

Mining Landmark Images for Scene Reconstruction from Weakly Annotated Video Collections

Helmut Neuschmied^[0000–0001–8153–6840] and Werner Bailer^[0000–0003–2442–4900]

JOANNEUM RESEARCH – DIGITAL, Graz, Austria

Abstract. Many XR productions require reconstructions of landmarks such as buildings or public spaces. Shooting content on demand is often not feasible, thus tapping into audiovisual archives for images and videos as input for reconstruction is a promising way. However, if annotated at all, videos in (broadcast) archives are annotated on item level, so that it is not known which frames contain the landmark of interest. We propose an approach to mine frames containing relevant content in order to train a fine-grained classifier that can then be applied to unlabeled data. To ensure the reproducibility of our results, we construct a weakly labelled video landmark dataset (WAVL) based on Google Landmarks v2. We show that our approach outperforms a state-of-the-art landmark recognition method in this weakly labeled input data setting on two large datasets.

Keywords: Landmark retrieval, image classification, few-shot learning, fine-grained classification, learning from weakly-labeled data

1 Introduction

The creation of XR applications depends on the availability of high-quality 3D models. For many XR use cases in culture and media, models of buildings or public spaces, commonly referred to as landmarks, are of great importance. They can be used as basis for augmenting them with information or educational content, or they can be used in virtual studies to enhance storytelling. Performing a high-quality capture process on demand is often not feasible due to cost reasons or time limitations (e.g., when producing for news). However, the archives of broadcasters and cultural institutions hold a vast amount of image and video data that is a valuable source for creating 3D models. Novel scene representation methods such as NERFs [11] also open new possibilities for creating such assets from 2D imagery.

Video data is a particularly interesting source for this purpose, as a moving camera shot may provide a number of views of the landmark of interest, taken at the same time under the same conditions. There are possibly also multiple shots from the same recording in a video. However, if the landmarks of interest are at all annotated, the annotation is weak, i.e., metadata exist on the level of a video

file, broadcast, or in the best case a story, but do not provide information about the exact temporal location of views of the landmarks in each frame. Despite their limitations, these weakly annotated videos offer a valuable resource, as they can be used to train models that are able to detect the same landmark in entirely unannotated content, replacing the labor-intensive and time-consuming process of manually labeling training samples. However, the success of deep learning models depends on the quality and size of annotated datasets used for training. The video thus contains “noise” from the point of view of this purpose, such as interviews, shots of anchorpersons, close-up interior views etc. In order to effectively use these data, the first task is to isolate the relevant ranges of frames showing the landmark of interest. Apart from very popular landmarks, the number of samples may still be small, thus requiring a method capable of learning from a small set of samples (few-shot learning).

In order to mine images for scene reconstruction, the problem to be solved can be defined as follows. Given is a collection of N videos $\mathcal{V} = \{v_1, \dots, v_N\}$, where each video v_k is composed of a set of M_k frames $v_k = \{f_1^k, \dots, f_{M_k}^k\}$, a set of P landmarks $\mathcal{L} = \{L_1, \dots, L_P\}$, and a set of landmarks contained in a video $A_k = \{L_p, L_q, \dots\}$ (with $|A_k| \ll P$, and likely $A_k = \{\}$ for some videos). The first step is to mine a training set $\mathcal{T}_p = \{f_j^k | L_p \in A_k \wedge \text{vis}(f, L_p) = 1\}$, where the function $\text{vis}(f, L_p)$ returns 1 if the landmark L_p is at least partly visible in frame f , and 0 otherwise. Then a landmark classifier can be trained from \mathcal{L} in order to annotate further videos and increase the set of images mined for reconstruction. By training our deep learning models on the training data mined from the weakly annotated video dataset, we aim to develop algorithms capable of recognizing landmarks with high precision, thus being able to collect an image set providing the basis for reliable scene reconstruction.

Although landmark recognition is a widely addressed research topic, there is a challenge of finding a dataset that is appropriate for our task setting. While there are a number of public landmark recognition datasets, they are image based with per image annotations (see discussion in Section 2.3). There are relevant weakly annotated video datasets [3], but those are not openly available.

The main contributions of this paper are as follows:

- We construct a novel weakly annotated video landmark dataset (WAVL), based on the widely used Google Landmarks v2 dataset [23] that is publicly available to the research community.
- We propose a pipeline for training data selection and training a fine-grained landmark classifier from a small set of training samples per class.
- In order to study the case of unannotated collections, we propose an alternative approach using data mined from web search to feed the same pipeline.
- We analyze the impact of weakly annotated video data on robustness in landmark recognition on two datasets, comparing against a state-of-the-art method.

The rest of this paper is organized as follows. In Section 2 we review related approaches and datasets. We describe the construction of our dataset in Section 3

and present the proposed method in Section 4. Section 5 provides evaluation results and Section 6 concludes the paper.

2 Related Work

2.1 Landmark Recognition

Earlier landmark recognition methods were based on the extraction of local image features, often represented as visual words. With the advent of deep learning, convolutional neural networks (CNNs) were introduced to extract features from images, enabling both landmark recognition and the use of similarity measures between pairs of images. Noh et al. [12] introduced DELF (DEep Local Features), a fusion of classical local feature methods with deep learning techniques. DELF leverages features from CNN layers and integrates an attention module to enhance recognition accuracy. Boiarov et al. [1] extended this by utilizing the center loss function to train CNNs, which penalizes the distances between image embedding and their corresponding class centers. In handling variations due to different viewpoints, they employed hierarchical clustering to compute centroids for each landmark, effectively managing the variability inherent in landmarks. Razali et al. [17] propose a lightweight landmark recognition model using a combination of Convolutional Neural Network (CNN) and Linear Discriminant Analysis (LDA) for feature size reduction. They compared different CNNs and showed that the EfficientNet architecture with a CNN classifier outperformed the other models evaluated. Yang et al. [24] train two different variants of the ResNet network architecture (ResNeSt269 and Res2Net200_vd) with an increasing image resolution (step by step) in order to improve the landmark recognition feature vectors obtained from a CNN. They normalize and merge the embedded layer descriptors of the two models above, doubling their size. As a post-processing step, the retrieval results can be significantly improved by re-ranking methods, e.g. by spatial verification [13] which is a method that checks the geometric consistency using local descriptors.

2.2 Learning from Weakly-labeled Data

The challenge of learning from weakly-labeled data has spurred innovative approaches to filter out potential temporal noise in annotations [19]. A different perspective is presented by Li et al. [8], who propose transforming potential noise in weakly-labeled videos into valuable supervision signals. This is achieved through the concept of sub-pseudo labels (SPL), where a new set of pseudo-labels is generated, expanding the original weak label space. This creative approach demonstrates a shift from noise filtering to converting noise into useful information, harnessing the power of weakly-labeled data for improved learning.

The rise of the internet and social media has simplified the acquisition of data relevant to specific classification tasks. In cases where supervision is incomplete and only a portion of training data carries labels, harnessing the abundance of

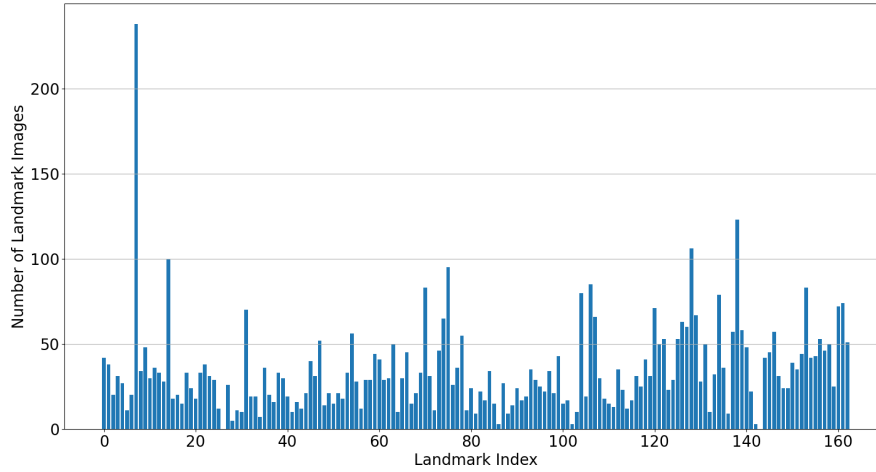
online data becomes crucial. The substantial information available from online sources contributes significantly to robust model development by augmenting labeled training data with additional instances from diverse contexts. But also images from web search engines like Google or DuckDuckGo tend to be biased toward images where a single object is centered with a clean background and a canonical viewpoint [10] and depending on the search term, there might also be pictures included that correspond to a completely different context than intended. Chen et al. [4] propose a two-stage CNN training approach for leveraging noisy web data. They initially employ simple images to train a baseline visual representation using a CNN. Subsequently, they adapt this representation to more challenging and realistic images by capitalizing on the inherent structure of the data and category.

2.3 Weakly Annotated Landmark Datasets

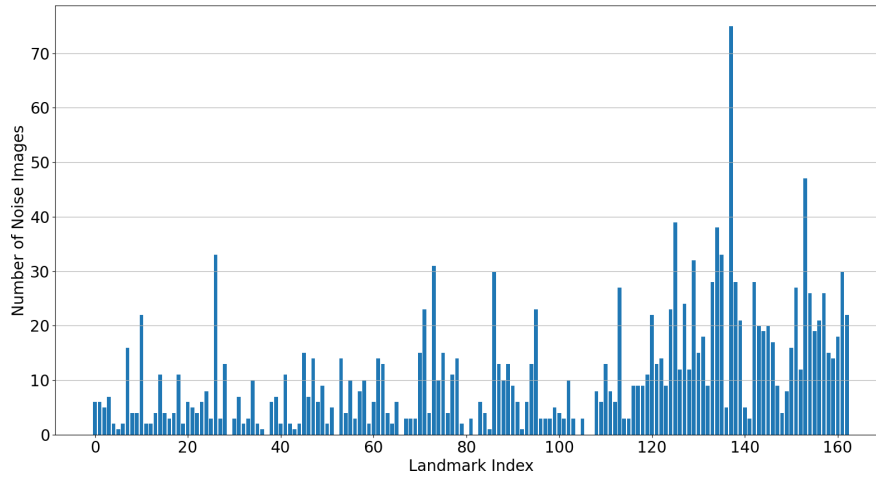
While there exist many datasets relevant to our task, most of them are fully labeled image datasets. Oxford [15] and Paris [16] buildings are early datasets used in landmark recognition and visual place recognition tasks. Over the years datasets grew in size, such as Pittsburg250k [22] and data diversity, e.g. including different captures times as in Tokyo 24/7 [21]. The Google Landmarks datasets v1 [12] and v2 [23] have become a widely adopted benchmark for this task. Datasets created for autonomous driving research such as BDD100k [26] provide location indexed videos and are thus sometimes used for visual geolocation. However, these datasets show landmarks always from a vehicle point of view, and in contrast to weakly annotated content from media archives the remaining content contains other street views.

The Italian public broadcaster RAI assembled a dataset with monuments of Italy [3]. The dataset contains about 2,000 clips depicting about 200 monuments from all regions of Italy, mainly acquired from RAI regional newscasts, collected for assessing similarity search in video. Annotations of the monument are provided on clip level. Each clip contains typically a news story, of which one or more shots contain an exterior view of the relevant monument, and in some cases also interior views. Some of the shots may show the monument occluded or in the background (e.g., as backdrop of an interview). In addition, the clips often contain other material of the story, e.g., the anchor in the studio introducing the topic (with an image that shows a view of the monument or something else), views of people in the street, close-up shots of people or interior items etc. As such, the dataset is typical for the type of content and the granularity of annotation to be found in a broadcast archive. The dataset is not available for public download but provided by RAI under a custom license agreement.

V3C (Vimeo Creative Commons Collection) [18] is a very large dataset (28,450 videos, about 3,800 hours) assembled for benchmarking video retrieval. We have considered amending the existing metadata with landmark annotations for a subset. However, a preliminary experiment found a too small number of clips (using landmark or city names as initial queries).



(a) Distribution of landmark images in the RAI-MI dataset.



(b) Distribution of noise images for each landmark in the RAI-MI dataset.

Fig. 1: Statistic information of the RAI-MI dataset.

The lack of a dataset that matches the characteristics of data and annotation granularity found in media archives and that is openly available led to the decision to construct such a dataset by combining commonly used datasets.

3 Weakly Annotated Dataset

We propose the Weakly Annotated Video Landmarks (WAVL) dataset. For evaluating landmark recognition, there is no need to use temporal correlations between successive video frames. This allowed us to simplify our dataset by focusing

Dataset	LM	LM	Noise	Mean	Std Dev.	Mean	Std Dev.
		Imgs	Imgs.	LM Imgs	LM Imgs	Noise Imgs.	Noise Imgs.
RAI-MI	163	5,620	1,871	34.55	26.92	10.63	10.60
WAVL	141	4,230	1,592	29.42	1.33	11.29	1.07

Table 1: Key figures for images per landmark from the RAI-MI and WAVL data sets (LM: landmarks, Imgs.: images, Std Dev.: standard deviation).

only on keyframes, which represent either a single frame per video sequence or frames extracted at defined temporal intervals. In order to mimic a keyframe dataset as extracted from archive video content, with similar characteristics as the RAI monuments of Italy (RAI-MI) dataset, we merged images from two different sources: the Google Landmarks v2 dataset [23] and the V3C [18] video dataset. These combined sources allowed us to create a dataset that contains sets of keyframes as they would be extracted from one video, containing both keyframes of a particular landmark as well as unrelated keyframes (noise). The video is annotated with the landmark visible in the subset of keyframes taken from Google Landmarks.

For each set of keyframes representing a video, we combined on average 30 associated images from the Google Landmarks v2 dataset with on average 11 keyframes (noise images) from one of the videos in the V3C1 dataset. This process has been done for 141 landmarks, resulting in a dataset of 5,770 images. The dataset has been made available at <https://github.com/XRecoEU/WAVL-Dataset>.

To perform a comparative evaluation on the RAI-MI dataset, we selected a subset of 163 different videos of landmarks representing buildings, and extract keyframes from them. The key figures for both datasets are shown in Table 1. The distributions of the landmark and noise images for the RAI-MI dataset vary more strongly, thus we plot them in Figure 1. For a number of cases the number of samples is in a range of 10 or less, so that the problem can at least partly be considered a few-shot learning problem.

4 Proposed Approach

We propose an approach to landmark recognition that treats the problem as a fine-grained image classification task. To realize this, we build on the Attentive Pairwise Interaction Network (API-Net) [27] with EfficientNet B3 [20] as the backbone architecture. However, the presence of weak annotations introduces a challenge as a substantial portion of training images contains incorrect labels. To address this issue, we have explored two different approaches.

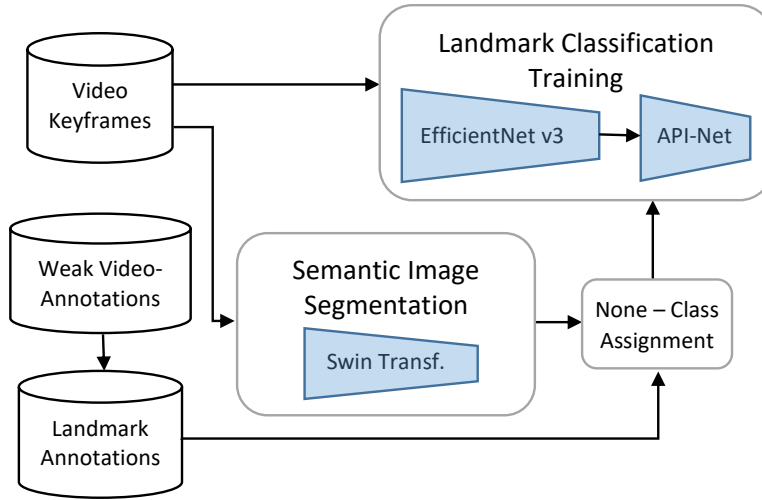


Fig. 2: Swin+API-Net: Training of the EfficientNet B3 network with API-Net from a filtered set of keyframes. The corresponding landmark class labels are changed to *None* class if no or only small building areas are detected in the images.

4.1 Semantic Segmentation Prefiltering

In the first approach (Swin+API-Net) (see Figure 2), given our focus on landmarks associated with buildings, we leverage a Swin transformer network for semantic image segmentation [9]. This process enables us to filter out images where structures like buildings or walls are observable, providing insights into the size of these areas. Additionally, we utilize information about the recognition of extensive regions where human figures are present, aiding in identifying anchorpersons or interior views with close-up shots, discarding these images as well. All images filtered in this process are put into a *None* class for the training process. This strategy of region-based filtering contributes to balancing the impact of keyframes unrelated to the landmark on the training process.

For inference, a simplified approach is adopted for landmark classification on images, using only the retrained EfficientNet backbone with the API-Net classification layer. This allows a straightforward evaluation of the performance of the trained model on individual test images, focusing solely on the ability of EfficientNet to recognise the trained landmarks.

4.2 Web Image Mining

In the second approach (CDVA+API-Net) (see Figure 3), we perform a targeted web search for relevant landmark training images using the landmark information embedded in the weak video annotations. To facilitate this, we employ

Dataset	Landmarks	Downloaded images	Rejected images	Landmarks w. ≥ 1 img.	Landmarks w. ≥ 10 img.
RAI-MI	163	6,130	1,177	162	153
WAVL	141	5,585	166	140	139

Table 2: Images obtained using the web image mining process for the RAI-MI and the WAVL datasets.

the DuckDuckGo¹ image search engine, retrieving a set of 40 images for each landmark term. Recognizing that web data can introduce considerable noise, our selection process is refined. We only retain images that show similarity to keyframes extracted from the weakly annotated videos. This similarity check is done by computing CDVA [5] descriptors from the images. In particular, we use the learned component of the descriptor, which is binarised to obtain 512 bit binary vector that can be efficiently matching using Hamming distance.

To filter the web images, we compare them to the keyframes and keep only those that have a similarity score above a certain threshold. We use a threshold of 0.6 to select web images in the first step. However, to account for the variability of web images, in a subsequent step we compare the remaining web images with the previously selected ones. If the similarity score exceeds a slightly lower threshold of 0.58, these previously rejected images are also used. These threshold values were determined experimentally with the help of a visualization of the filter results. Table 2 gives an overview of the web images obtained for the RAI-MI and the WAVL landmarks. Following the filtering of web images, we proceed to train an EfficientNet B3 using the API-Net framework, mirroring the process employed in our first approach. This strategy capitalizes on the wealth of online resources while maintaining a rigorous validation process to ensure the quality and relevance of the acquired images. Once the training has been completed, the images are again evaluated using only the EfficientNet B3 backbone and the classification layer of API-Net.

5 Evaluation

We evaluate the two proposed training approaches on both the RAI-MI and WAVL datasets. In the case of the WAVL dataset, we also compare our approach with a state of the art landmark recognition method in order to establish a baseline for the dataset. Given the composition of the WAVL dataset, which uses Google Landmarks v2 images, we chose to compare our approaches with a well-performing method on this benchmark dataset.

In this context, we considered the solution developed by the smlyaka team [25], for which the source code is publicly available². It has demonstrated exceptional

¹ <https://duckduckgo.com/>

² <https://github.com/lyakaap/Landmark2019-1st-and-3rd-Place-Solution>

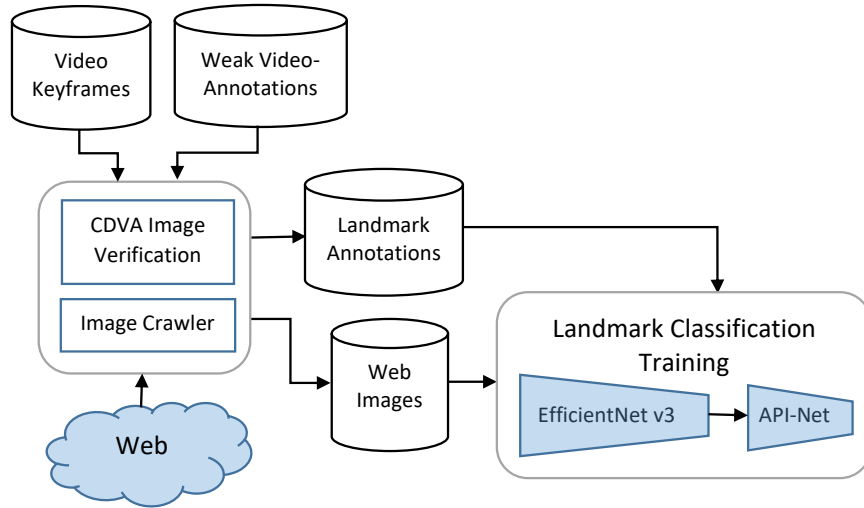


Fig. 3: CDVA+API-Net: Training of the EfficientNet B3 network with API-Net by using images which are crawled from the web. Only those web images are used which have sufficient similarity (based on CDVA descriptors) with one of the video keyframes.

performance, winning the first place in the Google Landmark Retrieval 2019 Challenge and the third place in the Google Landmark Recognition 2019 Challenge on Kaggle. This approach involves the initial pre-training of a ResNet-101 backbone on ImageNet and the Google Landmark Dataset v1 (GLD-v1) [12] training dataset. This pre-trained model has not been made available. We thus start from a model pre-trained on ImageNet. To compensate for pre-training on landmarks, we extended the number of training epochs on the WAVL dataset to 14 (instead of 5 as reported in [25]). The authors propose an automated data cleaning process to remove wrong annotations. The cleaning process involves a three-step approach that uses spatial verification to filter images by k-NN search. The authors use RANSAC [6] with affine transformation and deep local attentive features (DELF) [12] for spatial verification. If the count of verified images reaches a certain threshold, the image is added to the cleaned dataset. We use this approach once without (labelled smlyaka) and once with the data cleaning process (labelled smlyaka with data cleaning).

In evaluating our approaches, we employed a comprehensive suite of evaluation metrics. In addition to precision and recall we use balanced accuracy (BA), the mean of true positive and true negative rate [2] and symmetric balanced accuracy (SBA) [7], which aims to eliminate bias by the choice of the positive class. These metrics are related to a specific recognition threshold. In line with the Google Landmarks benchmark [23], we also use Global Average Precision (GAP, originally proposed as micro-AP in [14]) as a metric. GAP differs from the more commonly used mean average precision (MAP) in that mean of precision

Method	GAP	Threshold	Precision	Recall	BA	SBA
API-Net	54.67	7.5	86.80	41.53	63.67	63.41
Swin+API-Net	56.09	7.5	84.88	43.54	63.39	63.05
CDVA+API-Net	44.50	7.5	88.26	27.43	59.69	61.02

(a) Results on the RAI-MI dataset (best score in bold).

Method	GAP	Threshold	Precision	Recall	BA	SBA
smlyaka [25]	47.34	2.0	53.56	66.31	54.51	54.64
smlyaka with data cleaning. [25]	47.59	2.5	54.08	57.22	54.31	54.32
API-Net	37.43	9.0	56.27	32.22	53.59	53.99
Swin+API-Net	53.86	6.5	69.63	62.06	68.76	68.96
CDVA+API-Net	53.03	6.0	79.77	52.57	70.61	72.43

(b) Results on the WAVL dataset (best score in bold).

Table 3: Evaluation results for the RAI-MI and WAVL dataset.

at rank for all relevant returned results is determined, not taking the number of ground truth positives into account. Evaluation data are selected from videos which are not used for training data.

Table 3 lists the results of the two datasets. The values for precision, recall, BA and SBA are given for a specific threshold, which is determined with respect to a maximum value of BA. The variation of these metrics as a function of the threshold value can be seen in Figures 4 and 5 for methods Swin+API-Net and smlyaka with data cleaning, respectively.

Filtering out training images using image region classification (Swin+API-Net) has very different effects on the two datasets. In the RAI-MI dataset, the filtered images sometimes contain information that adds value to landmark recognition. These could be, for example, interior views of buildings or studio backgrounds where a newsreader is visible. On the other hand, selecting images by image region classification from the WAVL dataset significantly improves recognition results. In any case, filtering out images reduces the risk of learning unrelated information.

Using web images for training does not provide good results for the RAI-MI dataset. As can be seen in Table 2, one reason is that the web search for this dataset did not yield a sufficient number of usable images for all landmarks. For three of the 163 landmarks, the annotation is incorrect, which means that no landmark related images could be found on the web. Nevertheless, this approach achieves high precision values. We believe that this is due to the similarity constraint in the selection of web images, which will reduce the risk of false positives but also harm the diversity of samples. Another aspect is that in both datasets there are a relatively large number of images for each landmark. If this were not the case, the web-based approach would have the advantage that sufficient images for such landmarks could still be found on the web for training.

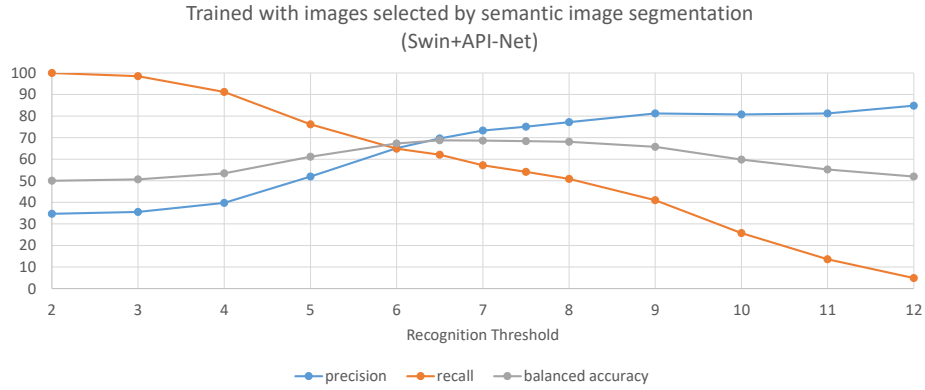


Fig. 4: Threshold dependent results on the WAVL dataset for the approach Swin+API-Net.

The comparison with the winning method of the Google Landmark Retrieval 2019 Challenge (smlyaka [25]) shows that while it performs best in terms of recall, it is otherwise outperformed by one of the proposed approaches on the WAVL dataset. It is particularly interesting that the data cleaning proposed in this approach does not provide any improvement under these conditions. This is because the frequency of similar images is not a good criterion for selecting training images in this use case.

For the purpose of mining images for 3D reconstruction the fact that the proposed approaches have higher precision than other methods is beneficial, as it improves the robustness of the reconstruction process, in particular for methods such as NERF, that do not include a feature matching and filtering step.

6 Conclusion

XR content creators need to mine existing archive content for landmark reconstruction. In order to address this problem, we have proposed two approaches for training landmark classifiers on weakly annotated video data. The training pipeline assumes weakly labelled input data, which represents the real world scenario of content available from audiovisual media archives. For this purpose we published the Weakly Annotated Video Landmarks (WAVL) dataset, constructed from the Google Landmarks v2 dataset and V3C1 as noise content. We show that the two proposed methods (either using semantic segmentation for pre-selection or mining web images) perform well, outperforming a state-of-the-art approach proposed for the Google Landmarks V2 dataset on our noisy version of it. A possible direction for future work is the integration of a re-ranking method as a post-processing step. We would also like to investigate approaches for incremental training of new landmarks.

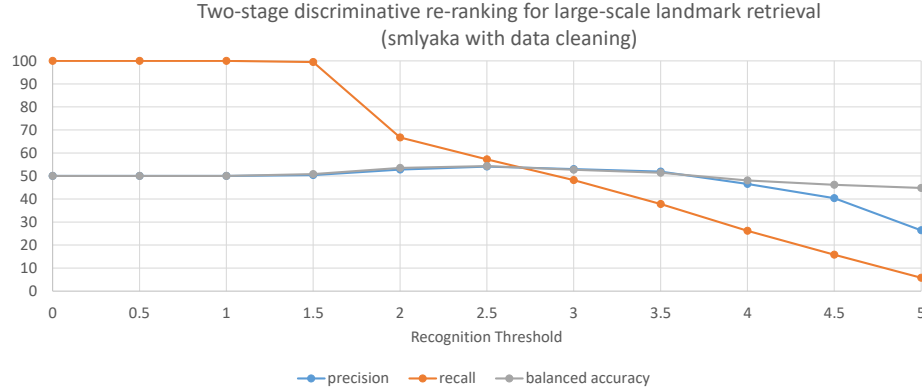


Fig. 5: Threshold dependent results on the WAVL dataset for the approach smlyaka with data cleaning.

Acknowledgements The authors would like to thank Stefanie Onsoni-Wechtitsch for providing the segmenter implementation.

The research leading to these results has been funded partially by the European Union’s Horizon 2020 research and innovation programme, under grant agreement n° 951911 AI4Media (<https://ai4media.eu>), and under Horizon Europe under grant agreement n° 101070250 XRECO (<https://xreco.eu/>).

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Boiarov, A., Tyantov, E.: Large scale landmark recognition via deep metric learning. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 169–178 (2019)
2. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition. pp. 3121–3124. IEEE (2010)
3. Caimotti, E., Montagnuolo, M., Messina, A.: An efficient visual search engine for cultural broadcast archives. In: AI* CH@ AI* IA. pp. 1–8 (2017)
4. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
5. Duan, L.Y., Lou, Y., Bai, Y., Huang, T., Gao, W., Chandrasekhar, V., Lin, J., Wang, S., Kot, A.C.: Compact descriptors for video analysis: The emerging mpeg standard. IEEE MultiMedia **26**(2), 44–54 (2019). <https://doi.org/10.1109/MMUL.2018.2873844>

6. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**(6), 381–395 (jun 1981). <https://doi.org/10.1145/358669.358692>, <https://doi.org/10.1145/358669.358692>
7. Gösgens, M., Zhiyanov, A., Tikhonov, A., Prokhorenkova, L.: Good classification measures and how to find them. *Advances in Neural Information Processing Systems* **34**, 17136–17147 (2021)
8. Li, K., Zhang, Z., Wu, G., Xiong, X., Lee, C., Lu, Z., Fu, Y., Pfister, T.: Learning from weakly-labeled web videos via exploring sub-concepts. *CoRR* **abs/2101.03713** (2021), <https://arxiv.org/abs/2101.03713>
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 10012–10022 (2021)
10. Mezuman, E., Weiss, Y.: Learning about canonical views from internet image collections. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*. vol. 25. Curran Associates, Inc. (2012)
11. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: *European Conference on Computer Vision*. pp. 405–421. Springer (2020)
12. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: *Proceedings of the IEEE international conference on computer vision*. pp. 3456–3465 (2017)
13. Perd’och, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9–16 (2009). <https://doi.org/10.1109/CVPR.2009.5206529>
14. Perronnin, F., Liu, Y., Renders, J.M.: A family of contextual measures of similarity between distributions with application to image retrieval. In: *2009 IEEE Conference on computer vision and pattern recognition*. pp. 2358–2365. IEEE (2009)
15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: *2007 IEEE conference on computer vision and pattern recognition*. pp. 1–8. IEEE (2007)
16. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: *2008 IEEE conference on computer vision and pattern recognition*. pp. 1–8. IEEE (2008)
17. Razali, M.N.B., Tony, E.O.N., Ibrahim, A.A.A., Hanapi, R., Iswandono, Z.: Landmark recognition model for smart tourism using lightweight deep learning and linear discriminant analysis. *International Journal of Advanced Computer Science and Applications* (2023), <https://api.semanticscholar.org/CorpusID:257386803>
18. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3c—a research video collection. In: *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I* 25. pp. 349–360. Springer (2019)
19. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022)
20. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International conference on machine learning*. pp. 6105–6114. PMLR (2019)
21. Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: *CVPR* (2015)

22. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 37, pp. 883–890 (06 2013). <https://doi.org/10.1109/CVPR.2013.119>
23. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2575–2584 (2020)
24. Yang, M., Cui, C., Xue, X., Ren, H., Wei, K.: 2nd place solution to google landmark retrieval 2020 (2022)
25. Yokoo, S., Ozaki, K., Simo-Serra, E., Iizuka, S.: Two-stage discriminative re-ranking for large-scale landmark retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1012–1013 (2020)
26. Yu, F.: Bdd100k: A large-scale diverse driving video database. BAIR (Berkeley Artificial Intelligence Research).[Online]. Available: <https://bair.berkeley.edu/blog/2018/05/30/bdd> (2018)
27. Zhuang, P., Wang, Y., Qiao, Y.: Learning attentive pairwise interaction for fine-grained classification. CoRR **abs/2002.10191** (2020), <https://arxiv.org/abs/2002.10191>