ECCV'20 ONLINE
23-28 AUGUST 2020

Self-supervising Fine-grained Region Similarities for Large-scale Image Localization

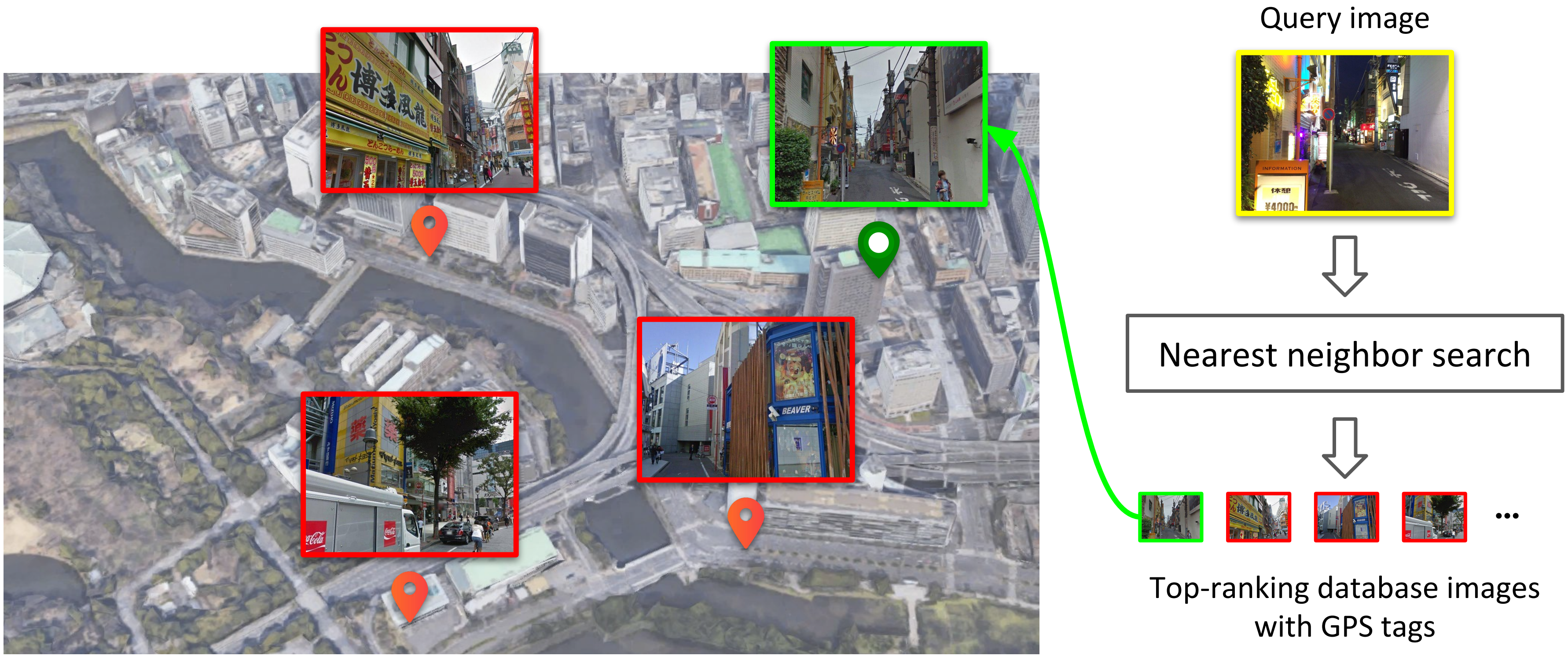




Yixiao Ge[1], Haibo Wang[3], Feng Zhu[2], Rui Zhao[2], Hongsheng Li[1]

[1] The Chinese University of Hong Kong,
[2] SenseTime Research,
[3] China University of Mining and Technology

Image Localization via Image Retrieval

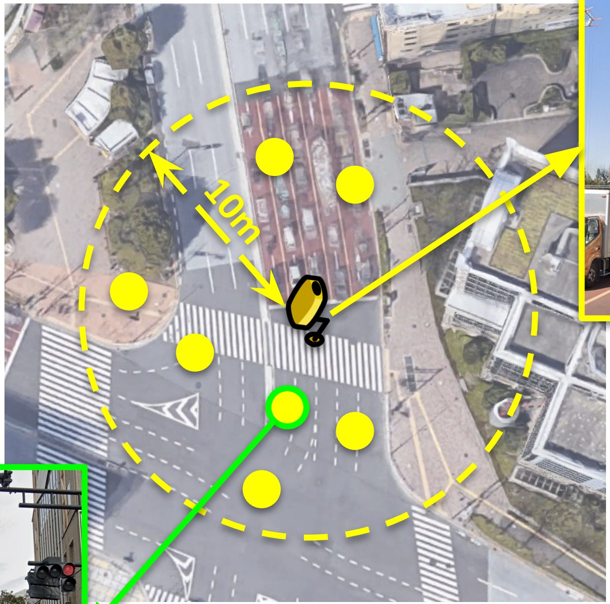


Challenge #1: Noisy Positives by Weak GPS Labels

Geographically close-by images may not depict the same scene when facing different directions.

Previous Solution: Train with Only the Easiest Positive

Potential positives filtered by GPS labels



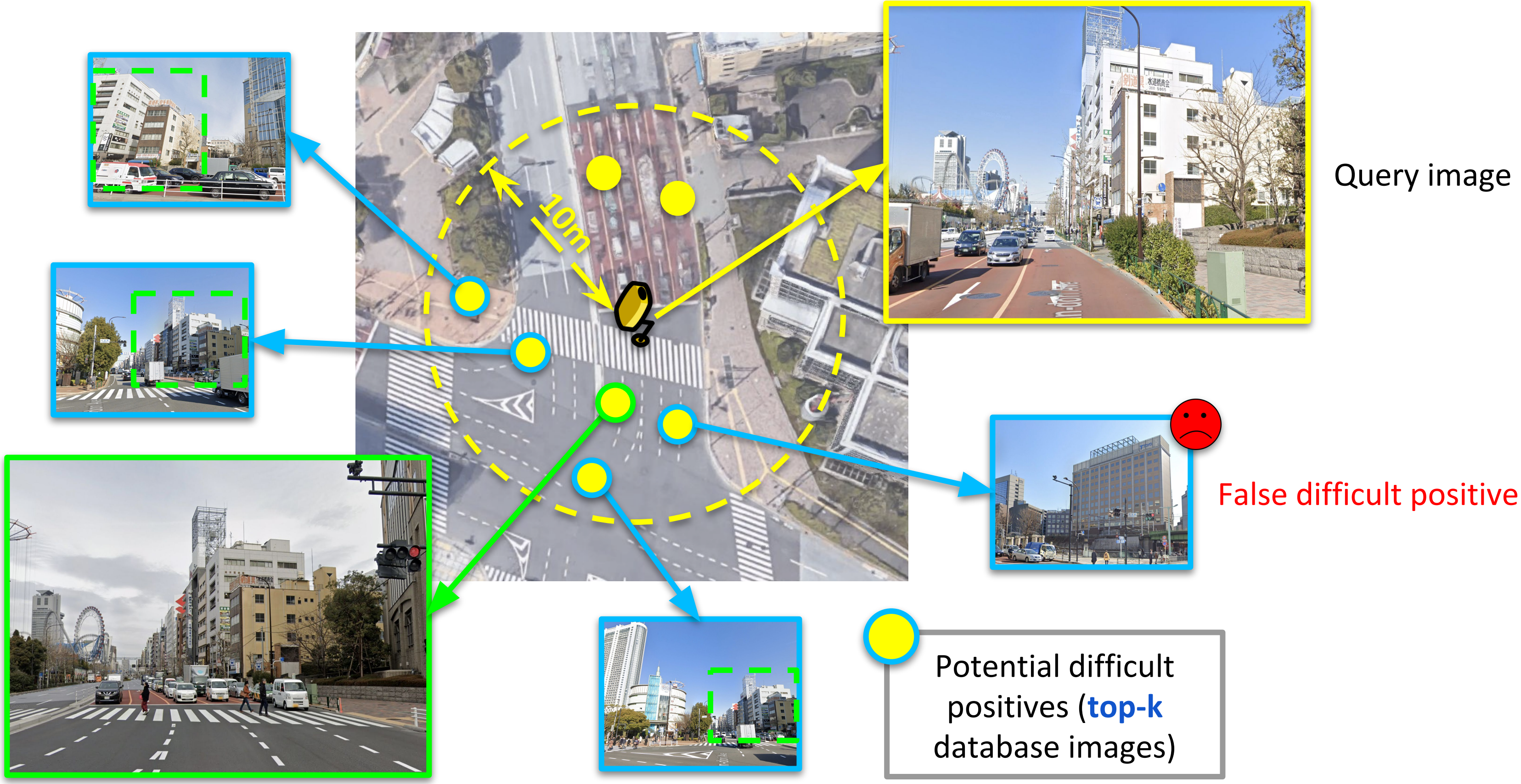
Query image

Top-1 database image

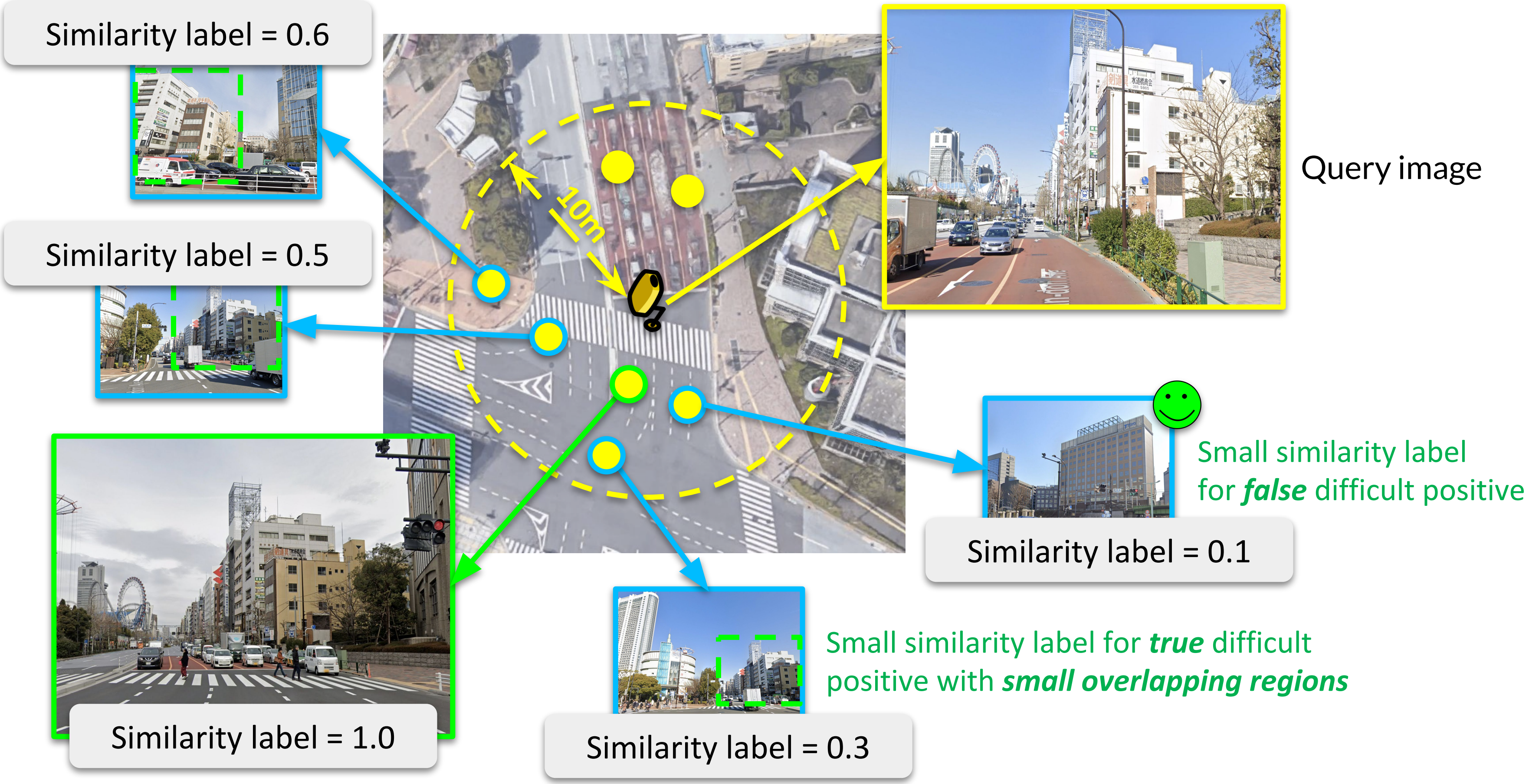
Forcing the queries to be closer to their already nearest neighbors results in a lack of robustness to varying conditions.

Difficult positives are needed!

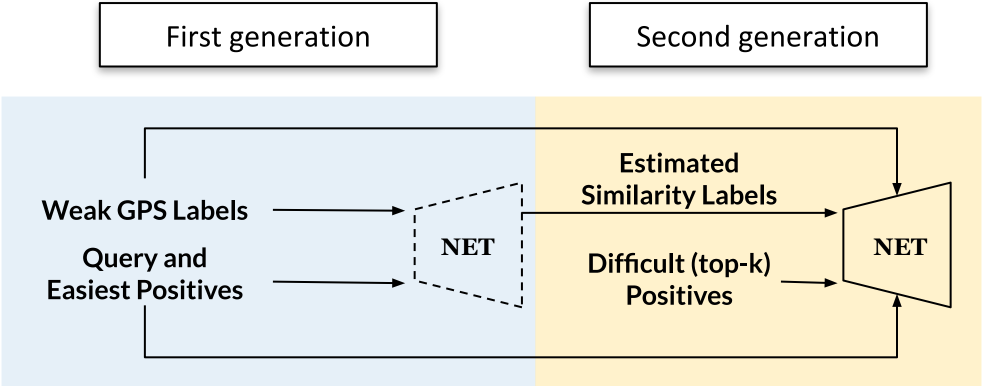
Motivation: Use Noisy Difficult Positives Properly



Our Solution: Image Similarities as Soft Supervisions

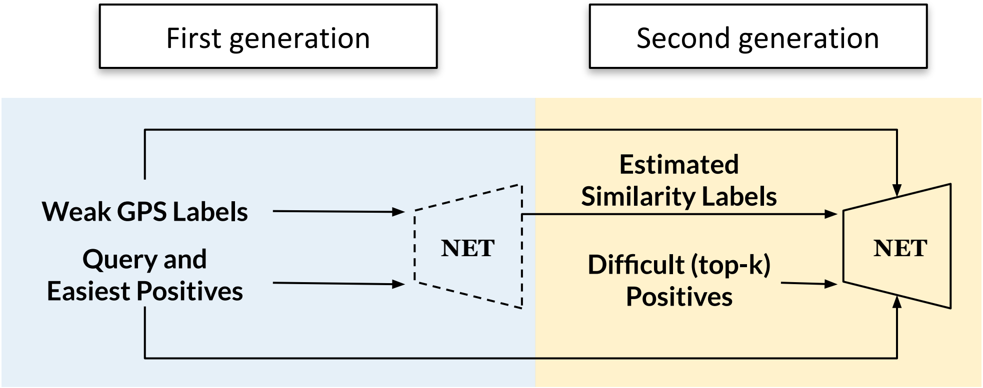


Our Solution: Similarity Labels



The first generation's query-gallery similarities serve as the soft supervision for training the network in the second generation.

Our Solution: Similarity Labels



Similarity labels: $\mathcal{S}_{\theta_1}(q, p_1, \cdots, p_k; \tau_1) = \text{softmax}\left(\left[\langle f^q_{\theta_1}, f^{p_1}_{\theta_1}\rangle/\tau_1, \cdots, \langle f^q_{\theta_1}, f^{p_k}_{\theta_1}\rangle/\tau_1\right]^\top\right)$

Query

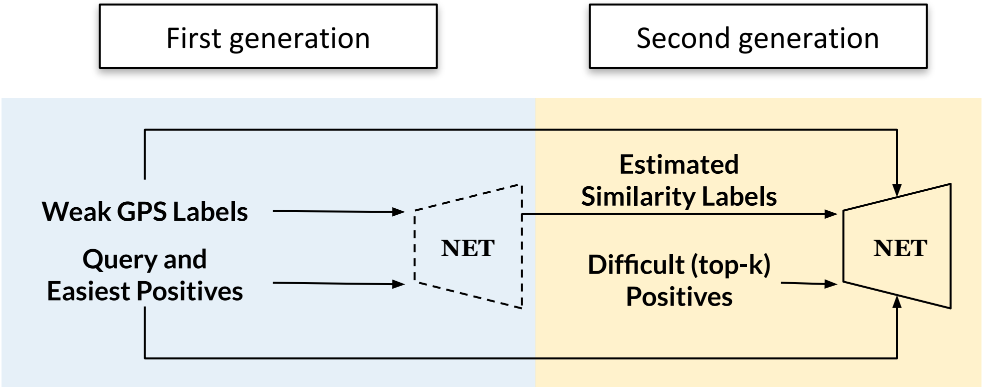
Positive #1

Temperature
for generation #1

Image similarity between
query and positive #1

Parameters of the network
in generation #1

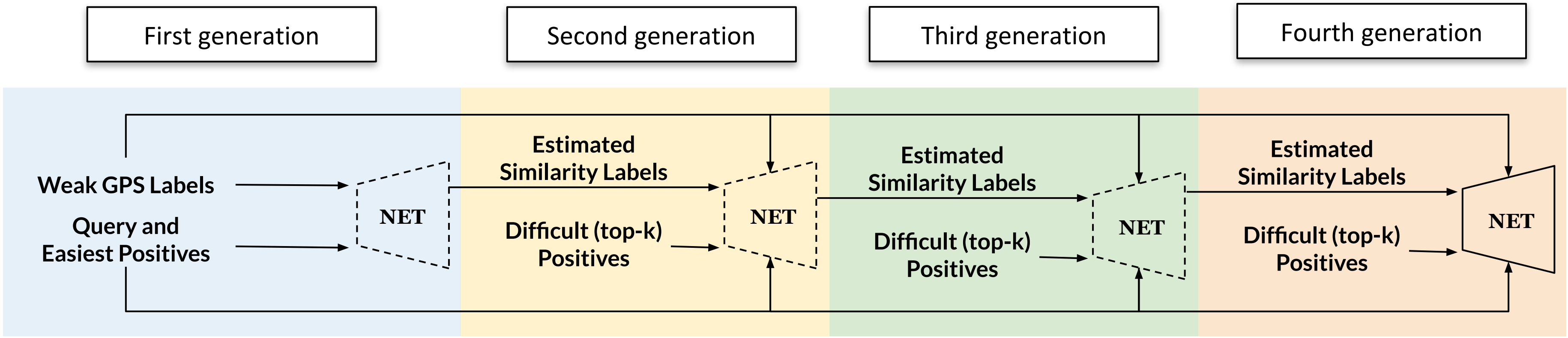
Our Solution: Similarity Labels



Similarity labels: $$\mathcal{S}_{\theta_1}(q, p_1, \cdots, p_k; \tau_1) = \mathrm{softmax}\left(\left[\langle f^q_{\theta_1}, f^{p_1}_{\theta_1}\rangle/\tau_1, \cdots, \langle f^q_{\theta_1}, f^{p_k}_{\theta_1}\rangle/\tau_1\right]^\top\right)$$

Soft-label loss: $$\mathcal{L}_{\mathrm{soft}}(\theta_2) = \underbrace{\ell_{ce}}_{\text{cross-entropy loss}}\left(\mathcal{S}_{\theta_2}(q, p_1, \cdots, p_k; 1), \underbrace{\mathcal{S}_{\theta_1}(q, p_1, \cdots, p_k; \tau_1)}_{\text{similarity labels (learning targets) estimated by the network in generation \#1}}\right)$$

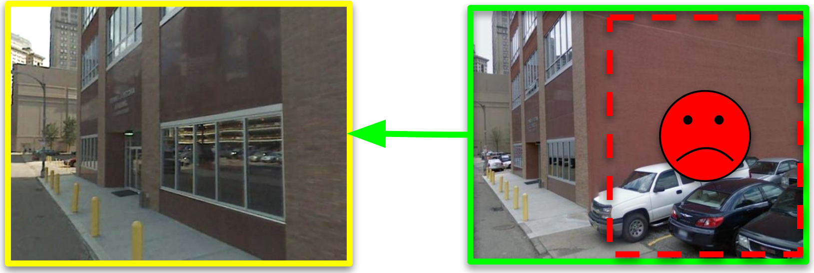
Our Solution: Self-enhanced Similarity Labels



The generated soft supervisions are gradually refined as the network generation progresses.

Challenge #2: Lack of Region-level Supervisions

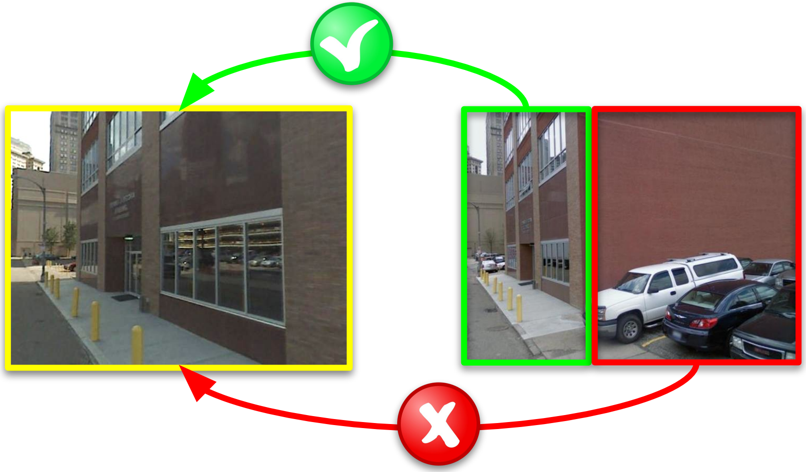
Only image-level labels



Query image

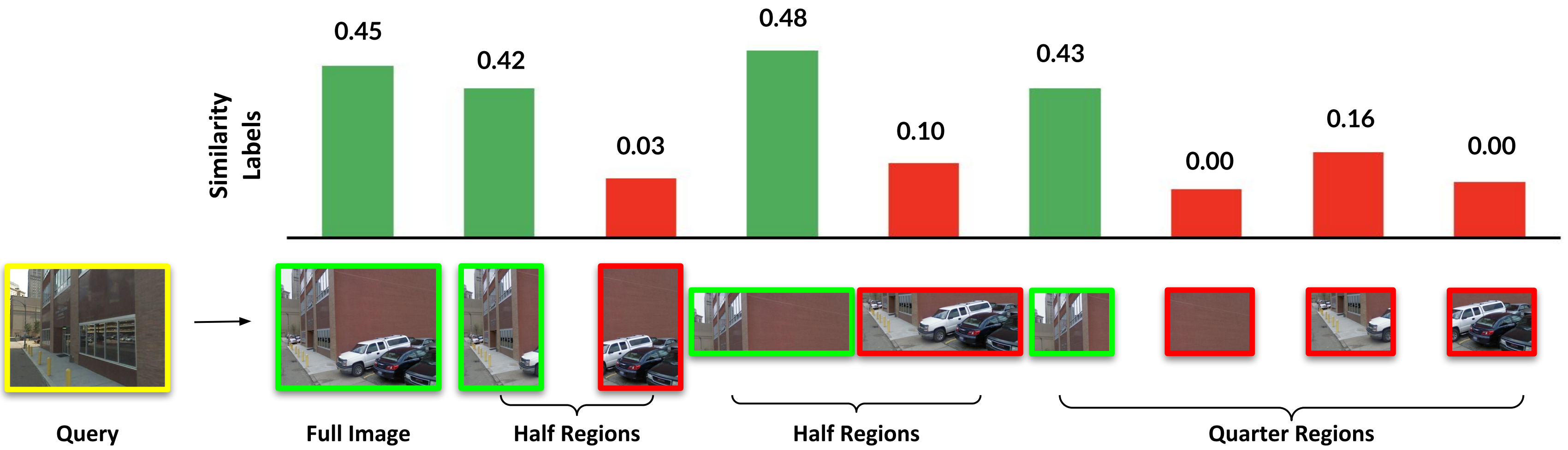
Positive sample

Ideal image-to-region labels



The correct image-level labels might not necessarily be the correct region-level labels.

Our Solution: Image-to-region Similarities as Soft Supervisions



Provide fine-grained image-to-region similarities to enhance the learning of local features.

Our Solution: Image-to-region Similarities as Soft Supervisions

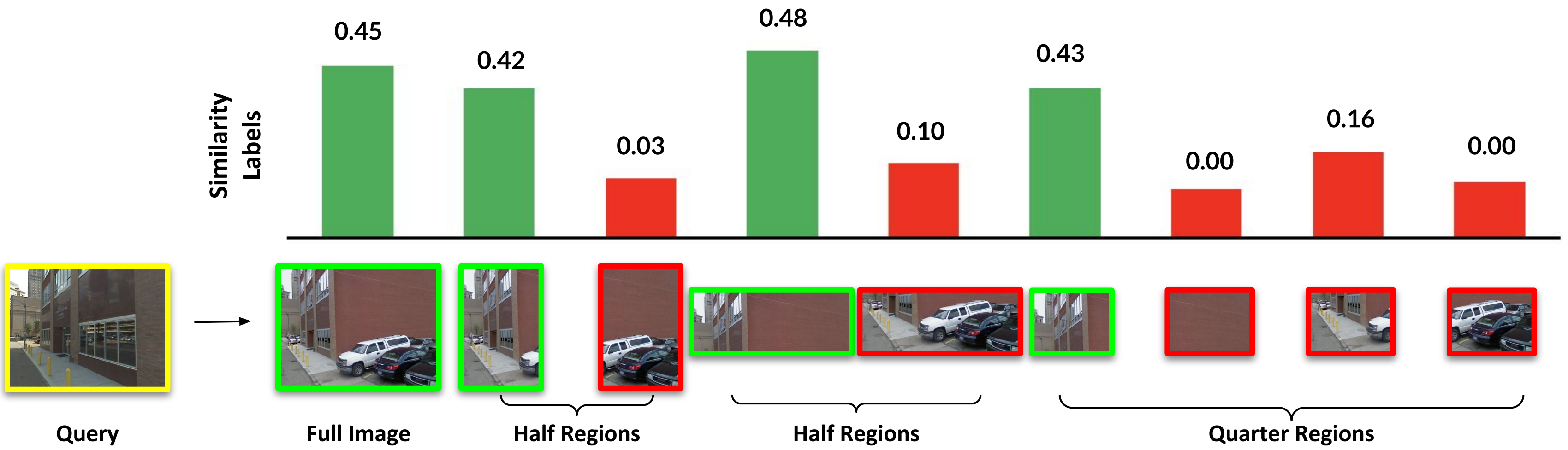
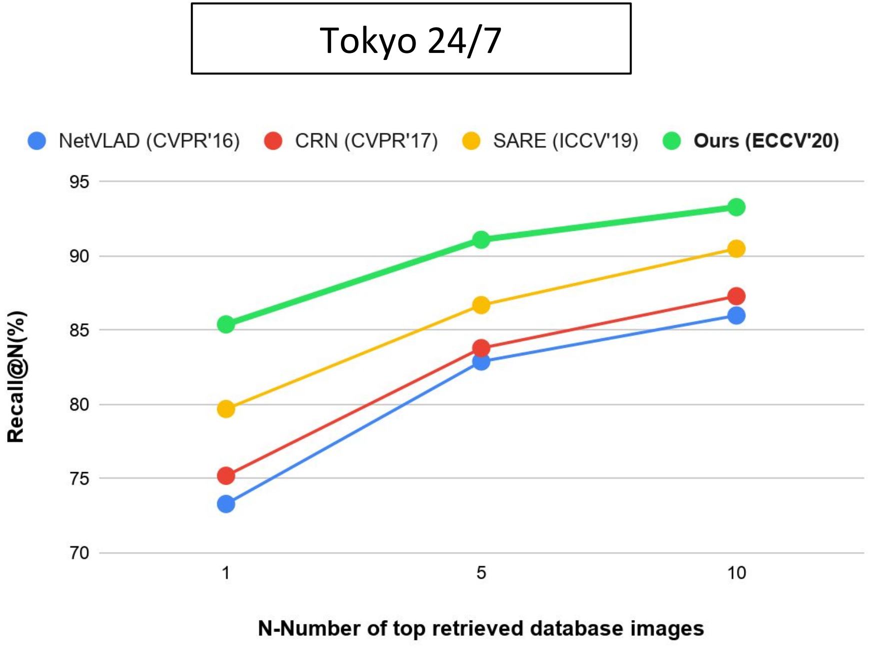


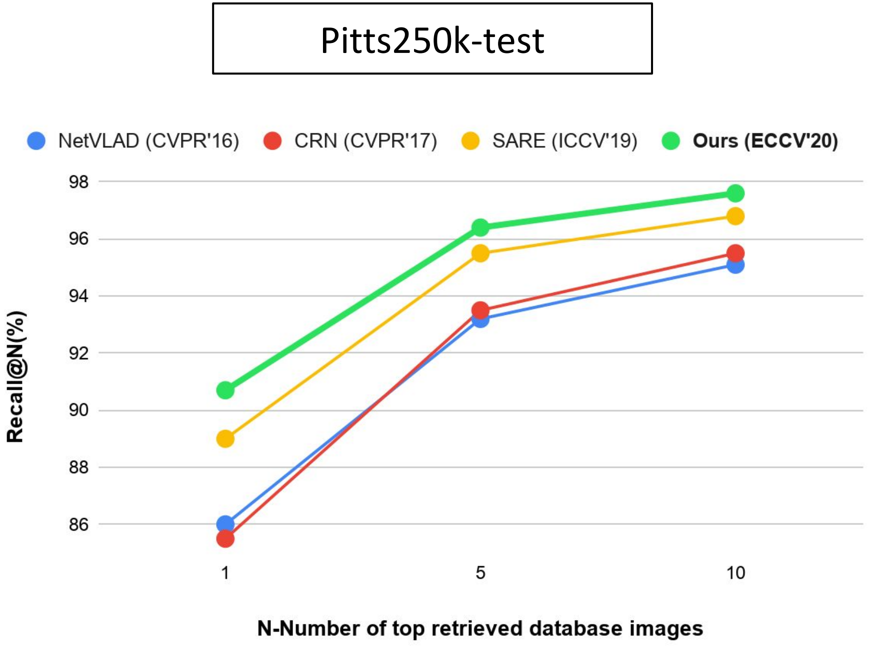
Image similarities between query and regions of positive #1

Fine-grained Similarity labels:

$$\mathcal{S}^{r}_{\theta_\omega}(q, p_1, \cdots, p_k; \tau_\omega) = \mathrm{softmax}\Big(\big[\langle f^{q}_{\theta_\omega}, f^{p_1}_{\theta_\omega}\rangle/\tau_\omega, \boxed{\langle f^{q}_{\theta_\omega}, f^{r_1^1}_{\theta_\omega}\rangle/\tau_\omega, \cdots, \langle f^{q}_{\theta_\omega}, f^{r_1^8}_{\theta_\omega}\rangle/\tau_\omega,} \\ \cdots, \langle f^{q}_{\theta_\omega}, f^{p_k}_{\theta_\omega}\rangle/\tau_\omega, \langle f^{q}_{\theta_\omega}, f^{r_k^1}_{\theta_\omega}\rangle/\tau_\omega, \cdots, \langle f^{q}_{\theta_\omega}, f^{r_k^8}_{\theta_\omega}\rangle/\tau_\omega\big]\Big)$$

Performances on Image Localization Benchmarks





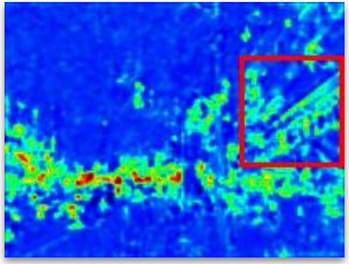
Comparison with State-of-the-art (#1)

Query image

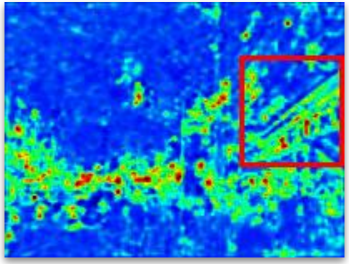
Query's heatmap

Retrieved top-1 image

SARE (ICCV'19)



Ours (ECCV'20)



Our method pays more attention on the discriminative shop signs than SARE.

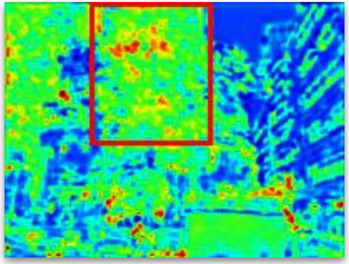
Comparison with State-of-the-art (#2)

Query image

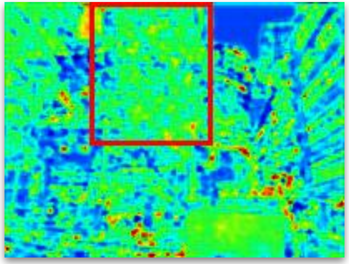
Query's heatmap

Retrieved top-1 image

SARE (ICCV'19)



Ours (ECCV'20)



SARE incorrectly focuses on the trees,
while our method learns to ignore such misleading regions.

Self-supervising Fine-grained Region Similarities for Large-scale Image Localization

Yixiao Ge[1], Haibo Wang[3], Feng Zhu[2], Rui Zhao[2], Hongsheng Li[1]

[1] The Chinese University of Hong Kong,
[2] SenseTime Research,
[3] China University of Mining and Technology

Code available at

https://github.com/yxgeee/SFRS