Tracking with Dense Correspondences

Jonáš Šerých

2024-12-10

FACULTY
OF ELECTRICAL
ENGINEERING
CTU IN PRAGUE

Motivation Applications: Video Editing, Video Style Transfer

Motivation Applications: Video Editing, Video Style Transfer

→ dense *(every pixel)*, long-term *(long video, through occlusions)* tracking

Optical Flow - Dense Tracking on Pairs of Consecutive Frames

Optical Flow = $(\Delta x, \Delta y)$ in each pixel

Optical flow $F_{(t-1)\to t}$ often works well.
Occlusions are usually neither handled nor benchmarked.

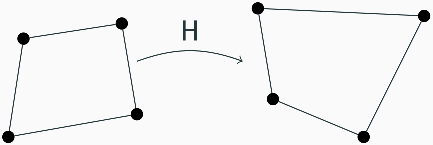
How to do long-term?

Planar Tracking

Homography Tracking

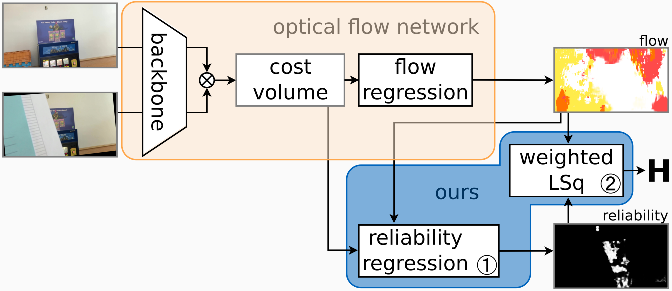
Known geometric model of the scene
$\rightarrow$ dense long-term tracking = sequence of geometric transformations

- Keypoints (e.g. Harris 1988) + tentative correspondences + RANSAC (Fischler, Bolles 1981)
- Intensity registration (e.g. Lucas-Kanade 1981; ESM 2004)
- CNN regressing 4 control points (DeTone et al. 2016)



$$\lambda \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \mathsf{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

WOFT: Weighted Optical Flow Tracker – Two-View Homography



0. Dense correspondences from Optical Flow
1. Reliability regression – "predict inlier/outlier"
2. Fit homography with weighted least squares
3. Failure detection via support set

+ The whole network is differentiable
+ Everything trained only with loss on **H**
+ Works on targets with few keypoints

Learned Correspondence Weights

Weight CNN trained **indirectly** by optimizing a loss on the weighted LSq homography

- Weights (yellow) focus on well-textured areas, corners
- Occlusions have zero weight (here "occlusion" by specular reflection)

WOFT: Weighted Optical Flow Tracker – Sequence Of Homographies

Large pose change $\rightarrow$ OF fails

Pre-Warp with previous pose $\rightarrow$ OF works on residual

WOFT = Pre-Warp $\rightarrow$ Weighted Flow Homography $\rightarrow$ Failure Detection

J. Šerých and J. Matas, “Planar object tracking via weighted optical flow,” in *WACV*, 2023

State-of-the-art Performance on Multiple Benchmarks

method	year	FPS	P@5 orig	P@5 rean	P@15 orig	P@15 rean
GOP-ESM	2019	4.95	42.9	–	49.7	–
SuperGlue	2020	3.7	39.1	42.1	58.0	55.7
Gracker	2017	4.8	39.2	–	63.2	–
SiamESM	2019	–	58.7	–	66.2	–
SOSNet	2019	1.5	56.6	60.9	69.9	67.0
SIFT	2004	0.8	62.2	65.8	71.3	69.6
OBD	2021	30	48.4	54.3	79.3	79.2
LISRD	2020	7	61.6	68.3	79.6	79.2
HDN	2022	10.6	61.3	70.9	91.5	92.4
$\text{WOFT}_{\downarrow 3}$		19.2	68.9	80.5	91.2	92.3
WOFT		3.5	**80.6**	**90.4**	**93.9**	**95.6**

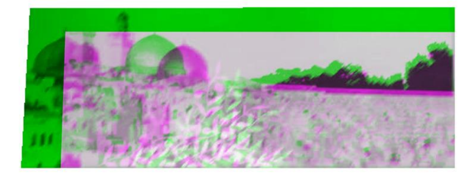
method	P@5	P@15
SIFT [40, 5]	43.8	54.5
SOL [83]	55.3	74.8
HDN [100]	74.4	94.5
Bit-Planes [107]	75.1	76.0
Gracker [84]	75.2	89.9
GOP-ESM [5]	90.8	93.1
SiamESM [93]	**96.1**	97.7
WOFT	**96.1**	**98.0**

New PlanarTrack 2023 benchmark

		WOFT [32]	HDN [41]	GIFT [24]	LISRD [29]	SIFT [25]
POT-210 [20]	PRE	0.805	0.612	0.553	0.617	0.692
	SUC	0.572	0.484	0.404	0.463	0.445
POT-210UC [20]	PRE	0.768	0.567	0.528	0.581	0.578
	SUC	0.536	0.442	0.379	0.419	0.378
PlanarTrackTst	PRE	0.433	0.263	0.254	0.167	0.142
	SUC	0.306	0.236	0.223	0.137	0.107

Improved GT – WOFT Score from 80.6 to 90.4

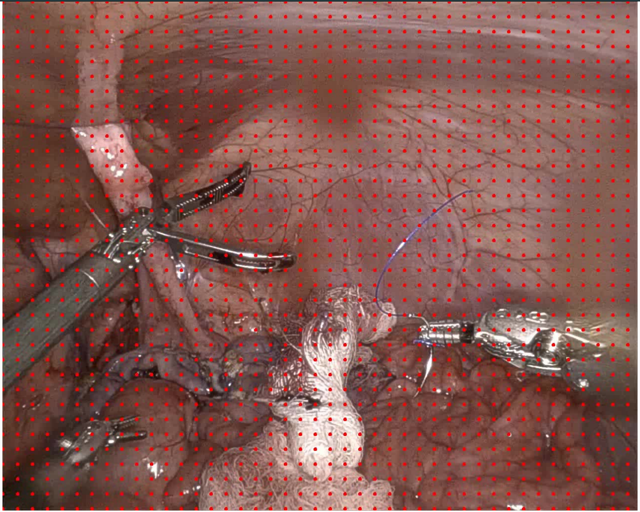
Original GT alignment



Improved GT alignment

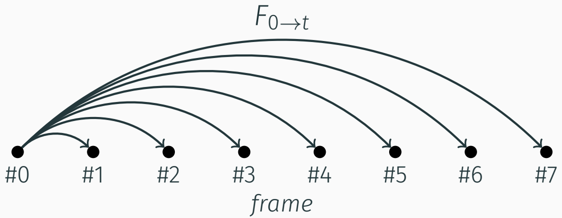
Dense Point Tracking on 3D Surfaces

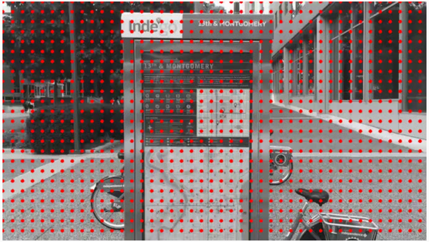
Dense Point Tracking



- Geometry unknown
- Non-rigid motion
- Not just one object
- Again use Optical Flow

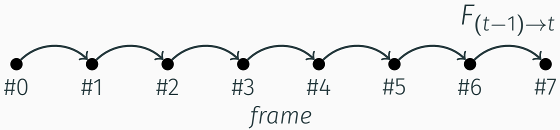
Template to Current Optical Flow Matching

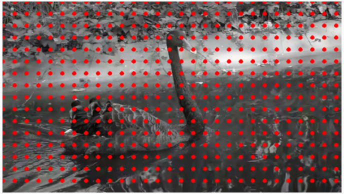




+ No error accumulation → no drift
+ Can recover after occlusions or failures
– But harder task - change of viewpoint, illumination, large motion

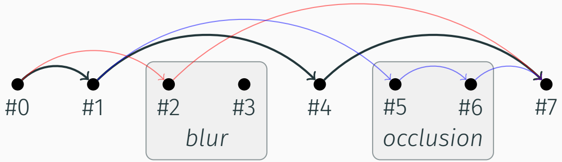
Optical Flow Chaining

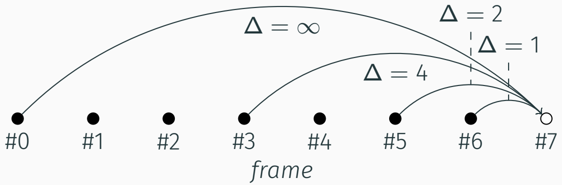




- – Cannot recover from temporary occlusions
- – Errors accumulate $\rightarrow$ drifting
- + Simple task
- + Optical Flow trained for this task

MFT - Multiple Flow Chain Candidates





1. Create several flow chain candidates
2. Pick the best one for each tracked point

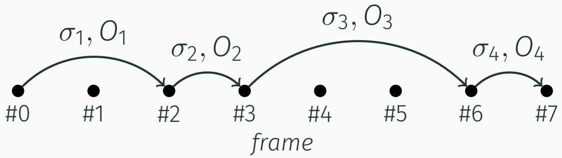
Keep the number of candidates small:

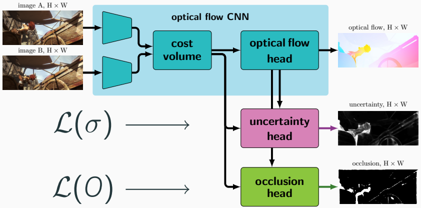
- Fix the best candidate on each previous frame
- Only consider chains ending with OF $F_{(t-\Delta)\to t}$, $\Delta \in \{1, 2, 4, 8, 16, \ldots, \infty\}$

MFT: Multi-Flow Tracker

M. Neoral, **J. Šerých**, and J. Matas, "MFT: Long-term tracking of every pixel," in *WACV*, 2024

Estimate uncertainty σ and occlusion O for each flow vector:





Chain the σ, O scores, pick the best:

$$\sigma^2_{0\to t} = \sum_i \sigma_i^2 \qquad O_{0\to t} = \max_i O_i$$

$$c^* = \arg\min_c \sigma^2_{c,0\to t}$$

$$\text{s.t. } O_{c,0\to t} < 0.5$$

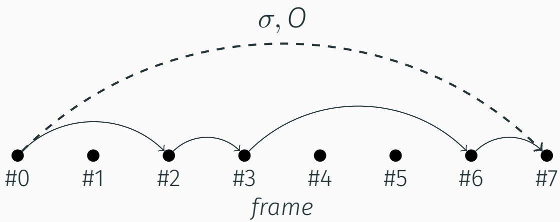
$$\mathcal{L}(\sigma) = \sum_{i=1}^{H\times W} \frac{l_h(||\vec{x}_i - \vec{x}_i^*||_2)}{2\sigma_i^2} + \frac{1}{2}\log(\sigma_i^2)$$

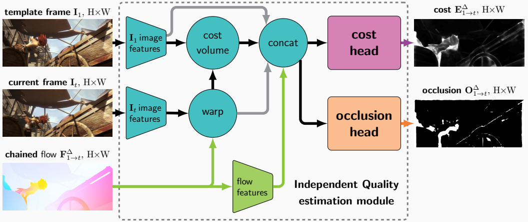
$$\mathcal{L}(O) = \text{Binary Cross-Entropy}$$

MFTIQ: Multi-Flow Tracker with Independent Quality Estimation

J. Šerých, M. Neoral, and J. Matas, “MFTIQ: Multi-flow tracker with independent matching quality estimation,” in *WACV*, 2025

Estimate uncertainty and occlusion for the whole chain, independently on the OF.





method	AJ ↑	$<\delta^x_{avg}$↑	OA ↑	OF runtime [ms] ↓ 512x512	720x1080
MFT	56.28	71.03	86.96	47	142
MFTIQ with					
RAFT	60.54	74.22	84.42	47	142
GMFLOW	55.28	69.83	83.55	24	137
NEUFLOW	55.73	70.26	80.87	10	18
GMFLOW-R	59.57	73.38	86.49	69	335
NEUFLOWV2	56.92	70.97	81.59	7	8
RAPIDFLOW	59.56	73.14	84.37	32	55
LLA-FLOW	61.78	75.18	85.44	117	475
MEMFLOW	62.30	75.97	85.95	121	610
FFORMER++	62.72	76.22	86.34	142	782
RPKNET	62.78	76.61	86.39	126	174
SEA-RAFT	63.51	77.18	86.22	34	105
ROMA	**65.67**	**79.82**	**87.75**	714	729

MFT and MFTIQ Results

method	speed	DAVIS strided			DAVIS first			RoboTAP first			Kinetics first		
	PPS↑	AJ↑	$<\delta^x_{avg}$↑	OA↑	AJ↑	$<\delta^x_{avg}$↑	OA↑	AJ↑	$<\delta^x_{avg}$↑	OA↑	AJ↑	$<\delta^x_{avg}$↑	OA↑
TAP-Net	555	38.4	53.1	82.3	33.0	48.6	78.8	45.1	62.1	82.9	38.5	54.4	80.6
CoTracker 0.8	64.8	79.1	88.7	60.6	75.4	**89.3**	54.0	65.5	78.8	48.7	64.3	**86.5**	
TAPIR	200	61.3	72.3	87.6	56.2	70.7	86.5	59.6	73.4	**87.0**	49.6	64.2	85.0
BootsTAP	200	**66.4**	78.5	**90.7**	**61.4**	74.0	88.4	**64.9**	**80.1**	86.3	**54.7**	**68.5**	86.3
MFT	10671	56.3	71.0	87.0	51.1	67.1	84.0	–	–	–	39.6	60.4	72.7
MFT RoMa	–	58.0	77.2	80.5	52.1	72.7	77.1	–	–	–	–	–	–
MFTIQ	709	65.7	**79.8**	87.8	59.9	**75.5**	84.5	60.0	77.5	85.2	48.7	65.9	85.2

MFTIQ performance close to BootsTAP (by DeepMind, tomorrow @ ACCV 2024), without needing 15M YouTube videos and 256 A100 GPUs.

MFTIQ RoMa $\approx$ 15$\times$ slower than MFT, but still fast compared to SOTA.

State-Of-The-Art Planar Tracking With MFTIQ

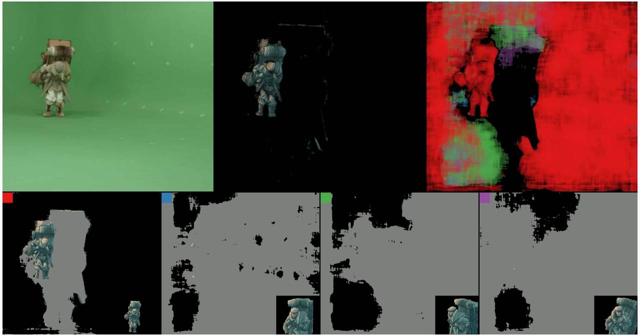
method	BL	OCCL	OOV	PERS	ROT	SC	UNC	all
LISRD	54.1	93.8	83.7	65.0	86.3	30.0	67.1	68.3
HDN	48.8	78.2	66.1	54.4	91.4	94.8	60.7	70.9
CGN	41.6	88.1	82.8	76.5	96.1	90.3	72.4	78.5
WOFT	60.4	**98.6**	96.3	95.4	99.3	94.0	88.2	90.4
HVC-Net	60.5	**98.6**	**97.2**	92.7	99.3	100.0	**90.1**	91.4
MFTIQ	**72.0**	**98.6**	95.0	**96.6**	**99.5**	**100.0**	89.1	**93.1**

STIR: Surgical Tracking Challenge @ MICCAI 2024

Accuracy Threshold (px)	Baselines			Submissions			
	RAFT	CSRT	MFT	MedTrack	CCG_DGIST	Jmees	ICVS_2AI
4	0.07258	0.22782	**0.4254**	0.38911	0.36089	0.26008	0.25403
8	0.19556	0.47379	**0.69355**	0.6754	0.63508	0.54435	0.51008
16	0.39919	0.67137	0.86492	**0.86694**	0.83871	0.77419	0.74194
32	0.64919	0.74798	**0.93347**	0.93145	0.90927	0.91734	0.8871
64	0.80444	0.81048	**0.96371**	0.95968	0.94355	0.95363	0.9254
Average	0.42419	0.58629	**0.77621**	0.76452	0.7375	0.68992	0.66371
Placement				**1st**	**2nd**	**3rd**	

MFT as a baseline submitted by challenge authors won.

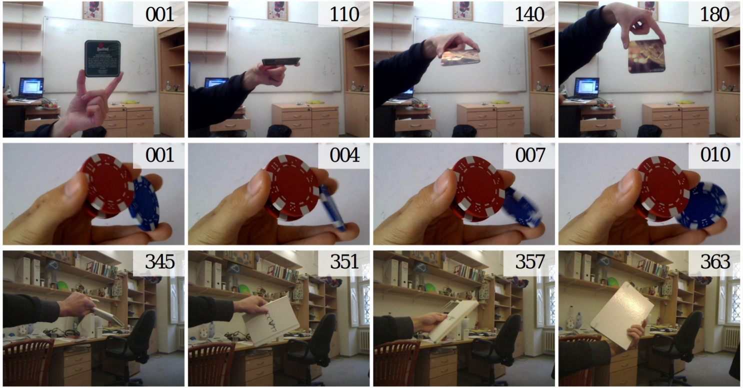
Video Style Transfer From Multiple Keyframes



Coin-Tracking

Coin-Tracking Task

J. **Šerých** and J. Matas, "Visual coin-tracking: Tracking of planar double-sided objects," in *GCPR*, 2019



- Sudden appearance change (side flip)
- Difficult motion blur (fast 3D rotations)
- Low textureness
- Strong illumination change
- Aspect ratio change

Segmentation

Pose estimation

Segmentation adaptation

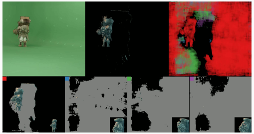
- Segmentation by k-NN classification in learned metric space - FASTVOS[5]
  - State-of-the-Art at that time
  - Worked well only on short videos
  - Simple online update

Coin-Tracking Results

Summary

- **J. Šerých** and J. Matas, "Visual coin-tracking: Tracking of planar double-sided objects," in *GCPR*, 2019. 1 GSC citation
- **J. Šerých** and J. Matas, "Planar object tracking via weighted optical flow," in *WACV*, 2023. 3 GSC citations
- M. Neoral, **J. Šerých**, and J. Matas, "MFT: Long-term tracking of every pixel," in *WACV*, 2024. 36 GSC citations
- **J. Šerých**, M. Neoral, and J. Matas, "MFTIQ: Multi-flow tracker with independent matching quality estimation," in *WACV*, 2025

Thanks for your attention.



Cost Function Optimized by MFT

Goal: minimize end-point-error on visible points

$$\mathcal{L}_G = \sum_{t=1}^{T} \sum_{i=1}^{H \times W} ||\vec{x}_{t,i} - \vec{x}^*_{t,i}||_2 \cdot [\![\text{visible}_{t,i}]\!]$$

MFT not trained end-to-end. Uncertainty loss:

$$\mathcal{L}_u = \frac{1}{2\,\sigma^2} l_H(\underbrace{||\vec{x} - \vec{x}^*||_2}_{\text{OF error}}) + \frac{1}{2}\log(\sigma^2)$$

selecting minimal uncertainty $\approx$ selecting minimal end-point-error

Cost Function Optimized by MFT

Goal: minimize end-point-error on visible points

$$\mathcal{L}_G = \sum_{t=1}^{T} \sum_{i=1}^{H \times W} ||\vec{x}_{t,i} - \vec{x}^*_{t,i}||_2 \cdot [\![\text{visible}_{t,i}]\!]$$

MFT not trained end-to-end. Uncertainty loss:

$$\mathcal{L}_u = \frac{1}{2\,\sigma^2} l_H(\,||\vec{x} - \vec{x}^*||_2\,) + \frac{1}{2}\,\log(\sigma^2)$$

allow incorrect flows

OF error

pay for large σ

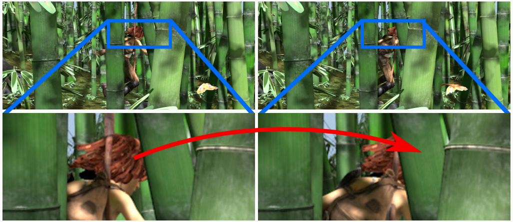
selecting minimal uncertainty $\approx$ selecting minimal end-point-error

Cost Function Optimized by MFT

Goal: minimize end-point-error on visible points

$$\mathcal{L}_G = \sum_{t=1}^{T} \sum_{i=1}^{H\times W} ||\vec{x}_{t,i} - \vec{x}^*_{t,i}||_2 \cdot [\![\text{visible}_{t,i}]\!]$$

MFT not trained end-to-end. Occlusion handling:



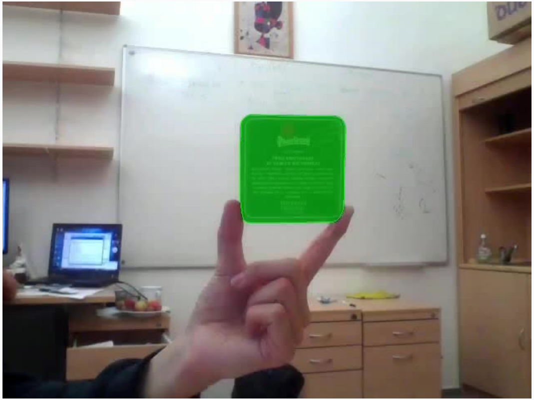
Even with perfect OF, must not start tracking the occluder.

$$c^* = \arg\min_c \sigma^2_{c,0\to t}$$

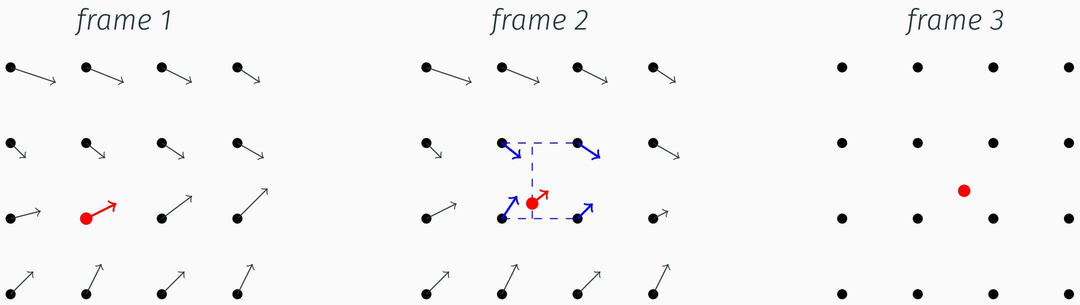
$$\text{s.t. } O_{c,0\to t} < 0.5$$

SAMv2 on Coin-Tracking

- \+ Surprisingly works almost perfectly
- \+ Even when initialized on single side - tracks both!
- – Cannot distinguish the two sides



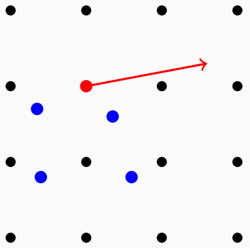
Optical Flow Chaining With Bilinear Interpolation



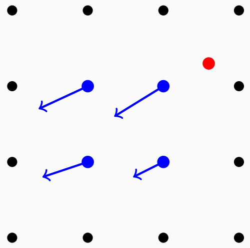
To chain $F_{1\to2}$ with $F_{2\to3}$, the later must be interpolated at the red point.
We use bilinear interpolation.

Optical Flow Direction

frame 1

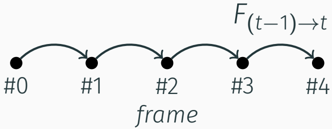


frame t

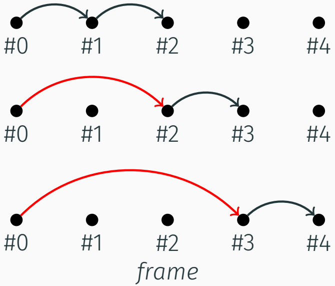


- Forward *(red)* Optical Flow suitable for point-tracking
  How did a query point from the first frame move?
- Backward *(blue)* Optical Flow suitable for texture transfer
  Each coordinate gets some color from the first frame

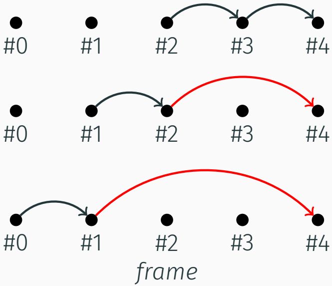
Optical Flow Chaining Direction



Forward chaining

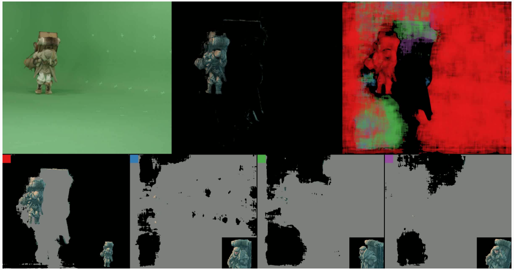


Backward chaining



Crivelli et al. (2014); Wu et al. (2023) AccFlow

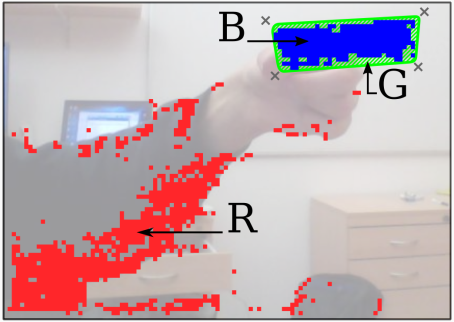
Video Stylization From Multiple Keyframes



CTR-BASE: Pose estimation

Perturb 4 homography control points, maximize product of:

1. Fraction of segmentation explained
2. Object visibility fraction
3. Previous visibility mask IoU
4. Appearance ZNCC



1. $\frac{|B|}{|B \cup R|}$

2. $\frac{|B|}{|G|}$

3. $\frac{|B \cap B_{t-1}|}{|B \cup B_{t-1}|}$

G - pose hypothesis, **B** - segmentation inside hypothesis
R - segmentation outside hypothesis