Combining global and local vision foundation models for explainable tattoo matching

Sabina B. van Rooij[a], Michalis Lazaridis[b], Stefanos Demertzis[b], Henri Bouma[a], and Petros Daras[b]

[a]TNO, The Hague, The Netherlands
[b]CERTH-ITI, Thessaloniki, Greece

ABSTRACT

Tattoo matching is important for criminal investigations. Recently, vision foundation models have shown increased performance for tasks like image classification and image retrieval. Global vision foundation models (e.g., CLIP) or local approaches (e.g., OmniGlue) show increased performance for image retrieval. In this paper, we show the added value of combining global and local approaches for explainable tattoo matching. We also investigate the use of a sketchification approach, facilitating the matching process for more abstract tattoos. We finally highlight the potential of local OmniGlue as a more explainable alternative to global image-based matching methods like CLIP.

Keywords: Tattoo matching, image retrieval, foundation models

1. INTRODUCTION

In the context of criminal investigations, the ability to efficiently identify matching tattoos within large databases can provide valuable investigative leads. To accelerate this process, AI-based tools can help by automatically retrieving visually similar tattoos. This reduces the manual effort required and increases the speed and scalability of the search.

Recent advances in vision foundation models offer new opportunities for improving tattoo retrieval. These models can operate at global level (e.g. CLIP), capturing overall image semantics, or at a local level, focusing on finer-grained details such as keypoints (e.g. OmniGlue). In this work, we explore how combining global and local vision foundation models can lead to better performance in the tattoo matching process. Preprocessing steps may also be required in special cases, for example when dealing with textual or abstract tattoos.

However, as with many AI tools, it is often unclear how the model arrives at its conclusions, i.e., which elements of the images have contributed to the match. Therefore, we explore ways to make this process more 'explainable' by gaining insight into how the matching tattoos were found. Additionally, we observed that when using image feature matching methods, it was challenging to define a robust similarity threshold to determine whether a retrieved match was correct.

Our findings show that the proposed position-guided method is able to provide a matching score that is more aligned with the performance of the matcher. The keypoint-based matching approach enhances the explainability of the retrieval process by explicitly indicating which regions of the images (i.e., matched keypoints) contribute to the final similarity decision. This may increase transparency and user trust in the system. Furthermore, we show that the method improves the robustness against partial occlusions. Finally, we observe that global vision foundation models can struggle to recognize clear concepts within the image due to variation in skin, shadow, body curvature, and background noise. To address this, we propose a sketchify preprocessing step, transforming the tattoo images into sketch-like, color-preserving representations optimized for deep feature extraction. We show that this sketchify approach has added value to concentrate on the ink of the actual tattoo.

E-mail: sabina.vanrooij@tno.nl, henri.bouma@tno.nl, lazar@iti.gr, demertzis@iti.gr, daras@iti.gr

S.B. van Rooij, M. Lazaridis, S. Demertzis, H. Bouma, P. Daras, Combining global and local vision foundation models for explainable tattoo matching,
Proc. SPIE, vol. 13679, (2025). https://doi.org/10.1117/12.3069997

2. RELATED WORK

The goal of content-based image retrieval (CBIR) is to extract and compare the visual content of images using low-level features such as color, texture, and shape, without relying on textual metadata. In the context of tattoo matching, CBIR methods aim to find visually similar tattoos to a given query image.[1]

Early image matching approaches primarily relied on local hand-crafted features, most notably Scale-Invariant Feature Transform (SIFT),[2] which proved effective at identifying keypoints invariant to scale and rotation. To improve retrieval efficiency, global representations such as the bag-of-words (BoW) model were later adopted,[3] which represent images by fixed-length vectors that count the occurrence of visual words. While this quantization introduces loss of accuracy, performance can be enhanced through techniques such as Hamming Embedding and Weak Geometric Consistency.[4] These improvements have been validated on large-scale tattoo databases, including over 300,000 images. This dataset is not publicly available. Some approaches also combine local and global descriptors to improve robustness for tattoo matching. For example, Kim et al.[5] integrates SIFT, shape context histograms, global shape histograms, and 2D Fourier Transforms to enhance robustness against translation, scale, rotation, and partial shape distortions. It was also shown that non-tattoo regions in the input image can negatively impact feature matching using local descriptors like SIFT and SURF,[6] highlighting the importance of accurate tattoo segmentation or region localization as a pre-processing step.

With the success of deep learning, CNN-based approaches emerged as a more powerful alternative than the early feature-matching approaches. For example, Di and Patel[7] proposed a Siamese CNN using AlexNet as a backbone for learning deep tattoo embeddings. They employed both triplet loss and contrastive loss to improve the discrimination between similar and dissimilar tattoos. These models were evaluated on the Tatt-C dataset,[8] which is currently not publicly available anymore. A comprehensive overview of tattoo matching approaches is provided by Han et al.,[9] covering developments in feature engineering, deep learning architectures, and sketch-based retrieval. The use of sketch-based image retrieval has also been surveyed more broadly by Pavithra and Sharath.[10]

Recently, even newer approaches such as CLIP and OmniGlue appear to outperform traditional CNNs on image matching tasks. CLIP[11] is an example of a vision-language foundation model that shows impressive zero-shot performance on many tasks and datasets. OmniGlue[12] enhances descriptor accuracy via a keypoint position-guided attention mechanism, leveraging the broad knowledge encoded in vision foundation models. OmniGlue demonstrates significant performance improvements over traditional methods like SIFT. In this work we combine these two methods to build a more robust tattoo matcher.

3. METHOD

The tattoo-matching method handles cropped tattoo images obtained through prior detection. The cropped tattoo images are used as query image to find the most similar cropped tattoo images in a database. In this section, two approaches for tattoo matching are described: the CLIP and OmniGlue method (Sec. 3.1) and the Sketchify method (Sec. 3.2).

3.1 CLIP and OmniGlue method

This method combines global and local vision foundation models for explainable tattoo matching. The method consists of three components: CLIP, OmniGlue and a combiner. The CLIP component generates global image-based features and CLIP-based matching score is based on the cosine similarity. The OmniGlue component generates local keypoints with descriptors. This makes the matching process more explainable by highlighting which parts (i.e. keypoints) of the images contribute to a match. The OmniGlue matching score is computed as the sum of confidences for all found keypoint matches. The keypoint-based approach increases transparency and user trust in the system. The combiner component is based on the assumptions that CLIP is a good baseline and that OmniGlue can have added value when there are many high-confidence matches. Therefore, for a high OmniGlue matching score, the decision is based on the OmniGlue prediction and for a low OmniGlue matching score, the decision is based on the CLIP prediction.

3.2 Sketchification method

Extracting deep features out of a real-world image is not always an easy task. Especially in the case of textual tattoos or abstract tattoos, global concept-based approaches, like CLIP, struggle to recognize clear concepts within the image. To enhance the performance of CLIP in such cases, we need to remove skin, shadows, body curvature and background noise, and concentrate on the actual tattoo ink. We introduce a multi-stage image processing pipeline designed to transform tattoo photographs into sketch-like, color-preserving representations optimized for deep feature extraction. The pipeline begins by detecting tattoo regions using a pre-trained RT-DETR (Real-Time Detection Transformer) model. To ensure consistent scale normalization, the image is resized such that the largest detected tattoo bounding box is scaled to a fixed size of 270 pixels. Object detection is then repeated on the resized image, yielding refined bounding boxes filtered by a confidence threshold of 0.7 (see Figure 1a).

Figure 1: (a) Tattoo detection and scale normalization. The input image is processed using the RT-DETR model to detect tattoo regions. Bounding boxes are drawn around detected areas, and the image is resized so that the largest detected region is scaled to 270 pixels on its longest side. (b) Sketch transformation via shadow removal. The resized image undergoes grayscale conversion and Gaussian blur-based background estimation. Division of the original image by the blurred background results in a shadow-free, high-contrast sketch representation that emphasizes edge structures.

Next, a sketch transformation is applied to the resized image using a shadow-removal process based on Gaussian blurring. Specifically, the image is converted to grayscale and an inverted Gaussian blur (kernel size 101×101) is applied to estimate the background. The original grayscale image is then divided by the blurred background with a scaling factor of 256. Let I be the input image after resizing, and let G be the grayscale version of I. We denote: $B = GaussianBlur(255 - G,\ \sigma)$, where σ corresponds to a kernel size of 101×101. Thus, the shadow-free sketch S is computed as:

$$S\,(x,\ y) = \frac{G(x,\ y)}{B(x,\ y)} \ \bullet \ 256 \tag{1}$$

where G(x, y) is the grayscale intensity at pixel (x, y), and B(x, y) is the corresponding value in the blurred inverted background. This operation produces a high-contrast pencil sketch-like result while suppressing shadows (see Figure 1b).

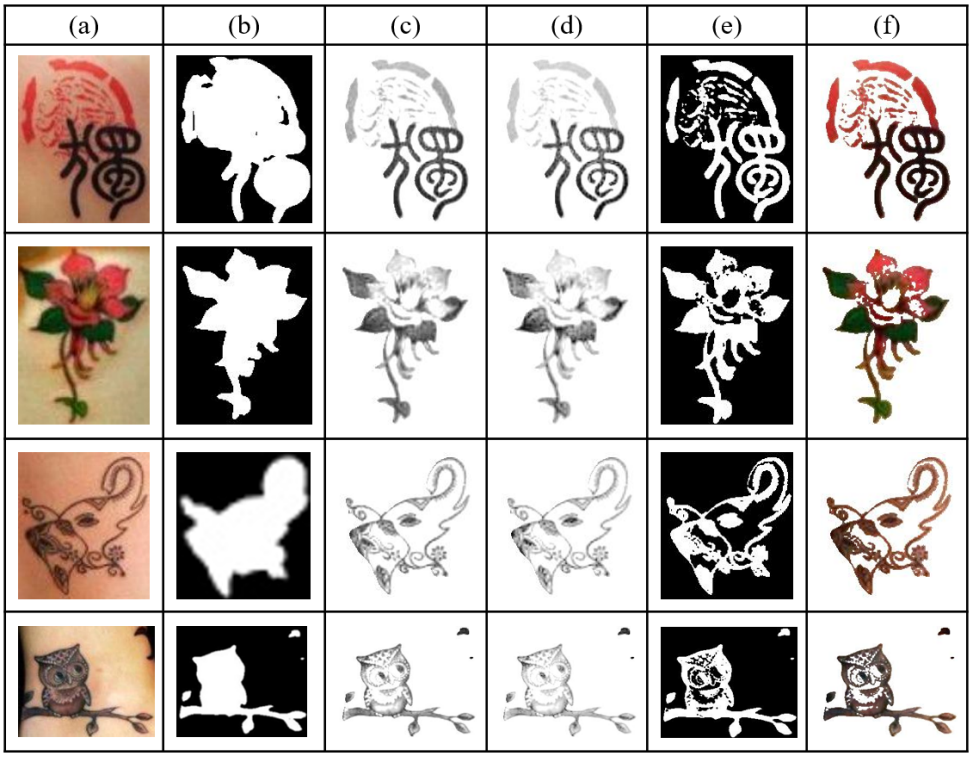


Figure 2: (a) **Original** crop is detected by RT-DETR. (b) Segmentation Mask is generated by SAM. (c) Sketch Region: The SAM mask is applied to the unshadowed grayscale sketch image, highlighting the tattoo's contours while suppressing background noise. (d) **Sketchify-Pencil**: The masked sketch undergoes multi-exposure fusion, grayscale conversion and normalization to enhance contrast. (e) Binary Mask is computed from the normalized sketch with thresholding (thr. = 254) to define the foreground region. (f) **Sketchify-Marker**: The binary mask is applied to the original color crop, producing the final color-preserved sketch where the tattoo retains its original color against a desaturated, sketch-like background.

Following the sketch generation, the pipeline isolates tattoo regions using semantic segmentation. The Segment Anything Model (SAM)[13] is used to produce binary masks corresponding to each detected bounding box (see Figure 2b). These masks are applied to the sketch image using alpha compositing to extract the segmented tattoo regions. Each masked region is then cropped according to its bounding box coordinates (see Figure 2c). To improve visibility under diverse lighting conditions, each cropped region is processed through a high-dynamic-range (HDR) fusion procedure (see Figure 2d), which will be called the 'Sketchify-Pencil' approach. Three exposure-adjusted versions $C_{0.5}$, $C_{1.0}$ and $C_{2.0}$ of each crop are created using scaling factors of 0.5, 1.0, and 2.0:

$$C_{0.5} = 0.5 \bullet\ C, \quad C_{1.0} = 1.0 \bullet\ C, \quad C_{2.0} = 2.0 \bullet\ C \tag{2}$$

and combined via pixel-wise averaging to form a contrast-enhanced image.

$$C_{HDR} = \frac{1}{3}\left(C_{0.5} + C_{1.0} + C_{2.0}\right) \tag{3}$$

This image is then converted to grayscale (single channel) and normalized via min-max scaling to the full 8-bit range [0,255] to ensure consistent dynamic range. Let G be grayscale version of C_{HDR}. The normalized image G_{norm} is computed as:

$$G_{norm}(x,y) = 255 \bullet \frac{G\left(x,y\right) - G_{min}}{G_{max} - G_{min}} \tag{4}$$

where $G_{min} = min_{x,y}G(x,y)$ and $G_{max} = max_{x,y}G(x,y)$. In the final processing stage, a binary threshold mask is created using the specified intensity threshold of T=254 (see Figure 2e). Pixels above this threshold are

designated as foreground (value 0), while all others become background (value 255).

$$M(x,y) = \begin{cases} 0, & if\ G_{norm}(x,y) > T \quad (foreground) \\ 255, & otherwise \quad (background) \end{cases} \tag{5}$$

This mask is applied to the original color crop through alpha compositing, thereby preserving color information in tattoo (see Figure 2f), which will be called the 'Sketchify-Marker' approach. The result is a hybrid visual representation that retains semantically meaningful color features while emphasizing structural characteristics. The sketchify Pencil and Marker approaches are analyzed in the experiments (Sec. 4).

4. EXPERIMENTS AND RESULTS

The purpose of the experiments is to analyze the added value of the proposed methods. In the matching experimental setup, we evaluate performance using a test set consisting exclusively of cropped tattoo images. The reference database similarly also contains only cropped images of tattoos. Each test image is treated as a query with the objective of retrieving the corresponding matching tattoo from the database. This setup isolates the matching task from detection, allowing a focused evaluation of retrieval performance based solely on the visual similarity of the tattoos.

4.1 Datasets

We use the WebTattoo dataset[9] to evaluate the performance of various tattoo matching and retrieval techniques. The dataset contains 400 query images of tattoos and 150 sketch queries, each with corresponding matches in the database. Examples of matching tattoos are shown in Figure 3.

(a) Example tattoo 1 (b) Example tattoo 2 (c) Example tattoo 3

(d) Matching tattoo 1 (e) Matching tattoo 2 (f) Matching tattoo 3

Figure 3: Three examples of tattoos from the WebTattoo dataset and their matching tattoos.

In our evaluation, we focus exclusively on the image-to-image matching task. To prepare the dataset, we separate the query images (test set) from the database and crop the tattoos using the provided bounding boxes. Since the evaluated methods (CLIP and OmniGlue) are used out-of-the-box without any fine-tuning, a training set is not required. The dataset contains tattoos captured under a variety of conditions, including changes in

viewpoint, resolution, lighting conditions, and tattoo aging, providing a realistic and challenging benchmark for retrieval.

Although other data sets have been proposed for tattoo detection and retrieval, they were not suitable for our study. The Tatt-C dataset,[8] although comprehensive, is no longer publicly available. The BIVTatt dataset[14] includes a small number of original tattoo images and a large set of synthetically augmented variants, which may not accurately reflect the complexity and variability encountered in real-world tattoo matching scenarios.

4.2 Metrics

For tattoo matching, we evaluate retrieval effectiveness using Recall@K (K=1, 5, 10), which quantifies the probability that the correct match appears within the top-K retrieved results. High recall values indicate the system's robustness in retrieving relevant tattoos.

4.3 Results with CLIP and OmniGlue

The aim of the experiment is to show that our Combined method performs better for tattoo matching than the CLIP component or the OmniGlue component alone.

Results for the CLIP component, the OmniGlue component and the Combined approach are shown in Table 1. The Combined approach uses a threshold value of 8 on the OmniGlue score as separator between CLIP and OmniGlue predictions. The table shows that the Combined approach outperforms both separate approaches. Furthermore, it shows that CLIP performs slightly better in terms of Recall@K than OmniGlue. This suggest while CLIP features may not provide the same level of interpretability as keypoint-based matching, they still capture meaningful information for successful tattoo identification. However, we see significant value in OmniGlue due to its greater explainability.

Model	Recall@1	Recall@5	Recall@10
CLIP	0.825	0.907	0.930
OmniGlue	0.734	0.782	0.812
Combined (ours)	**0.872**	**0.922**	**0.945**

Table 1: Comparison of Recall@K for different models. The best results are highlighted in bold.

To explore the potential benefits of combining both methods, Figure 4 presents a scatterplot comparing the CLIP scores with the OmniGlue scores. The colors and shapes indicate whether CLIP or OmniGlue ranks the correct match higher or whether their performance is equal. In particular, the red points, where CLIP outperforms OmniGlue in terms of ranking the correct match, are scattered along the x-axis, suggesting that the CLIP score is not a strong indicator of CLIP's reliability. In contrast, along the y-axis, we observe that above a certain threshold, most points are either blue squares or green triangles. The blue squares indicate cases where OmniGlue ranks the correct match higher than CLIP, while the green triangles represent instances where both methods assign the same rank to the correct match. This distribution suggests that applying a threshold to the OmniGlue matching score can be an effective strategy for identifying cases where OmniGlue is the more reliable method for retrieval. Selecting a threshold of 8 (indicated by the dashed line in Figure 4) on the OmniGlue score appears to improve the performance of the combined approach, particularly in terms of Recall@1, compared to using either CLIP or OmniGlue alone. This is shown in Table 1, where the last row indicates the scores when setting this threshold.

Figure 5a shows the performance of the Combined method as a function of the threshold applied to the OmniGlue matching score. When the threshold is set to 0, the Combined method defaults to the OmniGlue predictions, whereas at very high threshold values (i.e., above the maximum OmniGlue score), it defaults to the CLIP-based predictions. The plot reveals a performance peak at threshold values between 8 and 10, where the Recall@1 reaches 0.872. Notably, the peak is relatively broad, indicating that the method is not highly sensitive to the exact choice of threshold, particularly for values above the optimal range. This robustness is desirable, as it suggests that the combined approach remains effective even without fine-tuned threshold selection. It is

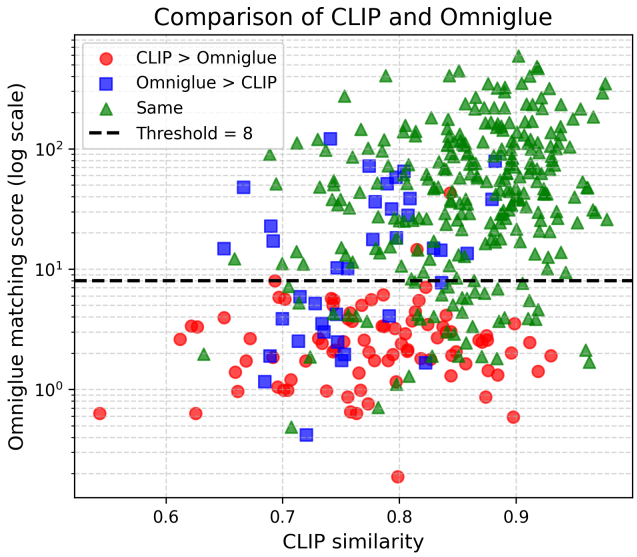
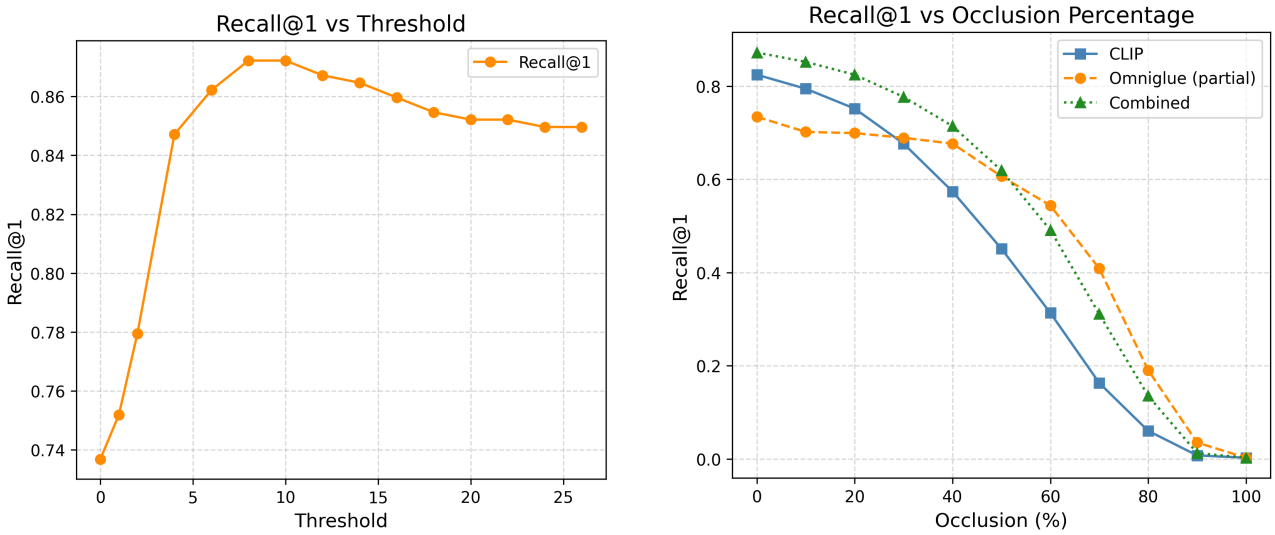


Figure 4: Scatter plot comparing CLIP similarity scores and OmniGlue matching scores. Each point represents a data sample, with colors indicating whether CLIP outperforms OmniGlue (red), OmniGlue outperforms CLIP (blue), or both models perform equally (green). The y-axis is displayed on a logarithmic scale to improve readability of dispersed values. The dashed line indicates the threshold of 8 on the OmniGlue matching score.



(a) Recall@1 vs. threshold

(b) Recall@1 vs. occlusion level

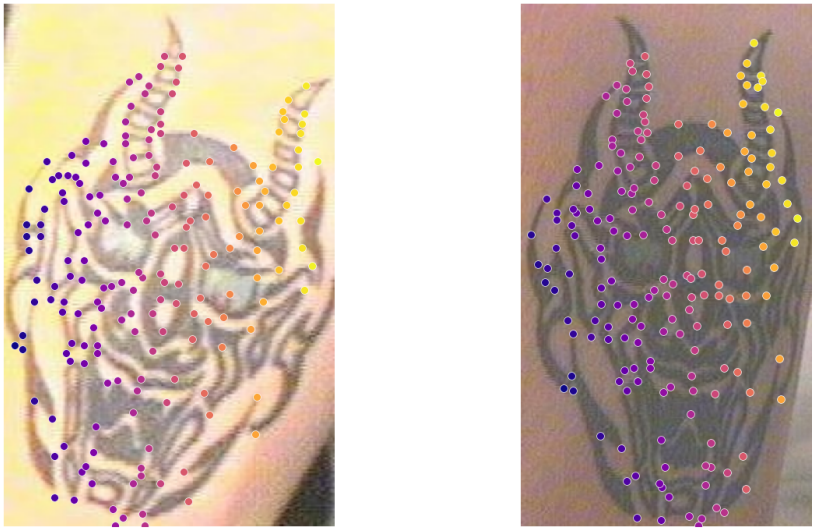
Figure 5: Comparison of Recall@1 across threshold values and occlusion levels.

important to note that this analysis was conducted on the test set. In practical scenarios, threshold selection should be based on a separate validation set to avoid overfitting.

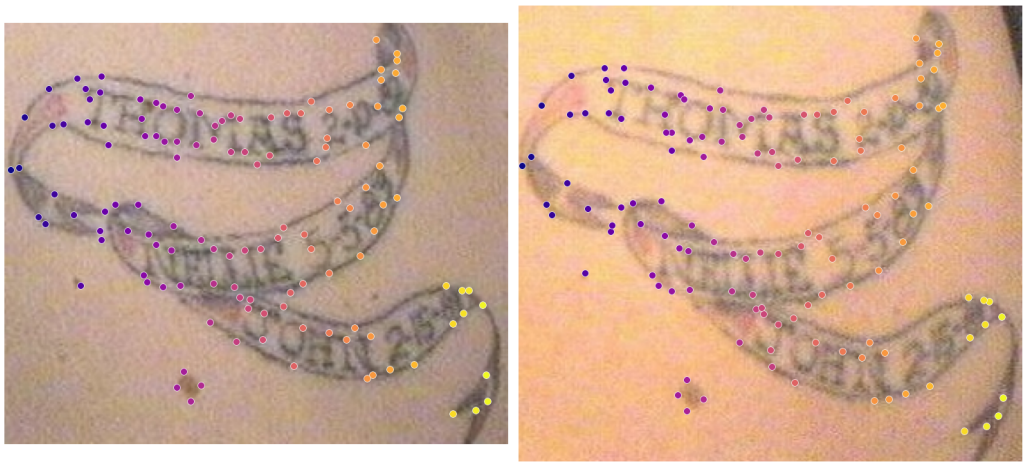
To further assess the robustness of the matching methods, we conduct an occlusion experiment in which increasing portions of the query images are occluded by blacking out one side of the image. The database images remain unchanged, reflecting the assumption that higher-quality images are typically available in reference databases. As the degree of occlusion increases, a decline in matching performance is expected for all methods. However, we hypothesize that OmniGlue will be more robust to occlusion due to its local, keypoint-based matching approach, and therefore its performance will degrade more gradually compared to CLIP. The results,

shown in Figure 5b, confirm this hypothesis. While CLIP performs better under low occlusion levels, OmniGlue surpasses CLIP at occlusion levels of 30% and above. This suggests that OmniGlue's local feature matching is more resilient to partial image loss. This is especially relevant for tattoo matching since parts of the tattoo could be occluded by clothing. Notably, the Combined method is outperformed by OmniGlue at occlusion levels above 50%. This is likely due to the fixed threshold of 8 used for the OmniGlue matching score. As occlusion increases, fewer keypoints are detected, leading to generally lower matching scores. Adaptive thresholding or dynamic calibration based on the degree of visible content could improve the performance of the Combined method in such scenarios.

The results in Figure 6 show that OmniGlue can successfully identify the correct tattoo from the database, even under difficult conditions, while also providing relevant keypoint matches. Furthermore, the figure shows that it gives explainable insight which locations in the image are related to each other.



(a) Query image (left) and top-1 match from the dataset (right)



(b) Query image (left) and top-1 match from the dataset (right)

Figure 6: Query images from the WebTattoo dataset and their top-1 matches from the database using OmniGlue patching. Colored points indicate the matching keypoints.

4.4 Sketchification results

An initial evaluation of CLIP on the whole WebTattoo dataset showed that it performs well overall. Nevertheless, a careful study of the results showed that CLIP does not perform well on textual tattoos or on tattoos with abstract shapes. This behaviour is expected. CLIP has primitive OCR capabilities and performs relatively well when the text is clean and digitally rendered.[15] When it is confronted with tiny, curved, low-contrast glyphs wrapped around skin, often in stylised or hand-drawn fonts, its performance drops dramatically. Similarly, very abstract shapes (tribal lines, mandalas, ink blots, etc.) with vague semantics, which probably ended up near one another during CLIP's training, also induce a significant performance drop. Thus, our concept targeted exactly those "hard cases". We selected among the initial database the tattoos that included text or abstract shapes, resulting in a subset of 73 images out of 400. Some representative tattoos can be seen in Figure 7.

Figure 7: Examples of the hard-case subset used for evaluating the sketchification method. These examples usually contain textual or abstract-shaped tattoos.

For the evaluation, both query images as well as database images were sketchified. Then, the 73 images have been used as queries against the 400 database images of the same modality (original, pencil or marker). The results can be seen in Table 2.

Method	Recall@1	Recall@5	Recall@10
Original	0.603	0.739	0.767
Sketchify-Pencil	**0.726**	**0.877**	**0.931**
Sketchify-Marker	0.657	0.781	0.822

Table 2: Comparison of Recall@K for different inputs. The best results are highlighted in bold.

Our sketchification methods (pencil and marker) clearly enhance the performance of CLIP. This is achieved by cropping and segmenting the tattoo region, which removes skin, shadows, body curvature and background noise, concepts that were out of CLIP's training domains. This gives CLIP the opportunity to concentrate to the actual, distinctive area of interest.

5. CONCLUSION AND DISCUSSION

Our results demonstrate the added value combining global and local vision foundation models for explainable tattoo matching. It also highlights the potential of local OmniGlue as an explainable alternative to global image-based matching methods like CLIP. While CLIP achieves slightly higher Recall@k, OmniGlue provides greater transparency by revealing which image regions contribute to a match. Moreover, we show that combining both methods can improve performance, particularly when leveraging a threshold on the OmniGlue matching score.

Results also show that the combined approach, which uses one fixed threshold, works better than the two separate approaches for occlusions up to 50%.

An advantage of the proposed approach is that no tattoo specific training data is necessary to use the method. However, despite its advantages, OmniGlue also has practical limitations, notably its long inference

times, which may hinder real-world deployment in time-sensitive forensic applications. Another disadvantage is that the combined approach contains one threshold parameter that needs to be optimized on a train set that is independent from the evaluation set. Future work should focus on optimizing efficiency to make keypoint-based matching more viable for large-scale databases.

Finally, our proposed sketchification method enhances the matching performance of CLIP, by removing distractions during the deep features extraction of "hard cases", like textual or abstract tattoos.

ACKNOWLEDGMENTS

The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

REFERENCES

[1] Jain, A. K., Lee, J.-E., Jin, R., and Gregg, N., "Content-based image retrieval: An application to tattoo images," in [*IEEE ICIP*], 2745–2748 (2009).

[2] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *International journal of computer vision* **60**, 91–110 (2004).

[3] Sivic and Zisserman, "Video google: A text retrieval approach to object matching in videos," in [*Proceedings ninth IEEE international conference on computer vision*], 1470–1477, IEEE (2003).

[4] Manger, D., "Large-scale tattoo image retrieval," in [*IEEE Conf. Computer and Robot Vision*], 454–459 (2012).

[5] Kim, J., Parra, A., Yue, J., Li, H., and Delp, E. J., "Robust local and global shape context for tattoo image matching," in [*IEEE ICIP*], 2194–2198 (2015).

[6] Yi, H., Yu, P., Xu, X., and Kong, A. W. K., "The impact of tattoo segmentation on the performance of tattoo matching," in [*IEEE Int. WIE Conf. Electrical and Computer Engineering*], 43–46 (2015).

[7] Di, X. and Patel, V. M., "Deep tattoo recognition," in [*IEEE CVPR Workshops*], 51–58 (2016).

[8] Ngan, M. and Grother, P., "Tattoo recognition technology-challenge (Tatt-C): an open tattoo database for developing tattoo recognition research," in [*IEEE Int. Conf. Identity, Security and Behavior Analysis ISBA*], 1–6 (2015).

[9] Han, H., Li, J., Jain, A. K., Shan, S., and Chen, X., "Tattoo image search at scale: Joint detection and compact representation learning," *IEEE Trans. PAMI* **41**(10), 2333–2348 (2019).

[10] Pavithra, N. and Sharath, K. Y., "Sketch-based image retrieval: Effectivity notion of recent approaches," in [*IEEE Int. Conf. Emerging Technology*], 1–6 (2020).

[11] 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*], 8748–8763, PmLR (2021).

[12] Jiang, H., Karpur, A., Cao, B., Huang, Q., and Araujo, A., "Omniglue: Generalizable feature matching with foundation model guidance," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 19865–19875 (2024).

[13] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., "Segment anything," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 4015–4026 (2023).

[14] Nicolás-Díaz, M., Morales-González, A., and Méndez-Vázquez, H., "Deep generic features for tattoo identification," in [*Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24*], 272–282, Springer (2019).

[15] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., "Learning transferable visual models from natural language supervision," in [*Proceedings of the 38th International Conference on Machine Learning*], 8748–8763 (2021). Extended version.