Graph Neural Networks for Knowledge Enhanced Visual Representation of Paintings

Efthymiou, A.; Rudinac, S.; Kackovic, M.; Worring, M.; Wijnberg, N.

DOI
10.1145/3474085.3475586

Publication date
2021

Document Version
Final published version

Published in
MM '21

License
Article 25fa Dutch Copyright Act

Citation for published version (APA):
Graph Neural Networks for Knowledge Enhanced Visual Representation of Paintings

Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Marcel Worrin, Nachoem Wijnberg
University of Amsterdam
Amsterdam, The Netherlands
{a.efthymiou,s.rudinac,m.kackovic,m.worring,n.wijnberg}@uva.nl

ABSTRACT
We propose ArtSAGENet, a novel multimodal architecture that integrates Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs), to jointly learn visual and semantic-based artistic representations. First, we illustrate the significant advantages of multi-task learning for fine art analysis and argue that it is conceptually a much more appropriate setting in the fine art domain than the single-task alternatives. We further demonstrate that several GNN architectures can outperform strong CNN baselines in a range of fine art analysis tasks, such as style classification, artist attribution, creation period estimation, and tag prediction, while training them requires an order of magnitude less computational time and only a small amount of labeled data. Finally, through extensive experimentation we show that our proposed ArtSAGENet captures and encodes valuable relational dependencies between the artists and the artworks, surpassing the performance of traditional methods that rely solely on the analysis of visual content. Our findings underline a great potential of integrating visual content and semantics for fine art analysis and curation.

CCS CONCEPTS
• Computing methodologies → Neural networks; Image representations; Multi-task learning; Semantic networks; • Applied computing → Fine arts.

KEYWORDS
multimodal modelling, multi-task learning, graph neural networks, automated art curation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION
Fine art analysis has been a subject of intensive research in recent years. Advances in multimedia and related disciplines and the vast amount of publicly available artistic data have encouraged research in many fine art analysis tasks, ranging from artistic style classification and creation period estimation [15, 38, 49, 50], style transfer [19, 25], object detection and retrieval in paintings [13, 20] to identification of semantic relationships between the artworks [6, 7]. However, most extant research concentrates almost exclusively on visual content analysis or in some cases on semantic associations. In this study, we focus on capturing and modeling the complex visual and semantic relationships between artists and artworks to gain a deeper and more comprehensive understanding of paintings.

Even though the visual content is a predominant characteristic of a visual artwork, automatic fine art analysis can significantly benefit from the well-established fundamental theories in the art domain, as well as the rich semantic information associated with the artworks. Personal links between painters, joint membership of artistic schools,
and shared exhibitions are possible ties in such networks. By connecting visual content and semantic information representing such network relations, it is possible to better understand the processes of artistic innovation and influence. For instance, our understanding of Pablo Picasso’s proto-cubist seminal painting *Les Demoiselles d’Avignon* (1907), depicted in Figure 1, is greatly enhanced by recognizing the influence of Doménikos Theotokópoulos’ (better known as El Greco) *Opening of the Fifth Seal* (1608–1614) (upper left node in Figure 1), or Édouard Manet’s *Olympia* (1863) (lower right node) [9].

Recent developments in structural modelling and the emergence of Graph Neural Networks (GNNs) [23, 31, 60] enable us to model these interesting properties and relationships. Nevertheless, most of the earlier GNNs fail to scale to domains that consist of large graphs with thousands of nodes and edges. In this work, we employ recently developed efficient GNN methods that can scale to large graphs with thousands of nodes and millions of edges [12, 23, 41, 60] and show that they are superior to traditional Convolutional Neural Networks (CNNs) in terms of predictive performance and computational efficiency for fine art analysis. To that end, we propose ArtSAGENet, a novel approach that extends GNNs and CNNs to jointly learn visual and semantic representations of fine art using the visual content of an artwork accompanied with its respective semantic relationships.

Following the rapidly growing body of research [32, 44, 51], we, first, employ a Multi-task Learning (MTL) [4] approach to learn visual representations of fine art. We argue that, next to the significant advantages of MTL in terms of computational resources, machines can naturally understand the influence of Doménikos Theotokópoulos’ (better known as El Greco) *Opening of the Fifth Seal* (1608–1614) (upper left node in Figure 1), or Édouard Manet’s *Olympia* (1863) (lower right node) [9].

The main contributions of this work are the following:

- We systematically evaluate multi-task learning for fine art analysis and confirm its significant advantages over single-task alternatives in a wide range of settings.
- We employ several Graph Neural Networks directly for a wide range of fine art analysis tasks by coupling visual features and homogeneous topological structure.
- We propose ArtSAGENet, a novel architecture for integrating semantic information and visual content.
- Through extensive experimentation we show that our proposed ArtSAGENet consistently outperforms strong traditional methods on style classification, artist attribution, creation period estimation and tag prediction.

The rest of this paper is structured as follows. In the next section, we review the relevant literature in automated fine art analysis, and thereafter, in Sections 3 and 4 we provide all the necessary details for our proposed methods and the experimental setup. Finally, in Section 5 we present the experimental results, and in Section 6 we conclude discussing our findings.

## 2 RELATED WORK

The recent successes in deep learning for multimedia have inspired research in fine art classification based on visual [8, 15, 37, 38, 48] and semantic content [7, 17, 18, 29]. In this section we briefly survey both research lines as well as multimodal approaches that integrate visual content and semantics.

### 2.1 Visual Content-Based Fine Art Analysis

Visual content-based fine art analysis has attracted a lot of attention following the recent advancements in computer vision and the emergence of large-scale visual art collections that contain paintings, sculptures, photographs and installations, such as the Rijksmuseum [38], DeepArt [37], BAM! [55] and OmniArt [49] datasets. Strezoski et al. [46] develop an interactive method to explore visual art collections using colors as visual cues, while the web-based platform proposed in [47] facilitates interactive exploration of the visual sentiment and emotion in paintings over time. Shen et al. [42] propose self-supervised spatially-consistent feature learning to discover near duplicate patterns in large collections of artworks. Chen and Yang [10] learn latent style representations inspired by the Gram matrix based correlation calculation using the VGG architecture [43] and evaluate their proposed methods using the Painting91 [26], arData [58] and Hipster Wars [27] datasets.

Bianco et al. [2] propose a Spatial Transformer Network to identify the most discriminative sub-regions of paintings by employing CNNs and evaluate the performance of their approach by constructing the MultiTaskPainting100k dataset. Similarly, Cetic et al. [6] utilize CNNs to study aesthetics, memorability and sentiment in paintings, while Cetinic et al. [7] and Elgammal et al. [15] explore the WikiArt collection to quantify stylistic properties according to Heinrich Wolfflin’s concepts [56]. Strezoski and Worring [48] employ a multi-task setting for several fine art classification tasks ranging from artist attribution and type prediction to period estimation. Tan et al. [50] utilize AlexNet and evaluate its performance on the artist attribution task amongst others. In contrast to these works, we don’t rely solely on traditional CNN architectures, but we model complex semantic-based artistic relationships that go beyond visual content to facilitate context-aware art analysis.

### 2.2 Semantic Content-Based Fine Art Analysis

There is an emerging interest in studying semantic principles of the fine art history [7, 29]. Kim et al. [28] utilize CNNs trained for style classification and study the intermediate layer’s activations for semantic interpretation. Garcia and Vogiatzis [18] develop SemArt and propose the Text2art challenge to evaluate semantic art understanding. In their work, they collect artistic data from the Web Gallery of Art (WGA) containing a textual description of each painting as well as information about the author.
2.3 Multimodal Representations

Recently, multimodal modeling for integrating visual content and semantics has gained a lot of traction. Our work is related to that of Chen et al. [11]. They propose multi-label Graph Convolutional Neural Networks by utilizing a multimodal architecture that jointly learns image representations and object label inter-dependencies using Graph Convolutional Networks [31], achieving state-of-the-art performance in object recognition tasks. In contrast to that, here we employ recently developed efficient Graph Neural Networks [31], which are trainable and fine-tuned to obtain the final visual representation. Finally, the obtained representations are merged and the final multimodal representation is passed to the last layer (classifier or regressor).

GraphSAGE (a-c) and ResNet-152 (e) are jointly trained for all three tasks in an MTL manner.

Figure 2: This figure depicts the ArtSAGENet architecture. Given a batch of images, the forward propagation is implemented as follows. GraphSAGE: (a) For each image in the batch we sample $k$ neighbors, e.g., $k = 3$, from $h$ hops, e.g., $h = 2$, (b) aggregate the node feature vectors within the neighborhood to (c) obtain the final node representation. ResNet: (a) Each image of the batch is passing through the frozen part of the pre-trained (on ImageNet) network and then (b) it passes through the network's last bottleneck which is trainable and fine-tuned to obtain the final visual representation. Finally, the obtained representations are merged and the final multimodal representation is passed to the last layer (classifier or regressor). GraphSAGE (a-c) and ResNet-152 (e) are jointly trained for all three tasks in an MTL manner.

Stefanini et al. [45] address the problem of cross-modal retrieval of paintings and their associated descriptions, and create the Artpedia dataset that consists of nearly three thousands paintings annotated with contextual sentences. Madhu et al. [36] employ traditional machine learning methods, and show that CNNs can efficiently learn art-specific domain knowledge to recognize human figures in paintings. Zhao et al. [63] propose ArtGCN to learn artistic node representations based on paintings' textual descriptions. They compare their ArtGCN with visual content CNN baselines [24] using SemArt [18], and show that their proposed method can surpass CNNs performance. In contrast, we do not employ explicit textual information, but we focus on the significant advantages of integrating visual content and semantic-based information.

3 LEARNING KNOWLEDGE-ENHANCED VISUAL REPRESENTATIONS

Traditional CNNs have been proven extremely effective in fine art analysis. However, the visual arts domain is characterized by an exceptional information richness, including e.g. the explicit and implicit social networks between artists. To that end, in this work we take a multimodal approach to enhance the performance of the conventional visual content-based CNNs. Figure 2 depicts ArtSAGENet-a two-branch architecture that jointly learns a CNN and a GNN-based model to learn multimodal deep context-aware visual representations of paintings.

3.1 Visual Representation Learning

Even though an arbitrary CNN can be deployed for learning visual representations, in this work we are adopting the ResNet-152 [24] architecture due to its excellent performance, choice which we will be further motivated in Section 5.1. Adopting the same notation as in [11], for a given painting, we can obtain a visual representation
where \( \mathbf{P} \) denotes the painting’s photographic reproduction, \( F_{\text{GAP}} \) denotes the global average pooling operation, \( \theta_{\text{CNN}} \) denotes the ResNet-152, or any other CNN-based model parameters, and \( D \) denotes the visual embedding dimensionality, i.e. 2.048. We fine-tune the last bottleneck of a pre-trained on ImageNet [14] ResNet-152, albeit we omit the classifier. For either classification or regression tasks, we merge the learned visual and contextual embeddings to obtain the final multimodal painting representation.

### 3.2 Graph Representation Learning

Recently, Graph Neural Networks have gained popularity, because of their outstanding performance in node classification, graph classification and link prediction tasks. Given the nature of the visual arts domain, semantic knowledge can be of great leverage when analyzing artists and artworks. We conjecture that information about, e.g. social networks between the artists, if encoded properly, can be exploited by GNNs for improved fine art analysis.

To this end, we propose a novel method of classifying artwork attributes following the node classification paradigm, as shown in Figure 1. That is, we treat each artwork as a node with links to other artworks/nodes based on their semantic properties.

Early GNN methods usually operate on an adjacency matrix \( \mathbf{A} \) that encodes these relations between nodes in a full-batch setting. This significantly increases the computation complexity that hinders scaling to large graph structures. Recent GNNs propose a number of sampling methods that alleviate the latter issue. Here, we employ such GNN architectures \([12, 23, 41, 60]\) that can scale to homogeneous graph structures with thousands of nodes and millions of edges.

**Graph construction.** Before training the aforementioned GNNs, we need to construct a predefined adjacency matrix \( \mathbf{A} \). We build the adjacency matrix \( \mathbf{A} \) using the following:

\[
\mathbf{A}_{ij} = \begin{cases} 
1, & \text{if } \text{property}(\text{artwork}_i) = \text{property}(\text{artwork}_j), \\
0, & \text{otherwise}.
\end{cases}
\]

Since, most of the GNN methods that we adopt were originally developed for learning in undirected homogeneous graph structures, we build an undirected homogeneous graph accordingly. We leave working with recently proposed GNNs \([53, 61]\) that can leverage structural heterogeneity for future research. We follow \([1, 17]\) and use artistic schools to link the artworks nodes, i.e. all of the available Paul Cézanne’s paintings will be linked with the ones of Édouard Manet’s given that both artists represent the French school of painting. We follow the same practice as in \([23]\) and downsample the edges in the original graph, so that nodes can have at most a degree of 128.

For our ArtSAGENet architecture we utilized the GraphSAGE architecture \([23]\) to obtain context-aware node representations. Figure 2 illustrates the proposed method. That is, given a node we uniformly sample \( k \) neighbors from \( h \) hops and aggregate the neighbourhood’s node feature vectors to obtain the node representation. We use a mean aggregator and obtain the node representation \( n_i \) as follows:

\[
n_i = \frac{1}{k} \sum_{j=1}^{k} f_{\text{AGG}} (x_j, \mathbf{A}) \forall j \in \mathcal{N}(i)
\]

where \( x_i \) is the node feature vector, \( \mathcal{N}(i) \) denotes the neighborhood for the node \( i \), \( x_j \) the node feature vector of the neighbor \( j \), \( f_{\text{AGG}} \) the neighbors feature vectors aggregator, while \( W_1 \) and \( W_2 \) the learned weights. We followed \([23]\) and set the number of hops \( h = 2 \) with neighborhood sample sizes \( k_1 = 25 \) and \( k_2 = 10 \) for the hops \( h_1 \) and \( h_2 \), respectively.

We considered two methods to obtain the node feature vectors. **Visual features.** First, we utilize visual representations as node features for the proposed GNN model in a multimodal fusion manner. That is, prior to training the GNNs, we utilize a pre-trained on ImageNet and frozen ResNet-34 architecture as a backbone to extract 512-dimensional paintings visual feature vectors. In particular, we extract the features after the last convolutional layer. Thereafter, we train the proposed GNNs using the image-level visual feature vector \( v_i \), as computed in Eq. (1).

**Bag-of-words tag feature vectors.** We also consider sparse input features. To this end, we use the painting’s tags as node feature vectors in a bag-of-words manner, i.e. by representing each node as an one-hot encoded vector based on its attributed tags. We collected the tags associated with the artworks from the WikiArt online collection. We considered only the 1,170 tags that appear more than 10 times in the WikiArt collection and introduce a special tag Unknown for paintings with no available tags.

### 3.3 Multimodal Embedding

Given the learned visual and context-aware embeddings, we use a merge operation to get the final knowledge enhanced visual representation. Even though we considered different merging operations, in this work we report results using only the concatenation of the visual and semantic embeddings as follows:

\[
x_i = v_i \odot n_i
\]

Comparative results using alternate algebraic merge operations can be found in supplementary materials.

### 3.4 Multi-Task Learning

Multi-task learning (MTL) is the setting of training an algorithm over multiple tasks and has been shown to be extremely suitable for fine art analysis \([17, 48]\). In this work, we propose and evaluate an MTL setting in order to attribute artworks to stylistic movements, artists and creation periods. We argue, that these three specific tasks are highly cooperative tasks, and therefore MTL can enhance the performance of our proposed methods. We use the following formula for training:

\[
L_T = \sum_{t=1}^{T} w_t L_t
\]

where \( w_t \) denotes the task-specific weight, \( L_t \) the loss for task \( t \) and \( L_T \) the combined loss across all tasks \( T \). In Section 5, we report results using the same task-specific weight \( w_t \) for each task \( t \).

For multi-class classification, we employ the categorical cross-entropy loss:

\[
L = -\frac{1}{N} \sum_{k=1}^{C} \sum_{t=1}^{T} y_{tk} \log \hat{y}_{tk}
\]

where \( N \) denotes the total number of the samples considered, \( C \) the number of total classes, \( y_{tk} \) the ground-truth label, and \( \hat{y}_{tk} \) the output given by the softmax function for the sample \( k \).
We follow the approach in [15] and evaluate AlexNet [33], VGG [43] and ResNet [24] architectures and their variants as CNN baselines. For model implementation, training and evaluation we used the PyTorch library [40]. For all CNN architectures, we are fine-tuning the last convolutional layer alongside the final fully-connected layer(s) by using pretrained versions on ImageNet.

We employ four recently proposed Graph Neural Networks that perform remarkably well in node classification tasks, namely Cluster-GCN [12], GraphSAGE [23], GraphSAINT [60] and SIGN [41], as GNN baselines. We use the same graph as constructed in Eq. (2). In addition, for all GNN baselines we use the paintings visual feature vectors as node representations. We extract these visual feature vectors from the last convolutional layer of a ResNet-34 pre-trained on ImageNet as in Section 3.2. An illustration of the aforementioned approach is shown in Figure 1. For GNN architectures implementation we used the PyTorch Geometric library [16]. We experimented with several configurations for each GNN and always selected the best performing variant.

### 5 RESULTS

In this section, we present the experimental and qualitative evaluation of our proposed method for fine art analysis.

#### 5.1 Fine Art Categorization

ResNet is the dominant CNN architecture. Traditional CNN architectures have been proven to be extremely powerful for visual arts analysis obtaining state-of-the-art performance in style classification [7, 15] and artist attribution tasks [62]. We observe a similar effect as in [15]. That is, for almost every task and dataset variant the dominant CNN is ResNet-152, followed by ResNet-34, while AlexNet seems to be the worst performer. Table 2 summarizes the evaluation results for all models and dataset variants.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>WikiArtFull</th>
<th>WikiArtModern</th>
<th>WikiArtArtists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artworks</td>
<td>75,921</td>
<td>45,869</td>
<td>17,785</td>
</tr>
<tr>
<td>Artists</td>
<td>750</td>
<td>462</td>
<td>23</td>
</tr>
<tr>
<td>Styles</td>
<td>20</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Dates</td>
<td>587</td>
<td>150</td>
<td>240</td>
</tr>
<tr>
<td>Timeframes</td>
<td>13</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Tags</td>
<td>4,879</td>
<td>3,652</td>
<td>2,370</td>
</tr>
</tbody>
</table>

For all models, we use early stopping by monitoring the validation error, e.g. we stop training if there is no improvement in validation loss after ten consecutive epochs. In addition to early stopping, we adopt a dynamic learning rate reducing strategy. That is, we monitor the validation error and reduce the learning rate by a factor of 10 once learning stagnates, i.e. if there is no improvement in validation loss after five consecutive epochs. We experimented with various settings for all models hyperparameters. We train all CNN architectures and our ArtSAGENet using Stochastic Gradient Descent (SGD) [34] with initial learning rate set to 0.001 and momentum to 0.9 with a mini-batch size of 16. For GNNs, we use the Adam optimizer [30] with a learning rate of 0.001 and a mini-batch size of 1,024, but we omit the dynamic learning rate scheduler.

### 4 EXPERIMENTAL SETUP

In this work, we utilize traditional visual content CNNs [24, 33, 43] and recently proposed GNNs [12, 23, 41, 60] as our baselines. In this section, we present the data collection and tasks that we use, and provide all necessary implementation details.

#### 4.1 Dataset Variants and Tasks

We evaluate our proposed methods using the WikiArt online user-editable visual arts collection on four downstream tasks, namely style classification, artist attribution, creation period estimation and tag prediction. For a fair comparison with related works [5, 15, 50] that used the WikiArt collection, we evaluate our proposed methods using the WikiArtFull dataset that consists of 75,921 paintings, and the WikiArtArtists subset that considers only the works of the 23 most representative artists in the WikiArt collection. In addition, we introduce another subset WikiArtModern, that consists of all the artworks from the year 1850 up to 1999, since we are interested in the modern art period as it incorporates the emergence of many major stylistic movements and several groundbreaking moments in the history of fine art. We make use of the WikiArtModern subset to evaluate the performance of our proposed methods in creation year estimation task. For this regression problem, we employ the Mean Absolute Error (MAE) [59] for training. Inspired by research in human age estimation [22, 35, 39], we report the Cumulative Score (CS) [22] evaluation metric, which is defined as follows:

\[
CS(\theta) = \frac{N_0}{N} \times 100, \quad (7)
\]

where \( N \) is the total number of paintings in the test set and \( N_0 \) denotes the number of paintings whose absolute error is less than \( \theta \) years.

Table 1 summarizes the dataset statistics for all the WikiArt dataset variants that we use in this work. Note that in Table 1 timeframes designate half-century periods, such as 1900-1950.

Finally, we evaluate the performance of all methods in a multi-label tag prediction task. To this end, we use a total of 54,919 paintings that are associate with at least one tag that, in turn, appears at least in 1,000 unique paintings. This results in 54 unique tags, ranging from face parts, i.e. forehead and lips, nature related tags, i.e. flowers and animals, urban objects, i.e. vehicles and boats, to themes, i.e. children portraits and famous people. A detailed description of the dataset collection is provided in supplementary materials.

#### 4.2 Baselines

We follow the approach in [15] and evaluate AlexNet [33], VGG [43] and ResNet [24] architectures and their variants as CNN baselines. Additionally, the dynamic learning scheduler that we have adopted seems to be extremely helpful, increasing the performance of our CNN baselines in all tasks. For example, single-task ResNet-152
achieves a 6% boost in performance compared to previous work [15] on style classification on WikiArt\textsuperscript{Full}. Furthermore, the multi-task learning approach enhances the performance of all CNNs across almost all tasks. We note that ResNet-152 performance on WikiArt\textsuperscript{Full} is the state-of-the-art performance using visual content-based CNNs. Finally, we have experimented with training the models both from-scratch and fine-tuning them. We observe that fine-tuned models that are pre-trained on ImageNet consistently outperform their from-scratch counterparts, while they converge extremely faster. For that reason, in this paper we do not report results of training from scratch.

**Graph Neural Networks for fine art analysis.** Table 2 reports the performance of our GNN baselines. Consistent with the other studies, we observe that multi-task learning in some cases deteriorates GNNs performance, thus, we report results using single-task GNNs. Each GNN seems to perform on-par with the traditional single-task CNN models, while they clearly outperform AlexNet and VGG and obtain state-of-the-art performance on the smaller dataset variant WikiArt\textsuperscript{Artists}.

We further note that GNNs achieve an outstanding performance on WikiArt\textsuperscript{Artists} artist attribution task. We hypothesize that this behavior is due to the simplicity of the task using the inherited artistic school attribute. Another interesting observation is that GraphSAGE requires only 20% of the available labelled data for training to surpass the performance of ResNet-152 - MTL trained on the WikiArt\textsuperscript{Artists} for artist attribution task.

**ArtSAGENet obtains the state-of-the-art performance in fine art analysis.** Finally, we evaluate our proposed ArtSAGENet using either visual features or tags as node feature vectors. Since we observe a slight boost in performance, we report results using multi-task learning. Table 2 illustrates that our proposed architecture is the dominant method for fine art analysis tasks.

It can be clearly seen that both ArtSAGENet variants consistently outperform all baselines in style classification and timeframe estimation. In addition, ArtSAGENet outperforms all baselines on the WikiArt\textsuperscript{Full} and WikiArt\textsuperscript{Modern} artist attribution task, while it obtains the second best performance on the WikiArt\textsuperscript{Artists} dataset variant. We also have to note that using dense node visual feature vectors yields better performance than its sparse counterpart that relies on tags.

Another major observation is that ArtSAGENet significantly outperforms the multi-task ResNet-152 in the artist attribution task across all dataset variants. We hypothesize that the art school attribute is in general an informative property, which ArtSAGENet leverages to accurately attribute artists to artworks. Nevertheless, it seems to be inferior to GraphSAINt on WikiArt\textsuperscript{Modern} creation year estimation task within a ±10 years period, albeit both methods perform comparably after a period of ±25 years. The cumulative accuracy curves for the creation year estimation task are shown in supplementary materials.

### 5.2 Multi-Label Tag Prediction

In order to evaluate the performance of our proposed methods for multi-label fine art categorization, we employ a fine art tag prediction task. Inspired by work in object recognition [11, 52], we report per-class F1-score (CF1), overall F1-score (OF1) and mean Average Precision (mAP) for the tag categories. Detailed definitions of these measures are provided in [57].

Results are summarized in Table 3. Once again, ArtSAGENet is clearly the best performer. Yet another interesting observation is that strong CNNs, i.e. VGG and ResNet, perform better than their GNN counterparts. This suggests that a simple homogeneous structural topology that encodes information about the artistic school can not

### Table 2: Accuracy of ArtSAGENet and the baselines. MTL denotes Multitask Learning. WikiArt\textsuperscript{Modern} - Date\textsuperscript{‡} reports the cumulative score as in Eq. (7) with \( \theta = 5 \) years. \( \heartsuit \) means using tags as node feature vectors. \( \spadesuit \) means using features extracted from a ResNet-34 model pre-trained on ImageNet as node feature vectors. Higher is better (best results in bold).
surpass the performance of powerful visual content-based methods in more complex tasks, as in tag prediction. However, it seems that all GNN baselines clearly outperform AlexNet in terms of mAP, while they need an order of magnitude less time for training and inference compared to CNN architectures.

In Table 4 we compare the time needed for training and inference in case of the CNN baselines, our ArtSAGENet and the GraphSAGE architecture. Since the computational runtime of GNNs is in the same order of magnitude, we highlight here only GraphSAGE, which we use as a building block for our ArtSAGENet. GraphSAGE clearly requires almost 50 times less time for training than any CNN. Finally, we observe that our ArtSAGENet requires only a small amount of time more than the ResNet-152 it relies on.

5.3 Qualitative Analysis on Paintings Retrieval
Further to fine art analysis tasks, we evaluate the performance of our ArtSAGENet in content-based fine art retrieval. That is, we extract the learned representations for both ResNet-152 and ArtSAGENet models and use the k-nearest neighbors (k-NN) algorithm to retrieve the top-5 nearest neighbors for each case. For ResNet-152 we extract the learned visual representations as in Eq. (1), while for our ArtSAGENet we use the multimodal embedding as obtained in Eq. (4). Then, for each reference painting we sort the paintings from the collection in ascending order based on their respective cosine distance to the query image.

Figure 3 illustrates the results for the ResNet-152 and the proposed ArtSAGENet models for each training setting, e.g. single-task versus MTL. The top-5 nearest neighbors retrieved using the Multi-task Learning model trained for style classification, artist attribution and timeframe estimation. The rest of the rows illustrate the top-5 nearest neighbors retrieved using the single-task classifiers.

For the single-task tag prediction classifier, bold means that tag is attributed to the query painting, too. ♠ means using visual features as node feature vectors.
Table 3: Performance comparison on tag prediction task. ♠ means using visual features as node feature vectors. Higher is better (best results in bold).

<table>
<thead>
<tr>
<th>Model</th>
<th>CF1</th>
<th>OP1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>54.1</td>
<td>59.8</td>
<td>60.9</td>
</tr>
<tr>
<td>VGG-16</td>
<td>57.4</td>
<td>63.6</td>
<td>65.3</td>
</tr>
<tr>
<td>VGG-19</td>
<td>56.8</td>
<td>62.9</td>
<td>64.9</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>59.4</td>
<td>64.6</td>
<td>66.3</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>60.3</td>
<td>65.2</td>
<td>66.7</td>
</tr>
<tr>
<td>Cluster-GCN</td>
<td>49.6</td>
<td>58.4</td>
<td>62.5</td>
</tr>
<tr>
<td>GraphSAINT</td>
<td>55.4</td>
<td>61.9</td>
<td>63.8</td>
</tr>
<tr>
<td>GraphSAGE</td>
<td>50.0</td>
<td>58.9</td>
<td>63.3</td>
</tr>
<tr>
<td>SIGN</td>
<td>51.7</td>
<td>59.6</td>
<td>63.7</td>
</tr>
<tr>
<td>ArtSAGENet</td>
<td>62.1</td>
<td>66.9</td>
<td>68.6</td>
</tr>
</tbody>
</table>

Table 4: Time analysis relative to GraphSAGE (fastest) on tag prediction task. ♠ means using visual features as node feature vectors. Train per epoch reports relative time difference for forward/backward operations for a single epoch for all models using the same batch size, e.g. 16. Inference reports the relative time difference for forward operations on test time for full test set. For each model and both training/inference we utilized an Nvidia GeForce 1080Ti 11GB GDDR5X GPU.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train (per epoch)</th>
<th>Inference</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphSAGE</td>
<td>×1</td>
<td>×1</td>
<td>5M</td>
</tr>
<tr>
<td>AlexNet</td>
<td>×47.11</td>
<td>×2.86</td>
<td>57M</td>
</tr>
<tr>
<td>VGG-16</td>
<td>×53.14</td>
<td>×3.11</td>
<td>134M</td>
</tr>
<tr>
<td>VGG-19</td>
<td>×54.41</td>
<td>×3.19</td>
<td>140M</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>×48.54</td>
<td>×2.89</td>
<td>21M</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>×52.60</td>
<td>×3.09</td>
<td>58M</td>
</tr>
<tr>
<td>ArtSAGENet</td>
<td>×52.74</td>
<td>×3.10</td>
<td>63M</td>
</tr>
</tbody>
</table>

Figure 4: Qualitative analysis of single and multi-task ArtSAGENet learned representations for painting retrieval.

6 CONCLUSION

In this work, we proposed ArtSAGENet, a multimodal approach that learns knowledge enhanced visual representations of fine art. Experimental results illustrate that our proposed method leverages both visual and semantic-based information achieving state-of-the-art performance in several fine art categorization tasks. Qualitative analysis of the learned representations shows that ArtSAGENet encodes interesting semantic-level properties, which suggest, for instance, that Pablo Picasso’s works in his Cubist period have a particularly close relation with those of Juan Gris. As this paper has demonstrated, integrating visual and semantic content provides a solid foundation for further research exploring such relationships in visual arts, and especially the dynamic processes of influence and innovation.
REFERENCES


