A Machine Learning Approach for Predicting Human Preference for Graph Drawings

Shijun Cai, Seok-Hee Hong, Jialiang Shen, and Tongliang Liu

University of Sydney, Australia

Abstract. Understanding what graph layout human prefer and why they prefer such graph layout is significant and challenging due to the highly complex visual perception and cognition system in the human brain. In this paper, we present the first machine learning approach for predicting human preference for graph layouts. Specifically, we propose a CNN-Siamese-based model to predict human preference from a pair of different layouts of the same graph. We employ a transfer learning method to overcome the insufficiency of the available ground truth human preference experiment data for training deep neural networks. Specifically, we exploit the quality metrics, which are correlated to human preference on graph layouts, to pre-train our model. Then, we fine-tune the model using the ground truth human preference experiment data.

Experimental results using the ground truth human preference data sets show that our model M+HP can successfully predict human preference for graph layouts, achieving the average test accuracy of 92.28% for large scale-free and mesh graphs. To our best knowledge, this is the first approach for predicting qualitative evaluation of graph layouts based on the ground truth human preference experiment data. Moreover, comparison experiments show that our model outperforms a simple baseline model and a previous Siamese-based model, demonstrating the importance of using graph layout images and the CNN-based model for predicting human preference.

This research is supported by ARC (Australian Research Council) Linkage Project (LP160100935) with Oracle Research lab. A preliminary version of this paper appeared in [3].

E-mail address: scai5619@uni.sydney.edu.au, seokhee.hong@sydney.edu.au, jshe9143@uni.sydney.edu.au, and tongliang.liu@sydney.edu.au (Shijun Cai, Seok-Hee Hong, Jialiang Shen, and Tongliang Liu)

This work is licensed under the terms of the CC-BY license.
1 Introduction

Evaluation of graph layouts is a significant problem in graph drawing. A number of quality metrics (or aesthetic criteria), such as edge crossings, bends, drawing area, total edge lengths, angular resolution, and stress, have been proposed for the quantitative evaluation of graph layouts [7]. Consequently, various graph drawing algorithms to optimize these metrics have been developed [7].

Qualitative evaluation on graph layouts is also available, using HCI (i.e., Human Computer Interaction) methodology, using human preference or specific task performance, measuring time and error. For example, edge crossings are shown to be important aesthetic criteria for performing human preference and shortest path tasks on graph layouts [34, 38]. Furthermore, large crossing angles are shown to be effective for shortest path tasks on graph layouts, when edge crossings are present [23].

Understanding what layout human prefer and why they prefer a specific graph layout over the others is significant, since it motivates researchers to design algorithms to compute such layouts, and guides users to choose specific algorithms to produce such layouts. However, it is extremely challenging due to the highly complex human visual perception and cognition system involving massively parallel processing using vision and memory in the human brain [37].

A series of human preference experiments have been conducted to better understand which graph layout human prefer. For example, Purchase [34] found the correlation between the human preference and fewer edge crossings in graph layouts. More recently, Chimani et al. [4] found the correlation between the human preference and lower stress in graph layouts, and Eades et al. [9] found the correlation between the human preference and higher shape-based metrics in graph layouts.

In this paper, we present the first deep learning approach for predicting human preference for graph layouts. Specifically, we propose a CNN-Siamese-based model that can be trained to predict human preference from a given pair of layouts of the same graph. Roughly speaking, CNN models read images rendered from graph layouts and convert them into feature vectors, which are inspired by the processing procedure of the human brain [14]. The Siamese model computes the difference between a pair of feature vectors.

To understand the procedure, we assume there exists a human preference measurement that measures a pair of layouts to a human preference label (that indicates which layout human prefer). The model can be regarded as a measurement from the input layout pair to the output prediction, which is trained to mimic the ground truth human preference measurement by fitting the training data which contains pairs of layouts and human preference labels.

The amount of the ground truth human preference data sets from existing human experiments [9] is relatively limited and insufficient for training deep neural networks. To address this, we train our model by employing the transfer learning method [33], which exploits data from related problems to help the original problem (e.g., human preference prediction).

Specifically, we train our deep neural networks by employing the layout pairs labeled by quality metrics (i.e., shape-based metrics, edge crossing and stress), shown to be correlated to human preference for graph layouts [4, 9]. More specifically, we first pre-train our model M+HP using quality-metrics-based pairs, which consists of two different graph layouts of the same graph and labels based on the correlated quality metrics. Then, we fine-tune the model using the human preference pairs, which consists of two different graph layouts of the same graph and human preference labels based on the ground truth human preference data. Extensive experiments show that our model M+HP successfully predicts human preference for graph layouts, achieving the average test accuracy of 92.28% for large scale-free and mesh graphs.
The main contribution of this paper is summarized as follows:

1. We present the first machine learning approach to predict human preference for graph layouts. Specifically, we propose a CNN-Siamese-based model to predict human preference from a given pair of different layouts of the same graph. To our best knowledge, this is the first approach for predicting qualitative evaluation of graph layouts by exploiting the ground truth human preference experiment data [9]. Note that our work differs from the other existing work using machine learning approaches for solving various problems in graph drawing [19, 25, 28], which mainly focus on quantitative evaluation, see Section 2.6 for the details.

2. We introduce a transfer learning method to overcome the insufficiency of the available ground truth human preference experiment data for training deep neural networks. Specifically, we pre-train our model M+HP by exploiting the quality metrics, which are correlated to human preference on graph layouts, and then fine-tune the model using the ground truth human preference experiment data [9].

3. Extensive experiments using the ground truth human preference data [9] show that our model M+HP successfully predicts human preference for graph layouts. Specifically, our model M+HP achieves average test accuracy of 92.28% for large scale-free and mesh graphs, and 63.77% for small sparse and biconnected graphs, significantly outperforming random guessing (i.e., greater than 50%) for the binary human preference problem. For large scale-free and mesh graphs, some layouts have much better quality than other layouts, and there was a strong preference among the layouts in the human preference data, resulting in high test accuracy. On the other hand, for small sparse and biconnected graphs, most layouts have similar good quality, and there was no strong preference among the layouts in the human preference data, which makes it more difficult to predict. Moreover, comparison experiments show that our model M+HP outperforms a simple baseline model B and a previous Siamese-based model DM [25] for all types of graphs, demonstrating the importance of using graph layout images and the CNN-based model for predicting human preference.

This paper is organized as follows. Section 2 describes the background, and Section 3 presents our CNN-Siamese-based machine learning model in detail. Section 4 describes experimental results with discussion, and Section 5 concludes with future work.

2 Background

2.1 Quantitative Evaluation for Graph Drawing

Various quality metrics for the evaluation of graph drawings, called aesthetic criteria, are available [7]. Traditional readability metrics include edge crossings, bends, area, total edge lengths and angular resolution. Consequently, many graph drawing algorithms have been designed to optimize these quality metrics [7]. However, most of these metrics consider the readability of graph drawings (i.e., how human better understand the graph drawings) and tend to focus on small graphs.

Recently, new faithful metrics have been developed, which measure how faithfully graph drawings visually display the ground truth structures of graphs. For example, Eades et al. [9] introduced
the shape-based metrics, by comparing the similarity between a graph $G$ with a proximity graph $G'$ computed from a drawing of $G$. The stress [7] is a distance faithful metrics, which compare the difference between graph theoretic distance of vertices and the Euclidean distance in a drawing.

Similarly, the cluster faithful metrics [31] compare the similarity between the ground truth clustering of a graph $G$ and the geometric clustering computed from a drawing of $G$. The symmetry faithful metrics [32] measure how the ground truth automorphisms of a graph are displayed as symmetries in a drawing, by computing exact/approximate geometric symmetry detection in $O(n \log n)$ time.

2.2 Qualitative Evaluation for Graph Drawing

Qualitative evaluation on graph layouts have been investigated by conducting the HCI-style human experiments, mostly associated with specific task performance, measuring time and error. For example, the seminal results by Purchase and Ware [34, 38] showed that small edge crossings are important aesthetic criteria for performing shortest path tasks in graph layouts.

Huang et al. [23] showed that large crossing angles are effective for graph reading performances (i.e., the shortest path task), which initiates new criteria of maximizing crossing angles [5, 10], and a new theory on RAC (Right Angle Crossing) graphs [8], as part of beyond planar graphs [21, 22]. Recent studies find that human untangling interaction task of hairball-like graph layouts [30] is positively correlated with the shape-based metrics, while surprisingly negatively correlated with the edge crossings and stress [9].

2.3 Human Preference Experiments in [4, 9]

More recently, a series of human preference experiments have been conducted [4, 9]. Specifically, in the human preference experiments, the system in Figure 1 showed two layouts of the same graph, randomly chosen from five different graph layouts, including force-directed layouts (such as FR [12]), stress minimization layouts and multi-level layouts (such as FM3 [17]). The task for participants was to choose their preferred layout from a pair of different layouts of the same graph, and select their preference score using a slider bar scaled from 0 to 5.

The data set used in the experiment includes well-known test suits such as Hachul’s library [17], Walshaw’s Graph Partitioning Archive 1, and randomly generated sparse and biconnected graphs.

The first experiment conducted at the University of Osnabrück [4] found the correlation between human preference for graph layouts and edge crossings and stress. Namely, humans prefer graph layouts with less stress and fewer crossings. The two follow-up experiments [9] conducted at the Graph Drawing conference 2014 and the University of Sydney, showed that the shape-based metrics are positively correlated with human preference, i.e., humans prefer graph layouts with high shape-based metrics.

2.4 Deep Learning

Recently, deep learning has achieved great success in various fields, such as computer vision, natural language processing, and speech recognition. The Convolutional Neural Network (CNN) is a representative deep neural network for image recognition and classification. CNNs are a type of multi-layer neural networks, designed and trained to recognize the nature of images by varying

\footnote{1https://chriswalshaw.co.uk/partition/}
the depth and breadth of a model [27]. CNNs can directly extract features from inputs of images by understanding the RGB values of pixels.

In the 1990s, LeCun et al. [29] introduced the first modern CNNs - LeNet-5 that can be successfully deployed for zip code and handwritten digit recognition. In 2012, Krizhevsky et al. [27] introduced the winning model – AlexNet that achieved outstanding performance in labeling natural images at the ImageNet challenge, which makes CNNs become the standard for image classification. After the AlexNet, much deeper and more complex CNNs has been developed, such as VGG (Visual Geometry Group) [35], GoogLeNet/Inception [24] and ResNet-50 [20].

Siamese neural networks were introduced by Bromley and LeCun to solve signature verification as an image discrimination problem [2]. Specifically, a Siamese neural network joins together the highest-level feature representations of twin inputs for image classification problems. For example, Koch et al. [26] used Siamese neural networks to rank the similarity between multiple inputs and discriminated input features.

2.5 Transfer Learning

Transfer learning [33] aims to improve the learning performance of a target task (or problem) by borrowing knowledge from related but different tasks, where the main idea is to learn task-invariant data representations [13]. Specifically, transfer learning transfers knowledge across different tasks to improve learning performance. Typically, if the target task has limited training examples, by using transfer learning, we could use the related tasks (called source tasks) that have sufficient training data. By exploiting the relationship between the source and target tasks, different assumptions have been proposed for transfer learning [13, 39], e.g., covariate shift and target shift.

For example, in computer vision, complex deep neural networks, e.g., AlexNet [27] and VGG [35], are often trained by employing the transfer learning technique to leverage the large-scale dataset ImageNet [6]. Specifically, the networks are usually pre-trained on ImageNet first and then are
fine-tuned on the datasets of the target tasks.

2.6 Deep Learning Approaches in Graph Drawing

A number of researchers used deep learning methods for problems in graph visualization, mainly focusing on quantitative evaluation, i.e., quality metrics [19, 25, 28]. For example, Haleem et al. [19] used a CNN model to predict multiple readability metrics, e.g., node spread, group overlap and edge crossings, using graph layout images with up to 600 vertices. Our work aims to predict qualitative evaluation, i.e., the human preference between two graph layouts.

Kwon and Ma [28] designed a GNN-based (Graph Neural Network) encoder-decoder neural network to generate good layouts from the test layouts. In our work, we choose the CNN models to read graph layout images and convert them into feature vectors (i.e., inspired by the processing procedure of the human brain), naturally following the original human preference experiments in [4, 9], where participants read graph layout images and then choose a preferred layout.

Klammler et al. [25] used a Siamese neural network DM for comparing a graph layout $D$ with its deformed layout $D'$. Specifically, the model input consists of quality metrics of $D$ and $D'$, and a multi-layer perceptron model learned the combined feature of the two sets of quality metrics. Therefore, their work predicts a better quality layout based on quantitative evaluation (i.e., quality metrics).

Note that our work utilizes the CNN-Siamese-based model based on the ground truth human preference experiment data [9], by comparing two graph layouts $D_1$ and $D_2$, computed using two different graph layout algorithms. Therefore, our work predicts qualitative evaluation (i.e., human preference). Furthermore, in Section 4, the experimental comparison shows that our model M+HP outperforms DM, demonstrating the importance of using graph layout images and the CNN-based model for predicting human preference.

3 A Machine Learning Approach

This Section presents our machine learning approach. Section 3.1 describes the CNN-Siamese-based Model in detail, and Section 3.2 explains how to employ transfer learning for predicting human preference for graph layouts. Section 3.3 introduces two labeled pairs, i.e., human preference pairs and quality-metrics-based pairs, and describes how to compute them.

3.1 A CNN-Siamese-based Model

We present a CNN-Siamese-based model that can predict which layout human prefer from a given pair of layouts. The notable advantage of CNNs is that they are powerful in extracting features from image inputs. The use of the Siamese model is natural since it deals with a pair of layouts to measure their difference. Specifically, Siamese neural network consists of twin feature extractors and a subtraction part to compute the difference between the input pair of layouts.

Figure 2 shows the pipeline of our model, including four essential parts: (a) Input data, (b) Twin CNN-based image feature extractors, (c) Subtraction part of the Siamese model, and (d) Output prediction. We now explain each part of the model in detail.
3.1.1 (a) Input data

To predict human preference using the machine learning model, we need input data for training and validation as well as testing. Specifically, the input data consists of a pair of layouts in color images, where a color image is a set of arrays with RGB pixel values, and a label (i.e., the human preference label $L_{HP}$ or the correlated quality metrics label $L_{M}$ described in Section 3.3), indicating which layout is better.

3.1.2 (b) Twin CNN-based image feature extractors

Our twin CNN-based image feature extractors, built on VGG [35], convert the input images into semantic feature vectors. Then the following parts (i.e., (c) and (d) in Figure 2) of the Siamese model output a prediction for human preference based on the semantic feature vectors. We now explain the details of the CNN-based feature extractor, as shown in Figure 2(b).

Convolutional layers are efficient in extracting semantic features of an image, which are inspired by the processing procedure of the human brain [14]. Multiple hidden layers are essential to increase the expressive ability of the deep neural network. Note that a large number of convolutional layers would increase the number of parameters and may cause the overfitting problem. In this paper, we set the number of the convolutional layers as two to demonstrate the proof of concept that predict human preference in graph layouts. A refined number of convolutional layers may further improve performance.

Max pooling layers retain the most significant features by down-sampling input features. Specifically, they down-sample the input feature by taking the maximum value over a window, defined...
by pool size (e.g., 2 \times 2 pool shown in Figure 2(b)). The **Fully connected layer** summarizes features for feature subtraction in the Siamese model (see Figure 2(c)).

### 3.1.3 (c) Subtraction part of the Siamese model

The subtraction part converts the pair of semantic feature vectors output by the twin feature extractors in part (b) into a single value in the range [0, 1] to predict human preference in part (d). Specifically, the two feature vectors are combined into a single vector by a **subtract layer**, which is reduced to a single value by employing a **fully connected layer** with a Sigmoid activation function, which ensures that the predicted value is in the range [0, 1].

### 3.1.4 (d) Output prediction

If the predicted value from (c) is smaller than 0.5, then we assign the predicted label \( L_P \) as 0 (i.e., the first layout is preferred by a human than the second layout); otherwise assign 1 (i.e., the first layout is less preferred by human than the second layout).

We then compare \( L_P \) with the corresponding label (i.e., the human preference label \( L_{HP} \) or the quality-metrics-based label \( L_M \) described in Section 3.3) when training, and compare \( L_P \) with \( L_{HP} \) when testing.

### 3.2 Transfer Learning

In general, deep neural networks have complex hypothesis classes. To train a deep neural network to understand human preference, we need a large amount of human-labeled pairs of layouts. However, annotating a large number of layout pairs is usually time-consuming and expensive. Fortunately, we can address the issue by employing the transfer learning technique [33], which helps us to reduce the complexity of the hypothesis class.

Intuitively, if we assume that the training data and the unknown test data are independent and identically distributed, a model well-trained (i.e., fit the training data well without overfitting) on the training data would generalize well on the test data (i.e., the test classification error will be similar to the training classification error). Increasing the training sample size and controlling the complexity of the hypothesis class are efficient ways to avoid overfitting and guarantee a good generalization property [36].

In this paper, we employ the transfer learning technique to reduce the hypothesis complexity of the deep neural networks used, since the training sample with human preference could be limited. Using transfer learning technique, we can train our model to minimize the difference between the predicted value and the ground truth label. By doing so, we hope that for a coming and unseen layout pair, the trained model can provide a prediction, which is close to the human preference label.

It has been shown that some quality metrics (i.e., shape-based metrics, edge crossing and stress) are correlated to human preference [4, 9]. Since such quality metrics are relatively easy to compute, we could easily obtain example pairs labeled by the quality metrics to help train our model. The mechanism that human use to decide more preferred layout could be very complicated due to the highly complex visual perception and cognition system in the human brain. Although some quality metrics are known to be correlated to human preference, the precise relationship between them remains unknown, which makes it difficult to introduce the covariate shift or target shift. We therefore transfer the hypothesis, using the related data to pre-train our model and then use the
target task data to fine-tune our model. This could be interpreted as putting some constraint on the hypothesis to learn, thus reducing the hypothesis complexity and training the model well.

Specifically, to employ the transfer learning technique, there are two stages in the training procedure:

1. the first stage M: we pre-train our model using layout pairs labeled by correlated quality-metrics-based label $L_M$;
2. the second stage HP: we fine-tune the model using layout pairs labeled by the human preference label $L_{HP}$.

### 3.3 Computing Labels

In this section, we introduce two labeled pairs, i.e., the human preference pairs and the quality-metrics-based pairs, and describe how to compute the human preference label $L_{HP}$ and the quality-metrics-based label $L_M$.

#### 3.3.1 Human preference pairs labeled by $L_{HP}$

Human preference pairs are processed from the ground truth data of the human preference experiments [9]. In the experiments, given a pair of layouts of the same graph, participants are required to choose their preferred layout with a preference score ranging from 0 to 5, where 0 means that the two layouts have the same preference and 5 means that the chosen layout is the most preferred.
Since human preference can be subjective, different participants may have different preferences for the same pair of layouts. Therefore, we use the average human preference scores to compute a human preference label. Specifically, for a layout pair $D_{i,j}$ and $D_{i,k}$ of the same graph $G_i$, where $j < k$, the human preference label $L_{HP}$ is computed as follows:

1. Let $n$ be the number of occurrences of the layout pair $D_{i,j}$ and $D_{i,k}$ in ground truth human preference data, where $P$ denote the preferred layout with the preference score $S$ (see Table 1(a), where $n = 5$).

2. For each occurrence, assign the weight $w$ using the preference score $S$: if $P$ is $D_{i,j}$, then assign $w = |S|$; otherwise (i.e., $P$ is $D_{i,k}$), set $w = -|S|$ (see Table 1(b)).

3. Compute the label $L_{HP}$ for the layout pair $D_{i,j}$ and $D_{i,k}$ as follows (see Table 1(c)):
   - if the average weight $\sum_{i=1}^{n} w_i/n > 0$, then assign the label $L_{HP} = 0$ (i.e., layout $D_{i,j}$ is more preferred by human than the $D_{i,k}$);
   - if $\sum_{i=1}^{n} w_i/n < 0$, then assign $L_{HP} = 1$ (i.e., $D_{i,k}$ is more preferred than $D_{i,j}$);
   - if $\sum_{i=1}^{n} w_i/n = 0$, then discard the layout pair without labeling.

After averaging human preference scores for each human preference pair, we have a human preference label 0 or 1. For example, Figure 3 and Table 1 show how to assign a human preference label. Specifically, for a layout pair $(D_{0,0}, D_{0,1})$, we label the pair $(D_{0,0}, D_{0,1})$ as $L_{HP} = 0$, as in Table 1(c) (i.e., human prefers layout $D_{0,0}$ than $D_{0,1}$).

### 3.3.2 Quality-metrics-based pairs labeled by $L_M$

Since the size of layout pairs with human preference labels can be small, we employ the transfer learning technique to pre-train our model, using the layout pairs labeled by correlated quality metrics. Specifically, human preference is positively correlated to shape-based metrics and negatively correlated to edge crossing and stress [4, 9], we compute three labels $M_{sh}$ (using shape-based metrics), $M_e$ (using edge crossing) and $M_{st}$ (using stress) for each pair of layouts.

Since the human preference experiments [9] use five different layouts for each graph, we compute the quality-metrics-based labels for all possible ten pairs of layouts. Specifically, the quality-metrics-based label for a layout pair $D_{i,j}$ and $D_{i,k}$, $j < k$, of graph $G_i$ is computed as follows:

1. Compute the quality metrics values for the layouts $D_{i,j}$ and $D_{i,k}$: Let $M_{sh,j}$, $M_{e,j}$ and $M_{st,j}$ (resp., $M_{sh,k}$, $M_{e,k}$ and $M_{st,k}$) denote the values of the shape-based metrics, edge crossings, and stress values of layout $D_{i,j}$ (resp., $D_{i,k}$).

2. Assign intermediate labels $L_{sh}$, $L_{e}$, and $L_{st}$ based on the three quality metrics:
   - if $M_{sh,j} > M_{sh,k}$ (resp., $M_{e,j} < M_{e,k}$ and $M_{st,j} < M_{st,k}$), then set $L_{sh} = 0$ (resp., $L_{e} = 0$ and $L_{st} = 0$);
   - if $M_{sh,j} < M_{sh,k}$ (resp., $M_{e,j} > M_{e,k}$ and $M_{st,j} > M_{st,k}$), then set $L_{sh} = 1$ (resp., $L_{e} = 1$ and $L_{st} = 1$);
   - if $M_{sh,j} = M_{sh,k}$ (resp., $M_{e,j} = M_{e,k}$ and $M_{st,j} = M_{st,k}$), then discard the layout pair without labeling.
Figure 4: Examples of the five layouts of a mesh graph $G_{10}$.

<table>
<thead>
<tr>
<th>$D_{1,0}$</th>
<th>$D_{1,1}$</th>
<th>$L_{sh}$</th>
<th>$L_{c}$</th>
<th>$L_{st}$</th>
<th>$L_{M}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{10,0}$</td>
<td>$D_{10,1}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$D_{10,0}$</td>
<td>$D_{10,2}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$D_{10,0}$</td>
<td>$D_{10,3}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$D_{10,0}$</td>
<td>$D_{10,4}$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$D_{10,1}$</td>
<td>$D_{10,2}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$D_{10,1}$</td>
<td>$D_{10,3}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$D_{10,1}$</td>
<td>$D_{10,4}$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$D_{10,2}$</td>
<td>$D_{10,3}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$D_{10,2}$</td>
<td>$D_{10,4}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$D_{10,3}$</td>
<td>$D_{10,4}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Example of assigning the quality-metrics-based label $L_{M}$ for ten layout pairs of a graph $G_{10}$ using intermediate labels $L_{sh}$, $L_{c}$, and $L_{st}$.

3. Compute the label $L_{M}$ for the layout pair $D_{i,j}$ and $D_{i,k}$ based on the majority voting using the intermediate labels:

- if the majority of intermediate labels is 0, then assign the final quality-metrics-based label $L_{M} = 0$ (i.e., layout $D_{i,j}$ is more preferred than $D_{i,k}$);
- if the majority of intermediate labels is 1, then assign $L_{M} = 1$ (i.e., $D_{i,k}$ is more preferred than $D_{i,j}$).

Table 2 shows examples of intermediate labels $L_{sh}$, $L_{c}$, and $L_{st}$, and the quality-metrics-based labels $L_{M}$ of the ten possible pairs of layouts of graph $G_{10}$ in Figure 4. For example, for the first layout pair $D_{10,0}$ and $D_{10,1}$, the intermediate labels are $L_{sh} = 1$, $L_{c} = 1$, and $L_{st} = 1$. By majority voting, we label the pair $(D_{10,0}, D_{10,1})$ as $L_{M} = 1$ (i.e., human prefers layout $D_{10,1}$ than $D_{10,0}$).

4 Experiments

This section presents the details of our experiment, including data sets, model training, prediction results, and discussion.
| Category | Type            | $|V|$      | $|E|$      | density  |
|----------|----------------|----------|----------|----------|
| small    | sparse         | 25 - 363 | 29 - 468 | 1.00 - 1.50 |
| small    | biconnected    | 34 - 240 | 78 - 477 | 1.92 - 2.94 |
| large    | mesh           | 397 - 8,000 | 729 - 15,580 | 1.41 - 1.95 |
| large    | scale-free     | 1,647 - 5,452 | 4,769 - 118,404 | 2.30 - 21.72 |

Table 3: The statistics of the data used in our experiments.

4.1 Data Sets

For our experiment, we use the data sets from the human experiments in [9] consisting of 146 graphs and their five layouts. Specifically, the graphs range in size from small (25 vertices and 29 edges) to large (8,000 vertices and 118,404 edges), and have different structures. More specifically, after we pre-process the ground truth human preference experiment data as described in Section 3.3.1, we obtain 511 human preference pairs. We also compute the ten quality-metrics-based pairs as described in Section 3.3.2, resulting in 1,460 quality-metrics-based pairs. Therefore, in total, we have 1,460 quality-metrics-based pairs for pre-training, and 511 human preference pairs for fine-tuning and testing.

In fact, we compare images rendered from graph layouts, and the CNN feature extractor extracts feature vectors from the image inputs, as described in Section 3.1.2. Specifically, we render all layouts using NetworkX [18], where the NetworkX.draw function was set with the vertex size as 0.6 in red color and the edge width as 0.2 in grey color. The image size is set as $320 \times 320$ in pixel when saved by employing the matplotlib.pylab function in Python.

To better examine the human preference on different types of graphs, we divide our data sets into four categories based on the size (i.e., small and large) and their structures (i.e., sparse, biconnected, mesh, and scale-free graphs). Table 3 shows the details of the human preference experiment data used in our experiment. Figures 3, 4, 5 show examples of mesh graphs (e.g., $G_{10}$ and $G_{6}$), scale-free graphs (e.g., $G_{13}$ and $G_{15}$), sparse graphs (e.g., $G_{0}$ and $G_{188}$), and biconnected graphs (e.g., $G_{18}$ and $G_{65}$).

4.2 Model Training

To demonstrate the effectiveness of our transfer learning approach presented in Section 3, we compare our model M+HP with two models M and HP as follows:

1. $M$: a model trained only on quality-metrics-based pairs.
2. $HP$: a model trained only on human preference pairs.
3. $M+HP$: our transfer learning model pre-trained using quality-metrics-based pairs labeled by quality-metrics-based label $L_M$, and then fine-tuned using human preference pairs labeled by human preference label $L_{HP}$.

We implement the models using the Keras library [16] in Python, and all experiments run on the Google Colab Pro [1]. To optimize the model, we use Adam optimizer with a learning rate of 0.01.

In the training phase, we aim to train our proposed deep neural networks to optimize the parameters of the deep neural networks by minimizing the difference between the predicted label $L_P$ and the corresponding labels (i.e., human preference label $L_{HP}$ or quality-metrics-based label $L_M$).
Figure 5: Examples of a sparse graph \((G_{188})\), biconnected graphs \((G_{18} \text{ and } G_{65})\), a mesh graph \((G_{6})\), and scale-free graphs \((G_{13} \text{ and } G_{15})\) with their five different layouts.

\(L_M\), where the difference can be minimized by the binary cross-entropy loss function. In the
testing phase, we compare $L_P$ with the human preference label $L_{HP}$ to evaluate the prediction results for the trained models.

Having a small training error but a large test error may cause overfitting, i.e., the model fits the training data well but cannot generalize well on the test data. To avoid overfitting, we use the following cross-validation method: randomly split the ground truth human experiment data into two data sets (i.e., a training data set and a test data set) with a ratio of 7:3, and then randomly select 30% pairs in the training set for validation, where the random split is repeated for five times employing the sklearn.model_selection.train_test_split function with the test_size = 0.3.

Moreover, to validate the effectiveness of our model using graph layout images and the CNN-based neural network, as well as the importance of using our quality metrics (e.g., shape-based metrics, edge crossings and stress) for training, we compare our model M+HP with three other models as follows:

2. $DM$ [25]: a Siamese neural network trained on a 57-dimensional feature vector of two layouts and a graph feature vector.
3. $DM2$: a variation of DM trained on our quality metrics instead of the 57-dimensional feature vector.

### 4.3 Prediction Results and Comparison of Models

Table 4 shows the experiment results on the test accuracy with our trained models (i.e., M, HP, M+HP) on each graph type, as well as the comparison with other models (i.e., B, DM, DM2). The number in each cell represents the average test accuracy with the standard deviation after five times of the random splitting.

<table>
<thead>
<tr>
<th>Type</th>
<th>$DM$</th>
<th>$DM2$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sparse</td>
<td>(50.84 ± 0.5)%</td>
<td>(53.25 ± 1.5)%</td>
<td>(56.02 ± 1.8)%</td>
</tr>
<tr>
<td>biconnected</td>
<td>(61.35 ± 2.1)%</td>
<td>(61.98 ± 2.7)%</td>
<td>(52.45 ± 1.4)%</td>
</tr>
<tr>
<td>Average small</td>
<td>(51.21 ± 1.3)%</td>
<td>(52.62 ± 2.1)%</td>
<td>(54.24 ± 1.6)%</td>
</tr>
<tr>
<td>mesh</td>
<td>(61.61 ± 2.5)%</td>
<td>(62.37 ± 2.8)%</td>
<td>(71.73 ± 3.6)%</td>
</tr>
<tr>
<td>scale-free</td>
<td>(57.81 ± 5.8)%</td>
<td>(57.88 ± 6.6)%</td>
<td>(58.23 ± 6.3)%</td>
</tr>
<tr>
<td>Average large</td>
<td>(59.71 ± 4.2)%</td>
<td>(60.13 ± 4.7)%</td>
<td>(64.99 ± 5.0)%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>$M$</th>
<th>$HP$</th>
<th>$M+HP$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sparse</td>
<td>(56.57 ± 3.1)%</td>
<td>(58.37 ± 3.1)%</td>
<td>(62.14 ± 2.6)%</td>
</tr>
<tr>
<td>biconnected</td>
<td>(52.49 ± 2.3)%</td>
<td>(61.39 ± 3.1)%</td>
<td>(65.40 ± 3.7)%</td>
</tr>
<tr>
<td>Average small</td>
<td>(54.53 ± 2.7)%</td>
<td>(59.84 ± 2.2)%</td>
<td>(63.77 ± 3.2)%</td>
</tr>
<tr>
<td>mesh</td>
<td>(76.49 ± 2.8)%</td>
<td>(82.31 ± 2.9)%</td>
<td>(86.55 ± 3.2)%</td>
</tr>
<tr>
<td>scale-free</td>
<td>(82.85 ± 4.7)%</td>
<td>(82.81 ± 3.7)%</td>
<td>(98.00 ± 4.5)%</td>
</tr>
<tr>
<td>Average large</td>
<td>(79.67 ± 3.7)%</td>
<td>(82.66 ± 3.3)%</td>
<td>(92.28 ± 3.8)%</td>
</tr>
</tbody>
</table>

Table 4: Test accuracy and standard deviation of six trained models: our model M+HP achieves the best test accuracy for all data types, demonstrating the effectiveness for predicting human preference for graph layouts.

The prediction results in Table 4 show that human preference for graph layouts can be predicted by a machine learning approach. Specifically, our model M+HP predicts human preference for a pair of graph layouts with an average test accuracy of 92.28% for large scale-free and mesh graphs, and 63.77% for small sparse and biconnected graphs, which significantly outperforms the random guessing for binary human preference problem.
4.3.1 Comparison between M, HP and M+HP

The test accuracy gradually increases along with M, HP and M+HP, as shown in Table 4. M+HP outperforms HP, esp. for large scale-free graphs, demonstrating the success of the transfer learning, namely, the importance of pre-training on quality-metrics-based pairs (i.e., layout images and $L_M$), and fine-tuning on human preference pairs (i.e., layout images and $L_{HP}$).

Note that HP performs better than M, esp. for small biconnected graphs, supporting the importance of using the ground truth human preference data over the quality metrics. Specifically, the average test accuracy for the model HP (resp., M) is 59.84% (resp., 54.53%) for small sparse and biconnected graphs, and 82.66% (resp., 79.67%) for large scale-free and mesh graphs.

4.3.2 Comparison with B, DM and DM2

The test accuracy gradually increases along with DM, DM2, B, M, HP and M+HP, as shown in Table 4. Note that our model M+HP outperforms B, DM and DM2 for all types of graphs, demonstrating the success of the transfer learning and the importance of using graph layout images with the CNN-based model for predicting human preference.

Similarly, M and HP also perform significantly better than B and DM, supporting the importance of using graph layout images and the CNN-based model for predicting human preference. Specifically, HP outperforms B and DM for all types of graphs, and M outperforms B and DM for large scale-free and mesh graphs.

Note that B performs significantly better than DM, esp. for small sparse graphs and large mesh graphs, indicating that a fully-connected neural network can be more effective than the Siamese neural network. Specifically, the average test accuracy for B (resp., DM and DM2) is 64.99% (resp., 59.71% and 60.13%) for large scale-free and mesh graphs, and 54.24% (resp., 51.21% and 52.62%) for small sparse and biconnected graphs. Furthermore, DM2 performs better than DM, esp. for small sparse graphs, showing the importance of using our quality metrics (i.e., shape-based metrics, edge crossings and stress).

4.3.3 Significance Test

To validate the comparison of the performance (i.e., statistically significant differences) among the six trained models, we conduct the Friedman test and the Wilcoxon signed-rank test. The Friedman test is a non-parametric statistical test that repeated measures ANOVA, and used to detect differences across multiple test sets [11]. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test method to compare the pair-wise models. Namely, it tests the location of a set of samples (e.g., each cross-validation testing set), and does not assume that the differences between paired samples are normally distributed [15].

Specifically, we run the significance test on the test accuracy of all the trained models employing the Friedman test (using scipy.stats.friedmanchisquare function), and then the Wilcoxon signed-rank test (using scipy.stats.wilcoxon function). More specifically, the significance test runs on a set of test accuracy values per random splitting and per graph type per model. Since there are four types of graphs and five times of random splitting for each type of graphs, we compare twenty test accuracy values of one model with the twenty accuracy values of the other model, respectively. The Friedman test result shows that $p$-value $= 0.0000000000000001 < 0.05$, i.e., there are statistically significant differences among the six models.

Table 5 shows the $p$-values of the Wilcoxon signed-rank tests for comparing the pair-wise trained models. The $p$-value depends on the median accuracy of the first model that is positive against
<table>
<thead>
<tr>
<th>Comparison</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>M+HP vs M</td>
<td>0.00000006</td>
</tr>
<tr>
<td>M+HP vs HP</td>
<td>0.00004286</td>
</tr>
<tr>
<td>HP vs M</td>
<td>0.00064817</td>
</tr>
<tr>
<td>M vs B</td>
<td>0.00153833</td>
</tr>
<tr>
<td>B vs DM2</td>
<td>0.00230330</td>
</tr>
<tr>
<td>DM2 vs DM</td>
<td>0.00268147</td>
</tr>
</tbody>
</table>

Table 5: The p-values of the Wilcoxon signed-rank tests for comparing six models. For each pair-wise comparison, the first model is significantly better than the second model, since the p-value < 0.05.

the median accuracy of the second model that is negative. The smaller the p-value, the better the first model, and p-value < 0.05 means that the difference is statistically significant.

Note that for all pair-wise comparisons, the first model is significantly better than the second model, since the p-value < 0.05. For example, the p-value = 0.00000006 (resp., 0.00004286) shows that our model M+HP is significantly better than M (resp., HP).

4.4 Summary and Discussion

4.4.1 Summary

The test accuracy in Table 4 and the Wilcoxon signed-rank tests in Table 5 (i.e., M+HP vs M and M+HP vs HP) show that our model M+HP can successfully predict the human preference for a pair of graph layouts, demonstrating the success of the transfer learning approach, which indicates that both the quality-metrics-based pairs and the human preference pairs are important for predicting human preference.

Note that HP trained by human preference pairs performs better than M trained by quality-metrics-based pairs, except for scale-free graphs, while the difference of accuracy over all types of graphs is significant (i.e., HP vs M). This shows the difference between the human preference pairs and the quality-metrics-based pairs, as well as the importance of the ground truth human preference experiment data, i.e., qualitative evaluation on graph layouts.

Moreover, M performs better than B, where the difference is significant (i.e., M vs B), demonstrating the importance of using graph layout images with the CNN-based model for predicting human preference. Similarly, B performs better than DM2, where the difference is significant (i.e., B vs DM2), suggesting that a fully-connected neural network can be more effective than a Siamese neural network for predicting human preference. Furthermore, DM2 performs better than DM, where the difference is significant (i.e., DM2 vs DM), demonstrating the importance of using our quality metrics (shape-based metrics, crossings and stress) for predicting human preference.

4.4.2 Large Scale-free and Mesh Graphs

For large scale-free and mesh graphs, some layouts have much better quality than other layouts visually and metric wise, leading to a high preference score in the ground truth human preference data. Therefore, the training and test data sets for the large scale-free and mesh graphs are more consistent without conflicts than the small sparse and biconnected graphs, resulting in much higher test accuracy for predicting the human preference label.
Figure 6: Examples of the test layout pairs for a mesh graph \( G_5 \) and a scale-free graph \( G_{13} \), where the layout on the left is more preferred by human than the layout on the right. Here, all three trained models succeed (\( \checkmark \)) to predict the human preference label.

Mesh graphs have distinct shapes, therefore, it is easier for participants to decide their preference with a high preference score. For example, Figure 6 shows two layouts of a mesh graph \( G_5 \), where \( D_{5,4} \) has visually better quality than \( D_{5,3} \). Therefore, \( D_{5,4} \) was the preferred layout with \( S = 5 \).

Large scale-free graphs have globally sparse and locally dense structures with small diameters, which often produce a tangled hairball drawing. Therefore, some graph layouts have much better quality than other layouts with poor quality, leading to participants easily choose their preference with a high preference score. For example, Figure 6 shows two layouts of of a scale-free graph \( G_{13} \), where \( D_{13,4} \) has visually better quality than \( D_{13,3} \). Therefore, \( D_{13,4} \) was preferred with \( S = 4 \).

Note that among the five graph layouts, \( D_{i,0} \), \( D_{i,1} \) and \( D_{i,4} \) are visually much better than \( D_{i,2} \) and \( D_{i,3} \). For example, see the five layouts of \( G_{10} \) in Figure 4, and \( G_6 \), \( G_{13} \) and \( G_{15} \) in Figure 5. Consequently, when a pair consists of layouts with different quality, i.e., \( D_{i,0} \) (resp., \( D_{i,1} \) and \( D_{i,4} \)) and \( D_{i,2} \) (resp., \( D_{i,3} \)), it is easy for participants to decide their preference consistently with high preference score. For example, Figure 6 shows layout pairs \( (D_{5,4}, D_{5,3}) \) and \( (D_{13,4}, D_{13,3}) \), where the three trained models all succeed to predict the human preference label.

On the other hand, when a pair consists of similar quality layouts, i.e., \( (D_{i,2}, D_{i,3}) \) or two layouts among \( D_{i,0} \), \( D_{i,1} \) and \( D_{i,4} \), it is difficult for participants to decide their preferences, resulting in low preference scores with conflicts between them. For example, Figure 7 shows a layout pair \( (D_{7,0}, D_{7,1}) \) with different preferences (i.e., \( D_{7,1} \) with \( S = 1 \) or \( D_{7,0} \) with \( S = 3 \)) among three participants, where M+HP and HP succeed to predict the human preference label, while M fails to predict. Similarly, a layout pair \( (D_{15,2}, D_{15,3}) \) has two low preference scores \( S = 1 \) or 2, where only M+HP succeeds to predict the ground truth human preference label, while M and HP fail to predict.
4.4.3 Small Sparse and Biconnected Graphs

Our three trained models M, HP, M+P all succeed to predict the human preference label for small sparse and biconnected graphs. For example, Figure 8 shows two layouts of a sparse graph $G_{185}$ and a biconnected graph $G_{42}$, where all three models succeed to predict the human preference label.

Note that for small sparse and biconnected graphs, all the five layout algorithms produced layouts with similar quality visually and metric wise. For example, see the sparse graphs $G_0$ in Figure 3 and $G_{188}$ in Figure 5, and biconnected graphs $G_{18}$ and $G_{65}$ in Figure 5. Therefore, participants tend to randomly choose a layout as their preference with a low preference score, leading to less consistency with possible conflicts in the training and test data sets.

Moreover, since the system used in the human preference experiments [9] randomly chooses two layouts of a graph, each participant answers his/her preference for a different set of layout pairs, and the average answers per each layout pair is quite small, around 2. These all together lead to possible conflicts in the training data and test data, resulting in much lower test accuracy than the large scale-free and mesh graphs.

For example, Figure 9 shows two layouts of a sparse graph $G_{185}$ and a biconnected graph $G_{66}$ with very low preference scores. For the layout pair $(D_{185,0}, D_{185,4})$ with the preference scores $S = 1, 2$, M+HP and HP succeed to predict the human preference label, while M fails to predict. For the layout pair $(D_{66,3}, D_{66,0})$ with the preference score $S = 1$, only M+HP succeeds to predict the human preference label, while M and HP fail to predict.

4.4.4 Implication and Limitation

Our trained models perform quite well on both large scale-free and mesh graphs as well as small sparse and biconnected graphs, however we found some limitations, which leaves room for further
improvement in the future.

The difficult cases along the decision boundary of similar quality layout pairs (i.e., average human preference score is close to 0) make the discriminative information insufficient to make a correct prediction. Moreover, the randomness of the set of layout pairs for each participant in the human preference experiments [9] as well as a small number of answers per layout pair lead to possible conflicts in the training data and test data, resulting in much lower test accuracy for the small sparse and biconnected graphs than the large scale-free and mesh graphs.

For example, all three trained models fail to predict the human preference label when a pair of layouts have similar visual quality and the preference score is very low, see a biconnected graph \((D_{113,2}, D_{113,1})\) with \(S = 1\) and a scale-free graph \((D_{13,4}, D_{13,0})\) with \(S = 1\) in Figure 10.

![Figure 8: Examples of the test layout pairs for a sparse graph \(G_{185}\) and a biconnected graph \(G_{42}\), where the layout on the left is more preferred by human than the layout on the right. Here, all three trained models succeed (\(\sqrt{\text{V}}\)) to predict the human preference label.](image)

Interestingly, there are some exceptional cases where only M or HP succeeds to predict the human preference label, while M+HP fails. For example, Figure 11 shows a layout pair \((D_{2,1}, D_{2,4})\) of a mesh graph, with \(P = D_{2,1}, S = 4\) or \(P = D_{2,4}, S = 3\), where only M succeeds to predict the human preference label. Here, the conflicting high preference scores may be due to subjective human preference. Similarly, Figure 11 also shows a layout pair of a sparse graph \((D_{187,0}, D_{187,4})\) with \(S = 2\), where only HP succeeds to predict the human preference label. Therefore, research on human preference for graph layouts deserves further investigation.

5 Conclusion

In this paper, we present the first deep learning approach, namely a CNN-Siamese-based neural network model, to predict human preference for graph layouts using the ground truth human preference data [9]. Due to the limited availability of the ground truth human preference data sets,
we also exploit the transfer learning technique and utilize correlated quality-metrics-based pairs for pre-training and human preference pairs for fine-tuning.

Experiments demonstrate that our model M+HP can successfully predict the binary human preference problem with an average test accuracy of 92.28% for large scale-free and mesh graphs,
Figure 11: Examples of the test layout pairs for a mesh graph $G_2$ and a sparse graph $G_{187}$, where the layout on the left is more preferred by human than the layout on the right. Here, only M (resp., HP) succeeds ($\sqrt{\times}$) to predict the ground truth human preference label.

and 63.77\% for small sparse and biconnected graphs. Moreover, comparison experiment results show that M+HP outperforms other models B and DM for all types of graphs, demonstrating the importance of using graph layout images with the CNN-based model for predicting human preference.

Note that the human preference experiment [9] used many small sparse and biconnected graphs, where all five graph layout algorithms produced visually similar good quality layouts, resulting in very low preference scores with possible conflicts. Moreover, it randomly chooses the set of layout pairs for each participant, leading to a small number of answers per layout pair and possible conflicts in the training data and test data, resulting in much lower test accuracy for the small sparse and biconnected graphs than the large scale-free and mesh graphs.

As a future work, we plan to conduct a new human preference experiment with different data sets with visually different layouts, to design a machine learning model for better predicting the human preference on graph layouts.

Acknowledgements

We thank anonymous reviewers for many constructive suggestions to improve the contribution and presentation of this paper.
References


