Understanding Art & Culture

Piera Riccio
University of Amsterdam
Amsterdam, Netherlands
p.riccio@uva.nl

Selina Khan
University of Amsterdam
Amsterdam, Netherlands
s.j.khan@uva.nl

Ludovica Schaerf
University of Zurich & Max Planck Society
Zurich, Switzerland
ludovica.schaerf@uzh.ch

Shuai Wang
University of Amsterdam
Amsterdam, Netherlands
s.wang3@uva.nl

Tiancheng Liu
Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
tcliu767@connect.hkust-gz.edu.cn

Athanasios Efthymiou
University of Amsterdam
Amsterdam, Netherlands
a.efthymiou@uva.nl

Noa Garcia
University of Osaka
Osaka, Japan
noagarcia@ids.osaka-u.ac.jp

Nanne van Noord
University of Amsterdam
Amsterdam, Netherlands
n.j.e.vannoord@uva.nl

Abstract

Art and cultural heritage objects carry visual, textual, relational, and symbolic meaning that cannot be reduced to standard image understanding tasks. This tutorial presents computational methods for studying fine art and cultural artifacts, organized around three themes: relationality, meaning, and recognizability. We cover multimodal and graph-based representation learning for fine art analysis, knowledge-retrieval and agentic reasoning frameworks for artwork interpretation, and instance-level recognition in cultural heritage settings, as well as large-scale museum benchmarks and synthetic data generation strategies. Beyond these technical contributions, the tutorial foregrounds the cultural dimension of multimedia research, inviting discussion on current approaches for studying art and culture and directions for future research.

CCS Concepts

• **Applied computing** → **Arts and humanities**; • **Computing methodologies** → **Computer vision**; *Natural language processing*; *Knowledge representation and reasoning*.

ACM Reference Format:
Piera Riccio, Selina Khan, Ludovica Schaerf, Shuai Wang, Tiancheng Liu, Athanasios Efthymiou, Noa Garcia, and Nanne van Noord. 2026. Understanding Art & Culture. In *International Conference on Multimedia Retrieval (ICMR '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3805622.3816071

1 Introduction

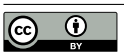
Within the realm of multimedia objects, a special role is often ascribed to objects with cultural and artistic significance [3]. Unlike generic visual content, art and cultural artifacts are shaped by processes of creation, historical context, and layers of symbolic meaning that resist purely data-driven interpretation [1, 27]. Yet, multimedia research has much to offer to these domains, and these domains, in turn, raise challenges that push the boundaries of standard multimedia methods.

In this tutorial, we demonstrate how multimodal representation learning, knowledge retrieval, and instance-level recognition can be adapted and extended to study fine art and cultural heritage. Organized around the themes of relationality, meaning, and recognizability, the tutorial aims to both equip attendees with concrete technical approaches and invite broader reflection on what it means to build multimedia systems that are sensitive to cultural context.

2 Relationality

Multimodal representation learning offers a principled approach to fine art analysis, where artworks are associated with visual, textual, and relational information [13, 14, 23]. Earlier work in computational art analysis focused primarily on image-centric tasks such as artist attribution, style recognition, and aesthetic analysis using deep visual representations [4, 5, 9, 26]. More recent research has increasingly explored semantic retrieval, contextual embeddings, and relational modeling for art understanding [2, 8, 12]. We present a line of work that develops representations of artworks by jointly modeling these complementary sources, moving beyond approaches based solely on visual features. We examine how such representations capture both visual characteristics and relationships between artworks and artists, enabling tasks such as similarity modeling [10], retrieval [14], and the analysis of artistic production over time [6]. By jointly modeling visual and relational structure, artworks can be analyzed not only through their visual properties, but also through the broader patterns that emerge across artistic production [7].

In addition, we extend the concept of multimodal representation learning for fine art analysis towards *relational* multimodal vision-language models. We frame artworks and captions not as isolated image-text pairs, but as entities embedded in structured networks of visual, textual, historical, and contextual relations. We explain strategies for integrating graph-based representations into inductive settings and for combining multimodal models with contextual learning. A central motivation is the distinction between standard VLM tasks and art-historical analysis. Art history scholarship, like VLMs, relies on textual analysis of visual content [34].



ICMR '26, Amsterdam, Netherlands

ACM ISBN 979-8-4007-2617-0/26/06
https://doi.org/10.1145/3805622.3816071

However, its analytical process is shaped by context, interpretative frameworks, archival research, and competing readings rather than by direct image-caption alignment alone. We therefore introduce key principles of art-historical research and interpretation, and show how these can inform the design of multimodal models that are more sensitive to context, relation, and meaning. Overall, the session aims to connect recent technical advances in computer vision, vision-language modeling, and graph-based learning with the methodological needs of fine art analysis.

In particular, Chinese calligraphy, also known as Shufa, is presented as a rich case for multimodal art understanding [31]. It encodes visual structure, linguistic content, and traces of the bodily movement that produced it, yet these modalities are rarely studied together. We examine computational approaches to calligraphy along three progressive layers. The first layer addresses the finished work, showing how structural aesthetic features of regular script and seal carving can be quantified and connected to human perceptual preferences [17, 19]. The second layer moves beneath the surface to the writing process, exploring how cross-modal methods can recover implicit micro-actions such as brush pressure, pause, and wrist angle from the visual trace alone [20, 35]. The third layer considers the machine as a participant in calligraphic practice, surveying style transfer and generative approaches from early pix2pix methods to recent diffusion models, and discussing emotion-driven generation that translates poetic sentiment into expressive strokes[11, 18, 30]. Together, these layers illustrate how calligraphy can serve as a testbed for multimodal representation, cross-modal translation, and human-AI co-creation in the cultural domain.

3 Meaning

Understanding fine art is a knowledge-intensive, multimodal task that requires reasoning over visual content, cultural history, artistic style, and symbolic meaning. It could be far beyond what standard image captioning or retrieval systems can provide. In this context, we present two complementary frameworks that tackle this challenge through structured knowledge retrieval and agentic reasoning. The first framework, ArtRAG [28], integrates a domain-specific art knowledge graph into a retrieval-augmented generation pipeline, enabling contextually grounded, multi-perspective artwork explanations without task-specific training. The second framework, A-MAR [29], extends this with an explicit reasoning and planning stage that conditions retrieval on structured, step-wise evidence requirements and moving beyond static retrieval toward interpretable, goal-driven multimodal reasoning.

Yet, artistic meaning often emerges through recurring motifs across images, stylistic evolution, shared symbolism, or historical context. This motivates considering also set-level visual understanding. In this direction, ImageSet2Text [21] explores how vision-language models combined with graph-based representations can generate coherent semantic interpretations for entire image collections. Instead of treating each image independently, ImageSet2Text summarizes visual elements that are in common across the majority of the images in a set. This technique is especially useful for enabling visual exploration of datasets in the context of cultural analytics.

While these frameworks improve contextualized reasoning and semantic interpretation, they rely on representations that ultimately attempt to structure or summarize artistic meaning. This raises the question: how can we capture the nuanced, evolving, and often subjective interpretations that emerge in art historical analysis? We therefore also discuss the challenges of representing and interpreting such context-rich information, particularly when combining structured metadata with the visual complexity of artworks themselves [15, 16].

4 Recognizability

Instance-level recognition (ILR) distinguishes individual objects rather than just broad categories, offering powerful applications for understanding art and culture. However, ILR faces two fundamental challenges: the scarcity of large-scale annotated datasets due to the fine-grained nature of the task, and the difficulty of handling distribution shifts between controlled gallery images and unstructured visitor photographs. We address these challenges by examining recent advances along two complementary fronts. First, we explore a large-scale benchmark built from The Met museum's open-access collection, featuring over 224k individual exhibits for training under studio conditions [33], while testing on noisy visitor-taken photos and out-of-distribution images. The benchmark shows that combining self-supervised and supervised contrastive learning for non-parametric classification is a promising direction. Second, we examine a novel synthetic data generation approach that bypasses manual annotation entirely: given only target domain names, it synthesizes diverse object instances without any real images [32]. Fine-tuning foundation vision models on this synthetic data improves retrieval across multiple ILR benchmarks, offering an efficient and scalable alternative to extensive data collection. Together, these perspectives show how the field is moving from curated benchmarks toward flexible, domain-independent ILR systems, unlocking real-world applications for cultural heritage and beyond.

Following the discussion on uniquely recognizable instances, the tutorial is continued with a talk on iconic images. Within Visual Culture, the notion of iconic images refers to particular images which have a unique position in collective memory [24], often very widely reproduced and recalled as well as strongly connected to a particular (often emotional) historical event. Because these images are widely circulated, they are frequently found in large-scale datasets, yet, from a technical perspective, they are often not treated any differently, which raises questions about a-culturality in how we learn from multimedia collections [27].

In this talk, the notion of iconic images is introduced, including methods for how to study them as well as open research questions [22, 25]. A central thesis that will be presented is that iconic images form an inviting research topic which can, and should, be studied through multimedia research, which naturally opens the direction for understanding that is better situated in art & cultural understanding.

References

[1] Taylor Arnold and Lauren Tilton. 2019. Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities* 34, Supplement_1 (12 2019), i3–i16. https://doi.org/10.1093/llc/fqz013

[2] Giovanna Castellano, Vincenzo Digeno, Giovanni Sansaro, and Gennaro Vessio. 2022. Leveraging Knowledge Graphs and Deep Learning for automatic art

analysis. *Knowledge-Based Systems* 248 (2022), 108859. doi:10.1016/j.knosys.2022.108859

[3] Eva Cetinic and James She. 2022. Understanding and Creating Art with AI: Review and Outlook. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 2, Article 66 (Feb. 2022), 22 pages. doi:10.1145/3475799

[4] Eva Cetinic and James She. 2022. Understanding and Creating Art with AI: Review and Outlook. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 2, Article 66 (Feb. 2022), 22 pages. doi:10.1145/3475799

[5] Elliot J. Crowley and Andrew Zisserman. 2016. The Art of Detection. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I (Lecture Notes in Computer Science)*, Gang Hua and Hervé Jégou (Eds.). Springer, 721–737. doi:10.1007/978-3-319-46604-0_50

[6] Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2026. Set2Seq Transformer: Temporal and Position-Aware Set Representations for Sequential Multiple-Instance Learning. *arXiv preprint arXiv:2408.03404* (2026). doi:10.48550/arxiv.2408.03404

[7] Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2026. VL-KGE: Vision–Language Models Meet Knowledge Graph Embeddings. In *Proceedings of the ACM Web Conference 2026* (United Arab Emirates) *(WWW '26)*. Association for Computing Machinery, New York, NY, USA, 7552–7563. doi:10.1145/3774904.3792677

[8] Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Marcel Worring, and Nachoem Wijnberg. 2021. Graph Neural Networks for Knowledge Enhanced Visual Representation of Paintings. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 3710–3719. doi:10.1145/3474085.3475586

[9] Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny, and Marian Mazzone. 2018. The Shape of Art History in the Eyes of the Machine. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). doi:10.1609/aaai.v32i1.11894

[10] Ahmed Elgammal and Babak Saleh. 2015. Quantifying Creativity in Art Networks. In *Proceedings of the 6th International Conference on Computational Creativity (ICCC)*. Brigham Young University, 39–46.

[11] Bin Fu, Zixuan Wang, Kainan Yan, Shitian Zhao, Qi Qin, Jie Wen, Junjun He, and Peng Gao. 2025. FontAnimate: High Quality Few-shot Font Generation via Animating Font Transfer Process. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16015–16025.

[12] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. 2019. Context-Aware Embeddings for Automatic Art Analysis. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (Ottawa ON, Canada) *(ICMR '19)*. Association for Computing Machinery, New York, NY, USA, 25–33. doi:10.1145/3323873.3325028

[13] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. 2020. ContextNet: Representation and exploration for painting classification and retrieval in context. *Int. J. Multim. Inf. Retr.* 9, 1 (2020), 17–30. doi:10.1007/S13735-019-00189-4

[14] Noa Garcia and George Vogiatzis. 2019. How to Read Paintings: Semantic Art Understanding with Multi-modal Retrieval. In *Computer Vision – ECCV 2018 Workshops*, Laura Leal-Taixé and Stefan Roth (Eds.). Springer International Publishing, Cham, 676–691. doi:10.1007/978-3-030-11012-3_52

[15] Selina Khan and Nanne van Noord. 2024. Context-Infused Visual Grounding for Art. arXiv:2410.12369 [cs.CV] https://arxiv.org/abs/2410.12369

[16] Selina Khan and Nanne van Noord. 2024. Stylistic Multi-Task Analysis of Ukiyo-e Woodblock Prints. arXiv:2410.12379 [cs.CV] https://arxiv.org/abs/2410.12379

[17] Tiancheng Liu, Ling Li, and Manling Yang. 2024. Chinese Seal Carving Aesthetic Evaluation. In *Proceedings of the 17th International Symposium on Visual Information Communication and Interaction (VINCI '24)*. Association for Computing Machinery, New York, NY, USA, Article 10, 5 pages. doi:10.1145/3678698.3687175

[18] Tiancheng Liu, Anqi Wang, Xinda Chen, Jing Yan, Yin Li, Pan Hui, and Kang Zhang. 2024. PoEmotion: Can AI Utilize Chinese Calligraphy to Express Emotion from Poems? In *Proceedings of the 29th International Symposium on Electronic Arts: ISEA2024 Everywhen Proceedings. Volume 1: Academic Papers*, Gavin Sade, Andrew Brown, Leah Barclay, Jen Seevinck, Anastasia Tyurina, and Rewa Wright (Eds.). Queensland University of Technology and ISEA International, Brisbane, Qld, 476–483. doi:10.5204/book.eprints.256296

[19] Tiancheng Liu, Jiayi Ye, Shumeng Zhang, Kang Zhang, and Chen Liang. 2025. Quantifying Structural Aesthetic Features and Personality Trait Preferences in Kai Shu Calligraphy. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 6730–6739. doi:10.1145/3746027.3754816

[20] Tiancheng Liu, Shumeng Zhang, and Nicolò Merendino. 2026. BruSHŪ: Cross-Modal Translation of Implicit Micro-Actions in Chinese Calligraphy. *Proc. ACM Comput. Graph. Interact. Tech.* doi:10.1145/3816090

[21] Piera Riccio, Francesco Galati, Kajetan Schweighofer, Noa Garcia, and Nuria M Oliver. 2026. Imageset2text: Describing sets of images through text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 8731–8739.

[22] Lisa Saleh and Nanne van Noord. 2022. The Computational Memorability of Iconic Images. *Computational Humanities Research Conference* (2022).

[23] Gjorgji Strezoski and Marcel Worring. 2018. OmniArt: A Large-scale Artistic Benchmark. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 4, Article 88 (Oct. 2018), 21 pages. doi:10.1145/3273022

[24] Nanne van Noord. 2022. A Survey of Computational Methods for Iconic Image Analysis. *Digital Scholarship in the Humanities* (2022).

[25] Nanne van Noord and Noa Garcia. 2025. The Iconicity of the Generated Image. arXiv:2509.16473 [cs] doi:10.48550/arXiv.2509.16473

[26] Nanne van Noord, Ella Hendriks, and Eric Postma. 2015. Toward Discovery of the Artist's Style: Learning to recognize artists by their artworks. *IEEE Signal Processing Magazine* 32, 4 (2015), 46–54. doi:10.1109/MSP.2015.2406955

[27] Nanne van Noord, Melvin Wevers, Tobias Blanke, Julia Noordegraaf, and Marcel Worring. 2022. An Analytics of Culture: Modeling Subjectivity, Scalability, Contextuality, and Temporality. In *NeurIPS Workshop on Cultures in AI/AI in Culture*. arXiv:2211.07460 [cs]

[28] Shuai Wang, Ivona Najdenkoska, Hongyi Zhu, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2025. ArtRAG: Retrieval-Augmented Generation with Structured Context for Visual Art Understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6700–6709.

[29] Shuai Wang, Hongyi Zhu, Jia-Hong Huang, Yixian Shen, Chengxi Zeng, Stevan Rudinac, Monika Kackovic, Nachoem Wijnberg, and Marcel Worring. 2026. A-MAR: Agent-based Multimodal Art Retrieval for Fine-Grained Artwork Understanding. *arXiv preprint arXiv:2604.19689* (2026).

[30] Qi Wen, Shuang Li, Bingfeng Han, and Yi Yuan. 2021. Zigan: Fine-grained chinese calligraphy font generation via a few-shot style transfer approach. In *Proceedings of the 29th ACM international conference on multimedia*. 621–629.

[31] Graffiti Works. [n. d.]. Art from Calligraphy. *Taylor & Francis* ([n. d.]), 165.

[32] Yankun Wu, Zakaria Laskar, Giorgos Kordopatis-Zilos, Noa Garcia, and Giorgos Tolias. 2025. Instance-Level Generation for Representation Learning. *arXiv preprint arXiv:2510.09171* (2025).

[33] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. 2021. The Met Dataset: Instance-level Recognition for Artworks. In *NeurIPS Track on Datasets and Benchmarks*.

[34] Pepe Ballesteros Zapata. 2024. Evolving Methodologies: Computation in Art History. *Hertziana Studies in Art History* 3 (2024).

[35] Aven Le Zhou, Jiayi Ye, Tiancheng Liu, and Kang Zhang. 2024. Archiving Body Movements: Collective Generation of Chinese Calligraphy. In *Proceedings of the 29th International Symposium on Electronic Arts: ISEA2024 Everywhen Proceedings. Volume 1: Academic Papers*, Gavin Sade, Andrew Brown, Leah Barclay, Jen Seevinck, Anastasia Tyurina, and Rewa Wright (Eds.). Queensland University of Technology and ISEA International, Brisbane, Qld, 862–870. doi:10.5204/book.eprints.256296