OCT 19-23, 2025 ICCV HONOLULU HAWAII

# Tracing Copied Pixels and Regularizing Patch Affinity in Copy Detection

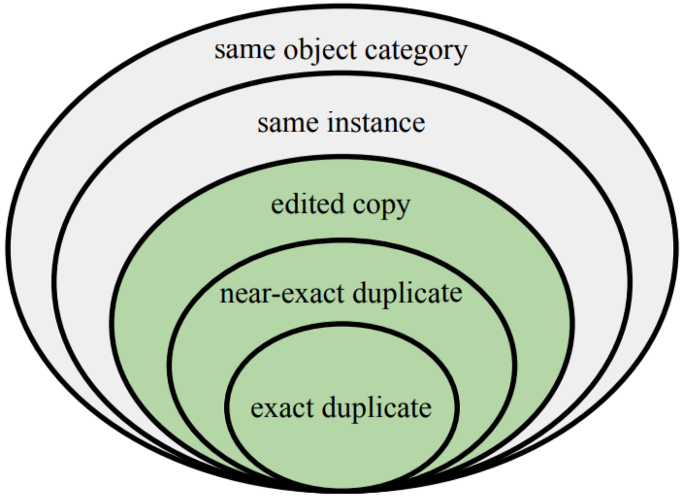
Yichen Lu · Siwei Nie · Minlong Lu · Xudong Yang · Xiaobo Zhang · Peng Zhang

ANT GROUP

# Background

## Fundamental Task

"determine whether a part of an image has been ***copied*** from another image"[1]




## Basic Copy Detection Pipeline

To determine whether a given query (Q) is a copy or an edited copy of an image within a reference database (R), a two-stage pipeline is employed:

- ***Coarse-grained Retrieval*** based on global feature similarity with ***descriptor***
- ***Fine-grained Matching*** based on detailed, one-to-one comparison with ***matcher***


[1]. Douze M, Tolias G, Pizzi E, et al. The 2021 image similarity dataset and challenge[J]. arXiv preprint arXiv:2106.09672, 2021.

# Background

## SSL Training in Copy Detection

Generate image pairs (original vs. edited copy) via Self-Supervised Learning (SSL).

### *Limitation 1: Coarse-grained Labels*

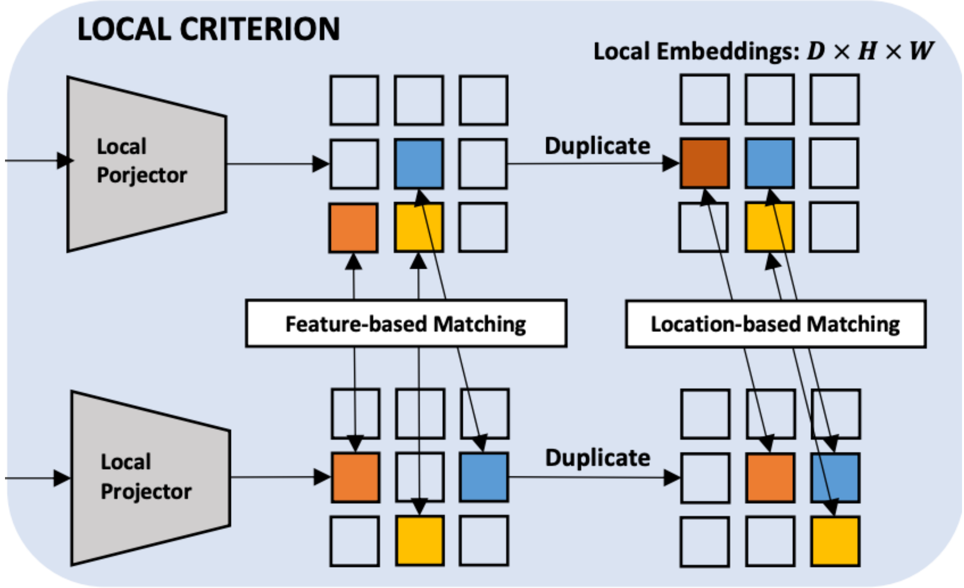
- Difficulty in detecting complex edited copies.
- Difficulty in detecting small, local copied regions.

### *SSL Strikes Back*[1, 2]

Create patch-level pseudo-labels with feature or location *matching, e.g. k*-NN

### *Limitation 2: Noise*

- False Positives
- False Negatives
- Partial Match
- Fixed-k Mismatch





[1]. Li C, Yang J, Zhang P, et al. Efficient self-supervised vision transformers for representation learning. ICLR 2022.
[2]. Bardes A, Ponce J, LeCun Y. Vicregl: Self-supervised learning of local visual features. NeurIPS 2022.

# Motivation

## Ideal Annotations for Copy Detection

- Fine-Grained: Pixel-level annotation of copied regions.
- Clean & Noise-Free: Exact coordinate correspondence.

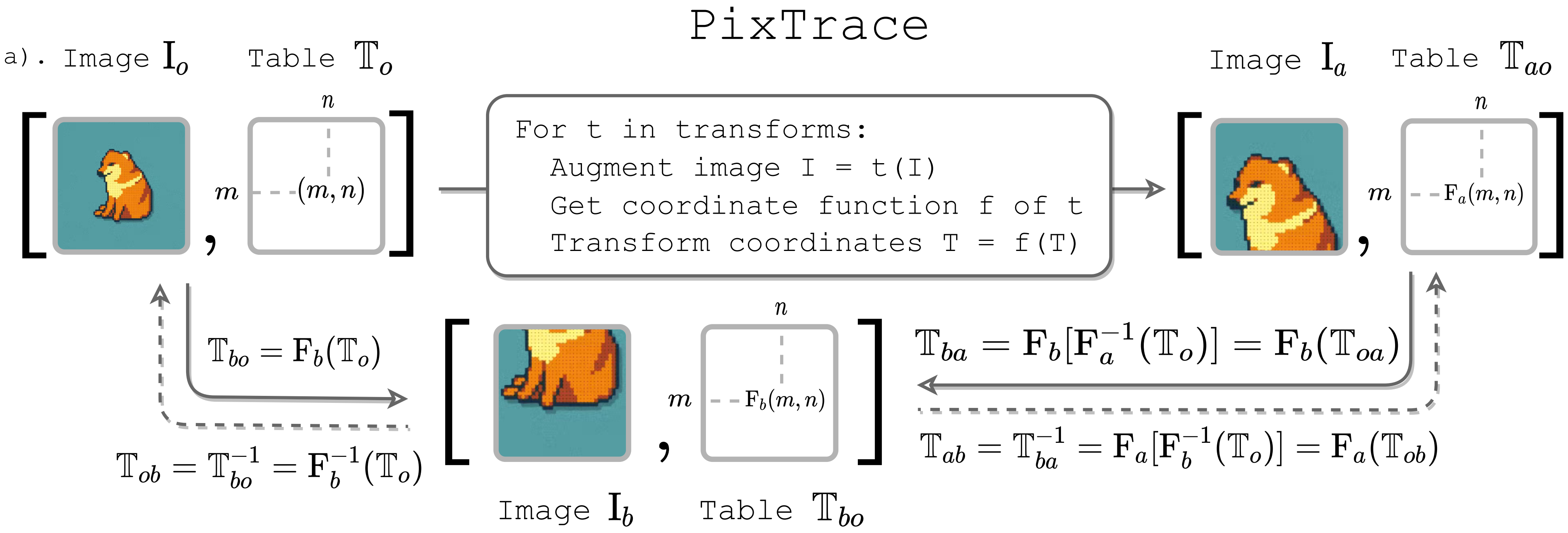
## Traceability of Copied Pixels

Pixel correspondences between original and edited regions can be traced through sequential editing operations.

# PixTrace

## Key Features

- Builds ***precise pixel-level*** correspondences between original and its copy edit images.
- Maintains pixel traceability even under multiple, complex transformations.
- Supports ***reversible (bi-directional)*** pixel tracing.
- Enables pixel tracing between ***different copies originating from a shared source***.



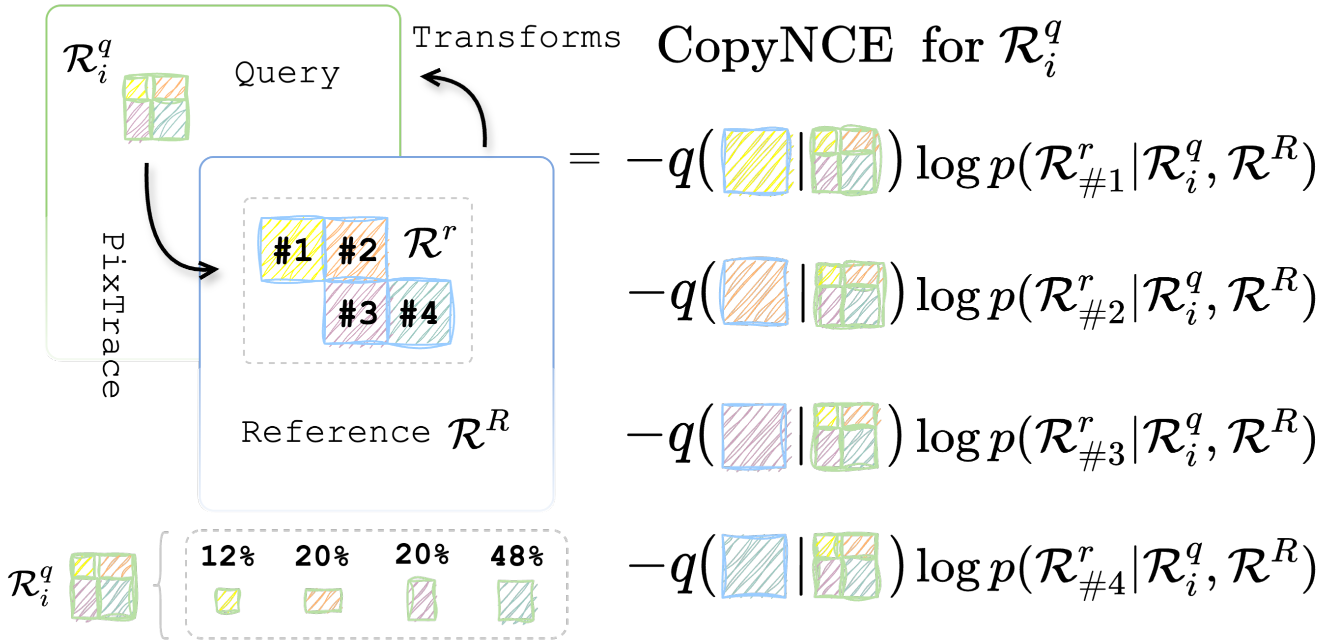
# CopyNCE

- Maximize mutual information between original and copy regions

$$\mathcal{L}^{\#1}_{\text{CopyNCE}} = -\log p(\mathcal{R}^r|\mathcal{R}^X, \mathcal{R}^q)$$
$$= -\log \frac{f_\theta(\mathcal{R}^q, \mathcal{R}^r)}{\sum_{\mathcal{R}^x \in \mathcal{R}^X} f_\theta(\mathcal{R}^q, \mathcal{R}^x)},$$

- Decompose regions into patches

$$\mathcal{L}^{\#2}_{\text{CopyNCE}} = \mathbb{E}_{\mathcal{R}^q_i}[-\log p(\mathcal{R}^r_{i+}|\mathcal{R}^X, \mathcal{R}^q_i)]$$
$$= \mathbb{E}_{\mathcal{R}^q_i}[-\log \frac{g_\theta(\mathcal{R}^q_i, \mathcal{R}^r_{i+})}{\sum_{\mathcal{R}^x \in \mathcal{R}^X} g_\theta(\mathcal{R}^q_i, \mathcal{R}^x)}].$$




- Regularize patch affinity with the patch overlap ratio

$$\mathcal{L}^{\#3}_{\text{CopyNCE}}(q, r, \mathbb{T}_{qr})$$
$$= \mathbb{E}_{\mathcal{R}^q_i}\Big[\sum_{\mathcal{R}^r_j \in \mathcal{R}^r} q(\mathcal{R}^r_j, \mathcal{R}^q_i)\,\underbrace{\big[-\log p(\mathcal{R}^r_j|\mathcal{R}^X, \mathcal{R}^q_i)\big]}_{\text{InfoNCE}}\Big]$$

$$q(\mathcal{R}^r_j|\mathcal{R}^q_i) = \frac{\hat{q}(\mathcal{R}^r_j|\mathcal{R}^q_i)^\gamma}{\sum_{\mathcal{R}^r_k \in \mathcal{R}^r} \hat{q}(\mathcal{R}^r_k|\mathcal{R}^q_i)^\gamma} \qquad \hat{q}(\mathcal{R}^r_j|\mathcal{R}^q_i) = \frac{|\{\mathbb{T}[c] \in \mathcal{R}^r_j | c \in \mathcal{R}^q_i\}|}{|\mathcal{R}^q_i|}$$

- Derive the final symmetric form

$$\boxed{\mathcal{L}_{\text{CopyNCE}} = \frac{1}{2}\big[\mathcal{L}^{\#3}_{\text{CopyNCE}}(q, r, \mathbb{T}_{qr}) + \mathcal{L}^{\#3}_{\text{CopyNCE}}(r, q, \mathbb{T}_{rq})\big]}$$

# Comparison with SOTAs

| Matcher | Settings: Arch | Settings: Res. | Settings: Local | Metrics: $\mu$AP | Metrics: RP90 |
|---|---|---|---|---|---|
| Separate‡ [24] | ViT-S | 224 × 112 | ✗ | 75.4 | 68.7 |
| | ViT-B | | ✗ | 78.4 | 72.9 |
| | ViT-L | | ✗ | 84.7 | 80.3 |
| **CopyNCE** | **ViT-S** | 224 × 224 | ✗ | **83.5** | **75.4** |
| **CopyNCE** | **ViT-S** | 336 × 336 | ✗ | **85.8** | **79.9** |
| ImgFp [42] | EsViT-B | 224 × 224 | ✓ | 61.2 | - |
| Separate‡ [24] | ViT-S | 224 × 112 | ✓ | 77.1 | 70.5 |
| | ViT-B | | ✓ | 80.7 | 75.6 |
| | ViT-L | | ✓ | 86.2 | 82.2 |
| $D^2LV$ [48] | Multi | 256 × 256 | ✓ | 88.6 | 80.1 |
| **CopyNCE** | **ViT-S** | 224 × 224 | ✓ | **87.4** | **81.3** |
| **CopyNCE** | **ViT-S** | 336 × 336 | ✓ | **88.7** | **83.9** |

| Descriptor | Settings: Arch | Settings: Res. | Settings: Pre/Post | Metrics: $\mu$AP | Metrics: RP90 |
|---|---|---|---|---|---|
| DINO [3] | ViT-S | 224 × 224 | ✗ | 20.0 | 6.8 |
| S-square† [33] | EffNet-B5 | 160 × 160 | ✗ | 66.4 | - |
| Lyakaap† [59] | EffNetV2-M | 512 × 512 | ✗ | 64.3 | 56.6 |
| SSCD [35] | R50 | Long×288 | ✗ | 61.5 | 38.3 |
| **CopyNCE** | **ViT-S** | 224 × 224 | ✗ | **70.5** | **63.6** |
| BoT [49] | R50 | 224 × 224 | Str | 70.5 | 61.6 |
| | | | YL / Str | 71.5 | 62.9 |
| SSCD [35] | R50 | Long×288 | SN | 72.5 | 63.1 |
| **CopyNCE** | **ViT-S** | 224 × 224 | **SN** | **72.6** | **68.4** |

Table 1. **Comparison with other SOTA methods. Left** is for matcher and **Right** is for descriptor. **Local** denotes inference ensembling with multiple local crops. **Pre/Post** is the pre-/post-processing, in which **SN** is score normalization, **YL** is YOLO pre-processing and **Str** is feature stretching. † denotes the method leverages extra data for training. ‡ means that we reproduce the results with its open-source code. **Multi** in $D^2LV$ stands for 11×R50 [16], 11×R152 [16] and 11×R50IBN [57].

Outperforms SOTAs ***across*** various ***resolutions***, ***pre/post-processing***, and enhancement ***tricks***.

# Ablation Studies & Param Analysis

## Descriptor

| Method | Parameter | $\mu$AP | Parameter | $\mu$AP |
|---|---|---|---|---|
| CopyNCE | **default** | **70.5** | $w_{\text{NCE}} = 0$ | 68.9 |
| | $w_{\text{NCE}} = 3$ | 70.5 | $w_{\text{NCE}} = 8$ | 69.9 |
| | $\gamma = 0$ | 67.9 | $\gamma = 0.5$ | 69.7 |
| | $\gamma = 1$ | 70.0 | $\gamma = 2$ | 70.4 |
| | $\gamma = 3$ | 70.5 | $\gamma = +\infty$ | 70.1 |
| | w/o NCE | 68.6 | layer=10 | 70.3 |
| | w/o GHNM $w_{\text{NCE}} = 0$ | 57.7 | w/o GHNM $w_{\text{NCE}} = 5$ | 61.8 |
| | R50 $w_{\text{NCE}} = 0$ | 62.7 | R50 $w_{\text{NCE}} = 5$ | 64.0 |
| FeatNN Cos | $k = 1$ | 56.5 | $k = 4$ | 48.1 |
| FeatNN NCE | $k = 1$ | 57.0 | $k = 4$ | 42.6 |
| LocNN Cos | $k = 1$ | 67.7 | $k = 4$ | 67.2 |
| LocNN NCE | $k = 1$ | 64.7 | $k = 4$ | 64.2 |
| Both Cos | $k = 1$ | 68.5 | $k = 4$ | 66.0 |
| Both NCE | $k = 1$ | 64.9 | $k = 4$ | 64.3 |

$w_{\text{NCE}} = 5$ brings ***+1.6% µAP*** gain over baseline

More significant over basic settings

Also ***effective with CNN-based architecture***

Alternative methods (e.g. *k-NN* on features or patch centers) *fail to surpass the baseline due to noise*, highlighting the ***necessity of noise-free supervision for copy detection***.

# Ablation Studies & Param Analysis

ICCV HONOLULU HAWAII OCT 19-23, 2025

ANT GROUP

## Matcher

| Method | Parameter | $\mu$AP | Parameter | $\mu$AP |
|---|---|---|---|---|
| CopyNCE | **default** | **83.5** | $w_{\text{NCE}} = 0$ | 70.9 |
| | $w_{\text{NCE}} = 1$ | 81.7 | $w_{\text{NCE}} = 5$ | 83.5 |
| | $\gamma = 0$ | 82.5 | $\gamma = 0.5$ | 82.6 |
| | $\gamma = 1$ | 83.5 | $\gamma = 2$ | 82.9 |
| | $\gamma = 3$ | 83.0 | $\gamma = +\infty$ | 82.6 |
| | enc-6-fus-6 | 84.0 | enc-10-fus-2 | 79.4 |
| FeatNN | Cos $k = 1$ | Fail | NCE $k = 1$ | Fail |
| LocNN | Cos $k = 1$ | Fail | NCE $k = 1$ | 78.7 |
| Both | Cos $k = 1$ | Fail | NCE $k = 1$ | 80.8 |

$w_{\text{NCE}} = 3$ brings ***+12.6% µAP*** gain over baseline

While feature *k*-NN collapses due to noise, patch-center *k*-NN offers a +7.8% **µ**AP gain. Even the best combination of ***these k-NN methods*** is still ***outperformed by CopyNCE***, with a remaining ***2.7% µAP gap***.

# More Experiments

## Comparison with DISC21 leaderboard

| Descriptor Track | | | Matching Track | | |
|---|---|---|---|---|---|
| Team | $\mu$AP | RP90 | Team | $\mu$AP | RP90 |
| **CopyNCE†** | **65.8** | **61.0** | **CopyNCE†** | **85.6** | **80.0** |
| lyakaap† | 63.5 | 55.4 | **CopyNCE** | **84.6** | **78.2** |
| **CopyNCE** | **60.9** | **56.7** | VisionForce | 83.3 | 73.1 |
| S-square | 59.1 | 50.9 | separate† | 82.9 | 79.2 |
| visionForce | 57.9 | 48.9 | imgFp† | 76.8 | 67.2 |

Table 4. **Leaderboard of DISC21 Phase 2.** † denotes the results achieved after finetuning on `dev set part I`. Note that finetuning is allowed by official rules [34].

## Results on VSC2022

| | SSCD SN [35] | ViT-S SN | ViT-B SN |
|---|---|---|---|
| Descriptor $\mu$AP | 64.99 | 70.59 | **71.57** |
| Matching $\mu$AP | 46.92 | **51.32** | 50.05 |

Table 8. **Results on VSC2022.** Results are produced by official baseline implementation of VSC2022 on its training set.

## Results on AnyPattern

| Method | $\mu$AP | R@1 | Method | $\mu$AP | R@1 |
|---|---|---|---|---|---|
| SSCD [35] | 14.22 | 20.24 | **ViT-S** | **27.07** | **34.68** |
| S-square [33] | 14.51 | 21.05 | **ViT-B** | **31.66** | **37.78** |
| Lyakaap [59] | 13.80 | 18.02 | ViT-S† | 25.38 | 31.57 |
| AnyPat. Base. [51] | 16.18 | 20.54 | ViT-B† | 28.05 | 34.36 |

Table 7. **Results on AnyPattern.** All methods are evaluated with "SmallPattern" protocol . "AnyPat. Base." denotes Baseline in AnyPattern. CopyNCE results are marked in blue and † means results achieved with augmentations that aligned with Lyakaap.

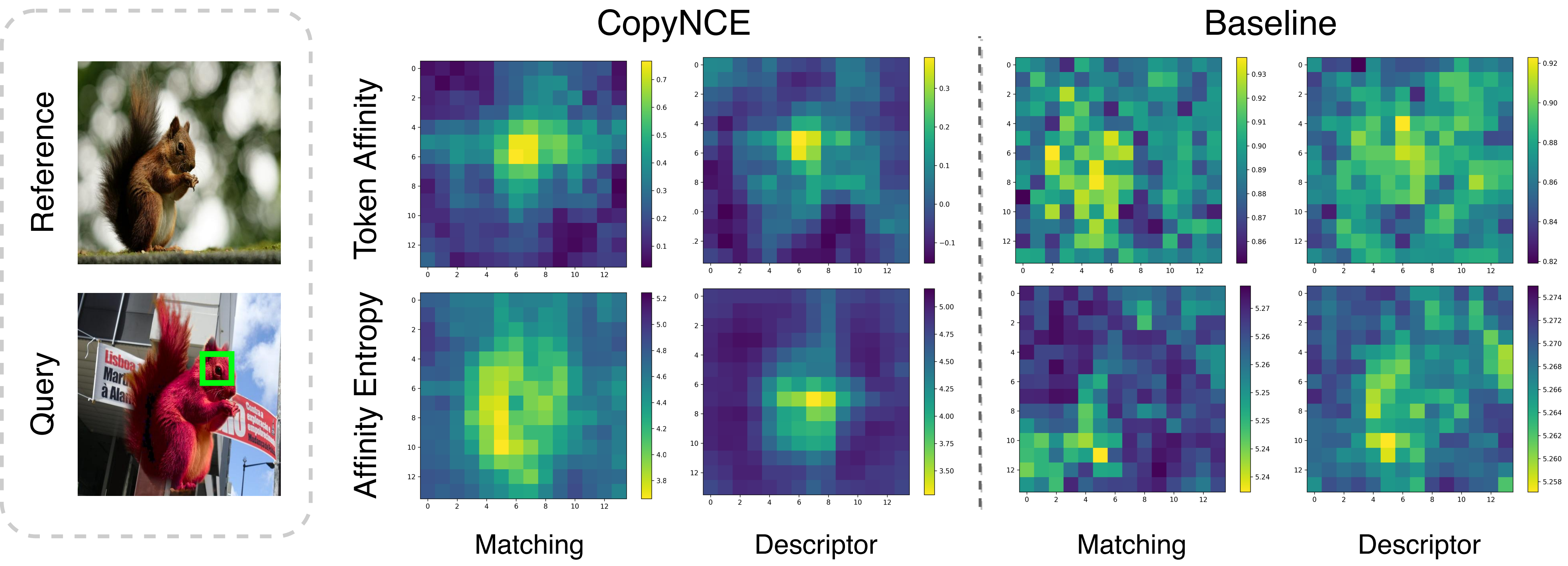
## Results on NDEC

without finetuning on NDEC

| Method | Model Arch. | $\mu$AP | RP90 |
|---|---|---|---|
| **CopyNCE** | **ViT-B+ViT-S** | **72.5** | **36.8** |
| Strong ASL | Multi | 64.1 | - |
| $D^2$LV ASL | Multi | 61.3 | - |

Table 5. **Results on NDEC. Multi** in "$D^2$LV ASL" and "Strong ASL" stands for 11×R50, 11×R152 and 11×R50IBN.

# Visualization




Affinity entropy is defined as: $\mathscr{E}_i = -\sum_j p_{ij} \log p_{ij}, \; p_{ij} = \frac{\exp(\cos(z_i^q, z_j^r)/\tau)}{\sum_k \exp(\cos(z_i^q, z_k^r)/\tau)}.$

Lower Affinity Entropy → Higher likelihood of an edited copy region.

# Key takeaways

| | Before | | After |
|---|---|---|---|
| **Annotation** | Global annotation<br>□ Coarse-grained<br>□ Noisy heuristic | **Traceability of Copied Pixels** → | ***PixTrace***<br>□ ***Fine-grained***: Pixel-level annotation of copied regions<br>□ ***Noise-free***: Precise coordinate correspondence |
| **Loss Function** | Baseline<br>□ Descriptor: Global contrastive and metric learning<br>□ Matcher: BCE | **Guidance from PixTrace** → | ***CopyNCE***<br>□ ***Regularization of Patch Affinity*** |