VisTCP: A Visualization Framework to Construct Knowledge-Graph-Based Representation for Traditional Chinese Painting

Zhiguang Zhou, Fengling Zheng, Miaoxin Hu, Lina You, Jin Wen, Huan Liu, Wei Zhang, Dekun Qian, Yuhua Liu, Wei Chen, Yigang Wang*, and Yong Wang*

***Abstract*—Structured representation can characterize semantic objects and relationships in images. It provides a possible effective way for the semantic understanding of Traditional Chinese Paintings (TCPs) to better support archaeology and art history research. However, most image-oriented structured representation methods perform poorly on TCPs, due to two major challenges: 1) the objects and events of TCPs exhibit substantial differences from modern natural images, which results in semantic misunderstandings of TCPs; and 2) it is difficult to achieve accurate identification of ancient objects and events in TCPs, even for domain experts. In this paper, we propose *VisTCP*, a visualization framework that combines a TCP-oriented intelligent model and expert knowledge, which enables art historians to achieve trustworthy structured representations of TCPs in a human-in-the-loop manner. Firstly, we conduct a pilot study with three domain experts to build a semantic taxonomy of TCPs. Then, expert-annotated data are used to train a TCP-oriented structured representation model, which can automatically extract meaningful objects and their relationships in TCPs. To inform users of the model uncertainty, we design a joint embedding visualization view to show the differences between expert annotations and model predictions. This allows users to refine the structured representation based on their domain knowledge, enabling iterative optimization of the model. Finally, we conduct a case study, a usage scenario, and expert interviews on a real dataset to demonstrate the effectiveness of *VisTCP* in supporting the structured representation and semantic understanding of TCPs.**

***Index Terms*—Traditional Chinese Paintings (TCPs), Visualization, Structured Representation, Image Understanding**

I. Introduction

With the digitization of traditional Chinese paintings(TCPs), computational methods have gained substantial attention in the fields of archaeology and art history, particularly for provenance analysis and authentication. However, art historians find that existing methods for structured understanding of TCPs, such as image classification, segmentation, and object detection, are inadequate for capturing the complex semantic content, which includes not only diverse objects, but also rich semantic events within the paintings. These methods often fail to support advanced semantic tasks such as semantic retrieval, question answering, and visual reasoning due to their inability to effectively represent semantic information.

Knowledge graphs or scene graphs have been proposed for semantic representation of natural and chart images due to their ability to encode entities and relationships within images [1]. It has been proven capable of supporting various high-level tasks, such as structured knowledge understanding [2], semantic retrieval [3], and visual question answering [4]. Inspired by this, it is highly desirable to characterize diverse objects of TCP and the inherent semantic relationships among them with a knowledge graph, thereby benefiting downstream tasks and empowering research in archaeology and art history.

Various automated methods have been developed based on large image datasets training, such as Visual Relationship Detection [5] and Scene Graph Generation (SGG) [6], to alleviate difficulties in repetitive and time-consuming knowledge extraction. However, in contrast to the large-scale data volume of natural images and the relatively simple visual elements and relationships in chart images, the sparse data volume, complexity of semantic content, and the diversity and professionalism of artistic styles in TCPs present significant challenges for knowledge graph construction. Specifically, **the objects and events depicted in TCPs differ markedly from those in natural images.** Moreover, **the accurate knowledge extraction is challenging due to significant variations in artistic styles across different TCPs.**

In this work, we aim to fill this gap by developing an interactive visualization framework to help users construct knowledge graph representations of TCPs. To meet comprehensive semantic representations of knowledge-graph-based structure for TCPs, we conducted an expert study with two art historians and analyzed insights from multiple rounds of expert interviews. Guided by the obtained high-level knowledge-graph-based representation structure, we designed and implemented *VisTCP*, an interactive visualization tool, that enables users to autonomously extract knowledge while enhancing their engagement in the construction of the TCP knowledge graph representation. We first conduct an expert study to develop a semantic taxonomy for TCP entities and relations

Equal contribution: Fengling Zheng and Zhiguang Zhou contributed equally to this work.

Zhiguang Zhou, Miaoxin Hu, Jin Wen, Dekun Qian, Yuhua Liu and Yigang Wang are with the School of Media and Design, Hangzhou Dianzi University (e-mail: zhgzhou@hdu.edu.cn, miaoxinhu@outlook.com, 231330023@hdu.edu.cn, qiandekun@hdu.edu.cn, liuyuhua@hdu.edu.cn, yigang.wang@hdu.edu.cn). Fengling Zheng and Lina You are with the School of Computer Science, Hangzhou Dianzi University (e-mail: fenglingzheng@hdu.edu.cn, linayou@hdu.edu.cn). Huan Liu, Wei Zhang and Wei Chen are with the State Key Lab of CAD&CG, Zhejiang University (e-mail: alisalh@zju.edu.cn, zw_yixian@zju.edu.cn, chenvis@zju.edu.cn). Yong Wang is with the College of Computing and Data Science, Nanyang Technological University (e-mail: yong-wang@ntu.edu.sg).

*Corresponding authors: Yigang Wang and Yong Wang

and to establish an initial set of expert-annotated samples. Building upon this, a structured representation model tailored for TCPs is constructed through transfer learning based on the object detection framework Mask R-CNN and the relationship inference model Total Direct Effect (TDE). Furthermore, we present a joint embedding visualization that aligns expert-annotated elements with model predictions, revealing uncertainty and incorporating expert knowledge for credible refinement of structured representations. In addition, feedback from art historians and newly annotated samples are collected for iterative model improvement, further reducing human effort in constructing structured representations of TCPs. Finally, a case study, a usage scenario, and expert interviews demonstrate that *VisTCP* effectively integrates intelligent models and prior expert knowledge to efficiently construct the structured representation for TCPs. This work mainly makes the following contributions:

- We conduct an expert study to derive a semantic taxonomy and construct an expert-annotated dataset for TCPs, upon which a TCP-oriented knowledge extraction model is developed to enable comprehensive semantic representations of TCPs.
- We develop a visualization framework, *VisTCP*, that tightly integrates joint embedding visualization with human–AI collaborative interaction, enabling art historians to refine structured representations and further optimize the intelligent model for structured representation.
- We conduct evaluations through a case study, a usage scenario, and expert interviews, demonstrating both the effectiveness of knowledge-graph-based representations in advancing TCP semantic understanding and the usefulness of *VisTCP*.

II. Related Work

A. Structure Representation for Traditional Chinese Painting understanding

Understanding Traditional Chinese Painting (TCP) requires interpreting multiple intertwined components, pictorial scenes, inscriptions, seals, and overall aesthetic composition. The pictorial scene, comprising *figures, objects, and landscapes* with their associated *events and spatial relationships*, constructs a narrative and symbolic world that reflects key dimensions of Chinese history and culture, forming the basis for the comprehensive understanding of TCPs [7].

Existing computational studies on TCP have explored various forms of *structured representation* to enhance digital understanding, supporting tasks such as content-based retrieval, classification, and visual exhibition. Early works extracted handcrafted visual features describing texture, layout, and composition for content-based retrieval [8], [9]. With the advent of deep learning, CNN-based methods were applied to painting classification and element detection [10]–[12], yielding higher-level representations of visual structure. Interactive visualization systems, such as ScrollTimes for element validation [13] and audio-visual annotation frameworks for immersive presentation [14], [15], further integrated visual and contextual information to improve interpretability. However, existing approaches mainly capture perceptual- and object-level structures, while the semantic and relational layers that embody contextual and symbolic meanings remain largely unexplored. Consequently, the absence of higher-level semantic representations hinders comprehensive understanding of TCPs.

Knowledge graphs (KGs) provide an interpretable framework for organizing entities, attributes, and relationships into an interconnected semantic network, enabling knowledge discovery and contextual inference [16], [17]. Adopting a knowledge-graph-based representation enables the structured encoding of symbolic and hierarchical relations in TCPs, supporting semantic understanding beyond conventional image-level analysis. However, constructing such representations demands domain expertise to accurately interpret the cultural semantics embedded in paintings. To this end, we construct a domain semantic taxonomy and an expert-annotated TCP knowledge graph that form the basis for expert-in-the-loop visual analysis and knowledge extraction.

B. Interactive Visualization for Human–AI collaboration

Interactive visualization has become an effective paradigm for integrating human expertise with machine intelligence. Through interactive exploration, it enables users to interpret and refine computational results, thereby bridging data-driven models with human reasoning. Such integration has demonstrated improvements in model reliability, interpretability, and data quality. Building on these advances, visualization research can be systematically categorized into *model-oriented* and *data-oriented* approaches. Model-oriented methods focus on visualizing and refining model behavior through interpretable representations. For instance, RetainVis [18] and ProtoSteer [19] allow users to incorporate domain knowledge into deep sequence models, while DRAVA [20] aligns semantic latent dimensions with human concepts via interactive manipulation. Data-oriented systems, on the other hand, focus on instance-level feedback and active learning. IRVINE [21] integrates interactive clustering and labeling for acoustic analysis, and RCLens [22] leverages user feedback to iteratively identify rare categories and improve labeling precision.

These studies collectively highlight that visualization not only supports model understanding but also enables interactive knowledge construction through uncertainty reasoning and active learning. Building upon these insights, our VisTCP framework employs uncertainty-aware visualization and expert feedback loops to facilitate human–AI collaboration in constructing and refining knowledge-graph representations of Traditional Chinese Paintings.

III. Requirement Analysis and Framework Overview

This section outlines the design requirements obtained through close cooperation with domain experts and provides an overview of the proposed visualization framework.

A. Requirement Analysis

This study involved two stages and included three art historians working on TCPs and one expert experienced in

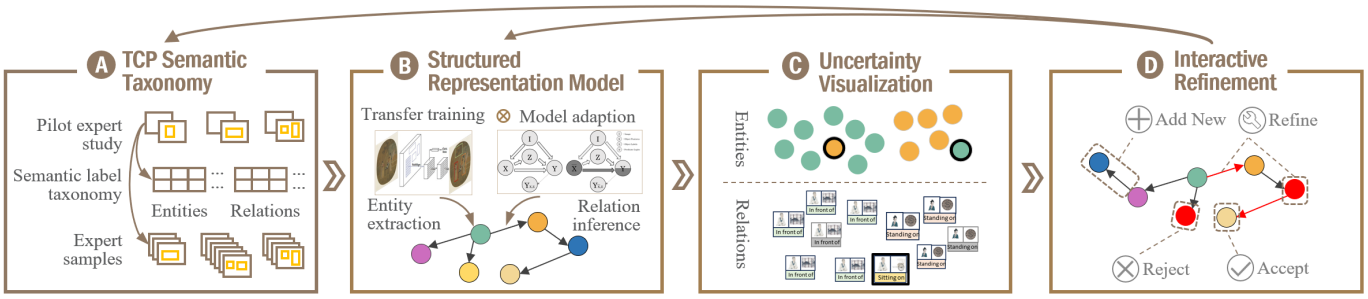

Fig. 1. The framework architecture of *VisTCP* contains four modules: a TCP semantic taxonomy module (A), a TCP-oriented model structured representation module (B), a visualization module (C), and an interactive refinement module (D).

the intersection of art imagery and computer vision, focusing on the construction and management of a digital TCP database. We collaborated with these four experts over two years. In the first stage, the art historians showed significant interest in developing a structured representation of TCPs for its powerful capabilities in supporting semantic retrieval and intelligent question-answering. We engaged with the experts' daily research activities for three months, encouraging them to describe any challenges they encountered in a think-aloud manner. Based on their feedback, we formed initial task requirements. Through multiple iterations of proposing solutions and receiving expert feedback, we ultimately defined the following requirements to empower art historians with enhanced knowledge interpretation capabilities, addressing their challenges in repetitive and uncertain knowledge extraction for the structured representation construction of TCPs.

R1. Build a structured representation model for TCP.

R1.1 Construct TCP semantic taxonomy. To enrich semantic comprehension in Chinese visual culture, numerous comprehensive classification systems, such as Iconclass and CIT [23], have been advanced for visual content description. While these established frameworks encompass a wide range of objects in Chinese visual culture, they overlook crucial semantic representations pertaining to ancient Chinese events. Consequently, experts advocate developing a specialized semantic taxonomy tailored for Traditional Chinese Paintings (TCPs) to foster semantic understanding of these cultural artifacts.

R1.2 Build structured representation model. Constructing structured representations is a repetitive and time-consuming process. Experts emphasized the need for a structured representation model aligned with the TCP semantic taxonomy to generate high-quality initial representations, accelerate knowledge extraction, and reduce human effort.

R2. Prompt uncertainty with prior expert knowledge for structured representation refinement.

R2.1 Visualize the uncertainty of knowledge extraction. Searching for prior references is a fundamental approach to attaining a precise comprehension of entities and semantic relationships in TCPs. While incorporating intelligent models can mitigate hurdles in knowledge extraction, both experts advocate for incorporating intuitive visual indicators alongside expert-driven prioritized knowledge to augment comprehension of extraction accuracy and signal uncertainties.

R2.2 Refine interactive structured representation. The consensus among experts underscores the necessity of integrating prior expert knowledge and user-friendly interactions to facilitate users in iteratively refining the initial structured representation of TCP. Additionally, experts highlight the need for intelligent models to efficiently adapt to the distinct visual characteristics of TCPs through user feedback. This adaptability is crucial for the accurate and efficient construction of the initial structured representations of TCPs.

B. Framework Overview

The aforementioned requirements motivated us to develop a visualization framework, *VisTCP*, that integrates intelligent models and prior expert knowledge into an intuitive visual interface, empowering art historians to construct a robustly structured representation for the semantic understanding of TCPs. *VisTCP* (Figure 1) includes four parts. First, a study with domain experts is conducted to create a TCP semantic taxonomy and an expert-annotated sample set (Figure 1(A)) for TCP-oriented structured representation model building (Figure 1(B)). Subsequently, we propose an uncertainty visualization that jointly maps semantic labels and image features from both model predictions and expert-annotated samples, to highlight uncertain entities and relationships (Figure 1(C)). Finally, the semantic label taxonomy undergoes iterative refinement through users' interactive inspection and updating of these entities or relationships (Figure 1(D)). Additionally, the model is iteratively fine-tuned based on user feedback to generate higher-quality structured representations. Through continual iterative optimization, the structured representation of TCP can be achieved more swiftly and accurately, with minimized manual intervention.

IV. Structure Representation Model

This section presents the construction of the structured representation model for TCPs, which involves (i) deriving a semantic taxonomy and constructing an expert-annotated dataset through an expert study (see Appendix A for the full description), (ii) instantiating an SGG-based model with Mask-RCNN for object detection and TDE for unbiased relation inference, and (iii) integrating an active-learning loop for iterative refinement.

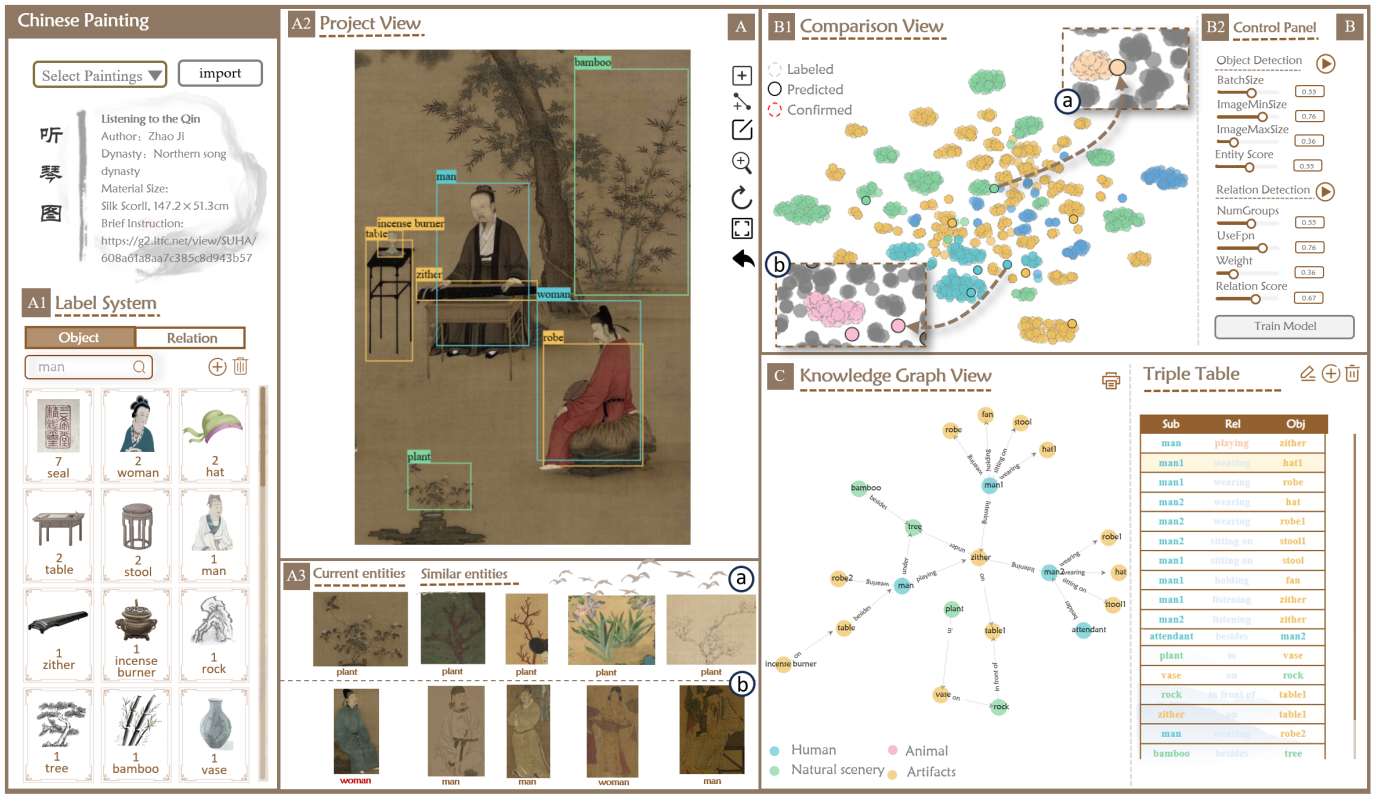

Fig. 2. The visualization interface of *VisTCP* consists of three components, including (A) an interactive refinement component provides the knowledge extraction results of a painting predicted by models and referenced samples for interactive refinement, (B) the uncertainty visualization component aligns the visual and semantic features of labeled and predicted elements to highlight ambiguities requiring expert verification, and (C) the structured representation component presents the obtained semantic graph from a TCP.

A. Semantic Taxonomy and Expert-Annotated Dataset Construction

We adopt a compact taxonomy with four entity classes (*human, natural scenery, animal, and artifact*) and two relation types (*event* and *location*).

Derivation method. The taxonomy was derived via an expert study with three TCP specialists using our annotation prototype: starting from a seed vocabulary compiled from museum thesauri and prior scholarship [23], experts annotated candidate entities and relations in a pilot study of 500 paintings (300 figure, 100 landscape, 100 flower-and-bird paintings), were allowed to add/modify labels, and then reconciled disagreements in a consensus meeting.

Annotated dataset construction. The same process produced a 500-image expert-annotated set that instantiates the taxonomy with object bounding boxes and relation triplets. Annotations were cross-checked across experts, conflicts were adjudicated, and each object/relation category was ensured to have at least 10 examples to support training and evaluation.

This set is used (i) to define detection/relationship categories for our Mask-RCNN [24] + TDE [6] pipeline and (ii) to drive active learning for iterative model refinement. Detailed study design, full label inventories, and per-type mappings are provided in Appendix A ($TablesI - II$).

B. Structured Representation Model Setup

The proposed structured representation model (R2) is built upon SGG [6], a widely used model for the structured representation of natural images, and consists of two parts: the advanced object detector Mask-RCNN [24] and unbiased relation inference method Total Direct Effect (TDE).

Object Detection. Mask-RCNN is a state-of-the-art deep learning algorithm for object detection, which is widely used for its ease of use and customization, accurate object detection, and faster processing. We adopt ResNet-101 as the backbone network architecture and the proposed TCP semantic taxonomy as the class labels, to build the object detection model for TCPs, to accommodate the characteristics of ancient China.

Relationship Inference. The Total Direct Effect (TDE) relation inference model, designed for unbiased scene graph generation, is utilized to mitigate the bias problem in SGG, achieving richer and more diverse relationship predictions. We also employ ResNet-101 as the foundational network architecture for the prediction of relationships in TCPs, while the semantic relationship labels of TCP taxonomy are utilized to enable the model to adapt to TCPs.

Loss Function. TTo supervise transfer learning in the network, we use the cross-entropy loss to train the structured representation model for TCPs, which measures the distance between the ground truth y and the predicted probability distribution p,

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(p_{i,c}\right) \quad (1)$$

N is the number of samples, C is the number of classes, $y_{i,c}$ represents the ground truth label of class c for sample i (1 if the sample belongs to class c, 0 otherwise), $p_{i,c}$ is the predicted probability of class c for sample i.

The expert-annotated sample set from the pilot expert study is used to train the model, and the experimental results confirm that the trained model is well adapted to TCPs, as demonstrated in Tables III–IV in Appendix B. Together, these two components form the structured representation model for TCPs.

C. Model Optimization

We refine our TCP-oriented knowledge extraction model through active learning (R4), based on the collections of user feedback for uncertainties and an increasing number of annotated samples. The loss function for optimizing the model is set the same as model transfer learning, to ensure iterative improvement. The important aspect of supporting model optimization is the selection of training samples. Here, we use annotated samples with joint embedding priority recommendations for model fine-tuning, which has been proven to be an effective active learning optimization strategy [25].

V. VisTCP Interface

In this section, we introduce the visualization and interaction designs in *VisTCP*. The interface of *VisTCP* is composed of three components, including Interactive Refinement Component (Fig. 2(A)), Uncertainty Visualization Component (Fig. 2(B)), and Structured Representation Component (Fig. 2(C)).

A. Uncertainty Visualization Component

The Uncertainty Visualization Component integrates a model panel to achieve knowledge extraction, and incorporates a Comparison View to enhance comprehension of the extracted knowledge.

1) Model Panel: **Support for automatic knowledge extraction.** As shown in Fig. 2(B2), the system supports two knowledge extraction models: (1) *Entity extraction*, detecting objects and outputting corresponding semantic labels and bounding boxes; and (2) *Relation inference*, predicting semantic relations between object pairs. The resulting structured representation is then integrated into the workflow.

2) Comparison View: This view presents a joint visualization of embedding vectors for model predictions and expert-annotated samples (Fig. 2(B1)), enabling art historians to intuitively assess results and pinpoint high-uncertainty elements for targeted validation.

Visual–Semantic Joint Embedding. We jointly embed predicted and expert-annotated entities/relations (R3) using visual and semantic features to improve interpretability and surface uncertain items for rapid refinement. Predicted elements are first paired with expert exemplars via text matching and feature similarity. Object and relation features are extracted by our TCP-oriented model and mapped into a shared space where Euclidean distances are re-weighted by semantic labels, pulling same-label items together and pushing conflicting

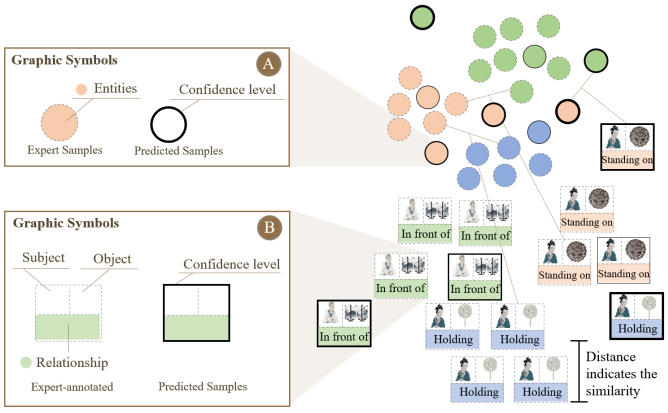

Fig. 3. The joint embedding visualizations of the predicted and expert-annotated samples in the TCPs for objects (A) and semantic relationships (B).

labels apart. The resulting label-aware embedding underpins the Comparison View.

Encodings. We concurrently expose three facets: (1) feature-driven distances, (2) per-label neighborhoods, and (3) the alignment/misalignment of predictions vs. expert annotations. To make these facets legible in one view, we project the embedding to 2D and encode aspects in distinct channels. *Entities* (Fig. 3(A)) are shown as a scatterplot: point color encodes the semantic label; solid vs. dashed borders distinguish prediction vs. expert exemplar; and border thickness encodes prediction confidence. *Relations* (Fig. 3(B)) use compact icons integrating subject–predicate–object; the same border scheme distinguishes sources.

Layout by feature similarity. Embedded positions are governed by feature proximity. We employ t-SNE (chosen for preserving local clusters) to maintain intra-class cohesion while separating inter-class neighborhoods, which facilitates side-by-side comparison of predictions and expert knowledge and highlights potential ambiguities.

Uncertainty cues. Elements that lie far from their own label cluster, fall near competing-label clusters, or have low prediction confidence are flagged as high-uncertainty candidates and routed to the refinement workflow.

B. Interactive Refinement Component

The Interactive Refinement Component provides a robust visual coordination environment and extensive interactive functionalities to streamline the iterative refinement of knowledge extraction, guided by prior expert knowledge. The Label System View, Painting View, and Expert View collectively present an overview of initially identified results, the original painting image annotated with semantic labels, and recommended expert-annotated samples, fostering a coherent and efficient refinement process.

1) Label System View: It provides an overview of the proposed TCP semantic taxonomy (Fig. 2(A1)), accompanied by corresponding label icons, to clarify the distinctive characteristics of objects and events pertaining to ancient China. Entity or relation selection and accuracy adjustment are also

provided in this overview for filtering, which allows users to focus on a certain class of interest for refinement.

2) Painting View: It presents the original painting image alongside the identified results with bounding boxes for objects, lines for relationships, and the corresponding predictive labels (Fig. 2(A)). To enhance user efficiency in refining entities and relationships, while simultaneously reducing visual clutter resulting from numerous entities bounding boxes and relation lines, each entity's bounding box will be transformed into a precise point marker once the user confirms its semantic label. Similarly, after entity extraction, further optimization or augmentation of relationships will be conducted in accordance with the visual cues provided in the Comparison View (Fig. 2(B)). In instances involving entity pairs with relationships, these connections are initially depicted as dashed lines, which are later changed to solid lines upon confirmation.

3) Expert View: It displays expert-annotated samples that have the most similar characteristics to the selected predictive sample, for refinement reference (Fig. 2(A3)). The labels of expert-annotated samples can be transferred to predictive samples through a double-click action. Herein lies an in-depth exposition of expert annotation samples, encompassing specific annotations of entities or events alongside corresponding illustrations to furnish additional contextual information.

Enabling interactive refinement. Collaborating with all other views, six operations are supported in here: 1) Modify entity. Users can change the label of an entity by clicking the corresponding bounding box. 2) Add entity. This action is initiated by a left-clicking arrow icon in the upper right corner, then draws a frame and selects a label to complete the entity add. The active input function is also provided for entity addition outside the label range to ensure the scalability of the label. 3) Delete entity. The user can remove the active bounding box corresponding to an entity via the *Delete* button. 4) Modify relation. Users can delete, modify, or confirm the relationship label by double-clicking on the relationship line between two entities. 5) Add relation. The user can add the relationship by clicking on the first and last entities in sequence. 6) Zooming and panning tools are also provided to facilitate knowledge extraction and enhance the exploration of paintings.

C. Structured Representation Component

The Structured Representation Component (Fig. 2(C)) offers a precision depiction of the current painting, delineating distinct entity types through node colors (blue: Human; pink: Animal; green: Natural Scenery; orange: Artifact) and representing distinct relationships through line colors (grey: location; yellow: event). When performing entity or relation editing operations, the interface provides real-time updates to promptly reflect changes in the graph's status. Furthermore, it offers two distinct layout methods: force-oriented layout and custom layout, facilitating exploration and analysis of TCPs. In addition, the Triple Table records various interactions to collect user feedback for model improvement.

VI. Evaluation

This section evaluates the effectiveness of the knowledge graph for TCP comprehensive semantic representation, followed by a usage scenario and expert interviews with 6 art historians to demonstrate the applicability of *VisTCP* in constructing structured representations.

A. Case Study

In this case, we primarily showcase the structured representations of three different scenes depicted in *Along the River During the Qingming Festival*: the passage of merchant caravan, pastoral activity, and theatrical spectacle, to illustrate the effectiveness of structured representations for TCP comprehensive semantic representation. Fig. 4 presents the structured representations, which show significant differences aligned with their visual scenes, as follows:

Passage of merchant caravans. Fig. 4(A) illustrates the structured representation of a bustling caravan, which presents a uniform distribution of animal nodes (pink), human nodes (blue), and artifact nodes (orange). A close observation reveals that all the animals are donkeys, all the humans are men, and the artifacts mainly consist of loads and two vehicles. This setup highlights the participatory features of the elements in the scene. The elements form a chain-like layout based on primary semantic relationships, predominantly unfolding in the same direction. The chain starts with a man towing a donkey, followed by the donkey pulling a vehicle, the vehicle following another vehicle, and a man riding a horse following the vehicle. These events collectively depict the visual scene of the merchant caravan's passage. Furthermore, the presence of numerous men shouldering loads around the carts enhances the visual scene of the caravan's passage, contributing to a comprehensive semantic understanding of the scene.

Pastoral activity. As illustrated in Fig. 4(B), the structured representation of a pastoral activity, primarily characterized by animal nodes (pink), natural scenery nodes (green), and human nodes (blue), thereby revealing a pastoral scene distanced from human settlements. A meticulous examination of this region reveals a plethora of structured representations, encompassing numerous boys and animals such as ox, sheep, and horses, engaged in various activities including playful interactions among the boys and animals eating grass. Conversely, the presence of artificial artifacts is notably sparse. This vividly portrays the semantic events in a pastoral setting, where boys are tending livestock in the countryside.

Theatergoing. The structured representation (Fig. 4(C)) of theatrical spectacle exhibits a unique aggregated structure, primarily centered around the semantic event of individuals watching a person perform on a theatre stage. Additionally, the "man95" node, representing the central figure, is connected to another local structure. Upon closer examination, it becomes apparent that this primarily describes individuals positioned on the same stage, involving events such as a man playing a drum, a table on the stage, and a flag on the stage. Together, they comprehensively depict the scene on the stage. Furthermore, this structured representation vividly delineates various interesting events in the theatrical scene, such as a

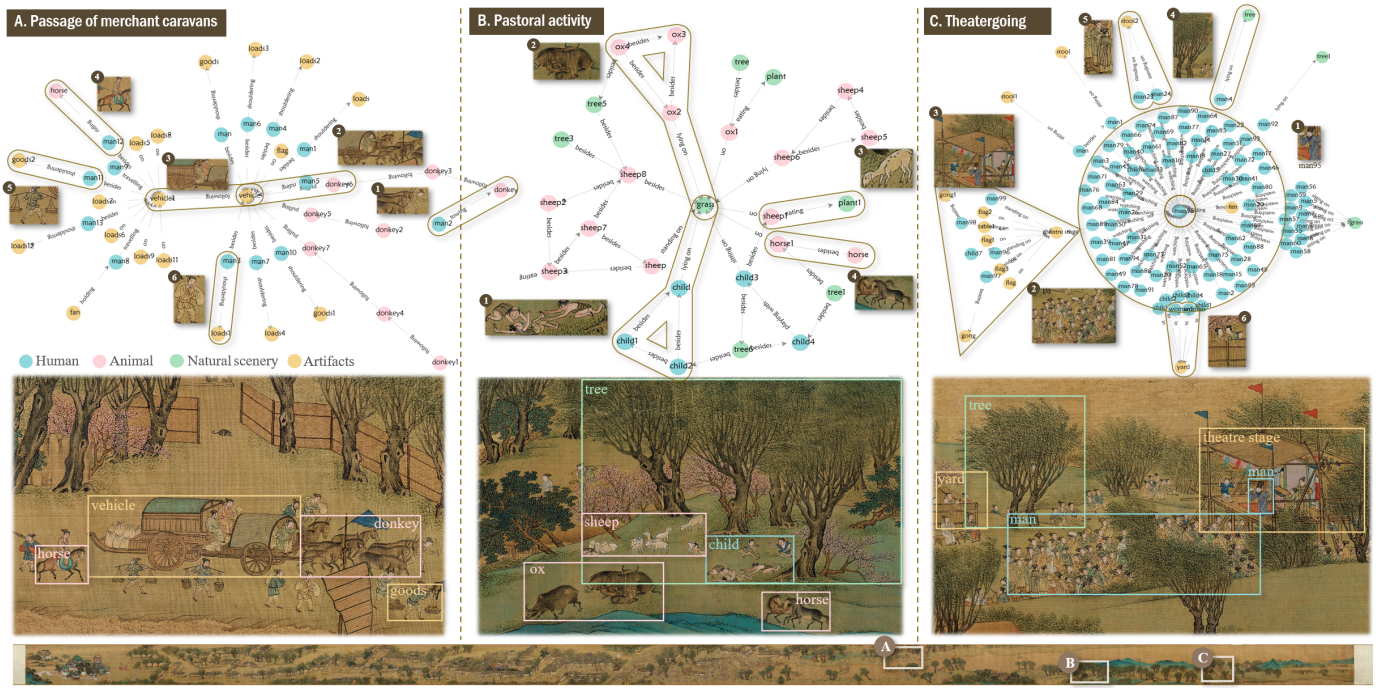

Fig. 4. Structured representation facilitates the understanding of the famous Chinese painting "Along the River During the Qingming Festival" across three distinct scenes: the passage of merchant caravans (A), pastoral activity (B), and theatergoing (C).

man watching the performance from a tree, a man standing on a stool to watch the performance, and a woman observing the performance in the courtyard. The triplet descriptions also suggest latent semantic events in the scene, facilitating a comprehensive semantic representation of the theatrical scene.

In summary, we observe that the structured representations of the three distinct visual scenes depicted in "Along the River During the Qingming Festival" exhibit markedly different characteristics, with the rich semantic events in each scene clearly delineated. Additionally, the visualization of these structured representations highlights less conspicuous semantic events within the visual scenes, thereby enhancing semantic understanding.

B. Usage Scenario

An art historian specializing in figure paintings and deeply interested in applying computer technology to analyze TCPs was invited to use *VisTCP* to construct the structured representation for the well-known Chinese painting *The Night Revels of Han Xizai.*

Understand the quality of automatic knowledge extraction. First, the art historian clicked the object detection button and examined the joint embedding visualization (Fig. 5(A)). He noticed that most predicted objects fell within the dense clusters of expert-annotated samples with the same class labels, indicating a high level of confidence in the detections. He then clicked on an orange predicted point and reviewed the expert samples in the expert view, confirming that its label was indeed a fan, consistent with the recommended label of the expert samples (Fig. 5(B-1)). The joint embedding visualization results garnered the art historian's trust. He then noticed some predicted points that were distant from the expert-annotated samples with the same class labels, suggesting a likelihood of incorrect identifications. The art historian concluded that the automatic object detection results required further refinement.

Explore and refine elements with uncertainty. The art historian reviewed the suspicious elements highlighted by the joint embedding visualization and found most were reasonable. For example, he noticed a blue point distant from its cluster (Fig. 5(A-3)) and, upon checking the expert view, saw that the object in the bounding box was incomplete, leading him to delete it. Exploring further, he observed two fans identified by the model in the Label System View. Clicking the icon, he found another fan far from the expert-annotated samples, suggesting misclassification. Examining the original painting, he found the fan identical to the previous one, with the deviation resulting from its plant-like illustrations (Fig. 5(B-2)). Finally, he confirmed that the identified women's labels aligned with a dense cluster of expert-annotated samples.

Identify missed elements by the model. After eliminating suspicious predictive elements, the art historian discovered that the structured representation model still missed certain elements, posing challenges to semantic understanding. To address this, the historian used entity and relationship addition in the painting view to capture the missed elements. With timely expert-annotated sample recommendations, the historian accurately completed the missing elements. For example, a cross flute was added, and the corresponding relationship to a woman playing the flute is also completed (Fig. 5(B)).

This usage scenario demonstrates that *VisTCP* can help art historians accurately and efficiently construct the structured representations of TCPs. It enables users to engage in the construction process, ensuring that the resulting structured expression achieves reliable TCP semantic understanding.

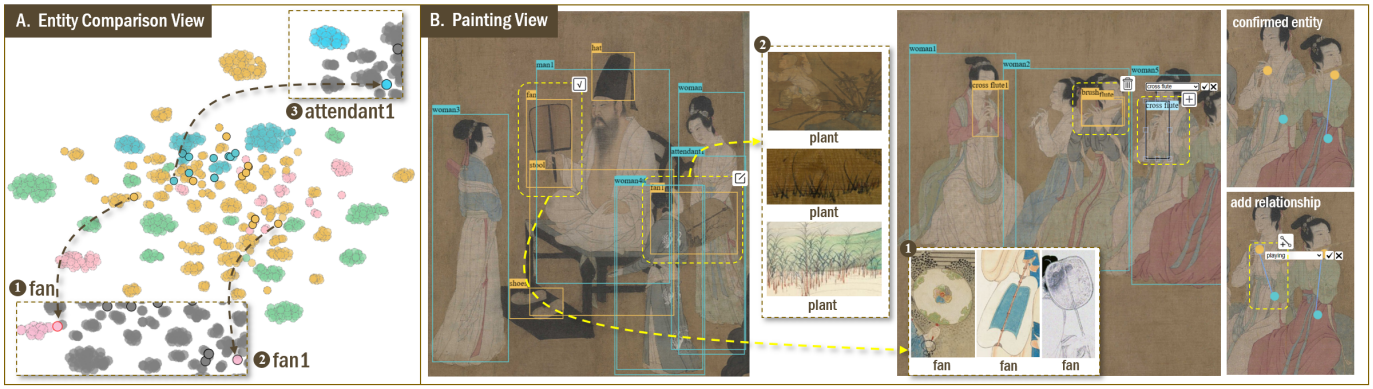

Fig. 5. Constructing the structured representation for "The Night Revels of Han Xizai". The visualization of automatically generated structured representations supports understanding of result quality (A). Then, the art historian examined and refined uncertain elements to produce a credible structured representation (B).

C. Expert Interviews

To evaluate the effectiveness of *VisTCP*, we interviewed six art historians, comprising two professors (AH1–AH2) with about 20 years of TCP research experience and four PhD students (AH3–AH6) focusing respectively on figure (AH3, AH5), landscape, and flower-and-bird paintings (AH4, AH6). None of the participants are co-authors of this work.

We first introduced our motivation to enable comprehensive semantic understanding through structured representation construction with *VisTCP*. Then, we showcased how to use *VisTCP* and asked them to construct structured representations of their interested paintings. During the process, we recorded their comments and interactions. Subsequently, the art historians were requested to describe the findings during the construction procedure verbally. Moreover, we asked them to rate *VisTCP* on a 7-point Likert scale based on the post-study questionnaire (Table I), focusing on the utility, visual design, interactions and usability of the system. The post-study interview lasted about 20 minutes each. Figure 6 shows the feedback from expert interviews. All participants provided informed consent prior to taking part in the interviews.

Utility. Most art historians ($rating_{\text{mean}}$ = 5.11, $rating_{\text{sd}}$ =1.28) offered positive feedback regarding the utility of the system in terms of the TCP-oriented knowledge extraction model($rating_{\text{mean}}$ = 4.83, $rating_{\text{sd}}$ =0.98), comparison view ($rating_{\text{mean}}$ = 5.00, $rating_{\text{sd}}$ =1.41) and expert sample recommendation ($rating_{\text{mean}}$ = 5.83, $rating_{\text{sd}}$ =0.75). First, the art historians agreed that "The proposed intelligent model can effectively support automatic object detection, which alleviates the repetitive and time-consuming knowledge extraction." Although other art historians also confirmed the value of the structured representation model, AH2 further said, "The intelligent model indeed can identify those more common and easy to understand relationships, it still produces some misunderstanding of semantic events belonging to ancient China." Second, most art historians are in favor of the joint embedding visualization. AH1, who used *VisTCP* to construct the structured representations of a complex figure painting, commented "Being familiar with current figure painting, I find that the automatic knowledge extraction results had some misidentification, but the uncertain visualization effectiveness hinted at these errors." Meanwhile, four art historians (AH3-AH6) also confirmed that the joint embedding of predicted elements and expert-annotated samples is useful, especially in "enhancing the credibility of automatic knowledge extraction." Although AH2 also expressed the advantage of the uncertainty recommendations on objects, he further emphasized the imperative for advancements in prompts related to uncertain relationships. Finally, two art historians praised the implementations of expert view in the interactive refinement component, which provides "credible expert-annotated references during the process of constructing structured representations."

Visual Design. The experts appreciated the visual design of *VisTCP* ($rating_{\text{mean}}$ = 5.58, $rating_{\text{sd}}$ =1.16), especially the design of Comparison View. Most experts praised the use of scatter plots to display discrepancies between model recognition results and expert-annotated samples, and commented "The glyph design for representing relationships effectively illustrates specific event information related to relationship triples, providing ample information for comparing and confirming relationships." They agreed that the design can present the distance of joint embedding vectors, and embedding vectors in respect to each identified element.

Usability and user interaction. The majority of the art historians provided positive feedback on user interactions ($rating_{\text{mean}}$ = 5.83, $rating_{\text{sd}}$ =0.83) and the usability ($rating_{\text{mean}}$ = 5.67, $rating_{\text{sd}}$ =1.03). Most experts agreed that the system has good usability and light workload, including ease of learning and no need for much physical or mental activity. Specifically, the comparison view retains the intuitive nature of scatter plots, making it easy to interpret without imposing additional cognitive load. AH1-6 confirmed that the system's interactions are smooth and easy to use. Meanwhile, the art historians also pointed out that "The linkage of multiple coordinated views provides more background knowledge for the interactive authentication of structured representations." Additionally, art historians are satisfied with the structured representation of TCPs constructed by *VisTCP*, as the system "supports a human-in-the-loop approach, allowing for the identification, confirmation, and modification of elements to ensure a reliable and semantically rich representation of the

TABLE I
THE QUESTIONNAIRE CONSISTS OF FOUR PARTS: THE UTILITY (Q1-3), VISUAL DESIGN (Q4-5), INTERACTIONS (Q6-7), AND THE USABILITY OF THE PROPOSED *VisTCP* (Q8-10).

Q1	TCP-oriented model is efficient for knowledge extraction.
Q2	Comparison View is effective for misunderstanding and credible results prompt.
Q3	Expert sample recommendation is useful for trustworthy results refinement.
Q4	The visual design is easy to understand.
Q5	Comparison View can enable intuitive representation of the automatic knowledge extraction results.
Q6	The interaction is smooth.
Q7	The interaction can support trustworthy structured representation confirmation and completion.
Q8	VisTCP is easy to learn and understand.
Q9	VisTCP is easy to use.
Q10	User is satisfied with constructed structured representation.

paintings."

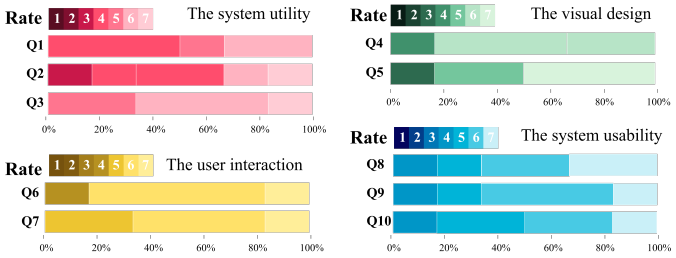

Fig. 6. The feedback of the expert interviews.

VII. DISCUSSION

A. Potential Usage Scenario

The construction of structured representation shows obvious benefits in TCP semantic representation. In addition, similar to the structured representation of natural images, scene graphs can support many high-level downstream tasks. On this basis, structured representations of TCPs show great potential in benefiting the following applications.

Semantic retrieval for TCPs. Scholars engaged in archaeological and art historical research often require extensive retrieval efforts to access the desired ancient paintings. As illustrated in the Case Study, the structured representation of TCPs can encapsulate a variety of objects and semantic events. Consequently, they enable art historians to explore TCPs through semantic retrieval. For instance, art historians can use structured representations to find paintings depicting specific semantic events, like a man playing the zither, facilitating in-depth content analysis. This approach reduces the repeated retrieval and screening involved in traditional methods.

TCP structured understanding by VLM. Existing literature has demonstrated that scene graph knowledge, which facilitates the structured representation of natural image content, can effectively enhance the multimodal structured understanding of large-scale visual-language models. Building upon this premise, integrating the structured representation of TCPs as knowledge-enhanced encodings into visual-language models can further augment the models' understanding of detailed structural knowledge inherent in ancient artworks, encompassing semantic relationships and attribute features.

B. Limitations and Future Work

Advancing TCP Semantic Taxonomy. Based on the expert study, we've developed a comprehensive TCP semantic taxonomy, which is specifically tailored to effectively describe the semantic content drawn in TCPs. Collaborating with art historians further reveals a demand for a more detailed and comprehensive label system. For instance, experts note the evolving nature of objects like hats which exhibit distinct stylistic characteristics across various historical periods and geographical regions. In the future, we will delve deeper into the research of TCPs making the semantic taxonomy accessible to more complex Chinese culture.

Incorporating diverse expert knowledge. In this study, we primarily furnish expert-annotated paintings as prior knowledge to facilitate the construction of credible structured representations. However, for specific understanding requirements of ancient artworks, art historians necessitate seeking additional research findings. Since the understanding of TCPs spans millennia, a plethora of documents and data records exist to aid art history analyses from various perspectives. These data, to a certain extent, assist art historians in multi-dimensional information understanding. In the future, we aim to further integrate multimodal data records to support more profound explorations of the semantic content in TCPs.

Improving model performance. *VisTCP* has developed an intelligent knowledge extraction model tailored for TCPs. However, the specialized nature of TCPs and the sparsity of annotated samples pose challenges to the current system's utility, especially uncertainty prompts and expert-annotated sample recommendations for relationship inference. Therefore, the system offers interactions and model iteration optimization based on expert feedback. We anticipate that through continual iterative improvements, the system's performance in constructing structured representations of TCPs will be continuously optimized.

Generalizability. In this work, we primarily collaborated with art historians to develop a human-machine collaborative system to facilitate the structured representation construction of TCP. Although this system is specifically designed for TCPs, the proposed framework can be extended to other domain-specific image types, particularly other art and natural images. We learned that the accuracy of knowledge extraction by the automatic intelligent method is far from usable. The scene graph of natural images constructed by our interactive system can fully guarantee the quality of the constructed graph, enabling its downstream roles. Furthermore, by appropriately defining domain-adapted semantic taxonomies and intelligent models, the system's comparison view and expert sample recommendation strategy can easily handle various professional image data, such as medical images, autonomous driving, and industrial inspection. For example, the comparison

view based on medical images can help doctors quickly identify abnormal results, and expert sample recommendations can aid doctors in achieving reliable disease diagnoses based on historical experience.

VIII. Conclusion

In this paper, we present *VisTCP*, a visualization framework that facilitates art historians to achieve structured representations of TCPs in an efficient and trustworthy way. The framework integrates the proposed TCP-oriented knowledge extraction model and prior expert knowledge, to support interactions between art historians and machines. We develop an interface with a joint embedding visualization to support art historians' understanding of structured representation quality and to highlight uncertainties for interactive refinement. The effectiveness and practical utility of *VisTCP* are demonstrated through a case study, a usage scenario, and expert interviews. The results reveal that the structured representation can effectively enable comprehensive TCP understanding and the framework interface offers art historians an efficient way to interact with intelligent models for improving structured representation.

IX. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. 62277013 and 62177040), the National Social Science Fund Major Project (No. 24&ZD075), the National Statistical Science Research Project (No. 2023LZ035), the Key R&D 'Jianbin' Tackling Plan Programme in Zhejiang Province, China (No. 2023C01119), the Public Welfare Plan Research Project of Zhejiang Provincial Science and Technology Department (Nos. LTGG23H260003, LGF22F020034 and LZ22F020015), the Open Project Programme of the State Key Laboratory of CAD&CG (No. A2301), and the Zhejiang Provincial Natural Science Foundation of China (No. LTGG24F020006).

References

[1] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[2] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2024.

[3] Y. Li, W. Ouyang, X. Wang, and X. Tang, "Vip-cnn: Visual phrase guided convolutional neural network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7244–7253.

[4] D. Teney, L. Liu, and A. Van Den Hengel, "Graph-structured representations for visual question answering," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3233–3241.

[5] M. Tajrobehkar, K. Tang, H. Zhang, and J.-H. Lim, "Align r-cnn: A pairwise head network for visual relationship detection," *IEEE transactions on multimedia*, vol. 24, pp. 1266–1276, 2021.

[6] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3716–3725.

[7] S. McCausland and Y. Hwang, *On Telling Images of China: Essays in Narrative Painting and Visual Culture*. Hong Kong University Press, 2013.

[8] D. Zhang, B. Pham, and Y. Li, "Modelling traditional chinese paintings for content-based image classification and retrieval," in *10th International Multimedia Modelling Conference, 2004. Proceedings.* IEEE, 2004, pp. 258–264.

[9] C.-C. Hung, "A study on a content-based image retrieval technique for chinese paintings," *The Electronic Library*, vol. 36, no. 1, pp. 172–188, Feb. 2018.

[10] Q. Meng, H. Zhang, M. Zhou, S. Zhao, and P. Zhou, "The classification of traditional chinese painting based on cnn," in *Cloud Computing and Security*, X. Sun, Z. Pan, and E. Bertino, Eds. Cham: Springer International Publishing, 2018, pp. 232–241.

[11] Q. Meng, K. Li, M. Zhou, and H. Zhang, "The elements extraction on traditional chinese paintings based on object detection," in *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 111–116.

[12] W. Jiang, Z. Wang, J. S. Jin, Y. Han, and M. Sun, "Dct–cnn-based classification method for the gongbi and xieyi techniques of chinese ink-wash paintings," *Neurocomputing*, vol. 330, pp. 280–286, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218313201

[13] W. Zhang, J. K. Wong, Y. Chen, A. Jia, L. Wang, J. Zhang, L. Cheng, and W. Chen, "Scrolltimes: Tracing the provenance of paintings as a window into history," *CoRR*, vol. abs/2306.08834, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.08834

[14] Y.-C. Lai, B.-A. Chen, K.-W. Chen, W.-L. Si, C.-Y. Yao, and E. Zhang, "Data-driven npr illustrations of natural flows in chinese painting," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 12, pp. 2535–2549, 2017.

[15] W. Ma, Y. Wang, Y.-Q. Xu, Q. Li, X. Ma, and W. Gao, "Annotating traditional chinese paintings for immersive virtual exhibition," *J. Comput. Cult. Herit.*, vol. 5, no. 2, aug 2012. [Online]. Available: https://doi.org/10.1145/2307723.2307725

[16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vision*, vol. 123, no. 1, pp. 32–73, may 2017. [Online]. Available: https://doi.org/10.1007/s11263-016-0981-7

[17] W. Zhang, J. Zhang, K. Wong, Y. Wang, Y. Feng, L. Wang, and W. Chen, "Computational approaches for traditional chinese painting: From the "six principles of painting" perspective," *CoRR*, vol. abs/2307.14227, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.14227

[18] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo, "Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 299–309, 2019.

[19] Y. Ming, P. Xu, F. Cheng, H. Qu, and L. Ren, "Protosteer: Steering deep sequence model with prototypes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 238–248, 2020.

[20] Q. Wang, S. L'Yi, and N. Gehlenborg, "Drava: Aligning human concepts with machine learning latent dimensions for the visual exploration of small multiples," ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3581127

[21] J. Eirich, J. Bonart, D. Jäckle, M. Sedlmair, U. Schmid, K. Fischbach, T. Schreck, and J. Bernard, "Irvine: A design study on analyzing correlation patterns of electrical engines," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 11–21, 2022.

[22] H. Lin, S. Gao, D. Gotz, F. Du, J. He, and N. Cao, "Rclens: Interactive rare category exploration and identification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 7, pp. 2223–2237, 2018.

[23] POSTHUMANITIES, "Chinese iconography thesaurus," https://chineseiconography.org/.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[25] C. Chen, Y. Wang, L. Liao, Y. Chen, and X. Du, "Real: A representative error-driven approach for active learning," in *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2023, pp. 20–37. [Online]. Available: https://doi.org/10.1007/978-3-031-43412-9_2