RIS: Region-to-Image Search using ViT-like Embeddings

Oussama Zayene [*1] **Lucas Genoud** [1] **Jean Hennebert** [1] **Houda Chabbi Drissi** [*1] **Benoît de Raemy** [2]

Abstract

We propose RIS (Region-to-Image Search), a two-stage framework for localized visual retrieval. RIS performs structural re-ranking directly within the latent embedding space of Vision Transformers, such as SigLIP2 and I-JEPA, bypassing traditional pixel-level verification. By matching a query Region of Interest through a spatially-consistent region-growing algorithm, the framework ensures geometric coherence across latent representations. Experimental results demonstrate that this embedding-based re-ranking improves Top-5 retrieval accuracy by at least 10% over standalone global methods, providing a robust and efficient mechanism for localized forensic search.

1. Introduction

Modern visual retrieval systems increasingly rely on powerful image encoders to produce semantically rich embeddings, enabling the organization of large-scale image collections without manual labels. However, in forensic and investigative scenarios, queries often consist of localized regions of interest (ROIs), such as specific objects or partial scenes that may occupy only a small fraction of the total image area. In such cases, global similarity metrics often prove insufficient, motivating the need for retrieval systems that explicitly exploit local visual correspondences.

For the past five years, Vision Transformers (ViTs) (Dosovitskiy, 2020) have emerged as a dominant paradigm in visual representation learning. By decomposing images into fixed-size patches and processing them as tokens, ViTs naturally provide both global image embeddings and dense patch-level representations. While global embeddings are standard for image-level search, patch embeddings offer a rich source of localized information that remains underutilized for fine-grained matching.

The objective of this study is to explore how these latent patch-level representations can support robust retrieval when only a small region of a query image is available. To achieve this, we propose a hierarchical retrieval framework that operates in two phases:

- **Offline Indexing:** Encoding the image corpus into a dual-representation latent space (global and per-patch), stored in a scalable vector index to facilitate rapid multi-level retrieval.
- **Online Inference:** A multi-stage process involving ROI selection, global filtering for scalability, and local patch matching, culminating in a final image re-ranking based on semantic similarity and spatial coherence.

We evaluate this framework across two distinct learning paradigms: contrastive (SigLIP2) and self-supervised (I-JEPA). Our qualitative results indicate that leveraging patch-level tokens substantially improves the robustness of image retrieval, particularly in scenarios where global context is limited and evidence is localized.

The paper is organized as follows: Section 2 reviews Vision Transformers and forensic retrieval. Section 3 details our proposed approach. Section 4 presents our experimental setup and a qualitative comparison between contrastive and self-supervised embedding models. Finally, Section 5 concludes the paper with a summary of our findings and and future research directions.

2. Related Work

Image retrieval has traditionally relied on hand-crafted features such as SIFT, HOG, or SURF (Azzopardi et al., 2021), which aimed to capture invariant local patterns under changes in viewpoint, illumination, or scale. Although effective for specific tasks, these descriptors were limited by their rigidity and struggled to generalize across diverse scenes. The emergence of deep learning transformed this landscape by replacing engineered features with learned representations that are optimized from large-scale datasets. The first wave of deep retrieval systems leveraged CNNs

[*]Equal contribution [1]iCoSys, HEIA-FR, University of Applied Sciences and Arts Western Switzerland [2]Morphean SA, Fribourg, Switzerland. Correspondence to: Oussama Zayene <oussama.zayene@hefr.ch>, Houda Chabbi Drissi <houda.chabbi@hefr.ch>.

Proceedings of the Swiss AI Days 2026, Martigny, VS & Fribourg, FR, Switzerland PMLR 309, 2026.

trained for image classification. Intermediate activations served as global descriptors, offering significantly improved semantic discrimination. However, classification-supervised models tend to produce embeddings sensitive to dataset biases and often fail to preserve fine-grained visual details that are crucial for retrieving images containing relatively small objects.

2.1. Vision Transformers and Patch-based Representations

ViTs (Dosovitskiy, 2020) introduced a paradigm shift in computer vision by replacing convolutional operations with self-attention mechanisms operating over patch tokens. By splitting an image into a grid of fixed-size patches and embedding each patch into a latent space, ViTs provide structured representations that preserve both global context and local information. This architectural design has made ViTs particularly attractive for tasks requiring fine-grained reasoning, such as localization and part-based matching. Across different training strategies, ViT-based models expose similar intermediate representations: patch embeddings corresponding to local image regions and a global embedding summarizing the entire image. These shared properties enable the development of model-agnostic retrieval pipelines that operate directly on patch and global embeddings.

2.2. Joint-embedding Models

Joint embedding approaches, exemplified by contrastive models like CLIP (Radford et al., 2021) and sigmoid-based methods like SigLIP (Zhai et al., 2023), learn robust visual representations by aligning images with corresponding textual signals. Unlike standard supervised learning, these models optimize a joint objective, such as InfoNCE or pairwise sigmoid loss, to project visual and textual features into a shared embedding space. This alignment encourages embeddings that capture high-level semantic similarity, enabling effective zero-shot retrieval and cross-domain generalization. In the context of retrieval, these ViT-based architectures have demonstrated strong performance for image-level similarity search and cross-modal tasks. While their patch embeddings are known to encode meaningful local semantics, standard applications rely primarily on the pooled global token (e.g., [CLS]). Consequently, the potential of using dense patch-level representations for fine-grained, ROI-based retrieval remains underutilized in standard pipelines.

2.3. Self-Supervised Learning Models

Self-supervised learning (SSL) has emerged as a dominant paradigm for learning general-purpose visual features without reliance on manual annotations. Early contrastive approaches, such as SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020), leverage data augmentation and instance-level discrimination to produce robust global embeddings. More recent methods, including DINO (Caron et al., 2021) and MaskFeat (Wei et al., 2022), shift focus towards learning spatially grounded features and capturing latent image structures. However, many SSL objectives prioritize invariance to augmentations, encouraging models to discard fine-grained spatial details. While effective for image-level classification, this invariance can be detrimental to tasks requiring precise localized understanding or region-based retrieval. The Joint Embedding Predictive Architecture (JEPA) (LeCun, 2022) proposes a fundamentally different approach. Unlike Masked Image Modeling (e.g., MAE (He et al., 2022)), which predicts raw pixels, or standard contrastive methods which maximize invariance, JEPA learns to predict missing information directly in the high-level embedding space. This strategy forces the model to capture semantic structure rather than low-level image statistics. I-JEPA (Assran et al., 2023) instantiates this framework for images, producing highly interpretable and spatially structured representations that are particularly well-suited for dense prediction tasks.

2.4. Global–Local Retrieval Strategies

Recent work has explored how global and local visual information should be combined for image retrieval. `ViT-GaL` (Phan et al., 2022) proposes a unified architecture that jointly learns global and local descriptors within a single ViT-based model, using additional convolutional modules and end-to-end supervised training to optimize retrieval performance. In parallel, (Aiger et al., 2025) has shown that the ordering of global and local matching stages is a critical design choice, contrasting conventional global-to-local pipelines with emerging local-to-global retrieval strategies enabled by efficient local search. Together, these studies highlight that both patch-level representations and global–local interaction strategies play a central role in modern retrieval systems.

In contrast to these approaches, RIS does not introduce a new retrieval architecture or training objective; instead, it leverages ViT-like embeddings and explicitly studies how frozen global and patch-level representations can be composed in a hierarchical, region-to-image retrieval pipeline. Our approach enables retrieval from user-defined query regions that corresponds to localized areas.

2.5. Summary and Synthesis

The evolution of visual representation learning has moved from global instance-level discrimination toward architectures that capture dense, localized information. Crucially, both the vision-language models (e.g., SigLIP) and self-supervised models (e.g., I-JEPA) discussed in this section are based on the ViT architecture. This shared foundation is critical for our work, as it provides direct access to

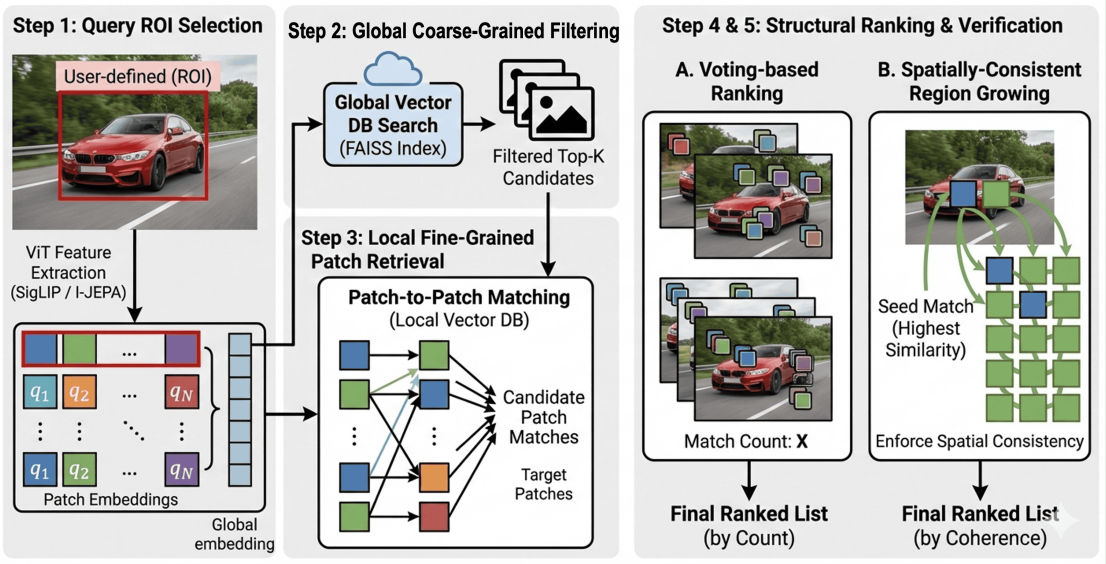

Figure 1. Overview of RIS inference pipeline. (1) **ROI Selection:** Extraction of dense patch embeddings q_i from a user-defined region. (2) **Global Filtering:** Coarse retrieval via FAISS to identify top-K candidates. (3) **Local Matching:** Comparative analysis of query patches against candidate target patches. (4) **Ranking:** Final scoring using either frequency-based voting or spatially-consistent region growing.

patch-level embeddings that preserve the spatial resolution necessary for region-based retrieval. Our proposed retrieval pipeline is designed to be model-agnostic. This allows us to rigorously evaluate and compare how different learning paradigms, namely language-supervised contrastive learning *versus* predictive self-supervised learning, impact retrieval precision and spatial coherence. Next section gives more technical details about the selected models.

3. Methodology

3.1. System Overview

The proposed RIS (Region-to-Image Search) framework is designed for localized visual search through a dual-phase architecture consisting of an offline indexing stage and an online inference stage. The core objective is to move beyond monolithic image descriptors by utilizing the inherent patch-level tokenization of ViTs. This allows the system to remain model-agnostic while ensuring privacy and scalability by indexing compact latent embeddings instead of raw pixel data. The end-to-end inference process is illustrated in Figure 1 and detailed in the following sections.

3.2. Feature Representation

To evaluate the impact of different learning paradigms on the RIS framework, we utilize two distinct ViT-based backbones for feature extraction.

I-JEPA (Assran et al., 2023) is a self-supervised model that learns by predicting latent representations of masked image regions. The architecture consists of a Context Encoder (E_C), a Target Encoder (E_T), and a Predictor (P). This predictive objective encourages the model to capture high-level semantic features relevant for localized matching without requiring manual supervision.

SigLIP2: Conversely, SigLIP2 (Zhai et al., 2023) is trained via a sigmoid-based contrastive objective to align image and text pairs. In this work, we employ only the visual encoder as a pure feature extractor. The encoder tokenizes each image into patches and outputs a global representation along with patch-level embeddings.

Both architectures operate by decomposing an input image I into a grid of non-overlapping patches of size $P \times P$. Each patch is linearly projected into a D-dimensional embedding space and augmented with positional encoding to yield a sequence of patch tokens:

$$P = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}, \quad \mathbf{p}_i \in \mathbb{R}^D, \quad i = 1, \ldots, N \quad (1)$$

Each embedding is ℓ_2-normalized to ensure scale-invariant similarity comparisons during retrieval:

$$\tilde{\mathbf{p}}_i = \frac{\mathbf{p}_i}{\|\mathbf{p}_i\|_2} \quad (2)$$

As depicted on the left side of Fig. 2, the model decomposes the input into a 16×16 grid of patches, each with a spatial resolution of 14×14 pixels. This discretization yields a dense set of $N = 256$ latent feature tokens, facilitating the localized matching required for the subsequent stages.

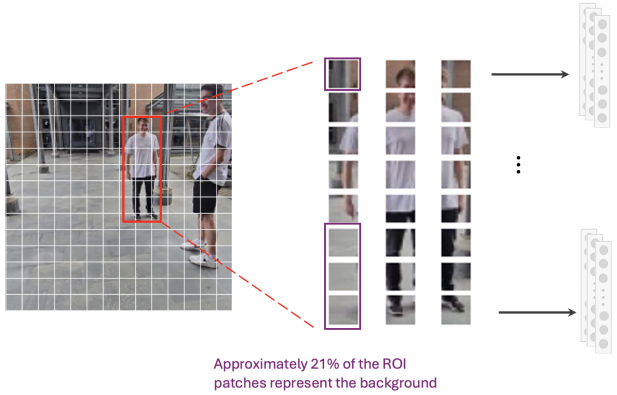

Figure 2. ROI Selection & Patch Extraction

3.3. Offline Indexing

The offline phase transforms the image database $\mathcal{D} = \{I_1, I_2, \ldots, I_M\}$ into a dual-level indexed structure. For each image $I_j \in \mathcal{D}$, the system extracts and stores two distinct levels of representation:

- **Global-Level Index:** A set of vectors $\mathbf{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_M\}$, where $\mathbf{g}_j \in \mathbb{R}^D$ represents the holistic context of image I_j. This index is used for rapid candidate pruning.
- **Patch-Level Index:** A collective repository $\mathcal{P} = \bigcup_{j=1}^{M} P_j$ containing all normalized patch tokens $\tilde{\mathbf{p}}_{j,i}$ from every image in the database.

Both levels are indexed using FAISS (Douze et al., 2024) to enable high-speed approximate nearest neighbor (ANN) search. In our experiments, We employ the following `IndexFlatIP` index, which allows ranking using inner product (equivalent to cosine similarity after normalization). Since the patch-level index $\mathcal{P}$ treats all tokens as a singular flat collection, a mapping function is required to retrieve the source metadata. We define the mapping f_{map} as:

$$f_{map}(i) \rightarrow (img\_id, patch\_id) \tag{3}$$

where i is the global index returned by FAISS, img_id identifies the parent image I_j, and $patch\_id$ denotes the specific patch index within that image (from 1 to N). This mapping is essential for the subsequent spatial verification phase, as it allows the system to reconstruct the geometric layout of matches.

3.4. Online Inference

During the inference phase, the system processes a query image I_Q where a user defines a region of interest (ROI_Q). The objective is to identify images $I_j \in \mathcal{D}$ that contain visual content semantically consistent with the ROI_Q by leveraging the dual-level index structure.

3.4.1. Query Representation

The ROI_Q is processed by the ViT backbone to extract a set of n patch embeddings $\mathcal{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ corresponding to the selected spatial region (see Fig. 2, right). Consistent with the indexing phase, all query vectors are ℓ_2-normalized to facilitate scale-invariant similarity comparisons.

3.4.2. Global Filtering

To maintain computational efficiency, the system first performs a coarse retrieval. The query descriptor $\mathbf{q}_g$ is matched against the global-level index $\mathbf{G}$ using cosine similarity:

$$S_{global}(j) = \mathbf{q}_g \cdot \mathbf{g}_j \tag{4}$$

This step identifies a candidate subset $\mathcal{C} \subset \mathcal{D}$ consisting of the top-K most similar images, effectively pruning the search space for the localized matching phase.

This global retrieval baseline provides a reference for the effectiveness of patch-wise and region-based retrieval strategies. While it captures coarse semantic similarity, it does not account for spatially localized content and is therefore insufficient for fine-grained region queries.

3.4.3. Local Patch Matching

For each candidate image $I_c \in \mathcal{C}$, the system retrieves the target patch set P_c from the index. We establish a set of local correspondences $\mathcal{M}$ where a match is identified if the cosine similarity between a query patch $\mathbf{q}_i \in \mathcal{Q}$ and a target patch $\tilde{\mathbf{p}}_{c,k} \in P_c$ exceeds a predefined threshold τ:

$$\mathcal{M} = \{(\mathbf{q}_i, \tilde{\mathbf{p}}_{c,k}) \mid \mathbf{q}_i \cdot \tilde{\mathbf{p}}_{c,k} > \tau\} \tag{5}$$

By invoking the mapping function f_{map} established in Section 3.3, the system resolves the global patch indices into local spatial coordinates. These localized correspondences serve as the foundational evidence for the re-ranking algorithms.

3.5. Ranking Strategies

The final stage of the RIS framework aggregates local correspondences $\mathcal{M}$ to produce a ranked list of images. We propose two alternative strategies for calculating the final matching score $S(I_Q, I_c)$: statistical voting and structural region growing. Figure 3 contrasts the fundamental mechanisms of the two proposed ranking strategies.

3.5.1. Voting-Based Image Ranking

In this approach, patch matches are treated as independent evidence. For each query patch $q \in \mathcal{Q}$, we retrieve its k-nearest neighbors in the latent space. To transform these local correspondences into a unified image-level score $S(j)$, we evaluated the four aggregation strategies detailed in Table 1. These methods range from simple frequency counting

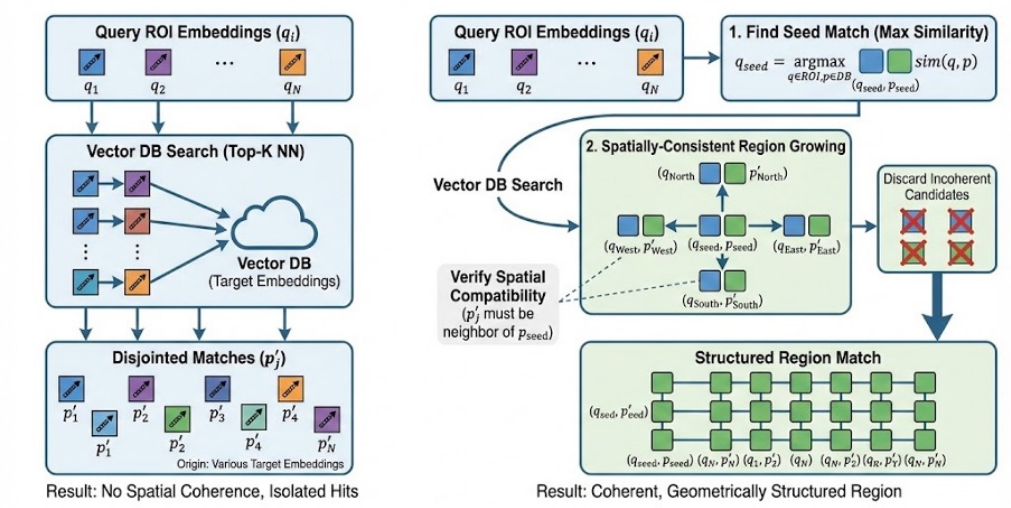

Figure 3. Comparison of retrieval strategies: (Left) Patch-wise retrieval, where independent searches result in disjointed matches across unrelated images; (Right) Spatially-consistent region growing, which propagates from a high-confidence seed match to build a structurally coherent region within a single target image.

to distance-weighted scoring, providing a baseline for understanding how local similarity translates to global relevance (without explicit spatial constraints).

Table 1. Summary of Voting-based Ranking Strategies

Strategy	Description
Majority Vote	Each neighbor contributes one vote: $V(j) = \sum_{p \in \mathcal{P}_{ROI}} \sum_{t=1}^{k} \mathbf{1}\{\gamma(i_{p,t}) = j\}$
Unique-image Vote	Only the best-ranked neighbor per image counts: $V(j) = \sum_{p \in \mathcal{P}_{ROI}} \mathbf{1}\{\exists t : \gamma(i_{p,t}) = j\}$
Weighted Distance Vote	Votes are weighted by inverse distance: $W(j) = \sum_{p,t:\gamma(i_{p,t})=j} \frac{1}{d_{p,t} + \epsilon}$
Weighted Unique Vote	For each patch, only the best-scoring neighbor per image contributes: $W(j) = \sum_{p \in \mathcal{P}_{ROI}} \max_{t:\gamma(i_{p,t})=j} \frac{1}{d_{p,t} + \epsilon}$

3.5.2. Region-Growing Based Ranking

As an alternative to voting, this strategy enforces local geometric consistency. It identifies the largest spatially-coherent cluster of matches between the ROI_Q and the target image.

Expansion Logic: Starting from the top K_s seeds, a region $\mathcal{G}$ is expanded by iteratively adding neighboring matches (i', j') that satisfy $i' \in \mathcal{Q}$ and $\mathbf{q}_{i'} \cdot \mathbf{p}_{j'} \geq \tau$. The framework retains the region with the largest spatial support, denoted as $\mathcal{G}^*$.

Scoring Function: The final score is computed by modulating the semantic similarity of the pooled region descriptors with the spatial coverage:

$$S(I_Q, I_c) = \left(\frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbf{q}_i \cdot \frac{1}{|\mathcal{G}^*|} \sum_{(i,j) \in \mathcal{G}^*} \mathbf{p}_j \right) \cdot \left(\frac{|\mathcal{G}^*|}{|\mathcal{Q}|} \right)^{\alpha} \tag{6}$$

where α is a hyperparameter controlling the influence of the region's size on the final ranking.

4. Experimental Evaluation

4.1. Experimental Setup

Dataset and Evaluation Protocol To the best of our knowledge, current public retrieval benchmarks focus primarily on holistic image-to-image similarity and lack specialized subsets for ROI to image matching. Consequently, we curated a controlled in-house dataset comprising 90 video frames captured via a mobile device. The dataset is partitioned into an index set of 81 images and a query set of 9 representative images (see Fig. 4), where each query contains a manually defined ROI_Q targeting a specific subject. To simulate challenging forensic scenarios, the index covers diverse indoor and outdoor scenes featuring subjects of varying heights and subjects wearing highly similar clothing. Furthermore, the subjects are recorded from multiple perspectives and at varying distances from the sensor. This introduces significant scale variability, requiring the RIS framework to robustly match regions across both proximal and distant subject-to-camera configurations.

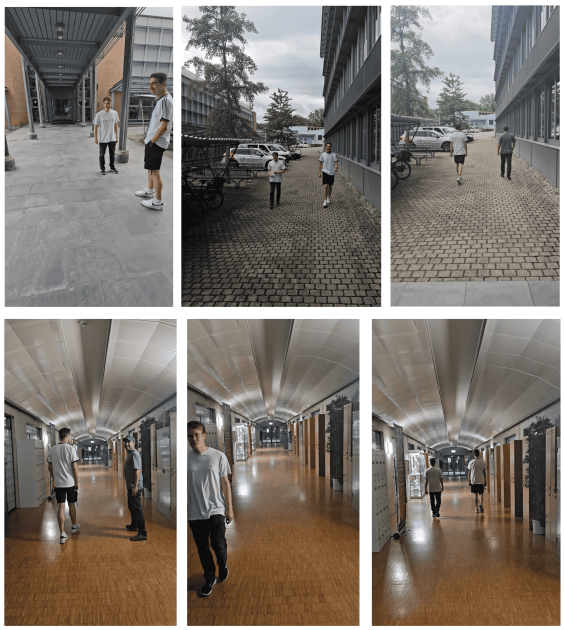

Figure 4. Representative samples of the collected dataset

Implementation Details The RIS framework is implemented in Python, leveraging the FAISS library for high-performance approximate nearest neighbor search. We benchmark two state-of-the-art ViT backbones:

- **SigLIP2:** The `siglip2-so400m-patch14-224` model, which produces 1280-dimensional embeddings.
- **I-JEPA:** The `ijepa_vith14_1k` variant, which also yields a 1280-dimensional latent space.

Since I-JEPA lacks a native $[CLS]$ token for global representation, we evaluated several aggregation techniques, including simple mean-pooling and Laplacian-based saliency weighting. We determined that **Saliency-Weighted Mean pooling** provided the most robust global descriptors for initial filtering. Images are pre-processed to a resolution of 224×224 pixels, resulting in a 16×16 patch grid. The offline stage generates a global index of 81 vectors and a local patch index of 20,655 vectors. All experiments were conducted on a workstation equipped with an NVIDIA RTX A6000 GPU.

Hyperparameters The following parameters were utilized to ensure consistent evaluation across both backbones:

- **Global Retrieval:** The top $N = 50$ candidate images are retrieved during the global filtering phase.
- **Seed Selection:** Each query patch considers $K = 5$ nearest neighbors, with the final seed determined by the maximum spatial consistency score (`argmax`).
- **Spatial Verification:** We employ a 4-connected neighborhood for region expansion with a similarity threshold $\tau = 0.6$.
- **Scoring:** Spatial coverage is set to $\alpha = 1.0$ to balance semantic similarity with geometric support.

4.2. Results and Discussion

Quantitative Impact of Hierarchical Retrieval The hierarchical architecture of the RIS framework is critical for high-precision retrieval. Across all queries, the local refinement stage consistently increased Top-5 accuracy by at least 10% compared to standalone global retrieval for both backbones. While the global stage acts as an efficient semantic filter, the region-growing phase is necessary to verify the spatial presence of the ROI_Q. We observed that SigLIP2 demonstrates superior performance in the initial global filtering, though final accuracy remains inherently bounded by the recall of the Top-N candidate set. Overall, both models achieved high retrieval robustness, correctly identifying approximately 90% of the subjects in the top results.

Structural Verification and Seed Sensitivity The region-growing strategy consistently outperformed independent voting methods by enforcing geometric consistency. However, our analysis reveals that the success of this process is highly sensitive to the initial seed selection; a high-confidence starting correspondence is vital for accurate expansion. In scenes with complex backgrounds, this structural constraint effectively filters out stochastic semantic matches. Conversely, the benefit of neighborhood support is attenuated in areas of high semantic homogeneity (e.g., a subject's face or uniform clothing). In such regions, the spatial support confirms the integrity of the object class but may not uniquely resolve fine-grained instance identity, reflecting a fundamental characteristic of ViT-based representations where semantic "concepts" dominate over unique biometric signatures.

Qualitative Analysis and Architectural Implications Figure 5 illustrates the retrieval performance, where green overlays denote the final grown region $\mathcal{G}^*$. Under identical hyperparameters ($\tau = 0.6$), a significant divergence in patch density is observed between the two backbones. I-JEPA generates more contiguous and expansive regions, capturing the subject's full silhouette, whereas SigLIP2 produces sparser patterns concentrated on highly discriminative features.

This behavior is rooted in the models' respective pre-training objectives. I-JEPA's masked image modeling objective prioritizes local geometric continuity, resulting in a "smooth" latent manifold where neighboring patches maintain high

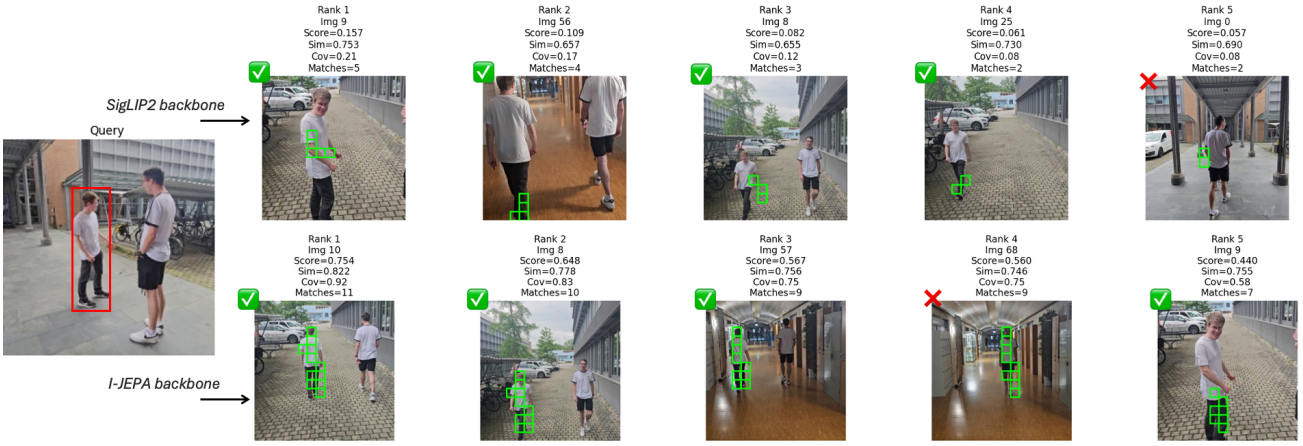

Figure 5. Qualitative comparison of localized retrieval results between SigLIP2 and I-JEPA. The green overlays represent the spatially-consistent regions ($\mathcal{G}^*$) grown from initial seeds.

mutual similarity. In contrast, SigLIP2's contrastive objective (image-text alignment) emphasizes high-level semantic distinctiveness, leading to a sharper similarity distribution. Minor variations in perspective cause SigLIP2 scores to drop below the threshold more rapidly than I-JEPA. Consequently, SigLIP2 excels at candidate pruning, while I-JEPA provides a more robust environment for dense, spatially-consistent local matching.

5. Conclusion and Future Work

In this paper, we introduced RIS, a hierarchical Region-to-Image Search framework designed for localized forensic retrieval. While standard image retrieval focuses on holistic similarity, our work highlights the significance of the ROI as a query primitive, a paradigm that remains relatively under-explored in current literature. By combining the semantic filtering of ViTs with a structural region-growing strategy, we demonstrated a consistent accuracy gain of at least 10% over standalone global methods. Our analysis further identified a fundamental trade-off: contrastive models like SigLIP2 offer superior global discriminative power, while predictive models such as I-JEPA provide superior local geometric continuity for spatial verification.

Future iterations will expand the dataset for rigorous quantitative evaluation, including metrics such as mAP and Rank-K accuracy. Furthermore, we will automate ROI selection by integrating YOLO (Wang et al., 2024) as an anchor generator. By indexing mean-pooled embeddings of detected bounding boxes, we aim to accelerate retrieval and refine initial candidate precision. Additionally, exploring adaptive similarity thresholds will be vital for maintaining robustness in high-occlusion forensic scenarios. Finally, we plan to incorporate the Segment Anything Model (SAM) (Kirillov et al., 2023) to isolate foreground patches, refining query purity and increasing the robustness of the spatial verification process.

Acknowledgements

This work was supported by Innosuisse under project 100.650 IP-ICT (VideoCognition), in partnership with iCoSyS and Morphean SA. We thank the entire team for their technical expertise and dedication.

Limitations

Despite the performance gains, certain limitations of the current RIS framework should be noted:

- **Dataset Scale:** The preliminary 90-frame dataset lacks the manual annotations required for a full statistical benchmarking (e.g., mAP), though qualitative results remain promising.
- **Scale and Quantization:** The 14×14 pixel token resolution limits the representation of fine-grained features. Additionally, a subject appearing significantly smaller in a target frame than the query ROI creates a patch-count disparity that hinders the region-growing connectivity score.
- **Operational Complexity:** Manual ROI selection limits autonomy, while the computational cost of iterative structural re-ranking introduces latency. Consequently, the framework is restricted to a top-N candidate set, making final accuracy dependent on the recall of the initial global stage.
- **Background Interference:** Rectangular ROIs often capture "distractor" patches (Fig. 2) that introduce

noise during global filtering and seed selection, potentially biasing matches toward environmental context rather than the subject.

Ethical Statement

The authors state that all third-party models, specifically SigLIP 2 and I-JEPA, were utilized in accordance with their respective open-source licenses for research purposes. The in-house dataset used for evaluation did not require human annotators, as it was utilized primarily for inference-based qualitative assessment. All computational experiments were limited to inference tasks and conducted on a single workstation; considering the small-scale test set and the absence of a training phase, the environmental impact and CO2 emissions associated with this work were negligible. This research is intended to advance localized retrieval techniques for forensic and search applications and was conducted with a commitment to transparency and ethical data use.

References

Aiger, D., Cao, B., Chen, K., and Araujo, A. Global-to-local or local-to-global? enhancing image retrieval with efficient local search and effective global re-ranking, 2025. URL https://arxiv.org/abs/2509.04351.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Azzopardi, G. et al. A comparative study of cfs, lbp, hog, sift, surf, and brief for security and face recognition. *ResearchGate Publication*, 2021.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, A., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library. 2024.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, C.-Y., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Phan, L., Nguyen, H. T. H., Warrier, H., and Gupta, Y. Patch embedding as local features: Unifying deep local and global features via vision transformer for image retrieval. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 2527–2544, December 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., and Ding, G. Yolov10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Wei, C., Fan, H., Xie, S., Wu, C.-Y., Yuille, A., and Feichtenhofer, C. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14668–14678, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

A. Appendix

Global Retrieval Results Figure 6 and Figure 7 depict global retrieval results for a sample query using the SigLIP2 and I-JEPA backbones, respectively. While both models achieve good semantic recall across varying perspectives, certain false positives are observed where the retrieved frames do not share the same environmental features as the query. These results underscore the limitations of relying solely on holistic embeddings and validate the requirement for a re-ranking stage.

Figure 6. Top-10 global retrieval results using SigLIP2

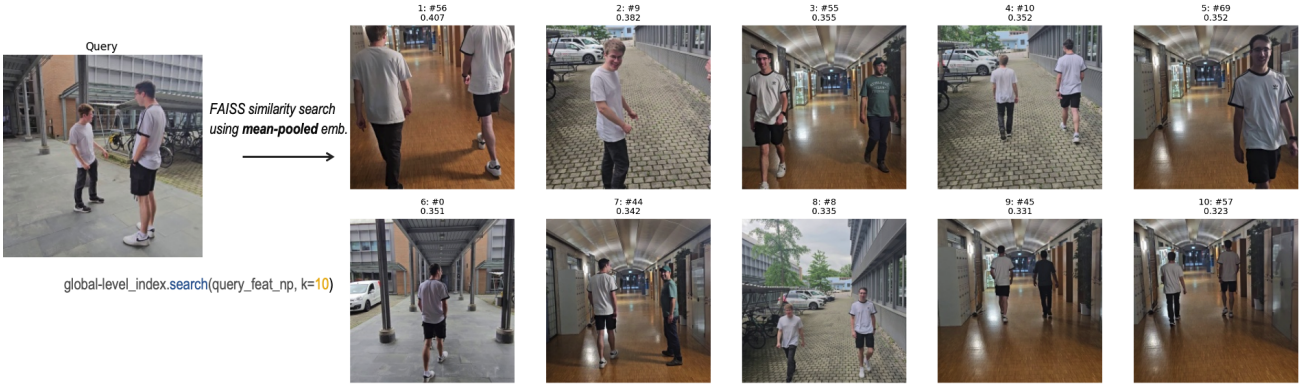

Figure 7. Top-10 global retrieval results using I-JEPA

Qualitative Analysis of Patch-Level Ambiguity Figure 8 highlights the performance limits of frequency-based patch voting, where typically only the top results maintain correct instance identity. It is worth noting that weighted distance approaches achieve higher discriminative precision than Majority Voting by employing inverse-distance weighting to penalize marginal matches and suppress the influence of ambiguous latent clusters. However, even with these refinements, high-similarity embeddings frequently correlate with unrelated subjects. This inherent semantic drift confirms that latent similarity alone is insufficient to uniquely resolve identity, necessitating the structural constraints provided by region growing.

Figure 8. Examples of independent patch-matching errors. Local query patches (white t-shirt) match high-similarity regions on unrelated subjects due to a lack of spatial constraints, necessitating the proposed region-growing verification.