MFTIQ: Multi-Flow Tracker with Independent Matching Quality Estimation

Supplementary Materials

A. Image Feature Extraction

For the DINOV2 features we use the author-provided `ViT-S/14-reg` network checkpoint. The ResNet50 [21] network, pre-trained on the ImageNet1K [9] dataset, is used to extract features from its first three blocks: the input block, residual block 1, and residual block 2. Each output feature is up-sampled to $\frac{H}{4} \times \frac{W}{4}$ and compressed to 32 channels using a convolutional layer.

The custom image features CNN is trained from scratch, and it is inspired by NEUFLOW's feature CNN [63]. Initially, an image pyramid is created by subsampling the input image at different scales (1/1, 1/2, 1/4). For each level of the image pyramid, a convolutional layer is applied with specific kernel sizes, strides, and padding to ensure the output resolution is $\frac{H}{4} \times \frac{W}{4}$ (k4:s4:p0 | k8:s2:p3 | k7:s1:p3). The outputs from each pyramid level are concatenated and compressed to 32 channels using an additional convolutional layer.

The features from all the feature providers (DINOV2, RESNET, custom CNN) are aggregated and compressed through a convolutional operation (from 5×32 channels down to 32 channels) to produce an additional *fused feature* for the cost-volume.

The impact of feature extractors on performance is demonstrated in Tab. 6. Excluding DINOV2 features causes a decrease in AJ from 65.7 to 64.6. Further removing both DINOV2 and RESNET features, leaving only the custom shallow CNN features, results in a more pronounced drop to AJ 61.5. Since the overall runtime is dominated by optical flow computation, it remains nearly unchanged (≈ -0.01 FPS) without the DINO and the RESNET backbones. Thus, we keep all three feature extractors.

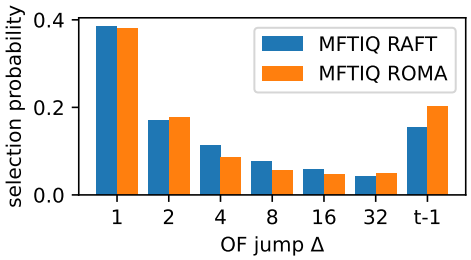

Figure 5. Probability of selecting OF with a given Δ on TAP-VID DAVIS [10], evaluated on frames more than 32 frames distant from template. Statistics are similar for RAFT and ROMA, but long jumps $\Delta = t - 1$ are selected more often with ROMA.

B. Feature Concatenation and Flow Features

The final featuremap contains 6 (DINO, $3\times$ RESNET, custom CNN, *fused*) cost-volumes, each flattened to 49 channels from the ± 3 range 7×7 cost-volume response maps, resulting in a total of 294 channels. In addition to that it contains 2×32 channels of the *fused features* from the template and the current frame (warped by the flow). Finally it has 64 channels of flow features derived from the input $\mathbf{F}_{1 \to t}$ flow chain by a small CNN, for a grand total of 422 channels.

C. Timing

As mentioned in Section 3.3, we implement caching for optical flow estimates and image features to improve efficiency. Table 5 reports the overall tracking timing for results shown in Tab. 1 and Tab. 2 in the paper both with standard caching during computation and with the caches

MFTIQ with	FPS ↑		PPS ↑		FPS pre-computed ↑		PPS pre-computed ↑	
	512×512	720×1080	512×512	720×1080	512×512	720×1080	512×512	720×1080
RAFT [51]	2.66	0.90	8234	8897	10.95	3.76	26944	33921
NEUFLOWV2 [62]	5.67	2.03	16446	19348	10.56	3.59	27589	32175
RAPIDFLOW [36]	3.06	1.35	9603	13058	10.65	3.49	28396	31960
GMFLOW [58]	3.63	0.76	11365	7638	10.31	3.47	27304	32075
SEA-RAFT [55]	2.93	0.93	9285	9195	10.24	3.40	27296	31591
MEMFLOW [13]	1.16	0.29	3836	2985	10.95	3.71	27412	32907
FFORMER++ [48]	1.04	0.24	3457	2437	10.47	3.76	27183	33303
ROMA [16]	0.21	0.19	709	1948	10.10	3.67	24986	32703

Table 5. **Runtime evaluation** of the whole MFTIQ tracker with various OF methods with *(right)* and without *(left)* OF and features pre-computed. All results shows processing speed in frames-per-second (FPS) and points-per-second (PPS) for two different resolutions of images. PPS were evaluated for a sequence of 80 images. In the case of pre-computed optical flow and image feature cache, speed is the same regardless of the OF method used up to a measurement noise.

method	AJ ↑	$<\delta^x_{avg}$↑	OA ↑
(1) Full MFTIQ (RoMa)	65.67	79.82	87.75
(2) -Dino	64.61	79.59	87.80
(3) -Dino -ResNet	61.54	78.58	85.02

Table 6. Influence of IQ feature extractors in the MFTIQ model. The table shows the performance variations when different backbones are omitted, with the remainder of the network held constant. All models followed identical training and evaluation protocols. The evaluation was conducted using the TAP-Vid DAVIS [10] (strided) dataset.

Δ-set hyper-parameter	AJ ↑	$<\delta^x_{avg}$↑	OA ↑	runtime [FPS] ↓ 512x512	720x1080
$\Delta \in \{1, 2, 4, 8, 16, 32, t-1\}$	65.67	79.82	87.75	0.21	0.19
$\Delta \in \{1, 4, 16, t-1\}$	65.50	79.57	87.42	0.35	0.32
$\Delta \in \{1, 8, 32, t-1\}$	59.03	72.79	82.34	0.35	0.32
$\Delta \in \{t-1\}$	57.46	70.08	78.73	1.31	1.14
$\Delta \in \{1\}$	54.67	70.99	73.35	1.31	1.14

Table 7. Ablation of different sets of Δ used for optical flow chaining. The default set of Δs *(first row)* (same as in MFT) performs the best. The base-4 *(second row)* set achieves a better speed / performance trade-off. MFTIQ RoMa evaluated on TAP-Vid DAVIS [10] (strided). Performance measured by average Jaccard (AJ), position accuracy ($<\delta^x_{avg}$), and occlusion accuracy (OA). Speed of tracking densely measured by average frames per second (FPS).

pre-computed offline. With optical flow and image features computed in advance, MFTIQ runs at 3.7 FPS on 720×1080 and at over 10 FPS on 512×512 video resolution.

D. Delta Set Ablation

Tab. 7 shows the effect of using different sets of Δs. Our default base-2 configuration, $\Delta \in \{1, 2, 4, 8, 16, 32, t-1\}$, follows the MFT setup [39]. However, we found that using a base-4 set, $\Delta \in \{1, 4, 16, t-1\}$, achieves a $1.6\times$ speedup with only a minimal performance decrease on the TAP-Vid DAVIS dataset [10]. Both direct matching between the template and the current frame ($\Delta \in \{t-1\}$) and consecutive frame chaining ($\Delta \in \{1\}$) result in a significant performance decrease across all evaluated metrics.

We have also evaluated (Fig. 5) the frequency of selection for each Δ in MFTIQ RAFT and MFTIQ RoMa in the default Δ-set. The results show similar statistics between the two OFs, though the direct jump ($\Delta = t-1$) is selected more frequently in RoMa. This is expected since the RoMa was trained on wide-baseline matching data, making it more reliable with more distant pairs of frames. Only frames beyond timestep 32 are evaluated to avoid biasing the results with smaller Δs at the beginning of the sequence, where longer Δs are not yet available for matching.