Robust Artwork Recognition on Mobile Devices: A Comparative Analysis of Preprocessing for Image Retrieval

Jiwon Lee
Hyper-Reality Metaverse Research Laboratory (ETRI)
Daejeon, Korea
ez1005@etri.re.kr

Sungwon Moon
Hyper-Reality Metaverse Research Laboratory (ETRI)
Daejeon, Korea
moonstarry@etri.re.kr

Jung-Jae Yu
Hyper-Reality Metaverse Research Laboratory (ETRI)
Daejeon, Korea
jungjae@etri.re.kr

***Abstract*— This paper proposes an optimal preprocessing technique to ensure accurate recognition of traditional Korean artworks captured in mobile environments. Based on experiments with a dataset of 380 artworks by Kim Hong-do, simulating real-world mobile capture conditions, we demonstrate that the 'cropping' technique not only improves Top-1 accuracy (cmc@1) by 8.85%*p* (percentage points) over the baseline but also outperforms the more complex 'segmentation' method. This proves that cropping is the most practical solution, offering an excellent balance between accuracy and computational efficiency. Furthermore, our *Grad-CAM* analysis on the Vision Transformer (*ViT*) model reveals the limitations of conventional *XAI* methods and suggests future research directions. The findings of this study provide an empirical guideline for designing the visual perception module of mobile AI systems that require real-time interaction.**

***Keywords*— Image Retrieval, Preprocessing, Computer Vision, Cultural Heritage, AI Persona, Vision Transformer**

I. Introduction

AI personas are emerging as a next-generation interface designed to foster emotional and intellectual interaction with users. In the field of cultural heritage, there is a growing demand for artist-based AI personas, for which a robust recognition technology for traditional artworks captured in mobile environments is a critical component.

However, real-world mobile environments introduce significant challenges, such as light reflections, cluttered backgrounds, and geometric distortions, all of which degrade recognition accuracy. Therefore, it is necessary to analyze the effects of preprocessing, which refines the artwork area from the input image and removes background information. As this preprocessing incurs computational costs, it is crucial to balance the gains in accuracy with the associated processing speed. For interactive mobile applications, the entire recognition pipeline should ideally operate within a real-time threshold of 100*ms*.

To address these challenges, this paper proposes an optimal preprocessing pipeline to maximize artwork recognition performance in mobile environments. The main contributions of this work are as follows: First, we systematically compare and verify the effects of different preprocessing techniques (baseline, cropped, and masked) on image retrieval performance using a dataset that reflects real-world mobile capture and data augmentation. Second, through a trade-off analysis of accuracy and efficiency, we propose 'area cropping' as the most suitable preprocessing pipeline for real-time mobile environments. Third, we experimentally investigate the limitations of applying *Grad-CAM* to *ViT*-based models and suggest future directions for *XAI* research.

II. Related Work

Artwork recognition and image retrieval have seen rapid advancements with the development of deep learning and transformer-based architectures. Convolutional neural network (*CNN*)-based approaches have dominated the image recognition domain, with techniques such as *ResNet*, *EfficientNet*, and *YOLO* being frequently employed for object detection and recognition tasks [1-3]. However, traditional *CNN* architectures encounter difficulties in capturing global context and relationships between image regions, which can be crucial in accurately identifying artworks with complex compositions and diverse visual attributes.

Recently, Vision Transformer (*ViT*)-based models have gained prominence due to their ability to model global context through self-attention mechanisms, which provide enhanced performance in various vision tasks, including image classification, object detection, and retrieval [4-5]. In particular, *UniCom* [6], a universal and compact ViT-based embedding extractor, demonstrates impressive retrieval performance by effectively learning global feature representations.

Moreover, preprocessing techniques significantly impact retrieval performance by refining images and removing irrelevant information. Recent studies have explored various preprocessing methods, such as bounding-box cropping and semantic segmentation, to enhance feature extraction accuracy [7-8]. However, there is limited systematic comparison on how these methods perform under realistic mobile conditions. Our research addresses this gap by providing a comprehensive comparative analysis, emphasizing practical considerations for mobile AI applications.

III. Proposed Method and System Configuration

A. Dataset Construction

To validate the performance of traditional artwork recognition, we constructed a dataset centered on the works of Kim Hong-do, a representative painter of the late Joseon Dynasty. For the training data, we collected 380 of his artworks from the *Gongu Madang* (Korea Copyright Commission) repository [9], all of which are free of copyright issues. To create the test data, 192 of these works were printed in color on *A4* paper and then photographed from a distance

of approximately 50*cm* using a Samsung *Galaxy S23 Ultra*. This process allowed us to build a realistic test set that includes factors such as light reflection, distortion, and background noise that can occur in a mobile viewing environment.

Fig. 1. Examples of Kim Hong-do's artworks

B. Preprocessing Techniques

To remove the complex backgrounds and visual noise present in mobile-captured images and to encourage feature extraction focused on the artwork itself, we applied three preprocessing techniques, as shown in *Fig. 1*. All preprocessing was performed based on the state-of-the-art segmentation model, *MobileSAM* [10].

- ***Baseline***: The original image captured by the mobile device, used as input without any preprocessing.
- ***Cropped***: An image containing only the area of the automatically detected bounding box of the artwork.
- ***Masked***: An image where the background area outside the artwork is masked in black using the segmentation results from *MobileSAM*. This is considered the most precise method for background removal.

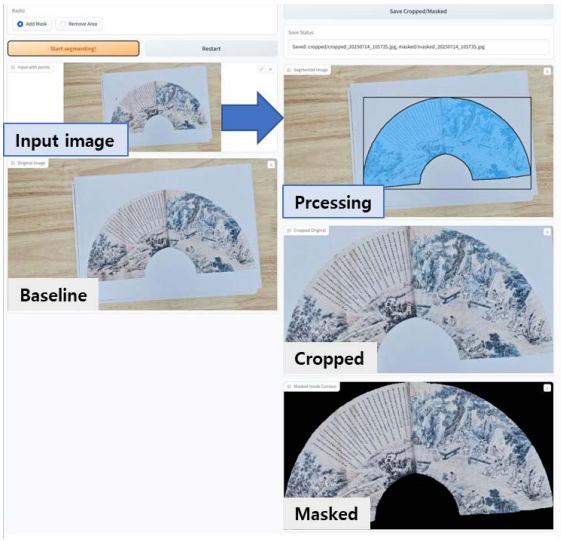

Fig. 2. Data processing example using MobileSAM

C. Retrieval Framework and Training Strategy

The experiments were conducted using the *Open-Metric-Learning (OML)* framework [11]. We employed the *UniCom* model, which has a *ViT-B/16* architecture, as the embedding extractor. The retrieval was performed using a nearest neighbor search based on an embedding space trained with a triplet loss with miner strategy. For training, we used the *all_triplets* miner to collect all possible triplet pairs from each minibatch and utilized a triplet loss function with a soft margin.

D. Data Augmentation

To accurately reflect the diverse and unpredictable conditions of a real mobile capture environment, we applied a series of standard image augmentation techniques to our test set. This approach was critical for building a robust evaluation framework that simulates real-world challenges such as varying distances, lighting conditions, and potential occlusions.

- **Resize**: Input images were resized to a fixed size of 512 pixels on one side. This standardizes input dimensions for the model while preserving aspect ratios to prevent unnecessary distortion.
- **Random Crop**: We applied random cropping, taking 85% of the total image area. This simulates variations in how users frame artworks, often including partial views or slightly off-center compositions, and helps the model generalize to different framing scenarios.
- **Color Jitter**: Color distortion was introduced using brightness and contrast factors of 0.2. This technique mimics the impact of inconsistent lighting, shadows, or screen glare frequently encountered when capturing images with mobile devices in various indoor or outdoor settings.
- **Additive Gaussian Noise**: Gaussian noise with a standard deviation of σ=1.0 was added. This simulates sensor noise and image degradation that can occur in low-light conditions or with lower-quality mobile cameras, enhancing the model's robustness to subtle image imperfections.

By incorporating these specific augmentation techniques, our aim was to create a test dataset that closely emulates the complexities and variability of user-generated mobile captures. This rigorous simulation ensures that our comparative analysis of preprocessing methods is based on conditions highly representative of the target application environment, allowing us to derive practical and generalizable insights for mobile AI systems.

IV. Experiments and Analysis

A. Quantitative Analysis

To analyze the impact of preprocessing on image retrieval performance, we utilized the following evaluation metrics:

- ***cmc@1***: The probability that the top-ranked result is correct.
- ***map@5***: The mean of the average precision scores for the top-5 results.
- ***precision@5***: The proportion of correct results within the top-5.

The retrieval performance for each preprocessing method on the augmented test set is shown in *Table I.*

TABLE I. IMAGE RETRIEVAL PERFORMANCE COMPARISON BY PREPROCESSING METHOD (ON AUGMENTED TEST SET)

Preprocessing Method	cmc@1	map@5	precision@5
Baseline	0.8542	0.8892	0.9427
Cropped	**0.9427**	**0.9529**	**0.9688**
Masked	0.9271	0.9429	0.9688

Our analysis reveals a substantial performance uplift across all preprocessing methods compared to the Baseline. Specifically, the cropped method achieved a *cmc@1* accuracy of 0.9427, marking an impressive 8.85%p improvement over the baseline's 0.8542. The masked method also demonstrated significant gains, with a *cmc@1* of 0.9271, an increase of 7.29%p from the baseline. This clearly highlights the critical role of preprocessing in enhancing artwork recognition accuracy in challenging mobile environments, effectively mitigating issues like cluttered backgrounds and distortions.

A particularly noteworthy finding is the superior performance of the cropped method over the more computationally intensive masked method. The cropped method approach achieved a 1.56%*p* higher *cmc@1* accuracy (0.9427 vs. 0.9271) and a 1%*p* higher *map@5* score (0.9529 vs. 0.9429) than the masked method. While both methods yielded the same *precision@5* (0.9688), the consistent lead of cropped method in *cmc@1* and *map@5* suggests that the precise pixel-level background removal performed by segmentation (masked method) might inadvertently introduce artifacts or subtle information loss that negatively impacts the global feature learning of the *ViT* model. This could be due to the artificial boundaries created or the removal of contextual information that, while seemingly irrelevant, might aid the *ViT* in forming robust representations. The consistent improvement across all metrics, including *map@5* and *precision@5*, further strengthens the conclusion that preprocessing, especially cropping, is highly effective for robust artwork recognition in mobile settings. The high *precision@5* values across the pre-processed methods indicate that when the model provides a *top-5* result, it is highly likely to contain the correct artwork, which is crucial for user experience in interactive AI persona applications.

This observed phenomenon, where the cropped method outperforms the masked method, is particularly intriguing for *ViT*-based architectures. Unlike Convolutional Neural Networks (*CNNs*), which primarily focus on local features, *ViTs* leverage global self-attention mechanisms. It's plausible that while segmentation aims for precise object isolation, the absolute removal of background pixels might deprive the *ViT* of subtle global contextual cues that it implicitly learns and utilizes, even if those cues appear to be "background noise" to a human observer. Cropping, by merely isolating the artwork within a bounding box, might retain just enough surrounding context for the *ViT* to perform optimally. This suggests that for *ViT* models, a less aggressive form of preprocessing that retains some peripheral information might be more beneficial than strict semantic segmentation, warranting further investigation into the interplay between *ViT*'s global attention and the nature of preprocessing.

B. *Efficiency Analysis and Practical Considerations*

The performance disparity between preprocessing methods becomes even more pronounced when considering processing efficiency, a critical factor for real-time mobile applications. Cropping via a lightweight object detector demands significantly less computational resources and time compared to precise semantic segmentation. For instance, while a segmentation model like *MobileSAM* [10] requires a certain inference time to generate detailed masks, extracting a bounding box for cropping can be achieved with substantially lower latency. From a practical application standpoint, the crop area can be rapidly extracted using only the region proposal module of a highly optimized, lightweight object detector such as *YOLOX* [12]. This modular approach effectively minimizes the processing time burden on the overall system, allowing the entire recognition pipeline to operate within the stringent real-time thresholds (e.g., 100*ms*) typically required for interactive mobile experiences.

To provide a clearer understanding, consider that the typical latency for object detection using a lightweight model for bounding box prediction can be orders of magnitude faster than full semantic segmentation on mobile hardware. While exact figures would depend on the specific hardware and model implementation, this difference is crucial for user perceived responsiveness. The ability to quickly isolate the region of interest means the subsequent embedding extraction and retrieval steps can commence with minimal delay. Therefore, when evaluating both the achieved accuracy and the imperative for computational efficiency, "*cropping*" emerges as the most optimized and practical technique for robust artwork recognition in real-time mobile environments. Its balance of high accuracy (as demonstrated in Section IV.A) and superior processing speed makes it an ideal candidate for deployment in resource-constrained mobile AI systems.

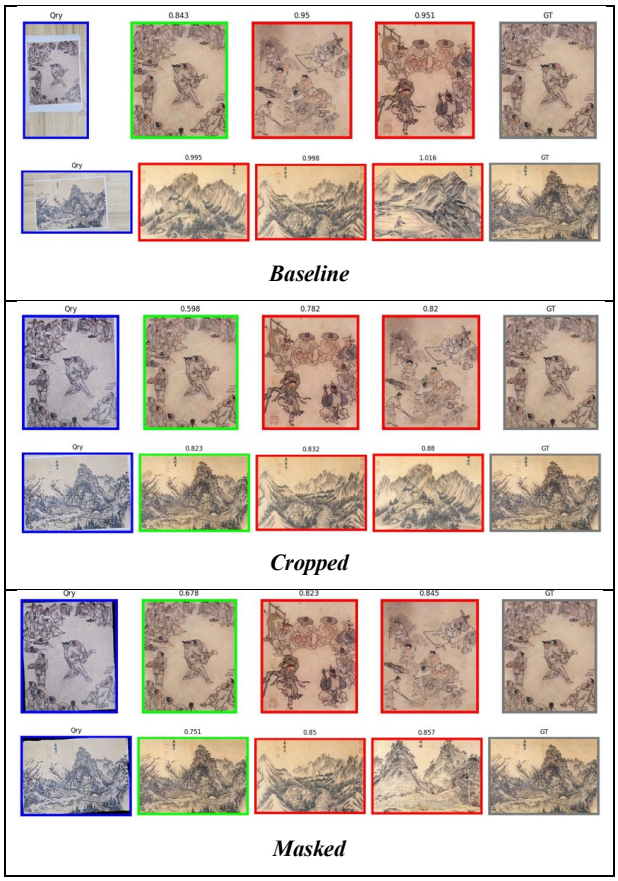

Fig. 3. Retrieval result comparison by preprocessing method. (Left: query, Middle: top search results, Right: ground truth).

C. Analysis of XAI Applicability

We applied *Grad-CAM* [13] to the *ViT-B/16* based model to visually analyze differences in attention areas, but observed that the model's primary attention locations showed little variation regardless of whether the input image was a baseline, cropped, or masked version. This is because, unlike *CNN*s, the *ViT* architecture considers relationships between all patches simultaneously through a *global self-attention* mechanism and concentrates information into the *CLS* token. As a result, even if the spatial structure of the input changes, the distinctiveness of the output attention map is diminished.

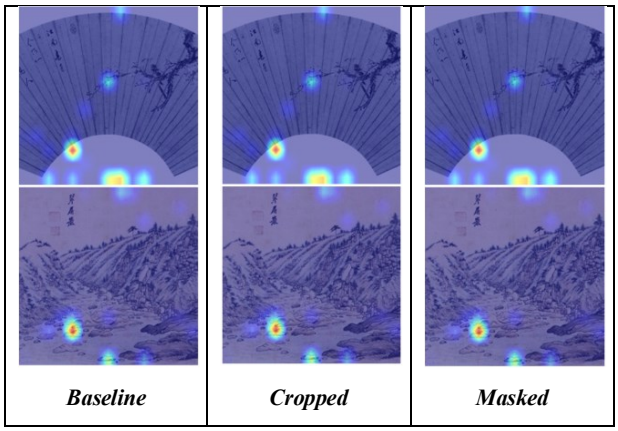

Fig. 4. Grad-CAM visualization results (showing minimal difference across preprocessing methods).

These results indicate that conventional *CNN*-based *XAI* techniques like *Grad-CAM* may not be suitable for *ViT*-family models. This shows the need to adopt alternative techniques better suited for *ViTs*, such as *Attention Rollout*, *Transformer Attribution*, or *Score-CAM*, in future work.

V. Discussion and Conclusion

This study conducted a comparative analysis of preprocessing techniques to improve the visual recognition accuracy of traditional artworks captured in mobile environments. We demonstrated that the area cropping method delivers superior performance in terms of both accuracy and computational efficiency. This finding indicates that integrating a lightweight object detector with a crop-based retrieval architecture is a practical and effective strategy for designing real-time visual perception modules in mobile-based AI personas.

However, practical limitations exist. The performance of the cropping method relies heavily on the accuracy of the object detection module, meaning inaccuracies or partial detections can negatively affect overall retrieval performance. Continuous advancements in lightweight detection models and their robustness under diverse environmental conditions are therefore critical.

Additionally, while cropping has proven more effective than precise segmentation methods in our experiments, further studies are required to generalize these findings across varying art styles and capture conditions. Complex artworks with intricate backgrounds or less-defined boundaries may still benefit from more sophisticated segmentation techniques.

Furthermore, by investigating the limitations of *Grad-CAM* on *ViT*-based models, this research highlights the necessity of developing more advanced *XAI* techniques tailored specifically for *ViT*s, such as *Attention Rollout* and *Transformer Attribution*. Such techniques could enhance model interpretability and foster user trust in mobile AI personas.

In conclusion, achieving an optimal balance between accuracy, computational efficiency, and interpretability remains an essential challenge in deploying robust artwork recognition systems in mobile environments. Continued research addressing these aspects is pivotal for the broader adoption of AI technologies in cultural heritage applications.

Acknowledgment

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025. (Project Name : Development of AI Agent Technology Based on Artists Unique Characteristics for Interactive Culture Creation, Project Number : RS-2025-02312732, Contribution Rate : 100%)

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, Las Vegas, NV, USA, 2016, pp. 770–778.

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, Long Beach, CA, USA, 2019, pp. 6105–6114.

[3] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv:2004.10934, 2020.

[4] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, Virtual Event, 2021.

[5] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. ICCV, Montreal, QC, Canada, 2021, pp. 10012–10022.

[6] X. An, J. Deng, K. Yang, J. Li, Z. Feng, J. Guo, J. Yang, and T. Liu, "Unicom: Universal and Compact Representation Learning for Image Retrieval," in Proc. ICLR. May. 2023

[7] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective Search for Object Recognition," IJCV, vol. 104, no. 2, pp. 154–171, 2013.

[8] A. Kirillov et al., "Segment Anything," in Proc. ICCV, Paris, France, 2023.

[9] *Gongu Madang* (Korea Copyright Commission) repository, [Online] Available : https://gongu. copyright.or.kr/gongu/main/main.do

[10] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster Segment Anything: Towards Lightweight SAM for Mobile Applications," arXiv preprint arXiv:2306.14289, 2023.

[11] OML-Team/open-metric-learning, [Online] Available: https://github.com/OML-Team/open-metric-learning

[12] J. Lee, B. Lee, and S.Jung, "YOLOX-based fast and lightweight single-class object detector," in Proc. ICTC, Oct. 2022, pp. 1635–1638.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. ICCV, 2017, pp. 618–626.