ECCV 2024

DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video

Narek Tumanyan, Assaf Singer*, Shai Bagon, Tali Dekel*

Weizmann Institute of Science

- **Problem/Objective**
  - tracking in video

- **Contribution/Key Idea**
  - tracking model

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video](#) ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **Tracking**

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video](#) ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **Tracking**

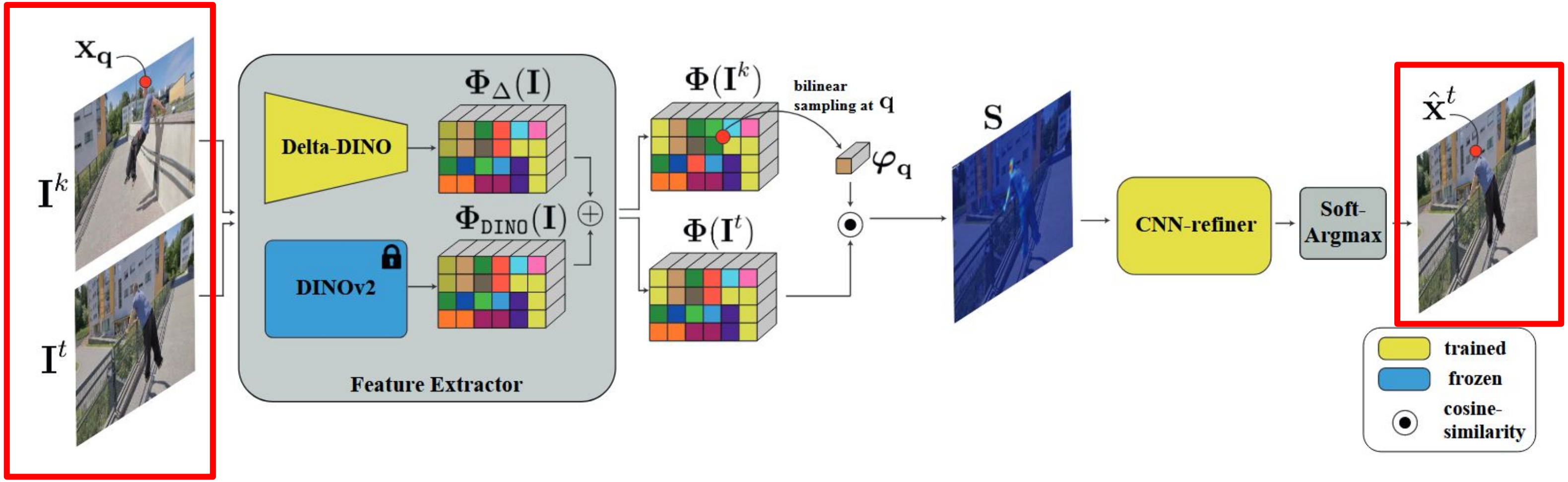
전유진

DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- Overview



Input : query point $\mathbf{x_q}$, video $\{\mathbf{I}^t\}_{t=1}^{T}$

Output : position estimates $\{\hat{\mathbf{x}}^t\}_{t=1}^{T}$

전유진

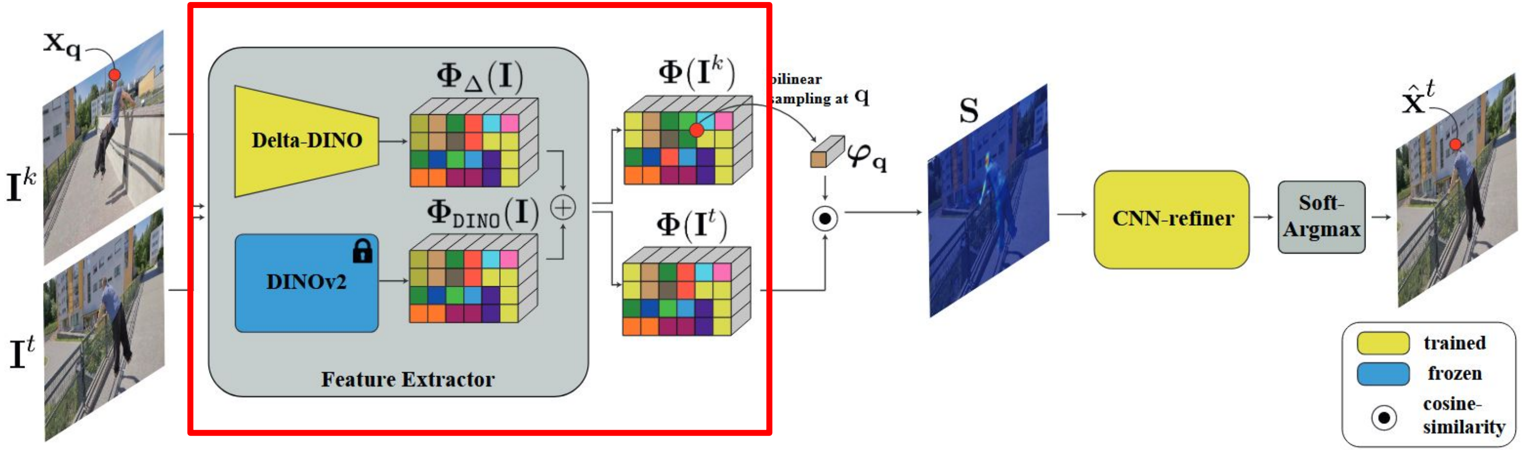
DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video

ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **DINO-Tracker**



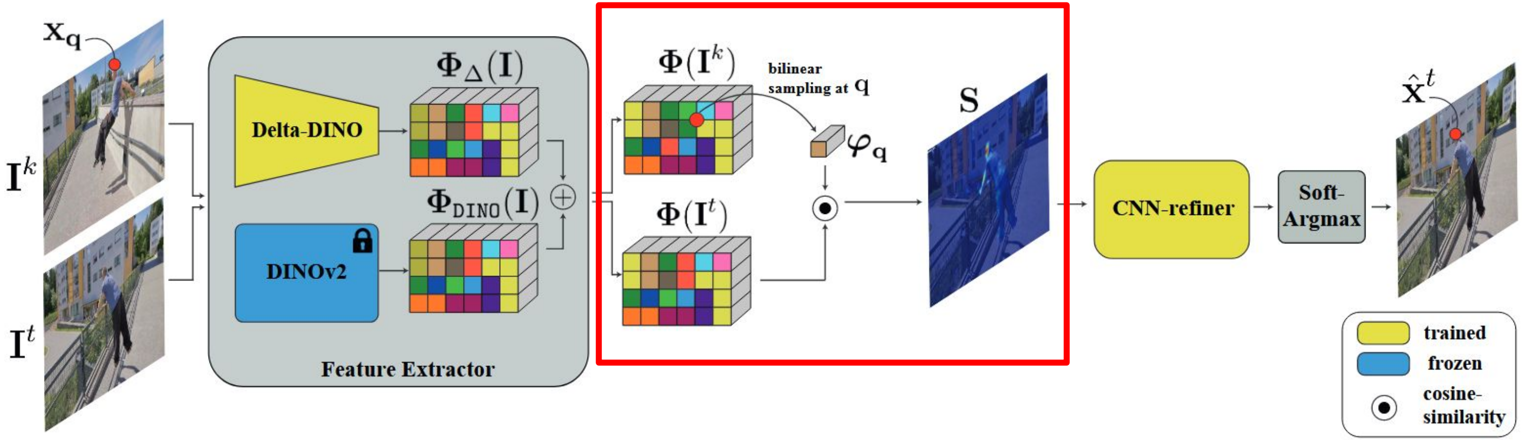
$$\Phi(\mathbf{I}) = \Phi_{\text{DINO}}(\mathbf{I}) + \Phi_\Delta(\mathbf{I}) \quad (1)$$

DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video ECCV 2024

Narek Tumanyan, Assaf Singer*, Shai Bagon, Tali Dekel*

Weizmann Institute of Science

- **DINO-Tracker**



$$\varphi_\mathbf{q} = \Phi(\mathbf{I}^k)[\mathbf{q}] \quad \Phi^t = \Phi(\mathbf{I}^t)$$

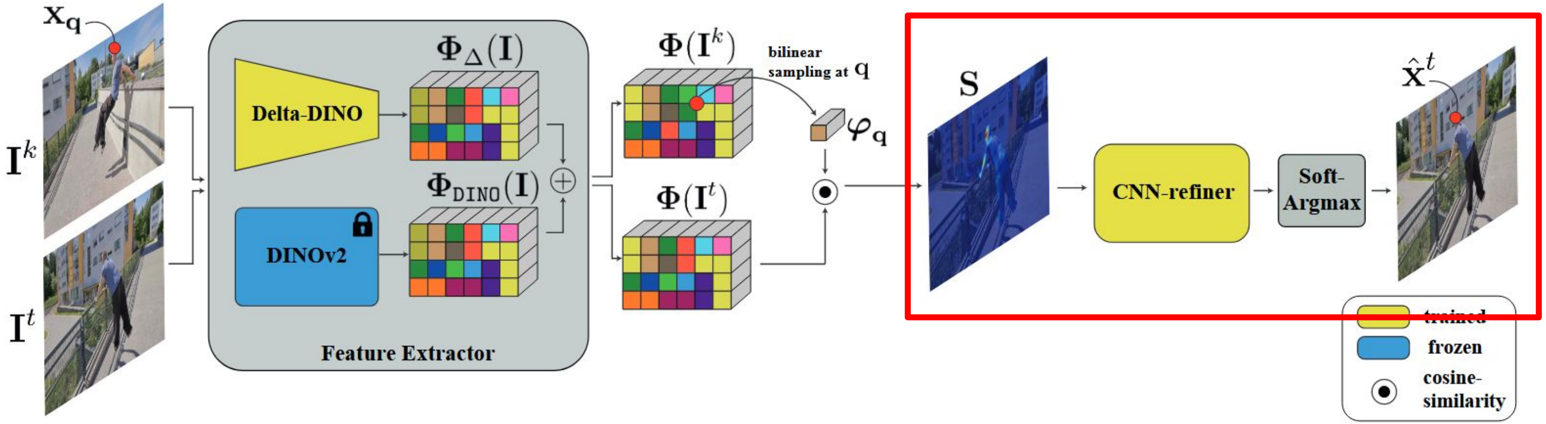
$$\mathbf{S}(\mathbf{p}) = \text{cos-sim}(\varphi_\mathbf{q}, \Phi^t(\mathbf{p})) \quad \text{where} \quad \text{cos-sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T \cdot \mathbf{b}}{||\mathbf{a}||_2 \cdot ||\mathbf{b}||_2}$$

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video](#) ECCV 2024

Narek Tumanyan, Assaf Singer*, Shai Bagon, Tali Dekel*

Weizmann Institute of Science

- **DINO-Tracker**



$$\hat{\mathbf{x}}^t = \frac{\sum_{\mathbf{p}\in\Omega} \mathbf{H}(\mathbf{p}) \cdot \mathbf{x}_\mathbf{p}}{\sum_{\mathbf{p}\in\Omega} \mathbf{H}(\mathbf{p})} \qquad (2)$$

$$\Omega = \{\mathbf{p} : ||\mathbf{x}_\mathbf{p} - \mathbf{x}_{\mathbf{p}_{max}}||_2 \leq R\}$$

$$\Pi(\mathbf{x}_\mathbf{q}, t) = \hat{\mathbf{x}}^t \qquad \mathcal{T}_q = \{\hat{\mathbf{x}}^t : \hat{\mathbf{x}}^t = \Pi(\mathbf{x}_\mathbf{q}, t), t = 1 \dots T\}$$

전유진

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video] ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **Self-Supervision**

- Optical flow

$$\Omega_{\texttt{flow}} = \{(\mathbf{x}^i, \mathbf{x}^j) \text{ cycle-consistent}\}$$

1) t 프레임에서 트랙 종료

$$||\mathbf{x}^t - (\mathbf{x}^{t+1} + \mathbf{f}_{t+1\to t}(\mathbf{x}^{t+1}))|| \geq \gamma_{\texttt{of}}$$

2) optical flow 와 일관되지 않는 correspondence $\mathbf{x}^j$ 제거

$$||\mathbf{x}^j - \mathbf{x}^{i\to j}||_2 \geq \gamma_{\texttt{of-lng}}$$

$$||\mathbf{x}^i - (\mathbf{x}^{i\to j} + \mathbf{f}_{j\to i}(\mathbf{x}^{i\to j}))||_2 \leq \gamma_{\texttt{of}}$$

$$\mathbf{x}^{i\to j} = \mathbf{x}^i + \mathbf{f}_{i\to j}(\mathbf{x}^i),\ \gamma_{\texttt{of-lng}} = 2\text{px}$$

전유진

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video](#) ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **Self-Supervision**

- Feature correspondences

1) best-buddies

$$\Omega_{\texttt{dino-bb}} = \{(\mathbf{p}^i, \mathbf{p}^j) \;\; \text{DINO bb}\}$$

$$NN(\varphi^i_{\texttt{DINO}}, \mathbf{\Phi}_{\texttt{DINO}}(\mathbf{I}^j)) = \varphi^j_{\texttt{DINO}} \wedge NN(\varphi^j_{\texttt{DINO}}, \mathbf{\Phi}_{\texttt{DINO}}(\mathbf{I}^i)) = \varphi^i_{\texttt{DINO}} \qquad (3)$$

2) refined best buddies

$$\Omega_{\texttt{rfn-bb}} = \{(\mathbf{p}^i, \mathbf{p}^j) \;\; \text{refined bb}\}$$

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video]

ECCV 2024

Narek Tumanyan, Assaf Singer*, Shai Bagon, Tali Dekel*

Weizmann Institute of Science

- **Self-Supervision**

- Feature correspondences

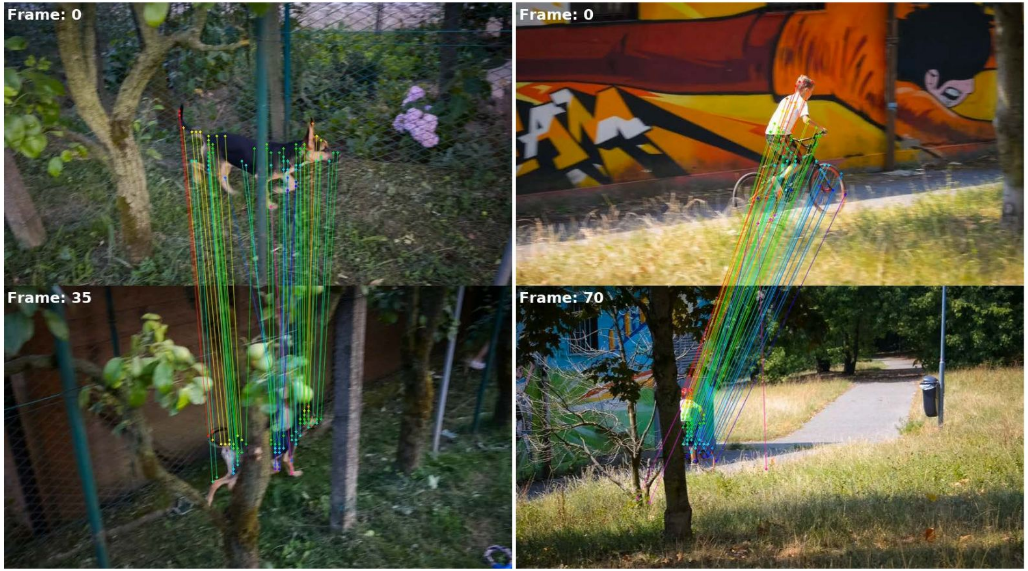


Fig. 8: *DINO best-buddies.* We visualize best-buddy pairs between distant frames. DINO best-buddies provide localized semantic correspondences, allowing the model to recover the object past repeating occlusions.

전유진

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video]

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **Objective**

$$\mathcal{L} = \mathcal{L}_{\texttt{flow}} + \lambda_1 \mathcal{L}_{\texttt{dino-bb}} + \lambda_2 \mathcal{L}_{\texttt{rfn-bb}} + \lambda_3 \mathcal{L}_{\texttt{rfn-cc}} + \lambda_4 \mathcal{L}_{\texttt{prior}} \tag{5}$$

$$\mathcal{L}_{\texttt{flow}} = \sum_{(\mathbf{x}^i, \mathbf{x}^j) \in \Omega_{\texttt{flow}}} L_H(\Pi(\mathbf{x}^i, j), \mathbf{x}^j) + L_H(\Pi(\mathbf{x}^j, i), \mathbf{x}^i)$$

$$\mathcal{L}_{\texttt{dino-bb}} = \frac{1}{|\Omega_{\texttt{dino-bb}}|} \sum_{(\boldsymbol{\varphi}^i, \boldsymbol{\varphi}^j) \in \Omega_{\texttt{dino-bb}}} \frac{1}{2} w^{ij}_{\texttt{dino-bb}} \left(l(\boldsymbol{\varphi}^i, \boldsymbol{\varphi}^j) + l(\boldsymbol{\varphi}^j, \boldsymbol{\varphi}^i) \right)$$

$$l(\boldsymbol{\varphi}^i, \boldsymbol{\varphi}^j) = -\log \frac{\exp(\text{cos-sim}(\boldsymbol{\varphi}^i, \boldsymbol{\varphi}^j)/\tau)}{\sum_{\mathbf{p}} \exp(\text{cos-sim}(\boldsymbol{\varphi}^i, \boldsymbol{\Phi}^j(\mathbf{p}))/\tau)}$$

$$\mathcal{L}_{\texttt{rfn-bb}} = \frac{1}{|\Omega_{\texttt{rfn-bb}}|} \sum_{(\boldsymbol{\varphi}^i, \boldsymbol{\varphi}^j) \in \Omega_{\texttt{rfn-bb}}} \frac{1}{2} w^{ij}_{\texttt{rfn-bb}} \left(l(\boldsymbol{\varphi}^i, \boldsymbol{\varphi}^j) + l(\boldsymbol{\varphi}^j, \boldsymbol{\varphi}^i) \right)$$

$$\mathcal{L}_{\texttt{rfn-cc}} = \sum_{(\mathbf{x}^i, \mathbf{x}^j) \in \Omega_{\texttt{rfn-cc}}} \frac{1}{2} w^{ij}_{\texttt{rfn-cc}} \left(L_H(\Pi(\mathbf{x}^i, j), \mathbf{x}^j) + L_H(\Pi(\mathbf{x}^j, i), \mathbf{x}^i) \right) \tag{4}$$

$$\mathcal{L}_{\texttt{prior}} = \frac{1}{H' \cdot W'} \cdot \sum_{\mathbf{p}} \underbrace{\left| 1 - \frac{||\boldsymbol{\Phi}(\mathbf{I})[\mathbf{p}]||_2}{||\boldsymbol{\Phi}_{\texttt{DINO}}(\mathbf{I})[\mathbf{p}]||_2} \right|}_{\mathcal{L}_{\text{norm}}} + \underbrace{\left| 1 - \text{cos-sim}\left(\boldsymbol{\Phi}(\mathbf{I})[\mathbf{p}], \boldsymbol{\Phi}_{\texttt{DINO}}(\mathbf{I})[\mathbf{p}] \right) \right|}_{\mathcal{L}_{\text{angle}}}$$

전유진

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video]

ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **Occlusion prediction**

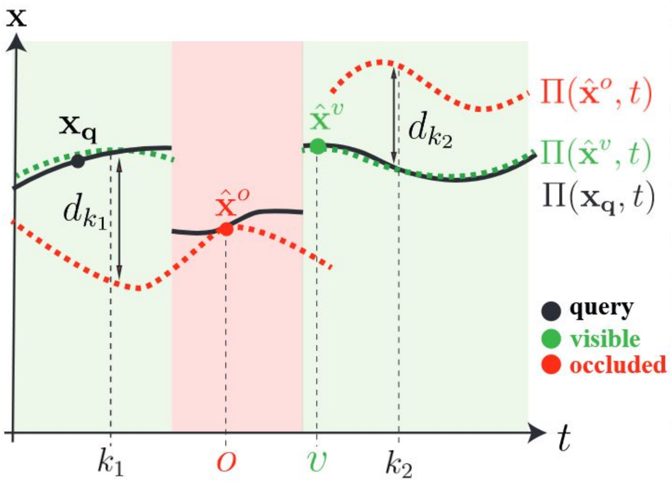


Fig. 3: *Visibility via trajectory agreement.* To determine the visibility of $\mathbf{x_q}$ at time $t = o$, we track $\hat{\mathbf{x}}^o$ across time and check the agreement between $\Pi(\hat{\mathbf{x}}^o, t)$ and $\Pi(\mathbf{x}, t)$. This is done by measuring d_{k_1}, d_{k_2} – displacements between the (black and red) tracks for anchor time steps k_1, k_2. Since these displacements are large, we classify $\mathbf{x_q}$ as occluded for $t = o$. For $t = v$, the track $\Pi(\hat{\mathbf{x}}^v, t)$ (green) agrees with $\Pi(\mathbf{x}, t)$, thus $\mathbf{x_q}$ is classified as visible for $t = v$.

전유진

DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video

ECCV 2024

Narek Tumanyan, Assaf Singer*, Shai Bagon, Tali Dekel*

Weizmann Institute of Science

- Experiments

Method	DAVIS-256			DAVIS-480			Kinetics-256			Kinetics-480			BADJA	
	δ^x_{avg}	OA	AJ	δ^x_{avg}	OA	AJ	δ^x_{avg}	OA	AJ	δ^x_{avg}	OA	AJ	δ^{seg}	δ^{3px}
RAFT [47]	56.7	–	–	66.7	–	–	50.4	–	–	60.5	–	–	45.0	5.8
DINOv2 [38]	61.4	–	–	64.7	–	–	60.3	–	–	61.0	–	–	62.8	8.4
TAP-Net* [12]	53.4	81.4	38.4	66.4	79.0	46.0	61.7	86.6	48.5	67.1	81.5	47.7	45.4	9.6
PIPs++* [63]	71.5	–	–	73.6	–	–	68.2	–	–	70.8	–	–	59.0	9.8
TAPIR* [13]	74.7	**89.4**	62.8	77.3	**89.5**	**65.7**	69.5	89.1	57.3	69.8	86.7	57.5	68.7	10.5
Co-Tracker* [26]	**79.2**	89.3	**65.1**	79.4	**89.5**	65.6	72.9	88.9	**59.9**	72.8	88.9	59.8	64.0	11.2
Omnimotion† [51]	67.5	85.3	51.7	74.1	84.5	58.4	69.2	**89.2**	55.0	–	–	–	45.2	6.9
Ours†	78.2	87.5	62.3	**80.4**	88.1	64.6	**73.3**	88.5	59.7	**74.3**	**89.2**	**60.9**	**72.4**	**14.3**

δ^x_{avg} : position accuracy

OA : occlusion accuracy

AJ : average jaccard : position + occlusion accuracy 둘다 고려

전유진

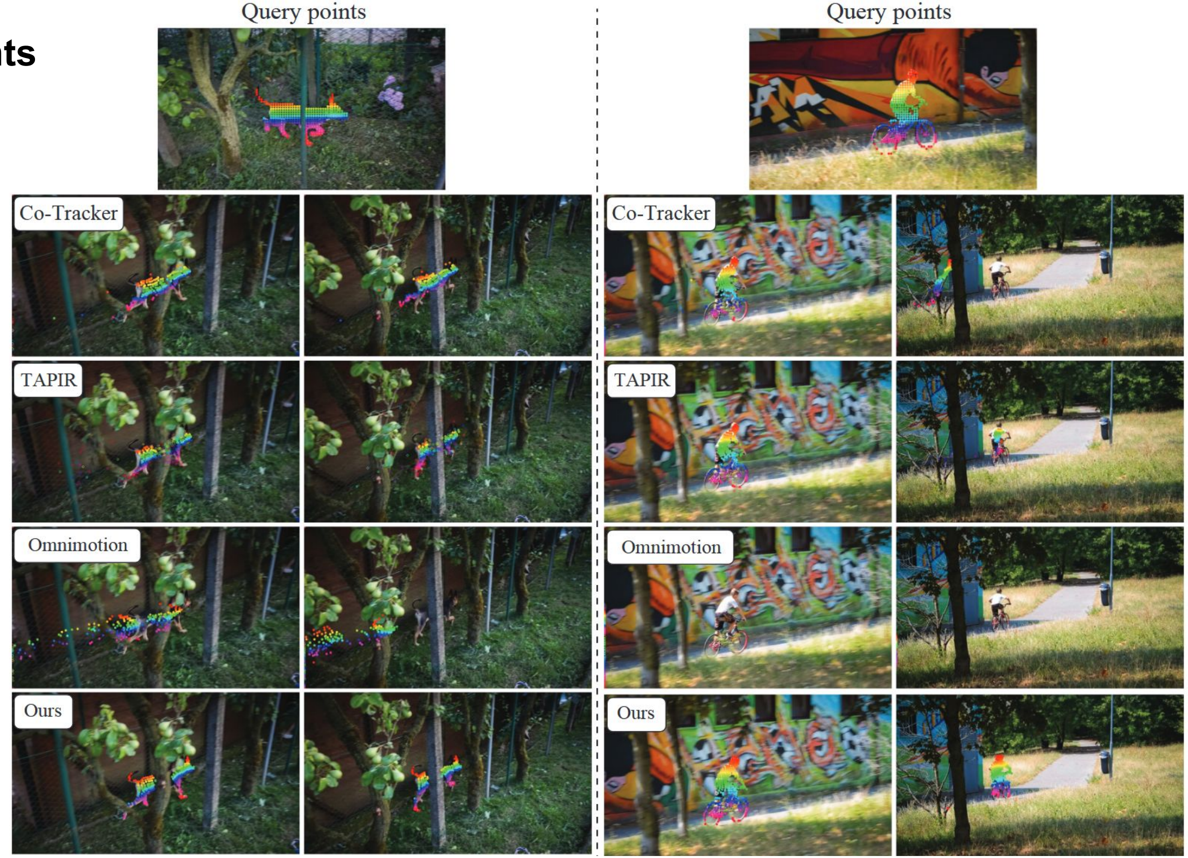
DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video

ECCV 2024

Narek Tumanyan, Assaf Singer*, Shai Bagon, Tali Dekel*

Weizmann Institute of Science

- Experiments



전유진

[DINO-Tracker: Taming DINO for Self-Supervised Point Tracking in a Single Video](#)

ECCV 2024

Narek Tumanyan∗, Assaf Singer∗, Shai Bagon, Tali Dekel

Weizmann Institute of Science

- **Experiments**

전유진