Self-supervised Video Instance Segmentation Can Boost Geographic Entity Alignment in Historical Maps

Xue Xia
ETH Zurich

Randall Balestriero
Brown University

Tao Zhang
Wuhan University

Lorenz Hurni
ETH Zurich

Abstract

Tracking geographic entities from historical maps, such as buildings, offers valuable insights into cultural heritage, urbanization patterns, environmental changes, and various historical research endeavors. However, linking these entities across diverse maps remains a persistent challenge for researchers. Traditionally, this has been addressed through a two-step process: detecting entities within individual maps and then associating them via a heuristic-based post-processing step. In this paper, we propose a novel approach that combines segmentation and association of geographic entities in historical maps using video instance segmentation (VIS). This method significantly streamlines geographic entity alignment and enhances automation. However, acquiring high-quality, video-format training data for VIS models is prohibitively expensive, especially for historical maps that often contain hundreds or thousands of geographic entities. To mitigate this challenge, we explore self-supervised learning (SSL) techniques to enhance VIS performance on historical maps. We evaluate the performance of VIS models under different pretraining configurations and introduce a novel method for generating synthetic videos from unlabeled historical map images for pretraining. Our proposed self-supervised VIS method substantially reduces the need for manual annotation. Experimental results demonstrate the superiority of the proposed self-supervised VIS approach, achieving a 24.9% improvement in AP and a 0.23 increase in F1 score compared to the model trained from scratch.

1 Introduction

Historical maps serve as unique and invaluable archives documenting the evolution of geographic features across diverse time periods [10]. Numerous historical research endeavors necessitate synthesizing information derived from multiple historical maps, such as identifying the construction year of buildings, analyzing the development of road networks and settlements, supporting ecological and hydrological preservation and restoration efforts, and exploring the evolution of toponyms [6]. Identifying and linking corresponding geographic entities from multiple historical maps enhances the structure and accessibility of historical map data [3]. In the cartographic domain, this issue is referred to as geographic entity alignment: when presented with a collection of historical maps $\{\mathrm{map}_1, \mathrm{map}_2, \ldots, \mathrm{map}_T\}$ illustrating the same geospatial location across various years, the goal is to identify matching entities across these maps that correspond to the same real-world features [8].

The conventional approach to geographic entity alignment involves two key steps (Figure 1(a)): 1. Extracting vector entities from scanned historical map images, typically achieved through image instance segmentation followed by vectorization; 2. Aligning corresponding entities from different maps using heuristic algorithms, including spatial distances, topological relations, and approximate
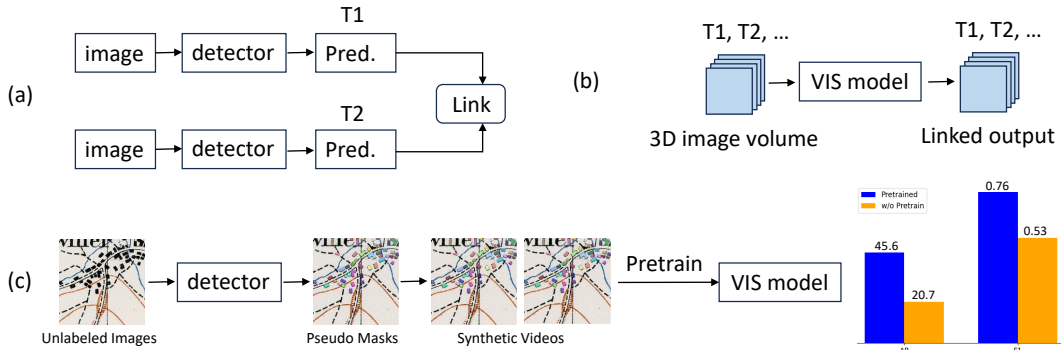
Figure 1: (a) Conventional approach for geographic entity alignment; (b) Proposed VIS method; (c) Proposed pretraining strategy for the VIS model.

topological relations [8]. The main limitation of the conventional approach is its low automation and reliance on handcrafted thresholds, compounded by varying distortions in historical maps across locations and time periods.

To overcome these challenges, we propose a novel approach that integrates the segmentation and linking steps through the use of VIS, generating a 3D volume of linked entities directly (Figure 1(b)). The benefits of using the VIS method for linking geographic entities in historical maps are threefold: it enhances generalizability by eliminating the need for heuristic linking; it naturally tracks slight displacements caused by map distortions; and it significantly improves automation over conventional multistep methods.

Due to the scarcity and high cost of quality training data for VIS-based methods, we further explore the integration of SSL into the proposed pipeline. Many SSL methods for videos rely on motion-based learning signals, such as optical flow [12, 11]. However, these methods are not well-suited for stationary or slow-moving objects—common characteristics in historical maps, making them inadequate for this context. Although models are often initialized with weights pretrained on video segmentation benchmarks like YouTubeVIS-2019 [13], the significant domain gap between historical maps and natural videos limits their effectiveness when transferred to out-of-distribution tasks. To address this challenge, we propose a pretraining strategy by generating synthetic videos from unlabeled historical map images (Figure 1(c)). Experimental results demonstrate that the proposed pretraining strategy boosts VIS model performance on the downstream task by 24.9% in AP and 0.23 in F1 score compared to training from scratch.

2   Self-supervised VIS for Geographic Entity Alignment

We present a self-supervised VIS approach for segmenting and linking geographic entities in historical maps. This method involves generating synthetic videos from unlabeled map images to pretrain the VIS model, which is then fine-tuned on downstream tasks. Further details are provided below.

Generating Synthetic Videos from Maps. The idea is to leverage large amounts of unlabeled historical map images to generate synthetic videos for pretraining. First, we use the small labeled dataset from the downstream task to train an image instance segmentation model, such as Mask R-CNN [5], to generate pseudo masks for the unlabeled images. We then create synthetic videos by duplicating the images and their pseudo masks to form two-frame sequences. Since the instance IDs are copied, objects remain naturally linked across the temporal dimension. Given that historical maps primarily feature static objects, this simple synthetic video approach effectively simulates real-world situations. Prior research shows that two-frame videos are sufficient to train a VIS model [9], so we limit our synthetic videos to two frames. This method enables us to generate nearly unlimited synthetic historical map videos for pretraining the VIS model without requiring additional manual annotations.

Video Instance Segmentation. Within the realm of VIS methods, two primary categories exist: the per-frame methods (a.k.a. online methods) and the per-clip methods (a.k.a. offline methods) [1]. Per-frame methods execute image segmentation independently on each frame and subsequently

2

Table 1: Mask2Former-VIS performance under different pretraining configurations

Pretraining Configurations	AP	AP50	AP75	Precision	Recall	F1
w/o Pretraining	20.7	40.5	21.1	0.47	0.60	0.53
COCO Images	**46.2**	**72.8**	**52.0**	**0.76**	**0.76**	**0.76**
YouTubeVIS-2019	44.8	69.8	49.8	0.74	**0.76**	0.75
ImageNet Synthetic Videos	32.0	56.5	33.5	0.64	0.68	0.66
Map Synthetic Videos	45.6	72.2	49.8	0.75	**0.76**	**0.76**

associate predicted instances across frames. In contrast, per-clip methods focus on modeling the spatio-temporal representations of individual video instances, enabling the direct prediction of a 3D mask for linked instances. In this paper, we adopt the per-clip VIS method, treating a historical map series $\{map_1, map_2, \ldots, map_T\}$ as a 3D spatio-temporal volume of dimension $T \times H \times W$, where T is the number of historical maps along the temporal axis, while H and W denote the height and width, respectively. To process the historical map sequence, we utilize Mask2Former-VIS [1], which outputs a 3D volume containing linked building instances across time, i.e. $\{e_1^i, e_2^i, \ldots, e_T^i\}$, where e_t^i is the binary mask of a building entity i in map_t. In Appendix A, we further explain the architecture of the VIS model.

Datasets and Evaluation Metrics. This study utilizes two datasets: an unlabeled image dataset for pretraining and a labeled video dataset for fine-tuning. Both are generated from the Swiss Siegfried map (©swisstopo), with a focus on building instances. Image tiles from both datasets have a fixed dimension of 256×256 pixels. The unlabeled dataset contains 5,832 images, which are fed into Mask R-CNN to generate pseudo masks for building instances. Images without buildings in the pseudo masks are excluded from the subsequent process of generating synthetic videos, resulting in 4,384 synthetic videos for pretraining the VIS model. The labeled video dataset includes 497 training, 63 validation, and 62 testing videos, with each video consisting of four frames, corresponding to four timestamps (year 1896, 1904, 1932, and 1945) from the Siegfried map. The VIS dataset follows the YouTube-VIS [13] format. We report Average Precision (AP) for videos, as used in YouTube-VIS [13], and F1 score as our evaluation metrics. Detailed definitions of these metrics are provided in Appendix B.

Implementation Details. *Supervised training of Mask R-CNN:* The model is initialized with pretrained weights from the COCO dataset [7]. Training runs for 400 epochs with a batch size of 8 and a learning rate of 1e-3, conducted on a single Nvidia Quadro RTX 5000 GPU. *Pretraining of Mask2Former-VIS:* We employ Mask2Former-VIS with a ResNet50 backbone, initializing the model using pretrained weights from the COCO dataset. The training process includes 4000 iterations with an initial learning rate of 1e-4, followed by a learning rate decay by a factor of 10 for an additional 2000 iterations. This training is executed on 2 Nvidia Titan RTX GPUs with a batch size of 16. *Fine-tuning on the downstream VIS task:* The Mask2Former-VIS model, pretrained on synthetic videos from the previous stage, is fine-tuned on the labeled video dataset. The fine-tuning follows the same configuration as the pretraining stage.

3 Experimental Results

To validate the effectiveness of the proposed self-supervised VIS method, we compare the performance of the Mask2Former-VIS model on the test set under different pretraining configurations. Table 1 presents the numerical results. The results show that pretraining consistently improves model performance. Even pretraining with synthetic videos, following the VideoCutLER method [9], which uses only unlabeled ImageNet data to generate synthetic videos as pretraining samples, outperforms training from scratch. However, pretraining with labeled natural videos from the YouTubeVIS-2019 dataset yields better results than using synthetic videos from ImageNet. The best performance is achieved by models pretrained on COCO images and synthetic historical map videos.

The success of synthetic historical map videos over other video datasets is likely due to their greater semantic similarity to the downstream task data. However, the performance is on par with models pretrained on the COCO image instance segmentation dataset. This can be explained by the fact that Mask2Former-VIS [1] is an extension of an image instance segmentation model Mask2Former
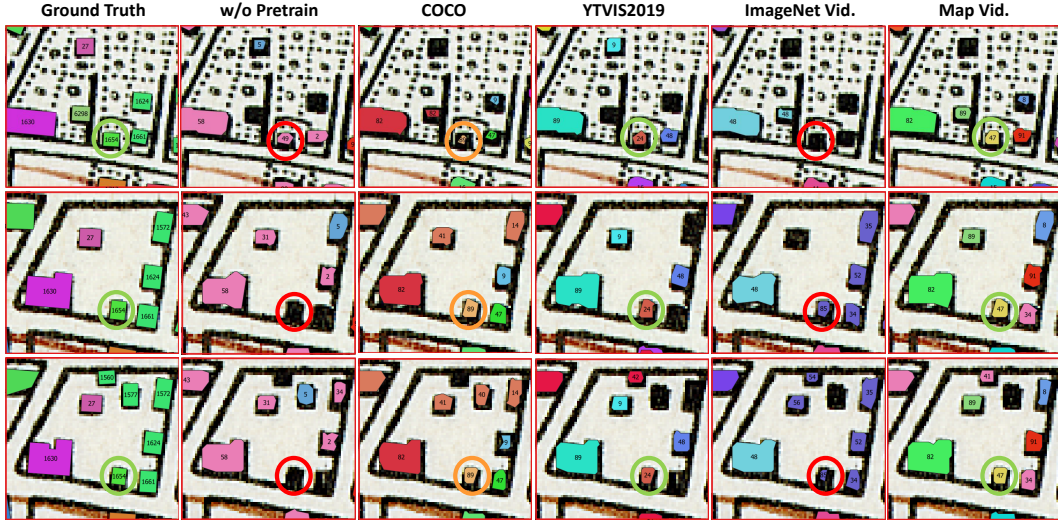
Figure 2: Segmentation and linking results for Mask2Former-VIS under different pretraining configurations. Results are shown for three frames corresponding to the years 1904, 1932, and 1945 (top to bottom).

[2]. For single-frame input, Mask2Former operates as a standard image segmentation architecture, and for multi-frame video data, it simply shares queries across frames, enabling object segmentation and tracking. Due to this architectural design, Mask2Former-VIS does not benefit substantially from video-based pretraining. However, we emphasize that the proposed pretraining strategy could still prove advantageous for other VIS models that require video-based pretraining when applied to out-of-distribution historical map data.

Figure 2 visualizes the segmentation and linking results. Models without pretraining or those pretrained on ImageNet synthetic videos or the YouTubeVIS-2019 dataset show more missing detections in their visual output. In contrast, models pretrained on the COCO dataset and synthetic historical map videos demonstrate better segmentation performance, with the latter showing slightly improved geometric accuracy (e.g., the instance marked inside the circle).

For instance tracking, all VIS models successfully track large instances or those without nearby confusing entities, regardless of pretraining configuration (e.g., the dark purple instance in the lower-left corner of the ground truth). However, in more complex scenarios—particularly when there are slight displacements of buildings (e.g., the light purple instance at the top of the ground truth)—the VIS models exhibit vulnerabilities, occasionally linking to the wrong entity. These scenarios can be challenging even for human annotators. Addressing this issue could be a key area for improvement, potentially by introducing slight displacements in the instances to simulate map distortions when generating synthetic historical map videos.

The numerical and visual results show that the proposed self-supervised VIS method offers an effective pipeline for automating geographic entity alignment. It addresses the inefficiencies associated with transferring pretrained weights to out-of-domain data, resulting in improved accuracy.

4   Conclusion and Outlook

This paper introduces a pioneering approach for segmenting and linking geographic entities in historical maps using self-supervised video instance segmentation. The proposed pretraining strategy generates large amounts of synthetic videos for use in the historical map domain, which differs significantly from natural videos and lacks sufficient labeled video data, without requiring additional manual annotations. It outperforms models pretrained on natural videos in linking building entities from Siegfried maps. Future work will focus on refining the strategy for generating synthetic historical map videos to better simulate real-world conditions, such as introducing slight displacements to mimic map distortions or merging smaller structures to reflect map generalization.

Acknowledgments and Disclosure of Funding

References

[1] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2Former for Video Instance Segmentation. *arXiv preprint arXiv:2112.10764*, pages 16–18, dec 2021.

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:1280–1289, 2022.

[3] Yao-Yi Chiang. Querying historical maps as a unified, structured, and linked spatiotemporal source. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, volume 03-06-Nove, pages 1–4, New York, NY, USA, nov 2015. ACM.

[4] Eliseo Clementini and Paolino Di Felice. Approximate topological relations. *International Journal of Approximate Reasoning*, 16(2):173–204, feb 1997.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2020.

[6] Magnus Heitzler and Lorenz Hurni. Unlocking the Geospatial Past with Deep Learning – Establishing a Hub for Historical Map Data in Switzerland. *Abstracts of the ICA*, 1:1–2, jul 2019.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8693 LNCS, pages 740–755. 2014.

[8] Kai Sun, Yingjie Hu, Jia Song, and Yunqiang Zhu. Aligning geographic entities from historical maps for building knowledge graphs. *International Journal of Geographical Information Science*, 35(10):2078–2107, 2021.

[9] Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, and Trevor Darrell. VideoCutLER: Surprisingly Simple Unsupervised Video Instance Segmentation. pages 22755–22764, aug 2023.

[10] Xue Xia, Chenjing Jiao, and Lorenz Hurni. Contrastive Pretraining for Railway Detection: Unveiling Historical Maps with Transformers. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 30–33, New York, NY, USA, nov 2023. ACM.

[11] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting Moving Objects via an Object-Centric Layered Representation. *Advances in Neural Information Processing Systems*, 35(NeurIPS), 2022.

[12] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised Video Object Segmentation by Motion Grouping. *Proceedings of the IEEE International Conference on Computer Vision*, pages 7157–7168, 2021.

[13] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:5187–5196, 2019.

A Video Segmentation Model

The architecture of Mask2Former-VIS [1] is presented in Figure 3. It contains three major components: a backbone feature extractor, a pixel decoder, and a Transformer decoder. The backbone feature extractor extracts low-resolution features, which are later upsampled by the pixel decoder to generate multi-scale, high-resolution features. The Transformer decoder operates on the multi-scale features to process object queries. A linear classifier is then applied to the object queries, resulting in class probability predictions. The mask predictions are derived through a dot product between the query and the high-resolution features. Given that all features possess a 3D spatio-temporal volume and the queries are shared across frames along the temporal axis, the architecture effectively segments and tracks object instances across time.
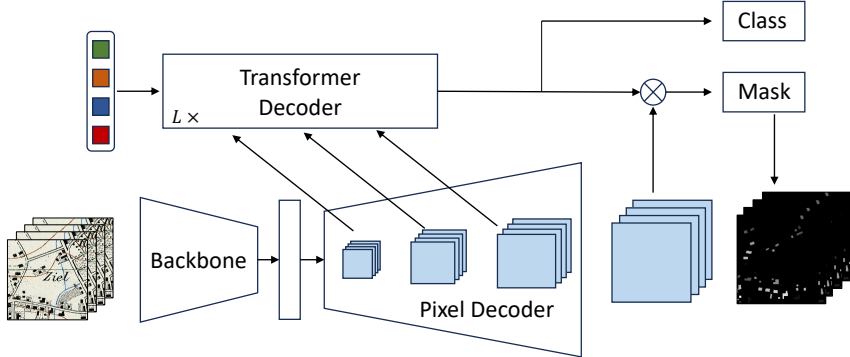

Figure 3: The simplified architecture of Mask2Former-VIS.

B Evaluation Metrics

Each linked instance $\{e_1^i, e_2^i, e_3^i, e_4^i\}$ — representing four years and incorporating empty masks for non-existent instances — is considered a true positive (TP) if its Intersection over Union (IoU) with a corresponding ground truth instance exceeds 0.5. Predicted instances without a match are categorized as false positives (FP), while ground truth instances that go undetected are labeled as false negatives (FN). Unlike image instance segmentation, video instances consist of a sequence of masks, so IoU is calculated both spatially and temporally by summing the intersections and unions across all frames:

$$\text{IoU}(i, j) = \frac{\sum_{t=1}^{4} \left| e_t^i \cap g_t^j \right|}{\sum_{t=1}^{4} \left| e_t^i \cup g_t^j \right|} \tag{1}$$

where g_t^j stands for ground truth instance. After identifying TP, FP, and FN, precision, recall, and F1-score can be computed accordingly. AP is evaluated at 10 IoU thresholds, ranging from 50% to 95% with a step of 5%.

C Traditional Approach

Table 2: Segmentation and linking results for Mask R-CNN

Image Instance Segmentation	AP	AP50	AP75
	62.6	87.5	76.0
Final Linking Results	P	R	F1
	0.49	0.70	0.58

Conventional geographic entity alignment approaches first extract vector entities from single map sheets separately and then link the extracted entities between two maps. For a collection of three or

more maps, the alignment is first performed between the maps of any two consecutive timestamps and then the results are combined [8]. Adhering to the two-step paradigm, we first use Mask R-CNN [5] to effectively segment building instances from historical maps. Subsequently, we leverage approximate topological relations [4] to establish links between these instances. More precisely, we employ the "approximately within" relation, defining it as applicable when the intersection area between two entities exceeds or equals 60% of the area of the smaller entity. The linked instances are then assigned a shared instance ID, ensuring uniqueness among different instances within a single image.

When exporting the Mask R-CNN output for building entity linking, the final result is fixed, and the confidence score is omitted. As a result, calculating the Average Precision (AP) for the final linked output of Mask R-CNN is not feasible. Therefore, we report only precision, recall, and F-score for the linked instances in Table 2. The results are visualized in Figure 4.

Figure 4: Segmentation and linking results for Mask R-CNN. Results are shown for three frames corresponding to the years 1904, 1932, and 1945 (left to right).

Based on the numerical results, the proposed self-supervised VIS method outperformed the conventional two-step approach, achieving a 0.18 higher F1-score. The diminished performance of the conventional method is primarily due to its heuristic-based linking procedure, as evidenced in the visual outputs. While it performs well in associating entities in neighboring frames (e.g., year 1932 and 1945), it often fails to link corresponding entities over long sequences.