delta_K,j)}$. Here, Δ_i represents logarithmic spacing and is defined as 2^{i-1} for $i < K$, with $\Delta_K = j - 1$. We limit the number of intermediate flows, denoted by K, to a maximum of 5 and ensure that $j - \Delta_{K-1} > 1$.

Additionally, MFT employs a scoring function for evaluating the quality of the intermediate flows chaining for each image point $\boldsymbol{p}_1$ in the reference frame 1. The scoring function $s^{(j-\Delta_k,j)}$ utilizes chaining of estimated flow variances and occlusion scores over an intermediate frame i:

$$\sigma^{(1,i,j)}(\boldsymbol{p}_1) = \sigma^{(1,i)}_{MFT}(\boldsymbol{p}_1) + \sigma^{(i,j)}(\boldsymbol{p}_i), \quad (5)$$

$$o^{(1,i,j)}(\boldsymbol{p}_1) = \max\{o^{(1,i)}_{\text{MFT}}(\boldsymbol{p}_1),\, o^{(i,j)}(\boldsymbol{p}_i)\}, \quad (6)$$

The point $\boldsymbol{p}_i$ is computed using $\mathcal{F}^{(1,i)}_{MFT}$ and the relation in Equation 1. The scoring function is then defined as $s^{(j-\Delta_k,j)}(\boldsymbol{p}_1) = -\sigma^{(1,j-\Delta_k,j)}(\boldsymbol{p}_1)$. If the chained occlusion score $o^{(1,j-\Delta_k,j)}(\boldsymbol{p}_1)$ exceeds an occlusion threshold θ_o, we set $s^{(j-\Delta_k,j)}(\boldsymbol{p}_1) = -\infty$. This score is used to select the best flow for every point $\boldsymbol{p}_1$, that is the flow with the lowest estimated variance computed on chains that do not contain occluded points.

MFT computes long-term flow for any point $\boldsymbol{p}_1$ in the reference frame 1 iteratively via chaining as

$$\mathcal{F}^{(1,j)}_{MFT}(\boldsymbol{p}_1) = \mathcal{F}^{(1,i_M)}_{MFT}(\boldsymbol{p}_1) + \mathcal{F}^{(i_M,j)}(\boldsymbol{p}_{i_M}), \quad (7)$$

where $i_M \in \{j - \Delta_k \mid 1 \leq k \leq K\}$ such that the score $s^{(i_M,j)}(\boldsymbol{p}_1)$ is maximal. Again, the point $\boldsymbol{p}_{i_M}$ is obtained using $\mathcal{F}^{(1,i_M)}_{MFT}(\boldsymbol{p}_1)$ and Equation 1. $\mathcal{F}^{(i_M,j)}$ is the flow obtained from an arbitrary method that can also estimate its variance $\sigma^{(i_M,j)}$ and occlusion score $o^{(i_M,j)}$. The flow chaining is visualized in Figure 1.

The estimated variance and occlusion score for frame j are then obtained from the chain over frame i_M as $\sigma^{(1,j)}_{MFT}(\boldsymbol{p}_1) = \sigma^{(1,i_M,j)}(\boldsymbol{p}_1)$, respectively $o^{(1,j)}_{\text{MFT}}(\boldsymbol{p}_1) = o^{(1,i_M,j)}(\boldsymbol{p}_1)$. A pixel observed in frame i is considered occluded in frame j if its value $o^{(i,j)}_{\text{MFT}}$ is above a threshold θ_o. In practice, we set different thresholds for different backbone networks as we discuss in Subsection 3.2.

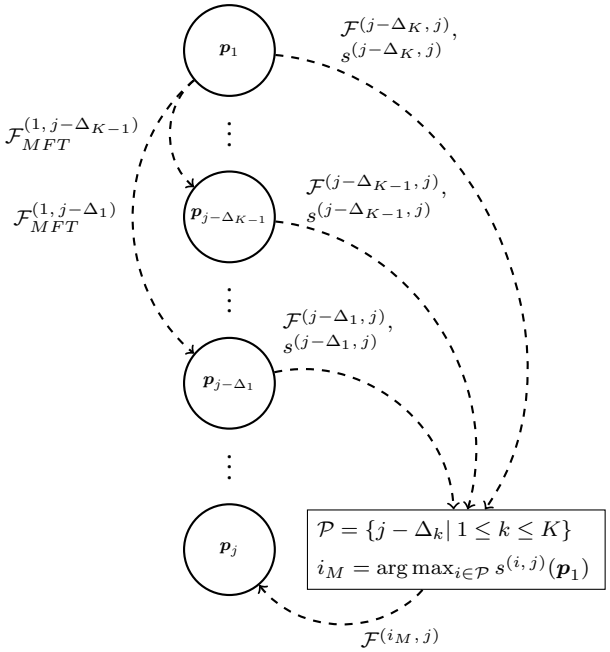


Figure 1: Illustration of the MFT flow chaining as defined in Equation 7. The optical flows are evaluated on the points in the outbound nodes of their respective arcs.

3.2. Integration of DKM and RoMa

As we mentioned in the Introduction, we make the conjecture that training RAFT for optical flow prediction on consecutive video frames is suboptimal for wide baselines. We therefore integrate DKM and RoMa, capable of handling wider baselines. However, integrating these methods with MFT poses certain challenges due to their incompatible outputs.

In the first place, neither RoMa nor DKM provides an occlusion score o, but only an estimate of the flow prediction certainty ρ. We therefore artificially set their occlusion scores as $o = 1 - \rho$. Furthermore, although σ and ρ both represent the quality of estimated optical flow, they are not directly comparable. But in order to integrate them into the MFT framework, we need to converse between them.

Through empirical analysis, we established a flow certainty threshold θ_ρ. When ρ exceeds this threshold, we deem the optical flow reliable, assigning $\sigma = 0$. Conversely, when ρ is below this threshold, σ is set to 1000, correlating higher uncertainties with increased variances in predicted flow. Additionally, we observed that while o_{MFT}, o_{DKM}, and o_{RoMa} fundamentally represent the same concept, their respective occlusion thresholds $\theta_{o_{\text{RAFT}}}$ and $\theta_{o_{\text{RoMa}}}$ vary.

In our experiments in Section 4, we use

$$\theta_{o_{\text{RAFT}}} = 0.02,\ \theta_{o_{\text{DKM}}} = \theta_{o_{\text{RoMa}}} = 0.95. \quad (8)$$

For a visual comparison between the original MFT and the integration of RoMa into MFT, see Figure 2.

3.3. Ensembling

We observed that, in terms of occlusion prediction, MFT's modification of RAFT achieves higher accuracy compared to RoMa. Conversely, RoMa exhibits better performance in optical flow prediction relative to RAFT. Based on these findings, we developed an integrated approach that combines the strengths of both methods. Specifically, our method utilizes occlusion data from RAFT, while RoMa is employed for position prediction, with both processes executed in parallel within the MFT framework. As detailed in Section 4, our most effective strategy involves employing RAFT for occlusion score prediction and RoMa for position prediction, provided the point is not predicted as occluded; in cases of occlusion, RAFT's predictions are preferred.

4. Experiments

In this section, we evaluate our proposed method. Initially, we compare the MFT framework with direct optical flow prediction and simple optical flow chaining. Subsequently, we explore RoMa's optical flow prediction performance within the MFT framework depending on whether it predicts the point as occluded or non-occluded, which serves as a foundational finding for our most effective ensembling strategy. The final part of our experimentation serves as a comparison of different ensembling strategies, justifying the design of our most effective architecture, and comparing it to other tracking methods.

Evaluation setup Our experiments were conducted on all 30 tracks of the TAP-Vid-DAVIS dataset [11] with a resolution of 512×512 using the *first* evaluation mode. This approach aligns with the methodology described in MFT [38]. It is important to stress that in the dataset, the tracks are annotated only sparsely with more focus on the foreground objects rather than the static background.

Evaluation metrics In assessing the performance of our approach, we employ three key metrics as defined by the TAP-Vid benchmark. The Occlusion Accuracy (OA) evaluates the accuracy of classifying the points as occluded. We measure the quality of the predicted positions, using average displacement error, denoted as $<\delta^x_{avg}$. This metric calculates the fraction of visible points with a positional error below specific thresholds, averaged over thresholds of 1, 2, 4, 8, and 16 pixels. These accuracies for individual thresholds are denoted as $< i$ with i representing the threshold. Additionally, the Average Jaccard (AJ) as defined in [11] index is used to collectively assess both occlusion and position accuracy.

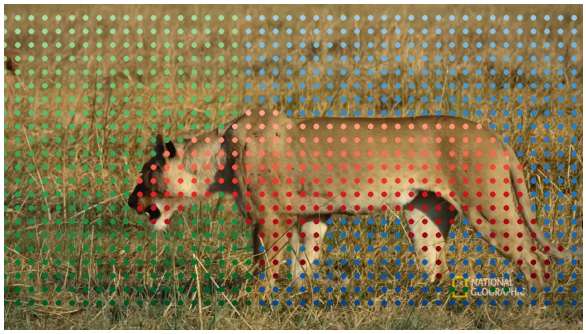
4.1. MFT Chaining

A key aspect of our analysis involves contrasting the performance of RAFT, DKM, and RoMa within the MFT framework against *direct* optical flow prediction with the first frame serving as a reference, and *chaining* of the optical flows computed on consecutive video frames. The results presented in Table 1 clearly show that for each base method (RAFT, DKM, RoMa), the MFT strategy consistently outperforms the other strategies in all metrics by a large margin. These results underscore the effectiveness of MFT in handling complex motion trajectories over extended periods, surpassing the limitations of direct prediction and simple chaining methods. A key observation exemplified in Figure 2 is that RoMa is substantially less prone to predict mismatches in the background than RAFT.

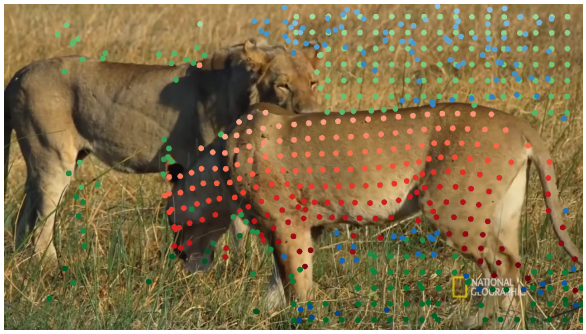
The results in Table 1 also show that RoMa within the MFT paradigm achieves arguably the best results in position prediction, while RAFT outperforms all other methods in the occlusion classification accuracy. This finding serves as a foundation for our ensemble strategies in Subsec. 3.3. Due to the consistently better performance of RoMa over DKM in the evaluation benchmark in all, average Jaccard, average displacement error, and occlusion accuracy we from now on focus our experiments on RoMa even if DKM runs slightly faster.

4.2. RoMa Visibility

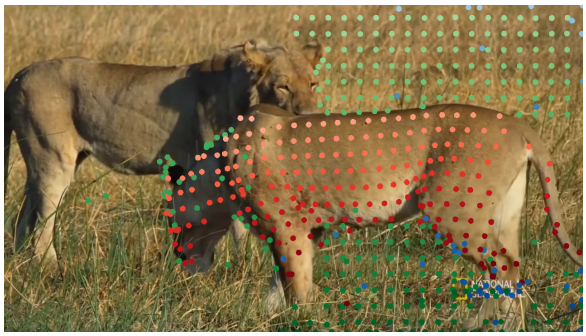
While RoMa demonstrates high accuracy in position prediction, its capability in occlusion detection is relatively limited in comparison to RAFT. However, the quality of occlusion prediction is vital for scoring the optical flows as described in Subsec. 3.1, and thus for computing new flows. We hence conjecture that if we only use the RoMa's optical flow predictions that are predicted as not occluded, we can achieve even better tracking results. The results, as shown



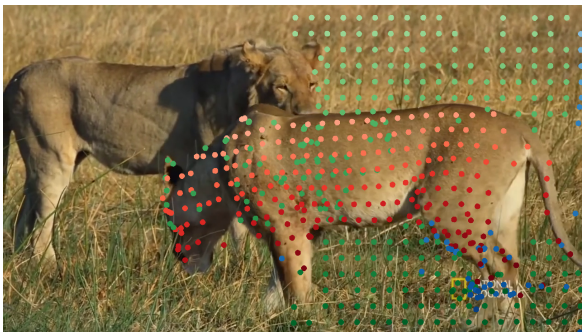
(a) Reference frame



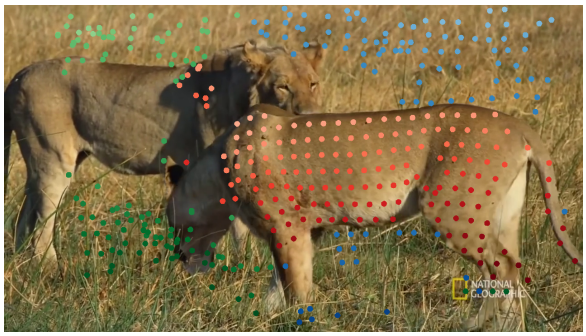
(b) RAFT-based MFT Strategy.



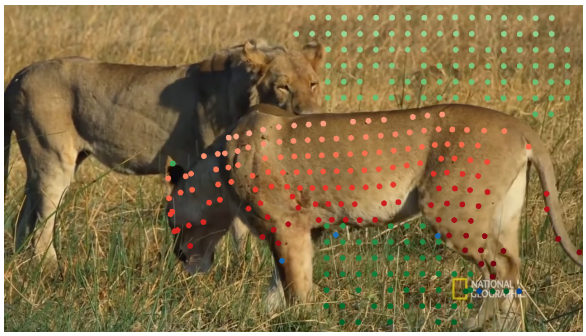
(c) RoMa-based MFT Strategy.



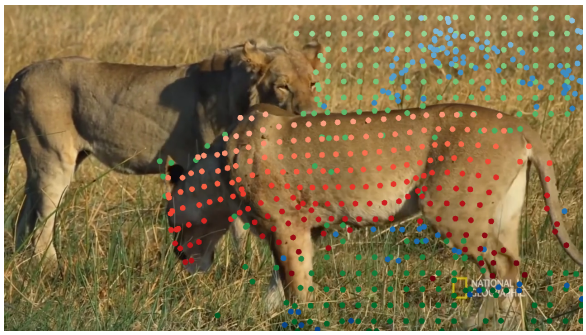
(d) DKM-based MFT Strategy.



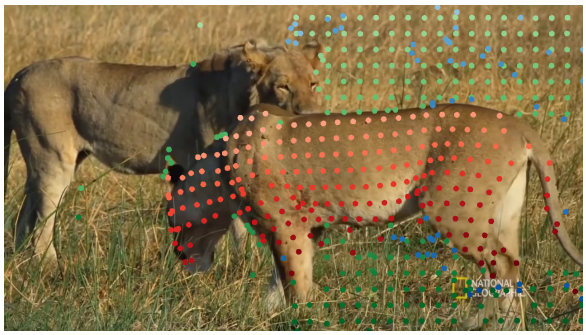
(e) Direct matching between frames #0 and #140 using RAFT.



(f) Direct matching between frames #0 and #140 using RoMa.



(g) Combined RAFT and RoMa strategy.



(h) Selective RoMa position prediction.

Figure 2: Visual comparison of selected dense tracking methods: (a) reference frame #0; (b)-(h) predicted positions of points in frame #140. All blue points are invisible in frame #140; blue points in (b)-(h) thus indicate false matches. Green points are visible both in frame #0 and frame #140. Red points highlight the points on the body of the lioness. Different shades are used to identify different points. The sequence is available at `https://cmp.felk.cvut.cz/~serycjon/MFT/visuals/ugsJtsO9w1A-00.00.24.457-00.00.29.462_HD.mp4`.

		main metrics							
base	strategy	AJ	$<\delta^x_{avg}$	OA	<1	<2	<4	<8	<16
	direct	38.4	50.8	65.6	29.0	44.1	54.6	60.4	65.7
RAFT	chain	38.7	55.0	69.5	25.2	43.8	59.4	70.4	76.3
	MFT	**47.4**	**67.1**	**77.7**	**34.0**	**57.3**	**74.3**	**82.8**	**86.9**
	chain	27.3	63.5	48.2	36.4	56.2	69.4	76.0	79.6
DKM	direct	34.0	60.7	52.8	37.0	54.5	65.3	70.9	76.0
	MFT	**47.8**	**72.0**	**70.2**	**43.0**	**65.8**	**79.0**	**84.5**	**87.8**
	direct	37.7	63.7	57.6	37.5	55.9	67.8	75.5	81.5
RoMa	chain	40.3	63.1	60.7	36.8	55.3	68.1	75.5	79.8
	MFT	**48.8**	**72.7**	**71.7**	**43.0**	**65.5**	**79.2**	**85.5**	**90.1**

Table 1: **TAP-Vid DAVIS evaluation of different optical flow combination strategies.** The MFT strategy outperforms both simple chaining and direct matching for all base optical flow methods on all the metrics.

predicted	$<\delta^x_{avg}$	<1	<2	<4	<8	<16
occluded	47.4	18.7	32.7	52.0	62.6	71.1
visible	**77.2**	**46.9**	**70.9**	**84.5**	**89.8**	**93.7**
any	72.7	43.0	65.5	79.2	85.5	90.1

Table 2: **TAP-Vid DAVIS evaluation of MFT-RoMa separated by the occlusion prediction.** Using only the points predicted as not occluded leads to improved position accuracy on all error thresholds.

in Tab. 2, indicate a marked improvement in tracking accuracy when measured only on points predicted as non-occluded.

4.3. Ensembling Strategies

In the concluding part of our experimental analysis, we compare various ensembling strategies within the MFT framework, building on the insights from the previous sections. The results, detailed in Table 3, demonstrate the effectiveness of the ensemble strategy.

RAFT-based MFT Strategy For comparison we show the original MFT strategy, utilizing RAFT for both position and occlusion predictions. This approach, while achieving the highest occlusion accuracy among all ensembling strategies tested, exhibits suboptimal performance in position precision.

RoMa-based MFT Strategy Substituting RAFT entirely with RoMa, we observed an improvement in position prediction accuracy. However, this modification led to a significant decrease in occlusion prediction accuracy, highlighting the trade-offs between these two aspects.

Combined RAFT and RoMa Strategy Our next strategy involved a simple combination of RAFT and RoMa: RAFT for occlusion prediction and RoMa for position prediction. This hybrid approach resulted in enhanced performance across all metrics, outperforming the aforementioned individual strategies.

Selective RoMa Position Prediction However, further refinement was achieved by integrating findings from Subsection 4.2. We found that RoMa's position predictions are more accurate for points it identifies as visible. Therefore, we devised a strategy where MFT-RoMa's position predictions are used only if the points are marked as visible; otherwise, RAFT's predictions are utilized. This selective strategy led to improvements in both position prediction accuracy and occlusion accuracy. We visually compare this strategy with other two best-performing strategies and MFT with RAFT in Figure 3.

Comparison with Point Trackers We observe that our approach closely rivals or exceeds the performance of established sparse point tracking methods like CoTracker and TAPIR in the average position accuracy while achieving worse performance in the occlusion prediction accuracy. It is noteworthy that our method attains these results within a strictly causal framework, contrasting with CoTracker and TAPIR, which utilize attention-based temporal refinement strategies. Moreover, it is important to highlight that, unlike our approach, CoTracker and TAPIR are designed as sparse trackers.

	MFT base		main metrics			visibility	
	position	occlusion	AJ	$<\delta^x_{avg}$	OA	precision	recall
(1)	RAFT	RAFT	47.4	67.1	**77.7**	**78.0**	**91.5**
(2)	RoMa	RoMa	**48.8**	**72.7**	71.7	74.5	85.3
(3)	RoMa	RAFT	50.2	72.7	77.7	78.0	91.5
(4)	RAFT/RoMa	RAFT	**51.6**	**73.4**	**77.7**	**78.0**	**91.5**
	TAPIR		56.2	70.0	86.5		
	CoTracker		61.0	75.9	89.4		

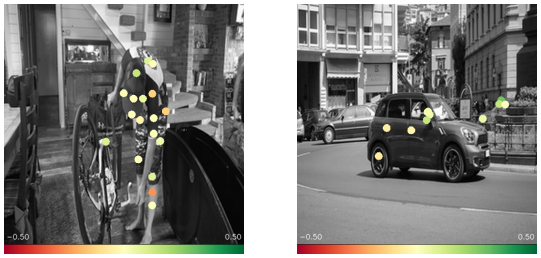
Table 3: **TAP-Vid DAVIS evaluation of combinations of two trackers.** We run MFT-RAFT and MFT-RoMa independently in parallel, using the two outputs for the final position and occlusion prediction. RAFT-based MFT (1) has good occlusion accuracy (OA), RoMa-based MFT (2) has good position accuracy $<\delta^x_{avg}$. Using MFT-RAFT to predict occlusion and MFT-RoMa to predict position (3) achieves better AJ. The best results (4) are achieved when the position is predicted by MFT-RoMa, but only when it predicts visible (see Tab. 2).

5. Conclusion

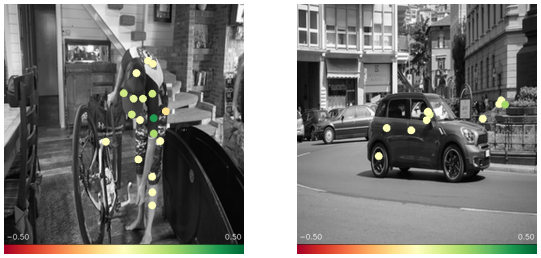
We have showcased the benefits of employing the MFT framework over direct optical flow computation and optical flow chaining. We have also demonstrated the flexibility of the MFT paradigm which can be readily used together with different optical flow computation methods. Without complex architectural modifications and using simple ensemble strategies, we were able to demonstrate position prediction accuracy on the Tap-Vid dataset competing with that of state-of-the-art sparse trackers that utilize non-causal tracking refinement.

Limitations and Future Work Our current approach does not take into account the speed of the baseline optical flow networks. The main limitation is the need for two optical flow networks to operate concurrently within the ensemble strategy. Exploring co-training strategies that enable a single network to deliver similar performance could be a viable solution. A key task is to bridge the existing gap in occlusion prediction accuracy between our method and the state-of-the-art. We also put forward the need for new datasets featuring dense annotations of point tracks in both the foreground and background.

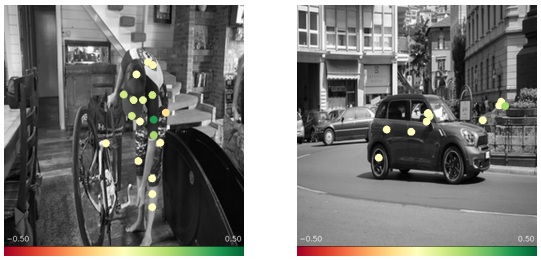
Acknowledgments This work was supported by Toyota Motor Europe and by the Grant Agency of the Czech Technical University in Prague, grant No.SGS23/173/OHK3/3T/13.



(a) RAFT-based MFT Strategy



(b) RoMa-based MFT Strategy



(c) Combined RAFT and RoMa Strategy

Figure 3: Images show the first frames of two selected TAP-Vid DAVIS sequences. Dots represent ground-truth tracking points, with shades of green showing the improvement in $<\delta^x_{avg}$ achieved by the Selective RoMa Position Prediction ensemble over methods (a)-(c), shades of red show the converse.

References

[1] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2544–2550. IEEE, 2010. 2

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 1

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 2

[4] P.-H. Conze, P. Robert, T. Crivelli, and L. Morin. Dense long-term motion estimation via statistical multi-step flow. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 3, pages 545–554. IEEE, 2014. 2

[5] P.-H. Conze, P. Robert, T. Crivelli, and L. Morin. Multi-reference combinatorial strategy towards longer long-term dense motion estimation. *Computer Vision and Image Understanding*, 150:66–80, 2016. 2

[6] T. Crivelli, P.-H. Conze, P. Robert, M. Fradet, and P. Pérez. Multi-step flow fusion: Towards accurate and dense correspondences in long video shots. In *British Machine Vision Conference*, 2012. 2

[7] T. Crivelli, P.-H. Conze, P. Robert, and P. Pérez. From optical flow to dense long term correspondences. In *2012 19th IEEE International Conference on Image Processing*, pages 61–64. IEEE, 2012. 2

[8] T. Crivelli, M. Fradet, P.-H. Conze, P. Robert, and P. Pérez. Robust optical flow integration. *IEEE Transactions on Image Processing*, 24(1):484–498, 2014. 2

[9] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[10] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019. 2

[11] C. Doersch, A. Gupta, L. Markeeva, A. R. Continente, L. Smaira, Y. Aytar, J. Carreira, A. Zisserman, and Y. Yang. TAP-Vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 2022. 1, 3, 5

[12] C. Doersch, Y. Yang, M. Vecerik, D. Gokay, A. Gupta, Y. Aytar, J. Carreira, and A. Zisserman. TAPIR: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10061–10072, October 2023. 1, 3

[13] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[14] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 1, 2, 3

[15] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg. RoMa: Revisiting robust losses for dense feature matching. *arXiv preprint arXiv:2305.15404*, 2023. 1, 2, 3

[16] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1456, 2013. 2

[17] M. Gladkova, N. Korobov, N. Demmel, A. Ošep, L. Leal-Taixé, and D. Cremers. Directtracker: 3d multi-object tracking using direct image alignment and photometric bundle adjustment. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3777–3784. IEEE, 2022. 2

[18] A. W. Harley, Z. Fang, and K. Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 59–75. Springer, 2022. 3

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[20] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981. 2

[21] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li. Flowformer: A transformer architecture for optical flow. *arXiv preprint arXiv:2203.16194*, 2022. 2

[22] J. Hur and S. Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 2

[23] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655, 2017. 2

[24] E. Ilg, T. Saikia, M. Keuper, and T. Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 614–630, 2018. 2

[25] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021. 2

[26] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1409–1422, 2011. 2

[27] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. CoTracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 1, 3

[28] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 3

[29] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L. Č. Zajc, A. Lukežič, O. Drbohlav, et al. The eighth visual object tracking vot2020 challenge results. In *European Conference on Computer Vision*, pages 547–601. Springer, 2020. 2

[30] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[31] S. Liu, K. Luo, N. Ye, C. Wang, J. Wang, and B. Zeng. Oiflow: Occlusion-inpainting optical flow estimation by unsupervised learning. *IEEE Transactions on Image Processing*, 30:6420–6433, 2021. 2

[32] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence-Volume 2*, pages 674–679, 1981. 2

[33] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan. Dynamic 3D gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2, 3

[34] A. Lukezic, J. Matas, and M. Kristan. D3S – a discriminative single shot segmentation tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7133–7142, 2020. 2

[35] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 2

[36] M. Menze, C. Heipke, and A. Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018. 1

[37] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[38] M. Neoral, J. Šerỳch, and J. Matas. MFT: Long-term tracking of every pixel. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6837–6847, 2024. 1, 2, 3, 4, 5

[39] M. Neoral, J. Šochman, and J. Matas. Continual occlusion and optical flow estimation. In *Asian Conference on Computer Vision*, pages 159–174. Springer, 2018. 2

[40] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3

[41] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 2

[42] A. Rosinol, J. J. Leonard, and L. Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3437–3444. IEEE, 2023. 2

[43] D. Rozumnyi, J. Matas, M. Pollefeys, V. Ferrari, and M. R. Oswald. Tracking by 3d model estimation of unknown objects in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14086–14096, 2023. 2

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[45] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *International Journal of Computer Vision*, 80:72–91, 2008. 1

[46] J. L. Schönberger and J.-M. Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[47] X. Shen, F. Darmon, A. A. Efros, and M. Aubry. Ransac-flow: generic two-stage image alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 618–637. Springer, 2020. 3

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[49] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[50] Z. Teed and J. Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1, 2, 3

[51] Z. Teed and J. Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 2

[52] P. Truong, M. Danelljan, and R. Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 3

[53] Q. Wang, Y.-Y. Chang, R. Cai, Z. Li, B. Hariharan, A. Holynski, and N. Snavely. Tracking everything everywhere all at once. *arXiv:2306.05422*, 2023. 1, 3

[54] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Muller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. *CVPR*, 2023. 2

[55] S. Wu, T. Jakab, C. Rupprecht, and A. Vedaldi. DOVE: Learning deformable 3d objects by watching videos. *IJCV*, 2023. 2

[56] G. Yang, D. Sun, V. Jampani, D. Vlasic, F. Cole, H. Chang, D. Ramanan, W. T. Freeman, and C. Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2

[57] G. Yang, M. Vo, N. Neverova, D. Ramanan, A. Vedaldi, and H. Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2863–2873, June 2022. 2

[58] C. Zhang, C. Feng, Z. Chen, W. Hu, and M. Li. Parallel multiscale context-based edge-preserving optical flow estimation with occlusion detection. *Signal Processing: Image Communication*, 101:116560, 2022. 2

[59] S. Zhao, Y. Sheng, Y. Dong, E. I. Chang, Y. Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6278–6287, 2020. 2