



|              |                                                                                          |
|--------------|------------------------------------------------------------------------------------------|
| Title        | GNNBoost: boosting artwork classification with graph embeddings                          |
| Author(s)    | El Vaigh, Cheikh Brahim; Garcia, Noa; Renoust, Benjamin et al.                           |
| Citation     | Multimedia Tools and Applications. 2025                                                  |
| Version Type | VoR                                                                                      |
| URL          | <a href="https://hdl.handle.net/11094/101001">https://hdl.handle.net/11094/101001</a>    |
| rights       | This article is licensed under a Creative Commons Attribution 4.0 International License. |
| Note         |                                                                                          |

*The University of Osaka Institutional Knowledge Archive : OUKA*

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka



# GNNBoost: boosting artwork classification with graph embeddings

Cheikh Brahim El Vaigh<sup>1</sup> · Noa Garcia<sup>2</sup> · Benjamin Renoust<sup>2,3</sup> · Chenhui Chu<sup>2,4</sup> · Yuta Nakashima<sup>2</sup> · Yiming Qian<sup>2,5</sup> · Hajime Nagahara<sup>2</sup>

Received: 16 April 2024 / Revised: 22 November 2024 / Accepted: 12 February 2025  
© The Author(s) 2025

## Abstract

The use of AI systems for managing large-scale cultural heritage artifacts has become possible due to the rise of digitization. To classify such content, machine learning is typically used, where contextual information is important to structure the data. One way to capture context is through a knowledge graph. In this study, we propose a new graph neural networks, we can improve artwork classification by utilizing the relationships between entities of the knowledge graph. Our experiments demonstrate that this approach achieves state-of-the-art results on multiple classification tasks across three datasets (SemArt paintings, Buddha statues, and Ukiyo-e woodblock prints). Moreover, our approach is effective in dealing with unbalanced data and we explore the use of both graph attention mechanisms and focal loss functions.

---

✉ Cheikh Brahim El Vaigh  
[cheikh-brahim.el-vaigh@u-bourgogne.fr](mailto:cheikh-brahim.el-vaigh@u-bourgogne.fr)

Noa Garcia  
[noagarcia@ids.osaka-u.ac.jp](mailto:noagarcia@ids.osaka-u.ac.jp)

Benjamin Renoust  
[renoust@ids.osaka-u.ac.jp](mailto:renoust@ids.osaka-u.ac.jp)

Chenhui Chu  
[chu@ids.osaka-u.ac.jp](mailto:chu@ids.osaka-u.ac.jp)

Yuta Nakashima  
[n-yuta@ids.osaka-u.ac.jp](mailto:n-yuta@ids.osaka-u.ac.jp)

Yiming Qian  
[qiany@ihpc.a-star.edu.sg](mailto:qiany@ihpc.a-star.edu.sg)

Hajime Nagahara  
[nagahara@ids.osaka-u.ac.jp](mailto:nagahara@ids.osaka-u.ac.jp)

<sup>1</sup> CIAD UR 7533, Université Bourgogne Europe, Dijon 21000, France

<sup>2</sup> IDS, Osaka University, Osaka 565-087, Japan

<sup>3</sup> Median Technologies, Valbonne 06560, France

<sup>4</sup> Kyoto University, Kyoto 606-8501, Japan

<sup>5</sup> Agency for Science, Technology and Research, 138632 Singapore, Singapore

**Keywords** GNN · Artwork classification · Knowledge graph · Label propagation · Focal Loss

## 1 Introduction

Art is a crystallization of human cultures, reflecting on people and society through their age, while viewers often consume Art in a context different from its creation. Recent technologies now enable artwork digitization, expanding their diffusion and preservation. Some museums even make public their digitized collections enriched with meta-data.

The progress in machine learning technologies makes it a new vista for artwork analysis, previously exclusive to human experts. Examples of analysis cover all sorts of classification, by author, origin, period, or also artwork retrieval from visual similarity [1–4].

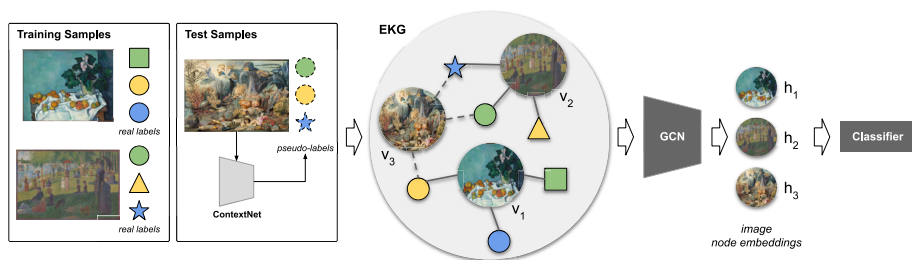
An artwork, beyond its visual aspect, contains contextual information (e.g. the time of creation and the author), which have a crucial role in defining the artwork itself, possibly placing it into trends. Contextual information adds important hints for modeling artwork. Jointly using both visual and contextual information has been proven to be an effective approach for artwork classification: In [1], a knowledge graph (KG) is built from the contextual information to model the interactions among pieces of artwork and their attributes, which include semantic proximity that might not reside in the visual information itself.

Existing methods for artwork classification are mostly based on *inductive learning* [5], which uses a model trained with plenty of annotated observations. This approach has been widely used in general. Still, it may not suit artwork classification when the number of samples (i.e., art pieces) with a certain attribute is limited. In contrast, *transductive learning* [6, 7] is applied to learn from smaller sample sets that contain missing labels. Label propagation assigns pseudo-labels to unlabeled samples (i.e. test samples) to effectively increase the number of samples available in training.

Knowledge graphs (KGs), which model the relationships among entities, such as images and even labels, facilitate this transductive process [6, 8, 9]. They explicitly provide auxiliary information on the relationships among labeled and unlabeled samples and serve as semantic proximity among samples and labels, enabling more informed label propagation.

In this work, we take a transductive learning approach over knowledge graphs for a multi-task learning problem, with the goal to predict various metadata of artwork. Our problem formulation treats all samples as nodes (whether labeled or not). Labels also are nodes. Assigned labels to a sample are represented by edges between the sample node and each label node. For transduction, we build an extended KG (EKG), which includes unlabeled samples in addition to labeled samples as nodes by assigning pseudo-labels to them (i.e. edges). This EKG captures (possibly inaccurate) relationships between unlabeled samples and labeled samples, and between unlabeled samples and labels themselves. The embeddings for all nodes in the EKG are learned through a graph neural network (GNN) and then these embeddings are used to predict multiple labels for unlabeled samples (see Fig. 1 for an overview). In order to evaluate our method, we rely on cultural and historical datasets used by art classification literature: SemArt dataset [3], the Buddha statues dataset [2], and the Ukiyo-e woodblock print dataset [10]. The evaluation shows our algorithm significantly surpasses previous works.

This paper is an extension of our previous work [11] where we initially studied the use of GCN embedding model in a mono-task setup. In this paper, we studied the impact of adding the attention mechanism in our embedding model using GAT [12]. We also study the



**Fig. 1** Overview of our GNNBoost framework. The input is artworks with their labels and unlabeled data with their pseudo-label predicted with a pre-trained model on artwork classification. We build an EKG on the artworks and their labels fed into a GNN embedding model and the output is used to devise the final classifier

impact of using our GNNboost model in a multi-task setting. Finally we use the focal loss function [13] to better take into account imbalanced data. We also extended our experiment to a new dataset namely Ukiyo-e woodblock prints dataset.

This paper builds on the work outlined in [11]. The main contributions of the paper are summarized as follows:

1. We devised an efficient strategy to build an EKG using their pseudo-labels. This EKG creates indirect links between the labeled and unlabeled samples, facilitating artwork classification as we show in the experimental evaluations.
2. We implemented a model for artwork analysis using GNNs to compute distinct artwork embeddings over the EKG. We experimentally show that this framework is robust to unbalanced label distribution, which is very typical for artwork datasets. In particular, we show that using graph attention mechanisms such as GAT [12] and focal loss [13] improved the robustness of unbalanced data.
3. Moreover, we compare our work to the state-of-the-art methods including the ones in [11] where our method scored higher performance on three different artistic domains: fine-art paintings, Ukiyo-e woodblock prints, and Buddha statues.

## 2 Related work

This section discusses how image classification with GNNs and label propagation can benefit the task of automatic artwork analysis.

### 2.1 Automatic artwork classification

The early work on automatic artwork analysis, artwork content is represented through manually crafted features in its digital form [4, 14, 15], such as color [16], brushwork [17], and the scale-invariant local features [18]. These handcrafted features were used to classify, often with SVM, artworks categorized based on their attributes, such as the artist, style, or historical period they belong to. However, the leading factor that limited the performance of the classifier is the quality of the handcrafted features.

The handcrafted features were then replaced by deep features, automatically learned by the family of convolutional neural networks (CNNs), such as VGG [19] or ResNet [20]. Those pre-trained CNNs acquire precise representation of diverse entities, such as natural images or works of art. They were thus mainly used as off-the-shelf features extractors to classify

artworks [1–3, 21–23]. CNNs can also be fine-tuned to handle multi-task art classifications [1, 2, 22, 24]. Meanwhile, features extracted from CNNs were limited to the visual aspect of images, neglecting their context (e.g. cultural and historical backdrop in which the artworks were created).

To solve this issue, Garcia and Vogiatzis [3] proposed to combine visual and textual information in the fine-art analysis, allowing a multimodal and semantical analysis of artworks. To further enrich the representation, visual information can be enhanced with specific information about the art pieces, such as the period of time, the style, or the artist. This allows the incorporation of the historical and social context of the artworks into their representation. A multi-task learning approach is then developed in [1], allowing different paintings to interact through their attributes. An accurate representation [1, 2] of artworks can be built with KGs with their attributes encoded as node embedding, such as node2vec [25]. Jointly with deep visual features [1, 2] greatly improves the performances in paintings and Buddha statues classification. Our research aligns with a comparable approach as we construct a semantic (KG) for art analysis. Furthermore, we explore the utilization of unlabeled data in a transductive setup to enhance the relationships between art pieces in the KG. We also tried a multi-task version of our GNNBoost model similar to the one proposed in [1].

## 2.2 Image classification with GNNs

GNNs have been used in image classification thanks to their ability to model relationships among entities in the knowledge graph in the tasks of link prediction [26] and node classification [26, 27]. Particularly, GNNs are applied to model the relations of images based on their labels [28–31]. The basic thought is to enhance visual features through a GNN learned over the KG. For example, GCNs that accounts for semantic links between different labels are deployed in [29] to improve the multi-label classification. Later, Wang et al. [30] improved this approach using the WordNet concepts hierarchy, composing zero-shot classification. Finally, Zhang et al. [31] model the inter-dependency between an image and its labels by leveraging a weighted adjacency matrix. We are following a similar concept but adding pseudo-labeled data into the training data pool to train a GNN following a standard label propagation process. However, these pseudo-labels are noisy, which imposes an additional challenge in training, as incorrect pseudo-labels have a negative impact on their neighborhood. To mitigate this impact, different types of pseudo-labels are applied for each artwork. We also incorporated the attention mechanism in the graph embedding using GAT [12] in order to improve it.

## 2.3 Label propagation

Semi-supervised learning techniques have effectively addressed the problem of label propagation across networks [32]. Transductive learning with GNNs has recently been introduced to incrementally label data. Douzde et al. [33] used an online pseudo-labeling scheme to annotate unlabeled samples, which updates pseudo-labels in the course of training, while Iscen et al. [6] performed the pseudo-labeling before train without changing it during the training process. Our approach is a fusion of these two: the unlabelled data is first annotated by pseudo-labeling with a pre-trained model, then improved during training.

### 3 Approach

We consider Our problem as a multi-label classification task. Formally, an image  $x$  of an art piece is associated with a set  $t = \{t_c | c \in \mathcal{C}\}$  of labels corresponding to the art piece's attributes, where  $\mathcal{C}$  is the set of attribute types and  $t_c \in \mathcal{L}_c$  is a label of attribute type  $c \in \mathcal{C}$ . Taking the Buddha statues [2] as an example, we have *Style*, *Size*, *Century*, and *Dimensions* in  $\mathcal{C}$ , and set  $\mathcal{L}_c$  of all possible labels for  $c$  has specific labels, e.g., *Nara* in  $\mathcal{L}_{style}$ , *Small* in  $\mathcal{L}_{size}$ , and *V* in  $\mathcal{L}_{century}$ . For each attribute type  $c$ , an image should have one and only one label in  $\mathcal{L}_c$ .

In this work we assume that we have two disjoint datasets: One is fully labeled set  $\mathcal{X}_F = \{x_i | i = 1, \dots, I_F\}$ , and the other is unlabeled set  $\mathcal{X}_U = \{x_i | i = 1, \dots, I_U\}$ , where  $I_F$  and  $I_U$  are the numbers of images in the respective sets. Each image  $x$  in  $\mathcal{X}_F$  has an associated label  $t$ , while the images in  $\mathcal{X}_U$  do not have labels and our goal is to be able to predict their labels.

One effective method for addressing this task is to train a specialized classifier for each attribute type  $c$  with  $\mathcal{X}_F$  and predict labels of  $x \in \mathcal{X}_U$  with the trained classifier.

As previously mentioned, the attributes (or their corresponding labels) in a real-world multi-label dataset are rarely standalone. There may be a relationship between two labels that is either included as additional information within the dataset or can be obtained from external sources such as DBpedia.<sup>1</sup> Particularly for artwork, such relationships include the time where an art piece was created and when its author was living, asserting that the author resided during the same timeframe (coincided with that period of time). The SemArt dataset, as described in [3], contains auxiliary information linking the labels together. For instance, it is stated that *Vincent van Gogh* belongs to the *Dutch* painting school, and all his works reflect this school's style. This relationship can be added in a knowledge graph where there is an edge connecting the author *Vincent van Gogh* to the school *Dutch*. This highlights the indirect associations between art pieces, leading to more effective image representation and classification.

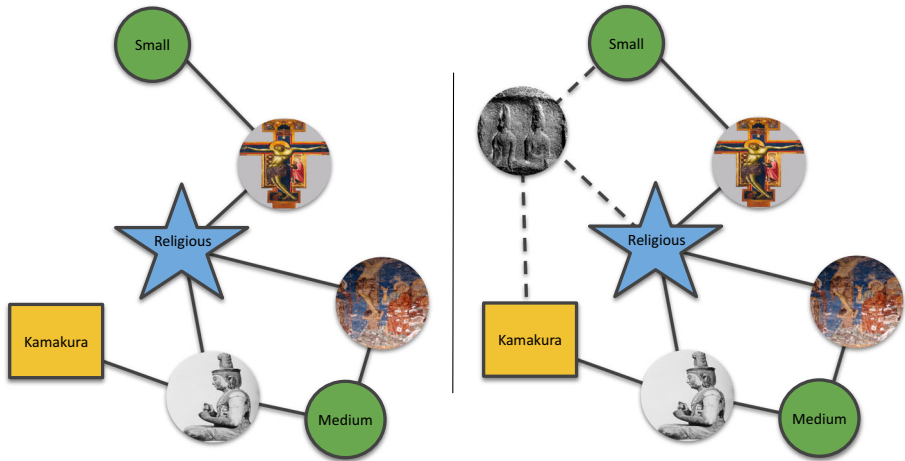
As a result, we approach our multi-label classification task using the transductive learning paradigm. This involves explicitly incorporating correlations among labels and art pieces from both the labeled and unlabeled sets through a KG, as depicted in Fig. 1. Specifically, on top of given relationships among labels  $\mathcal{L} = \bigcup_c \mathcal{L}_c$ , we build a KG whose nodes are labels in  $\mathcal{L}$  and images in  $\mathcal{X}_F$ . An edge in the KG (i) that links label  $t_c$  and image  $x$  represents label assignment, i.e.,  $x$  is associated with  $t_c$  and (ii) that links two labels represents the relationship given by auxiliary/external knowledge. We extend this KG by adding images in  $\mathcal{X}_U$  as in Fig. 2, referred to as an *extended knowledge graph* (EKG). Since these images have no labels, we assign *pseudo-labels* to make (noisy) edges in the graph. This allows to make a denser graph and facilitates to learn better representation of art pieces and labels through a GNN.

The following sections demonstrate how we built our KG, utilized image embeddings using the EKG, and conducted the training and prediction process incorporating pseudo-labels.

#### 3.1 KG construction

A KG is regarded as a graph denoted by  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  represents the entities, and  $\mathcal{E}$  denotes the edges between these entities. The definition of these edges rely on the

<sup>1</sup> <https://wiki.dbpedia.org/>



**Fig. 2** This is a toy example of a knowledge graph (KG) and its corresponding extended knowledge graph (EKG) based on a dataset of Buddha statues. The different nodes in the KG and EKG represent paintings, Buddha statues, or attributes. Plain lines in the KG represent existing relationships (left), while dashed lines in the EKG represent relationships that were inferred using pseudo-labels (right)

nodes they connect. For instance, if there is an edge connecting a specific art piece to a label representing an author, it indicates that the art piece was created by that author. An edge between two different labels means there are some relationships among these labels as exemplified above, which come from auxiliary/external knowledge. Therefore, our KG captures both the semantics of the art pieces and their labels as well as their contextual knowledge.

Formally, our KG's set of nodes are the union of the sets of the fully labeled art pieces and the labels themselves, i.e.  $\mathcal{V} = \mathcal{X}_F \cup \mathcal{L}$ . Label assignment makes edges between art pieces and labels, i.e.  $\mathcal{W} = \{(x, l) \mid l \in t\}$  where  $t$  is the set of labels associated with  $x$  and  $l \in \mathcal{L}$ . As mentioned above, some auxiliary/external knowledge can provide explicit relationships among pairs of labels, which adds corresponding edges to our KG, denoted by  $\mathcal{K} = \{(l, l') \mid l, l' \in \mathcal{L}\}$ . Using these two sets of edges,  $\mathcal{E}$  can be given by  $\mathcal{E} = \mathcal{W} \cup \mathcal{K}$ .

In transductive learning, we expand the KG  $\mathcal{G}$  to include the unlabeled set  $\mathcal{X}_U$ . This EKG is denoted as  $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ , where  $\mathcal{V}'$  is an extension of  $\mathcal{V}$  that incorporates the test set, specifically,

$$\mathcal{V}' = \mathcal{V} \cup \mathcal{X}_U. \quad (1)$$

The art pieces in  $\mathcal{X}_U$  have no corresponding labels, and therefore, no edges. To enhance image embeddings, we incorporate pseudo-labels into the graph. Let  $t'_c = g_c(x)$  denote the label in  $L_c$ , which is predicted by a classifier  $g_c$  for attribute type  $c$ . For every  $x \in \mathcal{X}_U$  and  $c \in \mathcal{C}$ , we can introduce new edges  $(x, t'_c)$  into the graph. Consequently, our extended graph is denoted as  $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ , where

$$\mathcal{E}' = \mathcal{E} \cup \{(x, t'_c) \mid x \in \mathcal{X}_U, c \in \mathcal{C}\}. \quad (2)$$

### 3.2 Image embedding with EKG

To encode the relationships provided by  $\mathcal{G}'$  into art piece (and label) embeddings, we adopted a GCN [34] model. Our GCN has  $N$  layers, and the hidden state of the  $n$ -th layer,  $H^{(n)}$ , is calculated based on the hidden state of the  $(n - 1)$ -th layer,  $H^{(n-1)}$ , using the following formula:

$$H^{(n)} = \text{ReLU}(D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}H^{(n-1)}W^{(n)} + b^{(n)}). \quad (3)$$

where  $A \in \mathbb{R}^{|\mathcal{V}'| \times |\mathcal{V}'|}$  represent the adjacency matrix obtained from  $\mathcal{G}'$ .  $D$  is a diagonal matrix containing node degrees, defined as  $D_{ii} = \sum_j A_{ij}$ , and  $I$  is the identity matrix. The matrix  $W^{(n)} \in \mathbb{R}^{d \times d}$  contains learnable parameters. The initial features of nodes are stored in  $H^{(0)} \in \mathbb{R}^{|\mathcal{V}'| \times d}$ , where each row vector  $h_i^{(0)} \in \mathbb{R}^d$  corresponds to the  $d$ -dimensional feature vector of node  $v_i$ . These initial feature vectors are obtained using ResNet50 [20] and node2vec [25] applied to the graph  $\mathcal{G}'$  for images and labels, respectively.

Using GCN, the latent connections between the nodes in the EKG can be efficiently represented. The adjacency matrix  $A$  establishes direct connections between the nodes in  $\mathcal{G}'$ , and the GCN then transfers information along these edges. This approach is particularly effective for labeling nodes that function as hubs—nodes with a high degree—as they have equal significance to their surrounding nodes and simplify the classification task by providing a direct path between image nodes. For example two images that are connected to the same author are very likely created by that author. The model relies on the direct relations between nodes (images or labels) as we can easily interpret such relationships.

In GCN, from a node-wise perspective, the update rule can be written as:

$$h_i = \text{ReLU}(\sum_{j \in N_i} c_{ij} W h_j). \quad (4)$$

where  $c_{ij} = \frac{1}{|N_i||N_j|}$  and  $N_i$  is the set of nodes neighbors to node  $i$ . In (4),  $h_i$  is computed with an explicit coefficient that solely depends on the node neighborhood. Therefore, we also use GAT [12] in order to improve node embedding. In GAT, a node embedding follows the same (4) but the coefficient  $c_{ij}$  is computed using attention mechanism:

$$c_{ij} = \text{softmax}(\text{attention}(h_i, h_j)). \quad (5)$$

where  $\text{attention}(h_i, h_j)$  is the attention coefficient based on the current embedding of the nodes  $i$  and  $j$ . In fact, in GAT  $c_{ij}$  is computed based on the embedding of the node that are linked to the node  $i$  and is a learnable parameter. In this manner, the coefficient  $c_{ij}$  is not fixed by the graph and is learned based on the actual data.

### 3.3 Training and inference

#### 3.3.1 Single-task learning

Given art piece embedding  $h_i = h_i^{(N)}$  for node  $x_i \in \mathcal{X}_F \cup \mathcal{X}_U$ , a classifier  $f_c$  that is fully connected (FC) with softmax generates predictions for each label set  $L_c$ . Formally,

$$y_{ic} = f_c(x_i) = \text{softmax}(W_c h_i + b_c), \quad (6)$$

where  $W_c \in \mathbb{R}^{|\mathcal{L}_c| \times d}$  and  $b_c \in \mathbb{R}^{|\mathcal{L}_c|}$  are learnable parameters. The loss function  $\ell$  is given by

$$\ell = - \sum_{x_i \in \mathcal{X}_F} \sum_{c \in \mathcal{C}} \sum_{k=1}^{|\mathcal{L}_c|} t_{ick} \log y_{ick}, \quad (7)$$

where  $t_{ick}$  denotes the  $k$ -th element of the one-hot vector of label  $t_{ic}$  for  $x_i$ . The evaluation of the loss function is applied solely on the labeled set  $\mathcal{X}_F$ . Upon completion of training, the  $h_i$  related to image  $x_i$  located in  $\mathcal{X}_U$  is input into  $f_c$  and softmax to determine its label.

We also implement a version of the single-task model using focal loss [13] to be efficient dealing with class imbalance. The main difference compared with our classical loss is that with focal loss we care about learning hard examples, which tends to reduce model overfitting on classes with large samples. The focal loss function is express as follow:

$$\mathcal{FL}(p_i) = (1 - p_i)^\gamma \log(p_i), \quad (8)$$

where  $p_i = CE(h_i, y)$  is the cross-entropy loss function applied to the feature vectors at the iteration  $i(h_i)$  and the ground truth labels  $y$ .  $\gamma$  is used to reduce the impact of misclassified samples in the loss function and smoothly adjusts the rate at which easy examples are down-weighted. We get the standard cross-entropy loss function for  $\gamma = 0$ . In our experiments we used  $\gamma = 2$  the best value reported for  $\gamma$  in [13].

### 3.3.2 Multi-task learning

In (7) we learn the classification of one category each time, assuming different tasks to be independent which is not necessary the case. Hence, we also tried a multi-task version of our model where we learn the classification of the different categories at once. The multi-task loss function can therefore be expressed as follow:

$$L = \sum_{t \in T} \alpha_t l_t, \quad (9)$$

where  $T$  is the set of the tasks at hand e.g., *Time classification*, *Author classification*,  $l_t$  is the loss function of a given task defined in (7) and  $\alpha_t$  is an hyper-parameter fixed for the task  $t$ . The main benefit of the multi-task setup is to use the mutual information that two tasks can share when classifying the same image. For example there maybe a link between an author of an image and the time when that image was created and this link can be valuable when it comes to predict the labels of this image.

## 4 Experiments

We extensively evaluated our approach on three different datasets namely: SemArt [3], Buddha statues [2], and Ukiyo-e [10]. *These three datasets are available online in their corresponding papers.* We showcase classification accuracy against state of the art and robustness with respect to noise in the initial guesses by  $f_c$ 's. After giving the details of the datasets (Sect. 4.1) and our implementation (Sect. 4.2), we devise an ablation studies on our approach (Sect. 4.3), comparing the importance of using different attributes of art pieces when building the KG as described in Sect. 3.1. Finally, we compare our approach against baseline competitors, showing significant improvement on the SemArt dataset, Buddha statues dataset, and Ukiyo-e dataset in Sects. 4.4 to 4.6.

## 4.1 Datasets

**SemArt dataset** It contains 19,244, 1,069, and 1,069 images for training, validation, and test, all of which are European fine-art paintings. Our fully-labeled set  $\mathcal{X}_F$  contains all training and validation images, while unlabeled set  $\mathcal{X}_U$  contains test images. A painting has the attributes Author, Title, Date, Technique, Type, School, and TimeFrame. Following previous work [1], we evaluate our models on four attribute types, i.e.  $\mathcal{C} = \{\text{Type, School, TimeFrame, Author}\}$ :

- **Type (T1.1)** classifying painting in 10 Types: portrait, landscape, religious, study, genre, still-life, mythological, interior, historical and other.
- **School (T1.2)** categorizing each painting into one of the 24 classes, which include Italian, Spanish, English, Dutch, French, Flemish, German, Netherlandish, Austrian, Hungarian, American, Danish, Swiss, Russian, Scottish, Greek, Catalan, Bohemian, Swedish, Irish, Norwegian, Polish, Other, and Unknown.
- **TimeFrame (T1.3)** devises painting classification according to 17 different period of time (e.g. 1001-1050, 1551-1600, 1701-1750) and one additional class (Unknown) for TimeFrames that have less than 10 paintings.
- **Author (T1.4)** aims at classifying painting according to 350 different painters. We evaluate two cases where all the authors are considered (even if some authors have only one painting in the dataset) and where only authors with at least ten paintings in the labeled set are kept.

For pseudo-labels assignment of the test images, we use the pre-trained ContextNet model [1] i.e. initial guess.

**Buddha statues dataset** This second dataset contains 2,665 images of faces of unique Buddha statues. We split them into 1,866, 266, and 533 images for training, validation, and test. Each images are labeled with the attributes types *Style*, *Size*, *Century of creation*, and *Dimensions*. The details on the attribute types are as follows:

- **Statue style (T2.1)** aims at classifying the images with the following styles: China, Kamakura period, Nara period, and Heian period.
- **Statue size (T2.2)** labels each image with one of these sizes: Small, Medium, and Big.
- **Century of creation (T2.3)** categorizes Buddha statues into seven distinct centuries: V, VI, VII, VIII, IX, XII, and XIII. Centuries with limited labeled images have been excluded from the classification process.
- **Statue dimensions (T2.4)** groups images according to 12 different dimensions ranging from 50 to 1,050 cm.

We used the model in [2], for computing the pseudo-labels which is also used as a baseline.

**Ukiyo-e dataset** We also validated our approach on a larger dataset namely Ukiyo-e dataset that is a website with over 220,000 Japanese woodblock prints dating from the early 17th century. These images are aggregated from a variety of sources from all around the world including museums, universities, libraries, auction houses, and dealers. After cleaning the dataset in order to remove corrupted images only 184,123 images are left. We split them into 136,091, 24,016, and 24,016 images for training, validation, and test. Each images are labeled with the attribute types *Era*, *Artist*, and *Date*. The details on the attribute types are as follows:

- **Era (T3.1)** consists in classifying Ukiyo-e images in one of the different Ukiyo-e seven Era e.g. Early Ukiyo-e (Early-Mid 1700s), 'Meiji Period (1868 to 1912) or Modern and Contemporary (1950s to Now).
- **Artist (T3.2)** performs image classification according to 153 different artists.
- **Date (T3.3)** aims at classifying images into 345 different numerical dates giving the image creation date.

To compute the pseudo-labels, we used the model ContextNet from [1], which is also used as a baseline. We first fine-tuned a pretrained ContextNet model on the Ukiyo-e images to build the first guess and used them to build the EKG for the Ukiyo-e dataset.

## 4.2 Implementation details

For the three datasets, we adopt a two-layer GNN for node classification. The GNN initial layer ( $H^{(0)}$ ) initialized using node2vec for the fully-labeled set and random for unlabeled set. We set the hidden states embedding dimension to  $d = 128$ . Each attribute type has a separate trained model. Our GNN architecture includes an initial embedding as part of its first layer, generating a hidden representation of size 16 according to the specified GNN hyperparameters [34]. This initial GNN embedding is then utilized in the second layer, producing an output size equal to the number of labels  $|L_C|$ .

During training, we employ the Adam optimization algorithm [35] with a learning rate of 0.001 and cap the maximum number of iterations at 5,000. Additionally, we incorporate an early stop mechanism, which halts training if the validation loss remains unchanged for a continuous period of 100 iterations. Our transductive learning approach processes the entire dataset at each iteration without partitioning it into mini-batches. Although global training typically demands a lot of memory, our advantage lies in having relatively small datasets that can be accommodated in memory. Consequently, this approach proves to be more convenient and faster, especially when training on GPUs.

## 4.3 Importance of pseudo-labels assignment

We evaluate on the SemArt dataset the benefit of building the EKG using pseudo-labels.

**Baselines** For comparison, we use two models: ContextNet [1], and pseudo-labels are randomly assigned to all attribute types ( $S_0$  Random initialization).

**Pseudo-label assignment** In the remaining variations of our model, we initiate the process of assigning pseudo-labels using the pre-trained ContextNet  $f_c$ . To analyze its contribution, we evaluate variants trained with pseudo-labels on one to four attribute types, referred to as models  $S_1$ ,  $S_2$ ,  $S_3$ , and  $S_{all}$ , respectively.  $S_1$  and  $S_3$  has four variants according to the choice of attribute types to be employed or excluded.  $S_2$  has six variants, which is combination of two out of four attribute types.

Table 1 presents the results for the SemArt dataset. In Rows 1 and 2, the accuracy of the baselines is displayed, indicating that 1) ContextNet [1] is a strong model, and 2) random initialization yields unsatisfactory results.

Rows 3 to 6 show results when pseudo-labels of only one attribute type is used ( $S_1$ ). This improves the classification accuracy over the original ContextNet, resulting in an improvement of 4 points for *School* and 18 points for *TimeFrame*. However, for *Author*, the data

**Table 1** SemArt classification results

|    | <i>Model</i>                         | <i>Type</i><br><i>T1.1</i> | <i>School</i><br><i>T1.2</i> | <i>TimeFrame</i><br><i>T1.3</i> | <i>Author</i><br><i>T1.4</i> |
|----|--------------------------------------|----------------------------|------------------------------|---------------------------------|------------------------------|
| 1  | ContextNet [1]                       | 0.815                      | 0.671                        | 0.613                           | <b>0.615</b>                 |
| 2  | $S_0$ Random initialization          | 0.100                      | 0.010                        | 0.050                           | 0.040                        |
| 3  | $S_1$ <i>Type</i>                    | 0.807                      | 0.368                        | 0.136                           | 0.013                        |
| 4  | $S_1$ <i>School</i>                  | 0.296                      | 0.718                        | 0.154                           | 0.035                        |
| 5  | $S_1$ <i>TimeFrame</i>               | 0.363                      | 0.392                        | 0.796                           | 0.047                        |
| 6  | $S_1$ <i>Author</i>                  | 0.352                      | 0.420                        | 0.271                           | 0.181                        |
| 7  | $S_2$ <i>School_Author</i>           | 0.397                      | 0.866                        | 0.314                           | 0.284                        |
| 8  | $S_2$ <i>School_Type</i>             | 0.861                      | 0.739                        | 0.213                           | 0.058                        |
| 9  | $S_2$ <i>School_TimeFrame</i>        | 0.363                      | 0.846                        | 0.843                           | 0.139                        |
| 10 | $S_2$ <i>Type_Author</i>             | 0.899                      | 0.486                        | 0.295                           | 0.302                        |
| 11 | $S_2$ <i>Type_TimeFrame</i>          | 0.915                      | 0.485                        | 0.853                           | 0.098                        |
| 12 | $S_2$ <i>Author_TimeFrame</i>        | 0.417                      | 0.504                        | 0.906                           | 0.354                        |
| 13 | $S_3$ <i>School_Author_TimeFrame</i> | 0.461                      | 0.882                        | <b>0.933</b>                    | 0.435                        |
| 14 | $S_3$ <i>Author_TimeFrame_Type</i>   | 0.930                      | 0.564                        | 0.932                           | 0.482                        |
| 15 | $S_3$ <i>Author_School_Type</i>      | 0.924                      | 0.827                        | 0.348                           | 0.394                        |
| 16 | $S_3$ <i>School_Type_TimeFrame</i>   | 0.929                      | 0.859                        | 0.877                           | 0.204                        |
| 17 | $S_{all}$                            | <b>0.939</b>               | <b>0.889</b>                 | 0.927                           | 0.479                        |

The initial part of the table shows the baseline, while subsequent sections display the classification accuracy when utilizing one, two, three, or all categories of pseudo-labels to construct the KG

imbalance in this attribute type caused a decline of 42 points, which is further explained in Sect. 4.4. While  $S_1$  improved accuracy for the attribute type used to construct the graph, it did not perform well on the other types. For example, utilizing pseudo-labels based on *School* (Row 4) only increased accuracy for the *School* attribute, while decreasing performance on *TimeFrame*, *Type*, and *Author*.

Rows 7 through 12 depict the outcomes with two attribute types ( $S_2$ ). Firstly, when the selected attribute types are employed, the accuracy is boosted. For instance, accuracy is elevated for both categories, such as in the case of *School* and *Type* (Row 8), surpassing the outcomes of  $S_1$  when solely *School* (Row 4) or *Type* (Row 3) is used. Secondly, the accuracy is enhanced for other categories as well, as is evident from the superior performance of *TimeFrame* and *Author* in Row 8, as opposed to Row 3 or Row 4. The enhanced accuracy underscores the benefits of leveraging the relationships between existing nodes to uncover hidden connections among them. Moreover, from a qualitative standpoint, it becomes evident that certain pairs of pseudo-labels align with the construction of the EKG. For instance, the pairing of *Type* and *TimeFrame* (Row 11) yield the best scores among all the  $S_2$  models, likely due to the prevalence of various painting types during specific historical time periods.

Rows 13 to 16 show results with pseudo-labels of three attribute types ( $S_3$ ), whereas Row 17 provides the accuracy when all the categories are used ( $S_{all}$ ). They further improve the performance, giving similar conclusions as  $S_2$ . For *TimeFrame*,  $S_3$  outperforms  $S_{all}$ . Although the difference in accuracy is only one point, this finding implies that excessive edge additions could be detrimental to performance, particularly when the added edges contain noise.

#### 4.4 Comparison over SemArt

In the subsequent analysis, we evaluate our approach against reported state-of-the-art approaches on the SemArt dataset:

- **Pre-trained Networks** [1]. employed VGG16 [19], ResNet50 [20], and ResNet152 [20] as pre-trained networks for image classification.
- **Fine-tuned Networks** [1]. Same as Pre-trained Networks but they are fine-tuned on the SemArt dataset.
- **ResNet50+Attributes** [1]. Combination of ResNet50 features with the predicted attributes by one of the three previous fine-tuned models.
- **ResNet50+Captions** [1]. Combination of ResNet50 features with a generated caption embeddings [36].
- **ContextNet** [1]. ContextNet, which utilizes a multi-task learning model (ContextNet MTL) that learns the classification of four attribute types of paintings as four shared tasks. ContextNet also employs a knowledge graph (ContextNet KG), similar to the one used in our approach. ContextNet KG uses ResNet50 features for visual features and node2vec for nodes embedding to model relations between nodes and perform classification.

The models mentioned above can be categorized into two groups: 1) Local models, which disregard context, include Pre-trained Networks, Fine-tuned Networks, and ResNet50+Captions. 2) Global models, on the other hand, aim to incorporate as much context as possible, as seen in ResNet50+Attributes and ContextNet.

Table 2 presents the results of our experiments. Our GCNBoost ( $S_{all}$ ) significantly outperforms existing methods in all attribute types, except for the *Author* attribute, where its accuracy is lower compared to ContextNet. It is worth noting that ContextNet MTL and ContextNet KG complement each other in modeling the global nodes interrelationships differently for the four categories of attributes. However, they both yield low scores for the *TimeFrame* and *Author* attributes, which could be attributed to the significant imbalance in the number of training samples in these attribute types (especially for *Author*). Despite GCNBoost's success in recovering the problem for *Type*, *School*, and *TimeFrame*, where the imbalance is not as severe, it failed to do so for the *Author* attribute.

We show in Fig. 3 the distribution of node degrees (i.e., number of edges) of labels in  $\mathcal{L}$  in EKG for different attribute types to understand the imbalance problem. We observed that all the attributes exhibit imbalance, but the attributes of type *Author* (as shown in Fig. 3d) display a significant imbalance. Specifically, there are approximately 2,000 *Author* nodes that have a small connectivity, indicating a low number of associated paintings, while only fewer than 200 nodes have medium to high connectivity, meaning they have more than 50 associated paintings.

To study the impact of data imbalance on our model's outcome, we exclude nodes labeled as *Author* with a degree lower than ten. The resulting output for this scenario is denoted as GCN  $S_{all}^*$  (*Author filter*) in the last row of Table 2. We observe that by eliminating painting with limited training samples, we can enhance the accuracy.

#### 4.5 Comparison over buddha statues

Using the same procedure as SemArt dataset, we utilized GCNBoost on the Buddha statues dataset. Our approach involved implementing a neural network (NN) classifier based on the one defined in the original study [2]. The NN classifier is a fully-connected layer followed

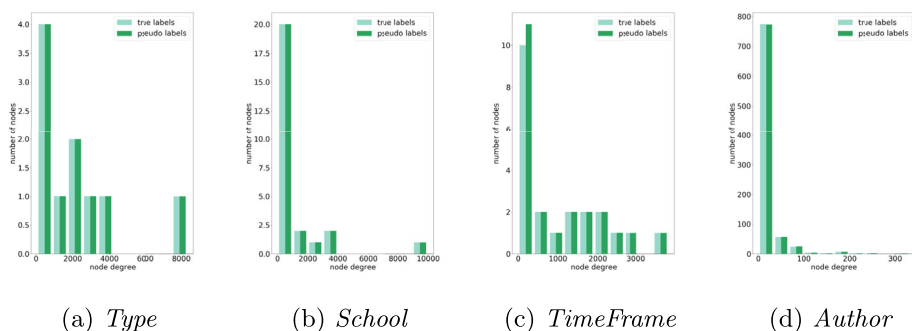
**Table 2** Comparison of different models over SemArt

| <i>Model</i>                         | <i>Type<br/>T1.1</i> | <i>School<br/>T1.2</i> | <i>TimeFrame<br/>T1.3</i> | <i>Author<br/>T1.4</i> |
|--------------------------------------|----------------------|------------------------|---------------------------|------------------------|
| VGG16 pre-trained [1]                | 0.706                | 0.502                  | 0.418                     | 0.482                  |
| ResNet50 pre-trained [1]             | 0.726                | 0.557                  | 0.456                     | 0.500                  |
| ResNet152 pre-trained [1]            | 0.740                | 0.540                  | 0.454                     | 0.489                  |
| VGG16 fine-tuned [1]                 | 0.768                | 0.616                  | 0.559                     | 0.520                  |
| ResNet50 fine-tuned [1]              | 0.765                | 0.655                  | 0.604                     | 0.515                  |
| ResNet152 fine-tuned [1]             | 0.790                | 0.653                  | 0.598                     | 0.573                  |
| ResNet50+Attributes [1]              | 0.785                | 0.667                  | 0.599                     | 0.561                  |
| ResNet50+Captions [1]                | 0.799                | 0.649                  | 0.598                     | 0.607                  |
| ContextNet MTL [1]                   | 0.791                | 0.691                  | 0.632                     | 0.603                  |
| ContextNet KG [1]                    | 0.815                | 0.671                  | 0.613                     | 0.615                  |
| GCNBoost $S_1$                       | 0.807                | 0.718                  | 0.796                     | 0.181                  |
| GCNBoost $S_2$                       | 0.915                | 0.866                  | 0.906                     | 0.354                  |
| GCNBoost $S_3$                       | 0.930                | 0.882                  | <b>0.933</b>              | 0.482                  |
| GCNBoost $S_{all}$                   | <b>0.939</b>         | <b>0.889</b>           | 0.927                     | 0.479                  |
| GCNBoost $S_{all}^*$ (Author filter) | —                    | —                      | —                         | <b>0.702</b>           |

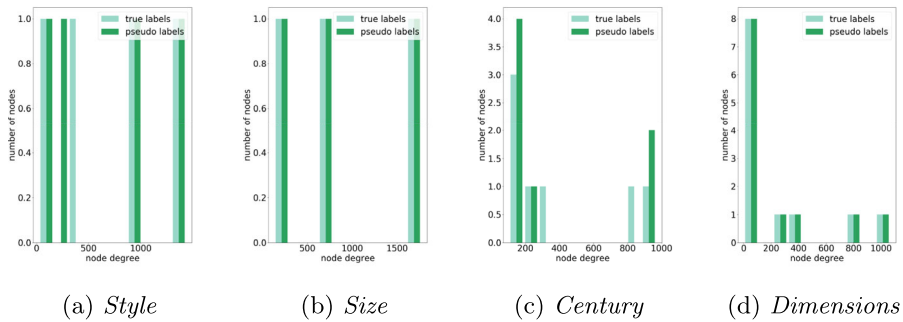
The rows GCNBoost  $S_1$  to  $S_{all}$  gives some variants of our model

by a softmax activation function and is based on the cross entropy loss. Two variants of the NN classifier were employed:

- **NN original** shares identical architecture and setup with [2]. It is trained on a dataset of 3,334 images, where certain images lack one or more of the four distinct attributes, resulting in sparse data. This version adopts a 70%-30% split of the dataset for 5-fold cross-validation.
- The **NN retrained** model shares the same architecture as [2], with a distinct configurations. The dataset of Buddha statues is divided into a 70%-20%-10% split for training, validation and test, respectively. Additionally, each statue considered in the experiment possesses all four attributes relevant to the tasks, resulting in a total of 2,665 images.



**Fig. 3** Distribution of nodes degrees for the four categories (*Type*, *School*, *TimeFrame*, *Author*) based on the true classes for test and train data, and the pseudo-labels for the test



**Fig. 4** Distribution of nodes degrees for the four categories (*Style*, *Size*, *Century*, *Dimensions*) based on the true classes for test and train data, and the pseudo-labels for the test

To create our EKG, we followed the same process as for SemArt, but with a distinct set of attributes. We utilized the NN retrained model to construct the pseudo-labels. The node degree distributions of this EKG are illustrated in Fig. 4, revealing some degree of imbalance, although less than the SemArt dataset. As previously discussed, GCNBoost can cope with imbalanced labels and mitigate the various attributes of a statue, ensuring robustness.

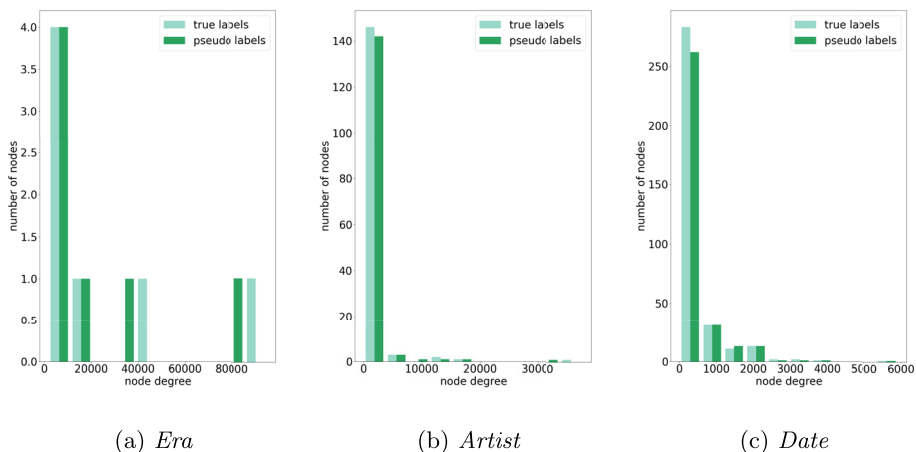
We gradually incorporating individual attribute types to thoroughly examine their impact. The outcomes are shown in Table 3, where the topmost row indicates the optimal results as reported in [2]. Furthermore, the second row represents our baseline results using the retrained NN, while the ensuing rows present our GCNBoost model's outcomes for each unique strategy. The performance of NN models differs significantly across various attribute types. However, GCNBoost can enhance classification accuracy for most attribute types, despite the limited data available (only 2,665 samples). Additionally, we observe a consistent increase in accuracy from the worst model ( $S_0$ ), which assigns random pseudo-labels, to the best model ( $S_{all}$ ), which includes all pseudo-labels and relationships among labels.

The original NN outperforms our approach in classifying *Style* ( $T2.1$ ) since it employs a 5-fold cross-validation technique, which uses more training data than our protocol. In contrast, GCNBoost and the retrained NN baseline leave 10% of the artworks as an unlabeled set, which was not done in the study by [2]. This difference in experimental setup could account for the discrepancy in performance between the two experiments.

**Table 3** Classification results on Buddha dataset

|   | <i>Model</i>                | <i>Style</i><br><i>T2.1</i> | <i>Size</i><br><i>T2.2</i> | <i>Century</i><br><i>T2.3</i> | <i>Dimensions</i><br><i>T2.4</i> |
|---|-----------------------------|-----------------------------|----------------------------|-------------------------------|----------------------------------|
| 1 | NN ( <i>original</i> ) [2]  | <b>0.98</b>                 | 0.78                       | 0.78                          | 0.78                             |
| 2 | NN ( <i>retrained</i> )     | 0.58                        | 0.65                       | 0.76                          | 0.46                             |
| 3 | $S_0$ Random initialization | 0.23                        | 0.30                       | 0.13                          | 0.08                             |
| 4 | GCNBoost $S_1$              | 0.57                        | 0.68                       | 0.74                          | 0.47                             |
| 5 | GCNBoost $S_2$              | 0.59                        | 0.85                       | 0.80                          | 0.76                             |
| 6 | GCNBoost $S_3$              | 0.88                        | 0.86                       | 0.86                          | 0.84                             |
| 7 | GCNBoost $S_{all}$          | 0.92                        | <b>0.94</b>                | <b>0.93</b>                   | <b>0.90</b>                      |

$S_{all}$  is using style, size, century, and dimensions to create the EKG



**Fig. 5** Distribution of nodes degrees for the three categories (*Era*, *Artist*, *Date*) based on the true classes for test and train data, and the pseudo-labels for the test

#### 4.6 Comparison over Ukiyo-e dataset

We apply GCNBoost to the Ukiyo-e dataset with the same process as the SemArt and Buddha datasets. We first fine-tuned the model ContextNet on the Ukiyo-e images and built first guess as pseudo-labels. This ContextNet fine-tuned on Ukiyo-e images is also used as a baseline. We also used this pseudo-labels to build an EKG in the same process as for SemArt and Buddha datasets but with a different set of attributes. We consider the classification of only three attributes as in the original paper [10]. In Fig. 5, we present the node degree distributions of the EKG and observe some level of imbalance, though not as pronounced as in the SemArt dataset. As we previously discussed, GCNBoost exhibits robustness to imbalanced labels, which is beneficial in handling such data distributions.

To further explore the impact of each attribute type, we adopt a similar approach of sequentially adding one attribute at a time. The corresponding results are summarized in Table 4. The top row displays the best outcomes achieved by the model reported in [10]. We used ContextNet in our experimental setup, it is presented in the second row. Subsequent rows exhibit the performance of our GCNBoost model for each distinct strategy.

The original models in the paper [10] exhibit remarkable performance. However, GCNBoost manages to enhance the majority of the classification accuracy. Notably, there is a

**Table 4** Classification results on Ukiyo-e dataset

|   |                                 | <i>Era</i>   | <i>Artist</i> | <i>Date</i>  |
|---|---------------------------------|--------------|---------------|--------------|
|   | <i>Model</i>                    | <i>T3.1</i>  | <i>T3.2</i>   | <i>T3.3</i>  |
| 1 | Ukiyo-e ( <i>original</i> ) [2] | 0.927        | 0.821         | 0.714        |
| 2 | $S_0$ Random initialization     | 0.618        | 0.248         | 0.215        |
| 3 | GCNBoost $S_1$                  | 0.702        | 0.338         | 0.255        |
| 4 | GCNBoost $S_2$                  | 0.789        | 0.343         | 0.241        |
| 5 | GCNBoost $S_{all}$              | <b>0.982</b> | <b>0.898</b>  | <b>0.732</b> |

$S_{all}$  is using Era, Artist, and Date to create the EKG

**Table 5** Classification results on SemArt comparing GCNBoost results with GAT [12], focal loss [13] and multi-task learning

|   | <i>Model</i>                   | <i>Type</i><br><i>T1.1</i> | <i>School</i><br><i>T1.2</i> | <i>TimeFrame</i><br><i>T1.3</i> | <i>Author</i><br><i>T1.4</i> |
|---|--------------------------------|----------------------------|------------------------------|---------------------------------|------------------------------|
| 1 | GCNBoost $S_{all}$             | 0.939                      | 0.889                        | 0.933                           | 0.479                        |
| 2 | GCNBoost MTL $\alpha_t = 1$    | <b>0.991</b>               | <b>0.992</b>                 | <b>0.971</b>                    | 0.532                        |
| 3 | GCNBoost MTL $\alpha_t = 0.25$ | 0.933                      | 0.964                        | 0.961                           | 0.524                        |
| 4 | GATBoost MTL $\alpha_t = 1$    | 0.954                      | 0.981                        | 0.961                           | <b>0.547</b>                 |
| 5 | GATBoost MTL $\alpha_t = 0.25$ | 0.933                      | 0.957                        | 0.960                           | 0.379                        |
| 6 | GCNBoost FL                    | 0.949                      | 0.966                        | 0.959                           | 0.504                        |

**Table 6** Classification results on Buddha dataset comparing GCNBoost results with GAT [12], focal loss [13] and multi-task learning

|   | <i>Model</i>                   | <i>Style</i><br><i>T2.1</i> | <i>Size</i><br><i>T2.2</i> | <i>Century</i><br><i>T2.3</i> | <i>Dimensions</i><br><i>T2.4</i> |
|---|--------------------------------|-----------------------------|----------------------------|-------------------------------|----------------------------------|
| 1 | GCNBoost $S_{all}$             | <b>0.92</b>                 | 0.94                       | 0.93                          | 0.90                             |
| 2 | GCNBoost MTL $\alpha_t = 1$    | 0.89                        | 0.97                       | 0.96                          | 0.93                             |
| 3 | GCNBoost MTL $\alpha_t = 0.25$ | 0.89                        | 0.97                       | 0.97                          | 0.93                             |
| 4 | GATBoost MTL $\alpha_t = 1$    | 0.89                        | <b>0.98</b>                | <b>0.99</b>                   | <b>0.94</b>                      |
| 5 | GATBoost MTL $\alpha_t = 0.25$ | 0.88                        | 0.91                       | 0.96                          | 0.93                             |
| 6 | GCNBoost FL                    | 0.89                        | <b>0.98</b>                | 0.94                          | 0.93                             |

consistent increase in accuracy from the weakest model,  $S_0$ , which employs random pseudo-label assignment, to the best model,  $S_{all}$ , which incorporates all pseudo-labels and their relationships. We were able to improve the classification accuracy on a new dataset that was not included in our initial study [11] which shows GCNBoost to be also working on different art analysis datasets.

#### 4.7 Improved GNN boosting

In this section we further improve upon GCNBoost by studying different implementations of GNN boosting using: attention mechanism, focal loss and multi tasking. First we trained our GCNBoost model in the multi-task setup using GCN [34]. Then, we replaced the GCN embedding model with GAT [12] which is an improved version of graph embedding that incorporate node attention mechanism as explained in Sec. 3.2. We also added the focal loss in order to deal with class imbalance. The results are shown in Table 5 for the SemArt dataset and in Table 6 for the Buddha status dataset. All our different new setting improved the classification results and they are better than the using GCN alone. For SemArt dataset the best results are obtained with the multi-task model with  $\alpha = 1$  (see line 2 of Table 5) using GCN except for Author classification where the best score is obtained when using GAT and  $\alpha = 1$  (see line 4 of Table 5). Very likely because of the multi-task setup and the size of the SemArt dataset the attention mechanism is not adding any improvement.

**Table 7** Classification results on Ukiyo-e dataset comparing GCNBoost results with GAT [12], focal loss [13] and multi-task learning

|   |                                | <i>Era</i>   | <i>Artist</i> | <i>Date</i>  |
|---|--------------------------------|--------------|---------------|--------------|
|   | <i>Model</i>                   | <i>T3.1</i>  | <i>T3.2</i>   | <i>T3.3</i>  |
| 1 | GCNBoost $S_{all}$             | 0.982        | 0.898         | <b>0.734</b> |
| 2 | GCNBoost MTL $\alpha_t = 1$    | 0.855        | 0.968         | 0.642        |
| 3 | GCNBoost MTL $\alpha_t = 0.25$ | 0.983        | 0.548         | 0.669        |
| 4 | GATBoost MTL $\alpha_t = 1$    | 0.959        | 0.875         | 0.665        |
| 5 | GATBoost MTL $\alpha_t = 0.25$ | 0.955        | 0.898         | 0.470        |
| 6 | GCNBoost FL                    | <b>0.987</b> | <b>0.983</b>  | 0712         |

For Buddha status classification we conducted the same experiments. We found similar performance for Style classification with the different models. Unlike SemArt dataset the best model for Buddha status classification is the multi-task model with  $\alpha = 1$  (see line 2 of Table 5) using GAT. For Status size classification both focal loss and the multi-task model with  $\alpha = 1$  using GAT obtain the best results (see lines 4 and 6). GCN is less accurate on Buddha status dataset with comparison to SemArt dataset since the former has less training samples and has more imbalanced classes.

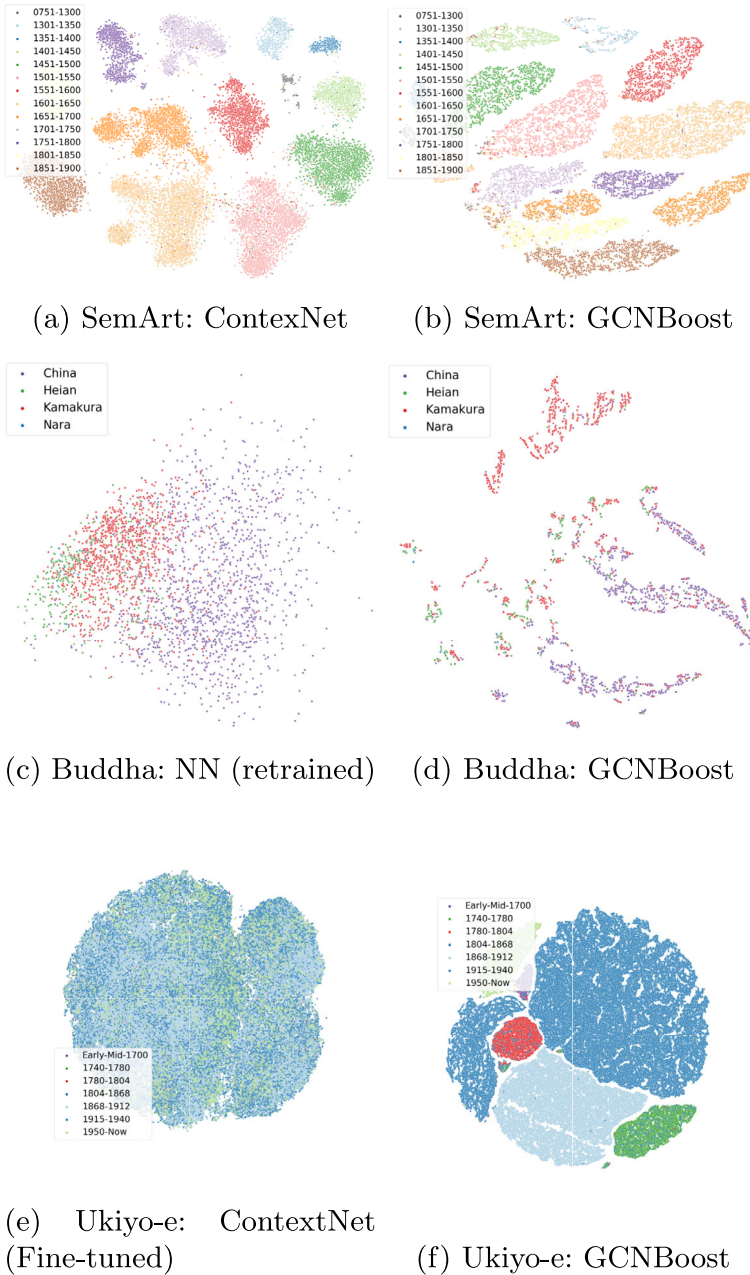
Finlay, we conducted the same experiments on the Ukiyo-e dataset and the results are reported in Table 7). We found the focal loss to be the most accurate for the classification of the categories **Era** and **Artist** and second best for **Date** classification (see lines 6 and 1). The multi-task models with GCN and GAT are able to improve the accuracy of the initial GCNBoost model on both **Era** and **Artist** classification.

With the new experiments using GAT, multi-tasking and Focal loss we were able to improve the accuracy of our GCNBoost model on the three different datasets. We showed the interest of using multi-tasking (see line 2 of Table 5), the interest of graph attention mechanism (see line 3 of Table 6) and the interest of the focal loss function as one can see in line 6 of Table 7). Thus we recommend GAT for small dataset such as the Buddha dataset, for middle-sized dataset multi-tasking seems to work the best as shown in [1] while for large datasets such as the Ukiyo-e dataset we recommend to use the focal loss model.

## 5 Discussion

To evaluate the effectiveness of our EKG and GNN approach, we conduct additional analysis to examine how node degree distributions impact the EKG and how well GNNBoost can cluster various art pieces with shared characteristics, akin to homophily in real-world networks.

Homophily [38], which is a well-studied concept in complex networks, suggests that entities with similar attributes tend to cluster together. This idea is observed in social ties [39] where people with shared characteristics tend to form cohesive groups, bringing cohesion to clusters of entities [40]. In the context of combining KG and pseudo-labels based on visual features, the notion of homophily is used to suggest that visually similar art pieces are likely to share similar semantic properties and context. While pseudo-labels can introduce noise, GNN can help mitigate this issue by considering the larger neighborhood, smoothing out the noise. However, this approach is limited when there is insufficient information due to the long-tail distribution of node degrees in real-world data [41]: a few observations tend to be



**Fig. 6** Visualization using tSNE [37] of the embeddings of ContextNet [1] (a) and our GCNBoost model (b), learned when predicting the category *TimeFrame* on SemArt [3], the NN [2] (c) and our GCNBoost (d) on Buddha statues [2] and finally the ContextNet fine-tuned on the Ukiyo-e dataset (e) and our GCNBoost (f)

extremely well connected, while many observations are little or not connected. As a result, some nodes have a disproportionate number of connections, leading to imbalanced label and attribute distributions. For example, the attribute type *Author* can have a skewed distribution (see Fig. 3d), making it challenging for a random walk to reach authors with low degrees and predict them to art pieces. We have explored the use of focal loss to address this issue, but it only helps with very small datasets. Therefore, we may consider using distribution-balanced loss [42] to mitigate this problem.

We used tSNE [37] to visualize the embeddings learned by ContextNet (Fig. 6a) and GCNBoost (Fig. 6b) when predicting the category *TimeFrame* on SemArt, as well as the embeddings learned by ContextNet (Fig. 6e) and GCNBoost (Fig. 6f) when predicting the category *Era* on Ukiyo-e, and the embeddings of the baseline NN model (Fig. 6c) and GCNBoost (Fig. 6d) when predicting the category *Style* on Buddha face. By following the idea of homophily [38], we observed that ContextNet (Fig. 6) generally clusters the different *TimeFrame* and *Era* categories, but still shows some ambiguities between the clusters. However, the Resnet50-based NN model for the Buddha face dataset did not produce significant clusters (Fig. 6c). In contrast, GNNBoost (Fig. 6b, d and f) tended to yield well-separated clusters (although not limited to separating art pieces based only on their *TimeFrame* or *Style*). For future work, we plan to investigate the combinations of visual and semantic attributes that can explain the observed clusters.

## 6 Conclusion and Perspectives

This work introduces a new approach to improving the use of a knowledge graph (KG) for multi-label classification using transductive learning. Our method involves incorporating unlabeled samples, which are assigned pseudo-labels through label propagation to create an enriched KG, or EKG. By utilizing GNNs on the EKG, we were able to enhance the performance of classification tasks across three distinct art datasets (SemArt [3],<sup>2</sup> Buddha statues [2], and Ukiyo-e [10]<sup>3</sup>). Our experiments also demonstrated the ability of our method to address data imbalance. Moving forward, we plan to explore various strategies for assigning pseudo-labels (e.g. incremental assignment) and implementing different GNN loss functions (e.g. distribution-balanced loss) to further address data imbalance. These results suggest that utilizing unlabeled samples can lead to improved automatic art analysis and guide our future work towards zero-shot learning for automatic art analysis.

**Funding** Open Access funding provided by Osaka University. This work is partly supported by JSPS KAKENHI Grant Numbers JP18H03571, JP18H03264, and JP20K19822.

**Data Availability** The datasets SemArt [3] and Ukiyo-e [10] are available online. Buddha dataset [2] will be made available on reasonable request.

## Declarations

**Conflict of Interest** The authors declare that they have no Conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence,

<sup>2</sup> <https://researchdata.aston.ac.uk/id/eprint/380/>

<sup>3</sup> <https://codh.rois.ac.jp/ukiyo-e/face-dataset/>

and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Garcia N, Renoust B, Nakashima Y (2019) Context-aware embeddings for automatic art analysis. In: Proceedings of the international conference on multimedia retrieval, pp 25–33
- Renoust B, Oliveira Franca M, Chan J, Garcia N, Le V, Uesaka A, Nakashima Y, Nagahara H, Wang J, Fujioka Y (2019) Historical and modern features for buddha statue classification. In: Workshop on structuring and understanding of multimedia heritage contents, pp 23–30
- Garcia N, Vogiatzis G (2018) How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the european conference on computer vision (ECCV), pp 0–0
- Khan FS, Beigpour S, Weijer J, Felsberg M (2014) Painting-91: a large scale database for computational painting categorization. *Mach Vis Appl* 25(6):1385–1397
- Provost F, Kolluri V (1999) A survey of methods for scaling up inductive algorithms. *Data Min Knowl Disc* 3(2):131–169
- Isen A, Tolias G, Avrithis Y, Chum O (2019) Label propagation for deep semi-supervised learning. In: Conference on computer vision and pattern recognition (CVPR)
- Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B (2003) Learning with local and global consistency. In: International conference on neural information processing systems. NIPS'03, pp 321–328
- El Vaigh CB, Torregrossa F, Allesiardo R, Gravier G, Sébillot P (2020) A correlation-based entity embedding approach for robust entity linking. In: 2020 IEEE 32nd International conference on tools with artificial intelligence (ICTAI). IEEE, pp 949–954
- Shi W, Gong Y, Ding C, Tao ZM, Zheng N (2018) Transductive semi-supervised deep learning using min-max features. In: Proceedings of the european conference on computer vision (ECCV), pp 299–315
- Khan SJ, Noord N (2021) Stylistic multi-task analysis of ukiyo-e woodblock prints. In: 32nd British machine vision conference 2021, BMVC 2021, Online, November 22–25, 2021. BMVA Press, pp 213. <https://www.bmvc2021-virtualconference.com/assets/papers/1514.pdf>
- El Vaigh CB, Garcia N, Renoust B, Chu C, Nakashima Y, Nagahara H (2021) Gcnboost: artwork classification by label propagation through a knowledge graph. In: Proceedings of the 2021 international conference on multimedia retrieval, pp 92–100
- Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. *Stat* 1050:20
- Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
- Carneiro G, Da Silva NP, Del Bue A, Costeira JP (2012) Artistic image classification: an analysis on the printart database. In: European conference on computer vision. Springer, pp 143–157
- Mensink T, Van Gemert J (2014) The rijksmuseum challenge: Museum-centered visual recognition. In: Proceedings of international conference on multimedia retrieval, pp 451–454
- Yelizaveta M, Tat-Seng C, Irina A (2005) Analysis and retrieval of paintings using artistic color concepts. In: 2005 IEEE international conference on multimedia and expo. IEEE, pp 1246–1249
- Li J, Yao L, Hendriks E, Wang JZ (2011) Rhythmic brushstrokes distinguish van gogh from his contemporaries: findings via automated brushstroke extraction. *IEEE Trans Pattern Anal Mach Intell* 34(6):1159–1176
- Shamir L, Macura T, Orlov N, Eckley DM, Goldberg IG (2010) Impressionism, expressionism, surrealism: automated recognition of painters and schools of art. *ACM Trans Appl Percept* 7(2):1–17
- Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds) Proceedings of international conference on learning representations
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- Bar Y, Levy N, Wolf L (2014) Classification of artistic styles using binarized features derived from a deep neural network. In: European conference on computer vision. Springer, pp 71–84
- Ma D, Gao F, Bai Y, Lou Y, Wang S, Huang T, Duan L-Y (2017) From part to whole: who is behind the painting? In: Proceedings of the 25th ACM international conference on multimedia, pp 1174–1182

23. Mao H, Cheung M, She J (2017) Deepart: learning joint representations of visual arts. In: Proceedings of the 25th ACM international conference on multimedia, pp 1183–1191
24. Strezoski G, Worring M (2018) Omniart: a large-scale artistic benchmark. *ACM Trans Multimed Comput Commun Appl (TOMM)* 14(4):1–21
25. Grover A, Leskovec J (2016) Node2vec: scalable feature learning for networks. In: Proceedings of the international conference on knowledge discovery and data mining, pp 855–864
26. Singh Dhami D, Yan S, Natarajan S (2021) Bridging graph neural networks and statistical relational learning: relational one-class gcn 2102
27. Rasool Izadi M, Fang Y, Stevenson R, Lin L (2020) Optimization of graph neural networks with natural gradient descent 2008
28. Liu Z, Jiang Z, Feng W, Feng H (2020) Od-gcn: object detection boosted by knowledge gcn. In: 2020 IEEE international conference on multimedia & expo workshops (ICMEW). IEEE, pp 1–6
29. Chen Z-M, Wei X-S, Wang P, Guo Y (2019) Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5177–5186
30. Wang X, Ye Y, Gupta A (2018) Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6857–6866
31. Zhang D, Cui M, Yang Y, Yang P, Xie C, Liu D, Yu B, Chen Z (2019) Knowledge graph-based image classification refinement. *IEEE Access* 7:57678–57690
32. Renoust B, Cheung G, Satoh S (2017) Estimating political leanings from mass media via graph-signal restoration with negative edges. In: 2017 IEEE international conference on multimedia and expo (ICME). IEEE, pp 1009–1014
33. Douze M, Szlam A, Hariharan B, Jégou H (2018) Low-shot learning with large-scale diffusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3349–3358
34. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: 5th International conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. <https://openreview.net/forum?id=SJU4ayYgl>
35. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Proceedings of the international conference on learning representations
36. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057
37. Maaten L, Hinton G (2008) Visualizing data using t-sne. *J Mach Learn Res* 9(11)
38. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. *Science* 323(5916):892–895
39. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. *Ann Rev Sociol* 27(1):415–444
40. Renoust B, Melançon G, Viaud M-L (2014) Entanglement in multiplex networks: understanding group cohesion in homophily networks. In: Social network analysis-community detection and evolution. Springer, pp 89–117
41. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. *SIAM Rev* 51(4):661–703
42. Wu T, Huang Q, Liu Z, Wang Y, Lin D (2020) Distribution-balanced loss for multi-label classification in long-tailed datasets. In: European conference on computer vision. Springer, pp 162–178