VORTEX: Vision Transformers for Interpretable Temporal Dating of Historical Paintings Through Ordinal Classification

David Song
University of Washington
1410 NE Campus Parkway Seattle, WA
davsong@uw.edu

Vibhav Peri
vperi@uw.edu

Abstract

This research addresses predicting exact creation years for Western paintings from 1600-1899. We propose VORTEX, combining Meta's DINOv2 Vision Transformer with Low-Rank Adaptation (LoRA) for efficient fine-tuning across three centuries of art. Our key contribution emphasizes interpretability through attention mechanism analysis, revealing how the model identifies period-specific visual cues that align with established art historical knowledge. The attention visualizations provide art historians with interpretable evidence for temporal attribution, bridging computational methods with traditional scholarly practices. We evaluate our approach on a dataset aggregated from Joconde, WikiArt, Web Gallery of Art, and Rijksmuseum collections. Results demonstrate significant improvements over traditional CNN approaches, OmniArt framework, and state-of-the-art multimodal models including Gemini 2.0 Flash, achieving mean absolute errors suitable for practical curatorial applications.

1. Introduction

The precise temporal placement of artworks represents a fundamental challenge in art history scholarship [14]. Accurate dating influences virtually every aspect of research, from understanding individual artistic development to tracing the evolution of broader cultural movements [6]. When researchers can confidently assign specific creation years to paintings, they unlock insights into how artistic innovations emerged, spread across geographic boundaries, and influenced subsequent generations of artists.

Despite centuries of scholarly effort and the recent acceleration of museum digitization initiatives, a substantial portion of the world's artistic heritage remains imprecisely dated. Major cultural institutions continue to struggle with temporal attribution, often cataloging significant works with vague temporal ranges. Even world-renowned institutions such as the Louvre routinely catalog paintings with broad temporal labels—for instance, "first half of the 19th century"—rather than specific years.[1] This imprecision creates cascading challenges for digital humanities research, hampering quantitative analyses of stylistic evolution, complicating provenance verification, and limiting our ability to understand the precise chronological relationships between artworks.

1.1. Current Challenges in Art Historical Dating

The traditional toolkit for artwork dating encompasses several complementary approaches, each with inherent limitations. Connoisseurship, which relies on deep expertise in period styles and individual artistic hands, remains subjective and often contentious among experts. Documentary evidence from contracts, correspondence, or contemporary records frequently proves incomplete or ambiguous. Scientific analysis techniques, including dendrochronology for panel paintings and chemical analysis of pigments, require significant resources and often yield broad date ranges rather than specific years. Moreover, curators still need to invest substantial manual effort, often requiring years of specialist training, to narrow broad time ranges into single years—a process that remains labor-intensive and difficult to scale across large collections.

The emergence of computational methods offers the potential to augment these traditional approaches with data-driven analysis. However, previous computational approaches to art analysis have primarily focused on related but distinct problems such as style classification, artist attribution, and visual similarity retrieval. While these efforts have yielded valuable tools for art historians, they have not directly addressed the challenge of precise temporal dating with the interpretability necessary for scholarly acceptance.

[1] See, for example, https://collections.louvre.fr/en/ark:/53355/cl010064491.

1.2. Objectives and Technical Approach

We introduce VORTEX (Vision Ordinal Regression Temporal EXtraction) to address the unique challenges of exact year prediction for historical paintings. Our approach targets Western paintings from 1600 to 1899, a period encompassing major artistic movements from Baroque through early Impressionism. We formulate three core objectives that guide our technical development:

(1) **Precise temporal prediction**: Develop a model capable of predicting exact creation years with mean absolute errors suitable for practical curatorial applications, targeting accuracy within ± 5 years for a significant portion of predictions [8].

(2) **Interpretable decision-making**: Implement attention-based visualization techniques that reveal which visual elements inform temporal predictions, providing art historians with evidence they can evaluate against established scholarly knowledge.

(3) **Computational efficiency**: Employ parameter-efficient training methods to make the approach practical for institutions with limited computational resources, enabling broader adoption across the cultural heritage sector.

To achieve these objectives, we leverage DINOv2 as our backbone, implement LoRA for parameter-efficient fine-tuning, and formulate year prediction as an ordinal classification problem using CORAL.

1.3. Expected Impact

From a technical perspective, we demonstrate that Vision Transformers achieve unprecedented precision in temporal art analysis when combined with appropriate task formulations. The integration of LoRA shows how large-scale pre-trained models can be efficiently specialized for domain-specific tasks.

From an art historical perspective, our emphasis on interpretability through attention analysis provides a crucial bridge between computational predictions and scholarly practices. By visualizing which visual elements inform temporal decisions, we enable art historians to evaluate and validate model predictions against their domain expertise. This transparency is essential for building trust in computational methods within humanistic disciplines.

We anticipate that successful demonstration of precise, interpretable temporal dating will convince museums and cultural institutions to employ such tools to assist with cataloging unlabeled works, validate existing attributions, and identify pieces warranting further scholarly investigation. The ability to perform large-scale temporal analysis across collections could reveal previously unrecognized patterns in artistic development and cultural exchange.

2. Related Work

The application of computer vision techniques to art historical problems has evolved significantly over the past two decades, with most research focusing on painting attribute prediction tasks. While attribute prediction has proven valuable for organizing and searching art databases, the challenge of precise temporal dating requires fundamentally different approaches that can capture subtle chronological progressions rather than broad categorical distinctions.

2.1. Evolution of Computer Vision in Art Analysis

Traditional Methods Early computational approaches to art analysis relied on hand-crafted features such as SIFT, HOG, and custom color space analyses for tasks including style classification and artist identification [10]. While these approaches demonstrated the feasibility of computational art analysis, they required extensive domain expertise and often failed to capture the subtle complexities that distinguish artistic styles and periods.

Convolutional Neural Networks Karayev et al. [10] demonstrated that deep features from AlexNet could effectively recognize artistic styles, establishing the foundation for neural network applications in art history. The OmniArt framework [17, 18] advanced this paradigm through multi-task learning architectures that could simultaneously analyze multiple artistic attributes including period classification. Complementary work demonstrated the effectiveness of transfer learning [4, 11] and developed specialized correlation features for period identification [3].

Visual Transformers The recent adoption of Transformer architectures in computer vision offers fundamental advantages for artistic analysis through global self-attention mechanisms. The DINOv2 models [15], trained through self-supervised learning on massive visual datasets, have demonstrated exceptional transfer capabilities particularly suitable for specialized domains where annotated examples are scarce.

2.2. Temporal Analysis in Art History Datasets

To date, no prior research has focused specifically on predicting exact creation years for historical paintings at the level of precision required for scholarly applications. However, several key datasets have included temporal year prediction as a subtask within broader art analysis challenges.

The Rijksmuseum Challenge [12], introduced in 2014, represented the first large-scale attempt to benchmark temporal prediction for artworks. This dataset provided 112,000 images from the Rijksmuseum collection spanning diverse media including paintings, sculptures, decorative arts, and prints, with various metadata including cre-

ation years, establishing baseline performance metrics for the field. Initial approaches using traditional computer vision methods achieved Mean Absolute Errors of approximately 72 years, highlighting the difficulty of precise temporal prediction across heterogeneous artistic media.

Building upon this in 2018, the next iteration OmniArt Challenge [18] significantly advanced the field by aggregating multiple sources into a unified multi-task learning framework. With 432,000 artworks annotated for various attributes including temporal information, OmniArt enabled researchers to explore how different visual tasks relate to temporal prediction. The best-performing models on OmniArt achieved MAEs of 70.1 years for temporal prediction, using ResNet-50 architectures within multi-task learning frameworks.

3. Dataset

Our dataset construction process aggregates art historical data from multiple institutions while maintaining consistent formatting and standards across them all.

3.1. Source Aggregation

To construct our dataset, we carefully evaluated prior work to discover datasets that meet our research scope and metadata requirements. We aggregated paintings from four major institutional and online sources 1.

(1) The Joconde database, maintained by the French Ministry of Culture, provides metadata for approximately 600,000 artworks from French museums with reliable scholarly annotations [13].

(2) WikiArt contributes broad stylistic diversity through its user-curated collection spanning multiple movements and geographic regions [10, 20].

(3) The Web Gallery of Art specializes in European paintings with particular strength in Renaissance through 19th-century works [16, 19].

(4) The Rijksmuseum collection offers meticulously documented Dutch and Flemish paintings with precise dating information derived from extensive provenance research [12].

This multi-source approach mitigates collection biases inherent in any single institution while maximizing coverage across our 300-year temporal range. It enables the model to learn more robust temporal features from diverse examples of how different regions and schools evolved during the same time periods.

3.2. Data Processing

Filtering Our dataset artwork inclusion criteria require that works be explicitly identified as paintings (encompassing oil, tempera, watercolor, and mixed media techniques), originate from European or North American contexts, include a digitalized image, and possess precise year annotations rather than period designations. We exclude drawings, prints, sculptures, and decorative arts to maintain focus on painted works where temporal stylistic evolution follows consistent patterns.

Standardization Our standardization process addresses both metadata consistency and visual uniformity across diverse source collections. For date standardization, we resolve the variety of dating conventions found in art historical documentation through the following strategies: single years are used directly; date ranges (e.g., "1650-1655") are resolved to their midpoint; circa dates (e.g., "c. 1700") use the specified year; and when both start and completion dates are provided, we use the completion date as most representative of the work's final appearance. Works with dating uncertainty exceeding ± 10 years are excluded to maintain precision in our ground truth labels. For image standardization, we resize all artworks to $XXX \times XXX$ pixel resolution, ensuring consistent input dimensions while preserving aspect ratios through appropriate padding or cropping strategies 1d.

Deduplication Following the methodology established by Mao et al. [11], we encode all images using MD5 hashing to identify and remove duplicate artworks that may appear across multiple source collections, ensuring that each painting appears only once in our final dataset.

3.3. Sampling Strategy

The temporal distribution of our aggregated dataset reveals expected patterns reflecting both historical factors and contemporary digitization priorities. Earlier centuries show lower representation due to fewer surviving works and selective digitization focusing on major artists. The 19th century demonstrates significantly higher representation, particularly for Impressionist and Post-Impressionist works that attract substantial public interest and digitization investment 1a.

Importantly, we make a deliberate decision not to artificially balance the temporal distribution through downsampling. Real-world applications will encounter similar biases, with cultural institutions holding proportionally more recent works. Training on this natural distribution ensures our model's performance metrics accurately reflect expected deployment scenarios. We partition the dataset using an 80/10/10 split for training, validation, and test-

Table 1. Dataset comparison across relevant features for year prediction of artworks.

Feature Dataset	Wiki Art	WGA	Joconde	Rijks	All Datasets
Number of Artworks	28997	10000	11954	1506	52457
Publicly Available	✓	✓	✓	✓	✓
Year Label Granularity †	0.2861	0.0096	0.0	0.0	0.1600
Earliest Year	1600	1600	1600	1600	1600
Latest Year	1899	1899	1899	1899	1899
Geographic Scope	Western Europe/America	Western Europe/America	Predominantly France	Western Europe	Western Europe/America

† Average year 'window' range

ing respectively. Stratified sampling by year ensures that each annual cohort maintains consistent proportions across all splits.

All dataset construction adheres to source licensing requirements, with metadata preserved to ensure reproducibility while respecting copyright constraints on image redistribution.

4. Methodology

Our technical approach combines vision transformer architectures with parameter-efficient training strategies and ordinal classification for precise year prediction.

4.1. Vision Transformer Foundation

We employ Meta's DINOv2-Base model [15] as our foundational feature extractor. The model processes 16×16 pixel patches through 12 transformer layers with 768-dimensional hidden states and 12 attention heads per layer.

Vision Transformers (ViTs) often beat CNNs on artwork analysis because their self-attention sees the whole canvas at once, capturing long-range stylistic relationships (composition, color balance, recurring motifs) that local convolutional filters can miss. Pre-trained on huge image corpora via self-supervision (e.g., MAE, DINO v2), ViTs transfer well to small art datasets and integrate seamlessly with text or metadata tokens for multi-modal work [5, 15].

4.2. Parameter-Efficient Fine-Tuning with LoRA

We implement LoRA [9] to address the computational challenges of fine-tuning on specialized art historical datasets. For each weight matrix $W_0 \in \mathbb{R}^{d\times k}$, we introduce low-rank adaptations:

$$W = W_0 + \Delta W = W_0 + BA$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with rank $r \ll \min(d, k)$.

4.3. Ordinal Classification Framework

We recognize year prediction as fundamentally an ordinal problem and adopt the CORAL framework [2]. For years $y_1 < y_2 < ... < y_{300}$ spanning 1600 to 1899, we predict:

$$P(Y > y_k|X) \text{ for } k = 1, ..., 299$$

CORAL ensures rank consistency through weight sharing across binary classifiers while maintaining individual bias terms, guaranteeing $P(Y > y_k) \geq P(Y > y_{k+1})$ for all k. During inference, we determine the predicted year by identifying the transition point where predictions shift from positive to negative.

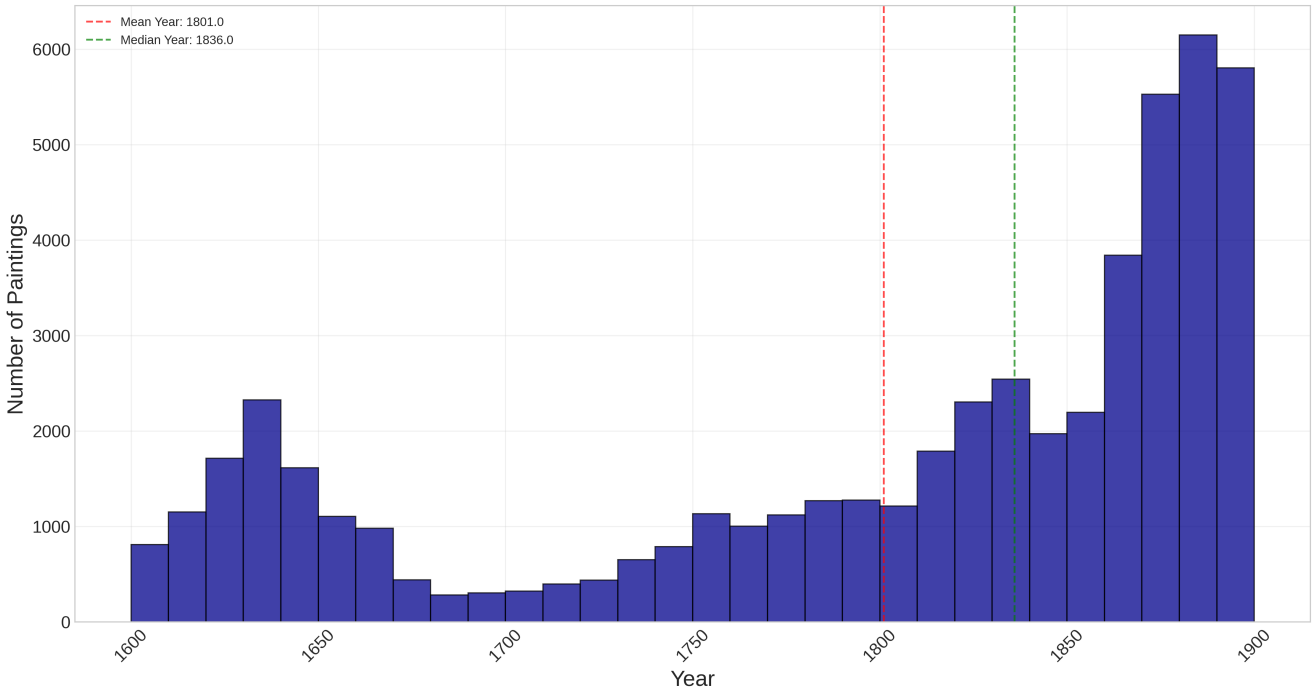
4.4. Training Configuration

To augment our dataset, we employ conservative augmentation strategies that avoid erasing any subtle stylistic cues. Spatial augmentations include random horizontal flips (many paintings have no inherent orientation) and slight rotations within ± 5 degrees to account for digitization variations. We specifically avoid aggressive color augmentations, as color palettes often provide strong temporal signals—the bright blues of early Baroque differ markedly from the earth tones of later Realism.

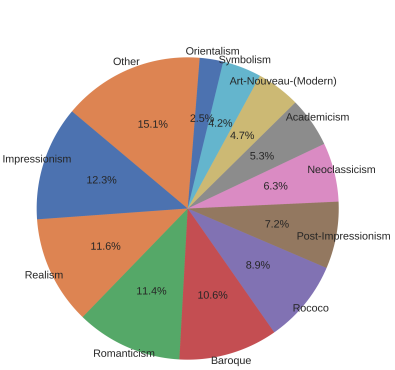
Optimization employs AdamW with cosine learning rate scheduling and linear warmup over 5% of training steps. We initialize at 1e-4 for LoRA parameters while keeping the base model frozen. Training continues for 50-100 epochs with early stopping based on validation MAE.

4.5. Interpretability Through Attention Analysis

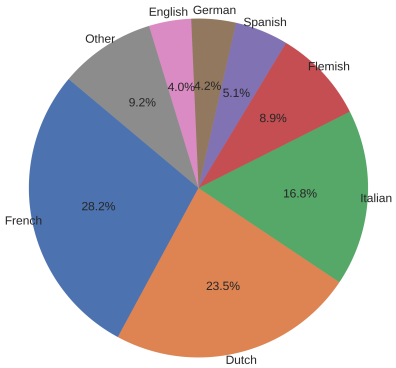
We employ Attention Rollout [1] to aggregate attention weights across transformer layers, producing interpretable attention maps. This addresses several key questions about model behavior. First, do attention patterns align with established art historical knowledge about period-specific characteristics? Second, does the model discover novel visual cues that might inform temporal classification beyond traditional scholarly analysis? Third, how do attention patterns differ between successful predictions and failure cases? By examining these patterns across diverse examples, we build understanding of both model capabilities and limitations, providing the transparency essential for scholarly acceptance of computational methods.



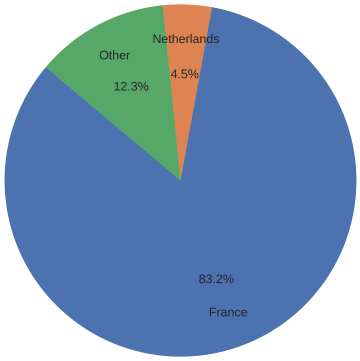
(a) Distribution of Western Paintings Collected Across all Datasets by Decade (1600-1900)



(b) WikiArt 'Style' Pie Chart



(c) WGA 'School' Pie Chart



(d) Joconde 'Ecode/Pays' Pie Chart

(e) Example Images Collected Over Multiple Years

Figure 1. (a) This shows the amount of images we have collected over all datasets per decade between 1600 and 1900. Also displays the mean year (1801) and the median year (1837) (b) Shows the 'style' classification of the images collected from WikiArt. The pie chart was created only considering images from WikiArt that fit our temporal and geographical filters. (c) Shows the 'school' classification of all images collected from WGA (d) Shows the school/country classification of all images collected from Joconde. and (e) An example of a few images collected in our dataset across the years. The year displayed is the year the painting was created or the median year if there was a range of < 10. You can clearly see a shift in style and subjects in these images as centuries pass.

5. Experimental Evaluation

Our experimental evaluation comprehensively assesses VORTEX's performance across multiple dimensions, from quantitative accuracy metrics to qualitative analysis of model behavior. We aim to determine whether VORTEX can perform better and provide better insights than existing commercial AI solutions.

5.1. Evaluation Metrics

Our experimental evaluation employs a comprehensive framework designed to assess both the practical utility and comparative performance of our approach. We adopt evaluation metrics that directly reflect the needs of art historical applications while enabling meaningful comparisons with existing methods.

The primary evaluation metric is Mean Absolute Error (MAE) measured in years, providing an intuitive measure of average dating precision. An accuracy within ± 5 years is considered acceptable for precise scholarly work.

5.2. Baseline Comparisons

We evaluate VORTEX against several strong baselines representing different methodological approaches:

OmniArt Challenge: We run the OmniArt model [18], using ResNet-50 with task-specific heads. While OmniArt originally included multiple prediction tasks across different artwork mediums, we focus solely on temporal prediction of paintings for fair comparison.

Gemini 2.0 Flash: We evaluate Google's state-of-the-art multimodal model [7] using the following prompt to request specific year predictions.

> *User: Closely examine this Western European painting. Only consider the painting itself. DO NOT USE ANY METADATA. Think carefully about what artistic movement it could be a part of and who the painter could be. Using these two attributes and any additional details about the painting, predict the exact year it was painted.*

This comparison benchmarks our specialized approach against general-purpose AI systems.

5.3. Quantitative Results

5.4. Interpretability Analysis

6. Discussion

6.1. Implications for Art History Scholarship

6.2. Limitations and Failure Modes

6.3. Future Work

7. Conclusion

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4190–4197, 2020. 4

[2] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. 4

[3] Wei-Ta Chu and Yi-Ling Wu. Image style classification based on learnt deep correlation features. *IEEE TMM*, 20(9):2491–2502, 2018. 2

[4] Elliot J. Crowley and Andrew Zisserman. The art of detection. In *ECCV*, pages 721–737, 2016. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. arXiv preprint arXiv:2010.11929, 2020. 4

[6] Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny, and Marian Mazzone. The shape of art history in the eyes of the machine. In *AAAI*, pages 2183–2191, 2018. 1

[7] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[8] Getty Vocabulary Program. Categories for the description of works of art (CDWA): Creation date. https://www.getty.edu/research/publications/electronic_publications/cdwa/definitions.pdf, Accessed 2024. Guidelines for documenting artwork creation dates. 2

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. arXiv preprint arXiv:2106.06823, 2021. 4

[10] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *BMVC*, 2014. Based on arXiv:1311.3715v3 [cs.CV] 23 Sep 2014. 2, 3

[11] Hui Mao, Ming Cheung, and James She. DeepArt: Learning joint representations of visual arts. In *ACM MM*, pages 1183–1191, 2017. 2, 3

[12] Thomas Mensink and Jan van Gemert. The Rijksmuseum Challenge: Museum-centered visual recognition. In *ICMR '14: Proceedings of the International Conference on Multimedia Retrieval*, pages 451–454. ACM, 2014. 2, 3

[13] Ministère de la Culture. Joconde: Conditions d'utilisation. https://www.culture.gouv.fr/Mentions-legales, Accessed 2024. Terms of use for the Joconde database. Actual specific terms page for data re-use should be confirmed. 3

[14] Alec Mishory. *Art history: an introduction*. Open University of Israel, 2000. 1

[15] Maxime Oquab, Timothée Durand, Jakob Verbeek, Hervé Jégou, and Armand Joulin. DINOv2: Learning robust visual features without supervision, 2023. arXiv:2304.07193. 2, 4

[16] Benoit Seguin, Carlotta Striolo, Isabella diLenardo, and Frederic Kaplan. Visual link retrieval in a database of paintings. In *ECCV*, volume 9913 of *Lecture Notes in Computer Science*, pages 753–767, 2016. ECCV 2016 Workshops, Part I. 3

[17] Gjorgji Strezoski and Marcel Worring. OmniArt: Multi-task deep learning for artistic data analysis. *arXiv preprint arXiv:1708.00684*, 2017. Version: v1 [cs.MM] 2 Aug 2017. 2

[18] Gjorgji Strezoski and Marcel Worring. OmniArt: A large-scale artistic benchmark. *IEEE TMM*, 14(4):88:1–88:21, 2018. This is likely the TOMM reference based on user's table. Actual TOMM page numbers would be different from an arXiv version. 2, 3, 6

[19] Web Gallery of Art. Web gallery of art: License and copyright information. https://www.wga.hu/licence.html, Accessed 2024. License information for the Web Gallery of Art. 3

[20] WikiArt. WikiArt: Terms of use. https://www.wikiart.org/en/terms-of-use, Accessed 2024. Terms of use for WikiArt.org. 3