Art Beyond Semantics: Sheaf-Informed Contrastive Learning for Multi-Relational Representations

Ludovica Schaerf[1,2,*], Antonio Purificato[3,4,*,†], Piera Riccio[5], Fabrizio Silvestri[3], and Noa Garcia[6]

[1] University of Zurich, Switzerland
[2] Max Planck Institute Bibliotheca Hertziana, Italy
[3] Sapienza University of Rome, Italy
[4] Amazon, Luxembourg
[5] University of Amsterdam, Netherlands
[6] The University of Osaka, Japan
ludovica.schaerf@uzh.ch
∗ Equal contribution † Work done outside of Amazon

🤗SemArt+ 🤗WikiArt+ Code

Abstract. Understanding a painting is never a single act. Art historians may analyze the same work through concepts of style, iconography, or historical context, dimensions that are not interchangeable, and each carries distinct semantic relationships between the visual and the textual. Vision-Language Models (VLMs) like CLIP, which learn a single shared embedding space, collapse this richness into a single homogeneous alignment, thereby losing the multi-relational structure that defines art-historical reasoning. We introduce **CANVAS** (**C**ontrastive **A**rt-aware **N**etwork for **V**ision-Language **A**lignment with **S**heaves), a framework for learning relation-aware multimodal representations inspired by sheaf theory. Each artwork is projected into multiple embeddings conditioned on the type of relation (*i.e.*, the context), and a novel contrastive loss encodes contextual information during training, with no dependency on external data at inference. We evaluate on three newly introduced benchmarks of artworks for multi-relational art understanding: WikiArt+, derived from WikiArt and Wikipedia, HertzianaDP, from the Bibliotheca Hertziana collection, and SemArt+, refined from the SemArt dataset. In multimodal retrieval and art understanding, **CANVAS** outperforms the baselines, supporting the view that multi-relational alignment is not just theoretically motivated but also practically essential.

Keywords: Image-Text Retrieval · Contextual Learning · Art Analysis · Sheaf Neural Networks

1 Introduction

Before computer vision had massive collections of images, it had art. Early classification algorithms were tested on digitized paintings, aiming to distinguish a

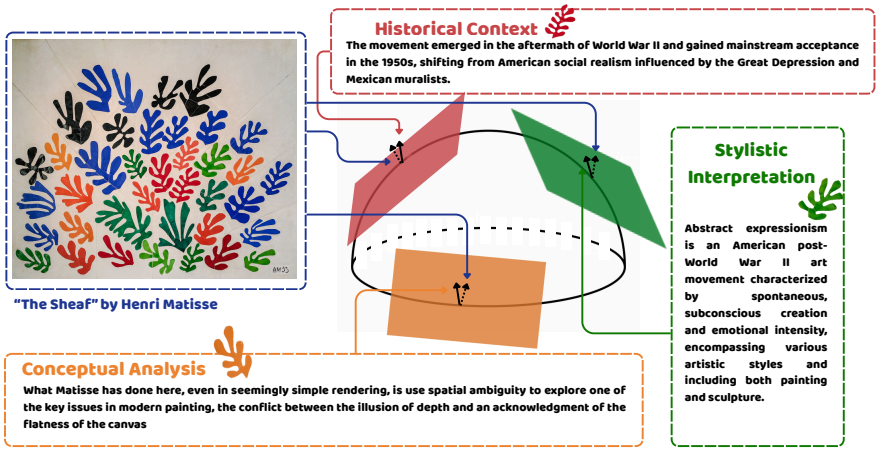

Fig. 1: Matisse's painting "La Gerbe" ("The Sheaf") from WikiArt+ dataset alongside three modes of art-historical understanding: Historical Context (*e.g.*, post-war artistic shifts), Stylistic Interpretation (*e.g.*, traits of the broader movement), and Conceptual Analysis (*e.g.*, the artist's use of spatial ambiguity). We show how **CANVAS** encodes the images and corresponding texts into subspaces corresponding to the different modes of interpretation. Artwork in the public domain.

master's brushstroke from a copyist's [20]. With the rise of large-scale online datasets, the field has since shifted focus to photorealistic images [6], which are easier to obtain and less open to interpretation. However, art remains a distinct challenge by itself [4]: while standard image analysis might focus on accurate object detection [38], the art domain is concerned with decoding the nuanced language of style, intent, meaning, and context that defines each artwork [4].

Current vision-language models (VLMs) [25] trained via contrastive learning fundamentally assume a single fixed relationship between images and their associated texts [14]. This limits the capabilities of these models in domains in which multiple valid interpretations per image can coexist. Art history is a canonical example: a single artwork can simultaneously engage with different types of descriptions. As shown in Fig. 1, Matisse's painting "La Gerbe" ("The Sheaf") can be, for instance, associated with texts referring to historical context (*e.g.*, situating the work within post-war European modernism and the broader shift of the avant-garde toward the United States), stylistic interpretation (*e.g.*, referring to recurrent elements characterizing the movement this painting belongs to), or conceptual analysis (*e.g.*, interpreting the meaning of the painting), each constituting a distinct yet interrelated mode of understanding the artwork itself [4].

In an attempt to capture the different interpretations of art in context, graph-based approaches [7, 29] propose encoding images across multiple aspects and modeling relational structures for classification and regression. Fundamentally, these approaches require access to the full graph structure at test time, including edge indices derived from test-set annotations [15]. This means that they operate

in a *transductive* setting, where predictions are only possible for images already integrated into the graph, and unseen images cannot be evaluated unless they are first annotated and added. This setting constrains their applicability in *inductive* scenarios, where previously unseen entities must be processed without known relational connections, a common requirement in real-world settings [32].

To overcome the limited capacity for multi-semanticity in VLMs and the constraints of graph-based approaches, we propose **C**ontrastive **A**rt-aware **N**etwork for **V**ision-Language **A**lignment with **S**heaves (**CANVAS**), a method that combines VLM representations with Sheaf graph formulations [3]. **CANVAS** encodes images and text using a pre-trained CLIP [25] model and projects them into relation-aware subspaces using graph-based distances. This allows the network to learn multiple representations within a single space without requiring the graph connectivity at test time. **CANVAS** works as a lightweight fine-tuning. Rather than representing each image and text as a single representation, it models them as a set, each capturing a distinct dimension of the link between an artwork and an associated text.

We validate **CANVAS** for cross-modal retrieval and classification accuracy on three art historical datasets: (1) the SemArt+ dataset [13] consisting of European paintings from the 3rd to 19th centuries, augmented with per-sentence annotations from Explain Me the Painting [1]; (2) the WikiArt+ dataset, an extension of WikiArt [31] consisting of worldwide art, enriched with Wikipedia-sourced explanations; and (3) a newly introduced dataset, HertzianaDP [19], consisting of a collection of paintings (P) and drawings (D) belonging to the Bibliotheca Hertziana, not publicly available prior to this work.[a] As, to the best of our knowledge, several large VLMs, such as OpenCLIP, have been exposed to the first two datasets during training [26], HertzianaDP serves as an ideal testbed for evaluating generalization to unseen artworks. Across all datasets, **CANVAS** consistently outperforms baseline VLMs and existing adaptation methods. In image-to-text retrieval, **CANVAS** outperforms the strongest baseline by a large margin across all datasets, with the biggest gains on HertzianaDP, which was not online at the time of CLIP pretraining. **CANVAS** also delivers strong results in text-to-image retrieval and relation-aware classification, with no other baseline matching its consistency.

2 Related work

This work combines multimodal representation learning with graph-based modeling in the context of art understanding. In the following, we first introduce how computer vision has been applied to art, and then we review relevant work on vision–language contrastive learning and graph-augmented representations.

[a] https://edmond.mpg.de/dataset.xhtml?persistentId=doi:10.17617/3.1GN3OL
https://edmond.mpg.de/dataset.xhtml?persistentId=doi:10.17617/3.Z8W2JR

2.1 Computer vision for art understanding

Early large-scale studies [5,18,34] demonstrate that deep convolutional networks trained on natural images can be effectively transferred to artistic domains, with a focus on a single domain-specific task. For example, Karayev *et al.* [18] explore style prediction across photographs and paintings, while Crowley *et al.* [5] investigate object retrieval and representation learning in artworks. Similarly, van Noord *et al.* [34] study artist attribution with deep features. Subsequent work [28, 30] instead builds on larger, more comprehensive benchmarks. Saleh *et al.* [28] analyze together style, genre, and artist classification on thousands of paintings, highlighting specific challenges of artistic imagery. Later, Strezoski *et al.* [30], introduce the OmniArt dataset, unifying multiple art-related prediction tasks and encouraging shared visual representations across them.

2.2 Vision-language contrastive learning

Vision-language contrastive models, such as CLIP [25], align images and texts in a joint embedding space. However, they incur several problems in the context of our work: (1) CLIP's contrastive loss treats all non-paired samples as negatives, pushing apart semantically related but non-matched image-text pairs within each batch, and (2) it enforces single global alignments instead of capturing multiple valid relationships between image regions or across images. Several works [22, 24, 35, 37] address the latter by refining the contrastive objective itself. Xie *et al.* [35] introduce a Main Semantics Consistency loss to prioritize learning the most semantically relevant aspect of an image or text, while Zhu *et al.* [37] and Pan *et al.* [22] move beyond single-semantic, exploring relation-level visual-semantic alignment within images through relational contrastive learning. Qiao *et al.* [24] extends the former by introducing a relation-guided cross-attention mechanism that modulates multimodal representations between images under each relation context, using natural-language relation descriptions, while Ruthardt*et al.* [27] use cross-attention to steer the CLIP representation toward the desired textual cue. A different extension of CLIP to multiple representations of the same image through text can be found in document retrieval models, such as ColPali and ColQwen [9]. These vision-language models are trained to produce multi-vector embeddings from document page images, directly optimizing the downstream retrieval objective via a late-interaction mechanism. These works allow for a coherent representation of textual additions within images. Despite their effectiveness, these contrastive approaches only represent within-image relational content and purely semantic relations. In the context of art, we wish to model multiple relations, including non-semantic ones.

2.3 Graph-augmented representations

Graph-augmented representation learning injects contextual knowledge, such as stylistic or historical connections, into learned embeddings, enriching them beyond simple image–text pairs. Some works have leveraged these possibilities in

the context of art. A common strategy employs graph neural networks (GNNs) to propagate information across entity neighborhoods and enrich per-node representations. ArtSAGENet [7] and ContextNet [11, 12] combine visual features with GNN representations that operate over artist-level relationships, thereby inserting contextual information into the embeddings. Furthermore, El Vaigh *et al.* [8] adopt a transductive GNN approach on a knowledge graph of artworks to jointly predict multiple metadata labels for unlabelled samples. Lastly, Scaringi *et al.* [29] replaces CLIP's text encoder with a GNN over a knowledge graph and learns a shared image-graph embedding space. While these graph-augmented methods leverage contextual information, they do not address multi-relational multimodal alignment. Some of the works [7, 8, 29], furthermore, require the full graph to be available at inference time, preventing their application to unseen entities.

3 Contrastive Art-aware Network for Vision-Language Alignment with Sheaves (CANVAS)

Our approach is based on an intuition: since an artwork and a text can be linked by different relationships that depend on both the visual and textual content, the way they are compared should adapt accordingly. For example, consider a painting paired with two different texts, one describing its *visual content* ("a portrait of a woman in Renaissance dress") and another describing its *historical context* ("commissioned by the Medici family in 1490"). We wish to learn distinct transformations for each relation: the content relationship should align the image embedding with visual-semantic features in the text space, while the context relationship should project the same image into a historical-factual subspace. Rather than centering the embeddings of each item, our model treats relationships as "first-class citizens": what matters most is not the item representations per se, but the maps that transport them into relation-specific subspaces.

This intuition can be naturally formalized as a graph through cellular sheaf theory [2]. In a cellular sheaf, nodes and edges of the graph are equipped with vector spaces (stalks), and restriction maps transform features between nodes' coordinate systems that pass through the edge, enabling the same data to be consistently represented in different local frames. Unlike standard GNNs, sheaf-inspired restriction maps provide a mechanism for learning how to transport both image and text representations into different edge-based subspaces, effectively deforming the embedding space to align with each relation.

3.1 Preliminaries

We first review cellular sheaf theory's key concepts. In a cellular sheaf, data is assigned to both the nodes and edges of a graph, enabling the joint modeling of local and global structure. Given an undirected graph $G = (V, E)$ [33], a cellular sheaf $\mathcal{F}$ associates to each node $v \in V$ a vector space $\mathcal{F}(v)$, and to each edge

$e \in E$ a vector space $\mathcal{F}(e)$. These spaces represent the local data supported on nodes and edges, respectively.

For every incident node-edge pair $v \trianglelefteq e$, the sheaf defines a linear restriction map $\mathcal{F}_{v \trianglelefteq e} : \mathcal{F}(v) \to \mathcal{F}(e)$, which encodes how information attached to a node is related to the information attached to the edge. The vector spaces $\mathcal{F}(v)$ and $\mathcal{F}(e)$ are referred to as *stalks*. Restriction maps are fundamental, as they determine how local data is consistently transported across the graph and thus provide the mechanism that connects local representations to global structure.

3.2 Method

We adopt *restriction maps* as learnable edge-conditioned transformations that project node representations into relation-specific edge subspaces, allowing a single image or text embedding to be compared differently depending on the relationship it participates in. Additionally, we introduce a context-aware soft contrastive loss derived from the graph structure. This objective enforces instance-level alignment between image–text pairs within a given relation type, while softening the negative term based on graph proximity: nodes that are nearby in the graph receive a weaker penalty than fully unconnected ones, reflecting the intuition that items sharing more relationships are semantically closer. We present the input graph formation and Sheaf-inspired relation-conditioned layer in Fig. 2, while the loss is schematized in Fig. 3.

3.3 Multimodal graph construction

As shown in Fig. 2, rather than treating each image–text pair in isolation, we model the dataset as a bipartite graph where edges encode typed semantic relations. This graph determines which pairs interact during training and provides the relational context that conditions the learned embeddings.

We model multimodal data as a bipartite graph $G = (U \cup V, E)$, where U denotes image nodes, V denotes text nodes, and $E \subseteq U \times V$ encodes semantic relations between images and texts. Each edge $e = (u, v) \in E$ is associated with a relation label r_e (*e.g.*, *content*, *context*), represented as the name of the relation.

CLIP initialization. We initialize node and edge features using a pretrained CLIP model [25]. Given an image $u \in U$, a text $v \in V$, and a relation description r_e for edge $e \in E$, we compute:

$$\mathbf{h}_u = f_{\text{img}}(u), \quad \mathbf{h}_v = f_{\text{text}}(v), \quad \mathbf{h}_{r_e} = f_{\text{text}}(r_e), \tag{1}$$

where f_{img} and f_{text} denote the CLIP image and text encoders.

After initialization, we fine-tune the projection layers and the last transformer blocks, while keeping earlier layers frozen. After the projection layer within the CLIP model, the $\mathbf{h}_u$ and $\mathbf{h}_v$ are embedded into a shared latent space of size d.

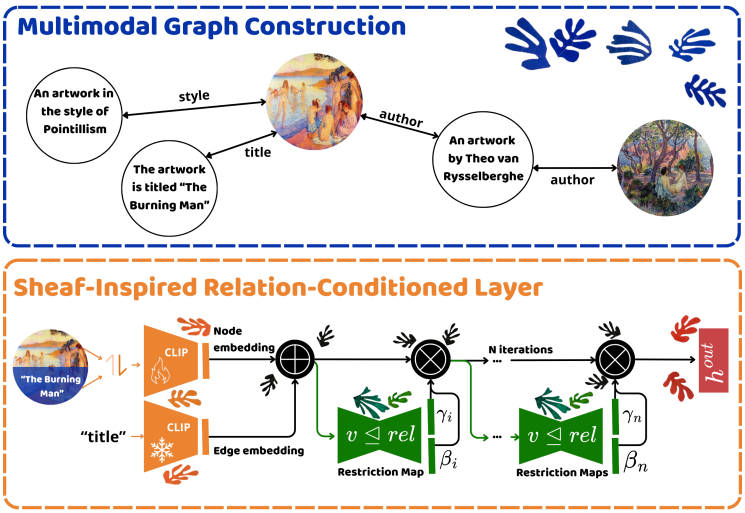

Fig. 2: Overview of CANVAS. Top: The input data is structured as a bipartite graph in which image nodes (artworks) and text nodes (descriptions) are connected via typed edges representing semantic relationships (*e.g.*, "style", "author"). Bottom: An image or text in a node is encoded using CLIP, alongside its relationship, the edge type. The node embedding is concatenated with the relation embedding and passed through learned restriction maps. These maps output relation-specific scale (γ_i) and shift (β_i) parameters that transform the node representations into relation-aware subspaces via affine modulation. This process is repeated N times. The $\otimes$ and $\oplus$ symbols denote the element-wise multiplication and concatenation operations, respectively.

3.4 Sheaf-inspired relation-conditioned layer

Nodes connected by different relation types should be compared in different representational subspaces. Inspired by cellular sheaf theory, we learn a relation-conditioned transformation that modulates each node embedding depending on the relation, producing distinct relation-aware views without separate encoders. We illustrate this mechanism in Fig. 3.

Given an edge $e = (u, v)$ with relation embedding $\mathbf{h}_{r_e}$, we want to produce a transformation that depends on both the node content and the relation type. We implement this process as a feature-wise linear modulation (FiLM) [23]: we first concatenate each node embedding (either the image embedding $\mathbf{h}_u$ or the text embedding $\mathbf{h}_v$) with the relation embedding $\mathbf{h}_{r_e}$.

$$\mathbf{z}_{u,e} = [\mathbf{h}_u \oplus \mathbf{h}_{r_e}], \quad \mathbf{z}_{v,e} = [\mathbf{h}_v \oplus \mathbf{h}_{r_e}], \tag{2}$$

where $\oplus$ denotes concatenation. We define a sheaf learner $\phi : \mathbb{R}^{d+d_r} \to \mathbb{R}^{2d}$ implemented as a three-layer MLP with LeakyReLU activations [3]. The output is split into feature-wise scale and shift parameters:

$$(\boldsymbol{\gamma}_e, \boldsymbol{\beta}_e) = \phi(\mathbf{z}_{\cdot,e}), \quad \boldsymbol{\gamma}_e, \boldsymbol{\beta}_e \in \mathbb{R}^d. \tag{3}$$

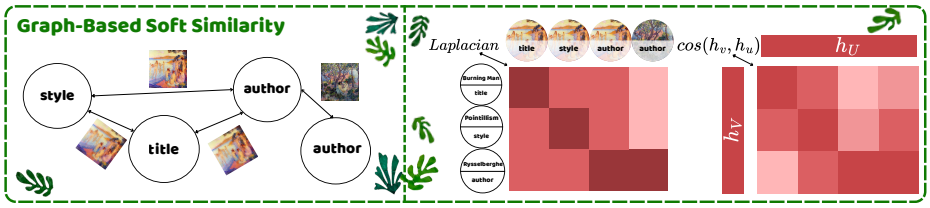

Fig. 3: Graph-based soft similarity computation via line graph construction and heat kernel diffusion. Left panel: Illustration of the line graph transformation $G' = L(G)$, where the relationships associated with the edges become the nodes and vice versa. Right panel: The heatmap on the left visualizes the soft target matrix computed via heat kernel diffusion on the line graph's Laplacian. Pairs involving the same image or text (*e.g.*, "Burning Man-title", "Burning Man-style") receive higher scores, reflecting their semantic proximity in the graph structure. On the right, the pairwise cosine similarities between image-relation and text-relation embeddings outputted by the sheaf layer in Fig. 2. The contrastive loss is computed between the soft target matrix and the cosine similarity matrix. Darker red indicates higher similarity.

The restriction map is applied as a FiLM modulation:

$$\tilde{\mathbf{h}}_{e,u} = \boldsymbol{\gamma}_e \otimes \mathbf{h_u} + \boldsymbol{\beta}_e, \quad \tilde{\mathbf{h}}_{e,v} = \boldsymbol{\gamma}_e \otimes \mathbf{h_v} + \boldsymbol{\beta}_e, \tag{4}$$

where $\otimes$ denotes element-wise multiplication. Intuitively, $\boldsymbol{\gamma}_e$ amplifies or suppresses dimensions of the embedding that are relevant to the relation, while $\boldsymbol{\beta}_e$ shifts the representation to align with the relation-specific subspace. The same node embedding is transformed differently for each relation, yielding a learned restriction map that locally deforms the embedding space for each relation type.

We stack N layers and average the outputs, as common practice in GNNs [16]:

$$\tilde{\mathbf{h}}^{\text{out}}_{e,u} = \frac{1}{N}\sum_{n=1}^{N} \tilde{\mathbf{h}}^{(n)}_{e,u}, \quad \tilde{\mathbf{h}}^{\text{out}}_{e,v} = \frac{1}{N}\sum_{n=1}^{N} \tilde{\mathbf{h}}^{(n)}_{e,v}. \tag{5}$$

Unlike standard GNN message passing, where each node aggregates information from all its neighbors, our formulation processes each image–text edge independently: the restriction map transforms a node embedding conditioned solely on the relation associated with that edge. This ensures that embeddings can be computed for individual image–text pairs at inference time, without access to the full graph structure.

3.5 Graph-based soft similarity

The InfoNCE contrastive objective aligns image–text pairs independently, without considering how different relationship types relate to one another. To capture higher-order dependencies between relations, we turn to the structure of the graph itself. We construct the line graph $G' = \mathcal{L}(G) = (V', E')$ [17], in which each node corresponds to an edge in the original graph $V' = E$, and two nodes

are connected whenever their corresponding edges $E' = \{e_i, e_j\} \subseteq E$ share a common endpoint $e_i \cap e_j \neq \emptyset$. Intuitively, the line graph makes the relationships between relationships explicit: two image–text pairs become neighbors in G' if they involve the same image or the same text. We illustrate this construction in Fig. 3. We then use the line graph to derive a soft structural prior that captures how closely related any two image–text pairs are. Let $L_{G'} = D' - A'$ be the combinatorial Laplacian of G', with degree matrix D' and adjacency matrix A' [33]. We compute contextual affinities via the heat kernel:

$$W = \exp(-\tau L_{G'}), \tag{6}$$

where $\tau > 0$ controls the extent of diffusion: small values preserve only immediate neighbors, while larger values propagate affinity across more distant pairs. To obtain a sharp, normalized distribution over neighbors, we raise each entry to a power α and row-normalize:

$$\tilde{W}ij = \frac{(Wij)^\alpha}{\sum_k (W_{ik})^\alpha}. \tag{7}$$

The resulting matrix $\tilde{W}$ serves as a soft target that modulates the contrastive loss: pairs that are structurally close in G' are penalized less strongly as negatives, reflecting the intuition that image–text pairs sharing common items are semantically closer.

3.6 Relation-aware contrastive learning

Our training objective combines two signals: a contrastive term that aligns matching image-relation-text pairs and a soft term that encourages the learned similarities to be consistent with the relational topology of the line graph.

Let $\mathbf{H}_U$ and $\mathbf{H}_V$ be the final image and text matrices, where $\mathbf{H}_U$ contains all the output representations $\tilde{\mathbf{h}}^{\text{out}}_{e,u}$ of the (relation type, image) pairs, while $\mathbf{H}_V$ contains all $\tilde{\mathbf{h}}^{\text{out}}_{e,v}$ of the (relation type, text) pairs.

We ℓ_2-normalize embeddings before computing similarities:

$$S_{UV} = \mathbf{H}_U \mathbf{H}_V^\top, \quad S_{VU} = \mathbf{H}_V \mathbf{H}_U^\top. \tag{8}$$

CLIP Loss. We adopt the symmetric InfoNCE objective [21]:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2} \left(\mathcal{L}_{\text{InfoNCE}}(S_{UV}) + \mathcal{L}_{\text{InfoNCE}}(S_{VU}) \right). \tag{9}$$

Graph-regularized KL loss. We construct a soft contextual target:

$$P = \lambda \tilde{W} + (1 - \lambda) I, \tag{10}$$

where $\lambda \in [0, 1]$ is a mixing coefficient that controls the trade-off between graph-based contextual similarity and strict identity matching, and I is the identity matrix. Following *Gao et al.* [10], we minimize the symmetric KL divergence to soften the CLIP loss:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2} \left(\text{KL}(P \,\|\, S_{UV}) + \text{KL}(P^\top \,\|\, S_{VU}) \right). \tag{11}$$

Final objective. The overall loss balances contrastive alignment and contextual relations:

$$\mathcal{L} = (1 - \eta)\mathcal{L}_{\text{CLIP}} + \eta\mathcal{L}_{\text{KL}}, \tag{12}$$

where $\eta \in [0, 1]$ balances the contrastive objective and the graph-regularized divergence. Overall, **CANVAS** combines relation-conditioned restriction maps with a context-informed contrastive objective. This unifies local relational transformations and global regularization for relation-aware multimodal alignment. Intuitively, the loss encourages the representations of each image-relation pair to align with the corresponding text-relation pair while addressing CLIP's hard-negatives problem when negative pairs are not uncorrelated (i.e., either different relations of the same item or items that are connected).

4 Experiments

Baselines We compare **CANVAS** against approaches that incorporate semantic or graph-structured information for multimodal retrieval and art understanding:

- ArtSAGENet [7]: it combines visual feature learning with a knowledge-enhanced graph component to improve multi-task visual representation learning in art. At test time, we remove the graph-connectivity dependency, as explained in [12], for a fair comparison.
- CLIP [25] and SigLIP [36]: it learns joint image-text embeddings using a contrastive objective, enabling zero-shot cross-modal retrieval and classification. SigLIP extends CLIP using a sigmoid-based loss. Throughout the paper, pre-trained models are denoted as CLIP and SigLIP, whereas fine-tuned variants are denoted as CLIP-ft and SigLIP-ft.
- ColPali and ColQwen [9]: These models produce multi-vector embeddings from document page images and rely on late-interaction matching mechanisms to directly optimize downstream retrieval tasks.
- GraphCLIP [29]: extends CLIP by aligning artwork images with knowledge-graph representations, enabling context-aware artwork classification.
- MSC [35]: it introduces a semantically optimized retrieval framework based on a Main Semantics Consistency loss. Their objective is to rank the semantically most relevant cross-modal matches during retrieval.
- RCML [24]: learns image–text embeddings by conditioning contrastive learning on semantic relations using relation-guided cross-attention.

Datasets We report results on 3 datasets for multi-relational art understanding:

- HertzianaDP: consisting of a collection of paintings and drawings belonging to the Bibliotheca HertzianaDP, not publicly available prior to this work, which was processed to obtain a mix of textual and metadata-based fields [19]; it consists of 45,819 images and 86,321 texts split into 125,267 nodes and 190,195 edges.

- SemArt+ [13]: we process the SemArt dataset augmented with per-sentence annotations from Explain Me the Paintings [1] through metadata cleaning and pronoun resolution (as documented in App. A.3); it contains 34,770 images and 62,289 texts split in 97,053 nodes and 151,430 edges.
- WikiArt+: we extend WikiArt [31] by enriching artworks, artists, and art movements with processed Wikipedia-sourced explanations; it consists of 22,028 images and 29,603 texts split into 51,631 nodes and 308,050 edges.

The data is split into train, validation, and test sets, ensuring each item appears only once to prevent leakage. Additional details are in App. A.2.

Implementation details All experiments are implemented in PyTorch. As CLIP implementation, we use OpenCLIP ViT-B/32 pre-trained on LAION-2B [25]. We apply a partial fine-tuning strategy: the last $L_{\text{ft}} = 3$ transformer blocks of the vision and text encoders, along with the projection heads, are unfrozen; all remaining parameters are kept frozen. Edge-relation embeddings are produced with the fully frozen text encoder. Computational cost is detailed in App. A.4.

The core of **CANVAS** consists of N stacked sheaf-inspired layers. Each layer learns restriction maps via a three-layer MLP Sheaf Learner, with hidden dimensions 512 and 256 and LeakyReLU activations. We set the latent dimension to $d = 512$ and stack $N = 3$ layers. We use the AdamW optimizer with a learning rate of 10^{-5} and a batch size of 256. We train for 50 epochs; we apply gradient clipping by norm with a maximum value of 1.0 and employ early stopping with a patience of 5 epochs, monitoring the validation loss. We report results on the test set using the model that achieved the best results on the validation set. The parameters we select are: $\tau = 0.7$, $\alpha = 1.2$, $\lambda = 0.7$, and $\eta = 0.5$. At inference time, **CANVAS** only requires the image or text and, optionally, a relation type. The relation type can be inferred from text as reported in **??**.

Evaluation Following prior work [7,35], we report Recall@K (fraction of queries with a correct match in the top-K), Precision@K (fraction of relevant items in the top-K), and NDCG@K (which gives higher credit to correct matches appearing earlier in the ranked list) for retrieval, and Accuracy for classification. For Recall, Precision, and NDCG, we report image-to-text (I2T) and text-to-image (T2I) scores. Due to space constraints, only Recall results are shown; additional results are in App. B.

5 Results

We evaluate whether multi-relational alignment improves cross-modal retrieval over single-relation VLMs, and whether relation-aware learning benefits artwork classification across diverse metadata types.

5.1 Cross-modal retrieval

Table 1 reports retrieval results across the three datasets. **CANVAS** achieves the best performance in image-to-text retrieval across all datasets by a large

Table 1: Retrieval results in terms of Recall (R) in text-to-image (T2I) and image-to-text (I2T) across the selected datasets ($K = \{5, 10\}$). **Bold** denotes the best model for a dataset, underlined the second best.

	HertzianaDP				SemArt+				WikiArt+			
Method	T2I R@5	T2I R@10	I2T R@5	I2T R@10	T2I R@5	T2I R@10	I2T R@5	I2T R@10	T2I R@5	T2I R@10	I2T R@5	I2T R@10
ArtSAGENet	0.001	0.001	0.000	0.001	0.003	0.004	0.002	0.003	0.001	0.002	0.000	0.000
CLIP	0.041	0.063	0.013	0.018	0.305	0.385	0.078	0.111	0.131	0.174	0.042	0.059
CLIP-ft	0.046	0.069	0.057	0.084	0.189	0.216	0.093	0.131	0.137	0.149	0.082	0.158
ColPali	0.011	0.017	0.003	0.006	0.123	0.177	0.021	0.028	0.043	0.061	0.008	0.011
ColQwen	0.007	0.011	0.001	0.002	0.154	0.206	0.025	0.037	0.037	0.054	0.008	0.011
MSC	0.000	0.001	0.000	0.000	0.006	0.010	0.012	0.023	0.002	0.003	0.011	0.024
RCML	0.026	0.041	0.044	0.071	0.134	0.163	0.076	0.105	0.017	0.025	0.080	0.148
SigLIP	0.061	0.083	0.016	0.022	0.334	0.423	0.095	0.133	**0.239**	0.301	0.067	0.091
SigLIP-ft	0.038	0.058	0.051	0.076	0.337	0.446	0.102	0.150	0.137	**0.307**	0.115	0.209
CANVAS	**0.099**	**0.158**	**0.386**	**0.514**	**0.376**	**0.489**	**0.645**	**0.714**	0.144	0.217	**0.703**	**0.781**

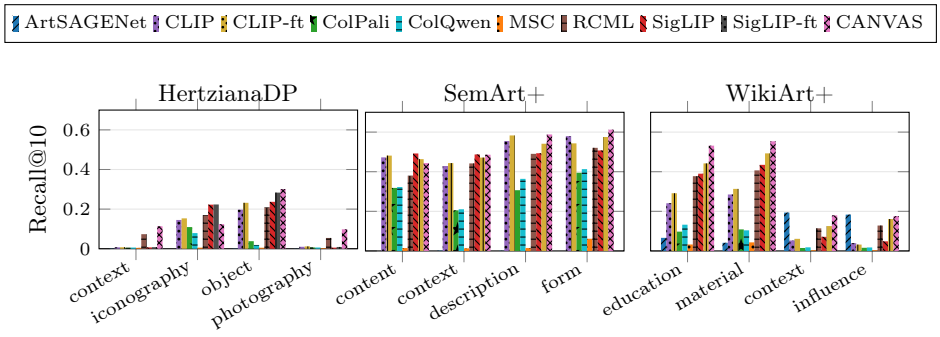

Fig. 4: Text-to-Image Recall@10 by relation type across the three datasets.

margin, with image-to-text Recall@10 reaching 0.514 on HertzianaDP, 0.781 on WikiArt+, and 0.714 on SemArt+. In text-to-image retrieval, **CANVAS** achieves the best results on HertzianaDP and SemArt+, while remaining competitive on WikiArt+. These results confirm that modeling multiple relational subspaces yields richer representations that consistently benefit cross-modal alignment. The advantage is evident on HertzianaDP, which is absent from CLIP's pretraining data, where **CANVAS** reaches an I2T Recall@10 of 0.514, compared to 0.022 for SigLIP. We note that, unlike CLIP, SigLIP, and MSC, which do not accept the relationship type, our textual prompt includes information about the relationship. This can be inferred by the model itself, as shown in App. B1, and does not require extra input information in practice. The I2T/T2I performance gap reflects a property of the data: images have multiple valid textual matches, giving I2T queries more correct targets, whereas a single text must retrieve a single image from among many similar alternatives. For instance, in HertzianaDP the test split has 6,874 images but 12,145 texts.

Figure 4 provides a finer-grained view, breaking down T2I R@10 by relation type. **CANVAS** achieves strong performance across all relation types, whereas VLM baselines exhibit high variance, performing well on descriptive relations but degrading on more complex ones (*e.g.*, context).

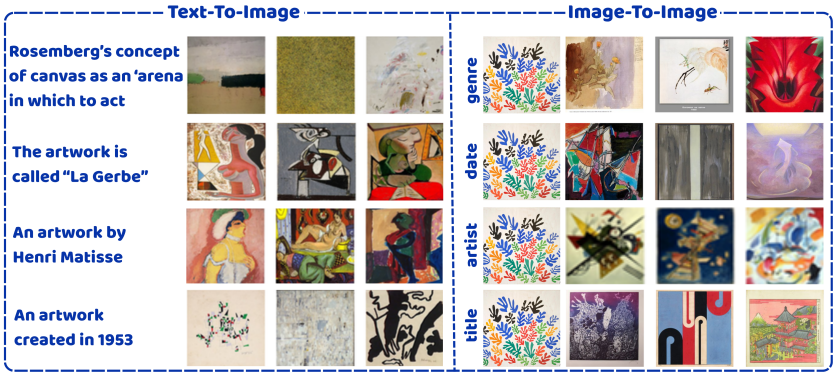

Fig. 5: Qualitative examples of relation-aware multimodal retrieval using CANVAS. Left panel (text-to-image): Each text query activates a different relation-specific subspace, yielding diverse retrieval results: Rosenberg's observation obtains gestural and materic artworks, the title is associated with an early cubist tradition, while artist queries return other Matisse works that are stylistically similar to his Fauvist period, and date queries return contemporaneous artworks from the early 50's. Right panel (image-to-image): Given "The Sheaf" as a visual query, **CANVAS** retrieves relevant images across different relation types: genre finds decorative works featuring plants, date retrieves artworks from the 40s and 50s, artist finds abstract expressionist works with similar compositions, and title recovers thematically related pieces.

Ablation We ablate on the main components of **CANVAS** and assess their impact: (a) KL and InfoNCE losses, (b) graph-regularization without relation conditioning, and (c) simple multi-head projections per relation type. Results are in Tab. 2. To test (a), we turn off first the KL and later the InfoNCE loss. We show that combining the two losses allows us to leverage the benefits of both (whereby the InfoNCE improves image-to-text retrieval, while the contextual-KL loss improves T2I). For (b), we remove the relationship-conditioned modulation and notice that it degrades the model's performance. For (c), we show the need for the restriction maps by substituting them with simple per-relation linear heads and notice that the substitution penalizes the model.

5.2 Relation-aware classification

Table 3 reports classification accuracy across relation types. **CANVAS** achieves the best or second-best result on every task. Crucially, no single baseline matches this consistency: when **CANVAS** ranks second, the top-performing method varies across tasks, meaning that no alternative baseline can be selected as a uniformly better choice. **CANVAS** consistently outperforms all VLM baselines on relationally complex attributes such as Style and Period, where single-embedding methods struggle. Compared to ArtSAGENet and MSC, the graph-based baselines, **CANVAS** achieves superior performance on the majority of tasks while operating inductively, without requiring graph connectivity at test time.

Table 2: Ablations on HertzianaDP. The results are reported on HertzianaDP because it is the only dataset without potential data leakage from the base CLIP.

Variant	T2I@5	T2I@10	I2T@5	I2T@10
CANVAS	**0.099**	**0.158**	0.386	0.514
(a) only InfoNCE loss	0.002	0.004	**0.525**	**0.541**
(a) only KL loss	0.098	0.156	0.385	0.490
(b) no relation modulation	0.002	0.003	0.001	0.002
(c) simple multi-head projections	0.021	0.040	0.140	0.170

Table 3: Classification accuracy across relation types on the selected datasets. **Bold** denotes the best model for a dataset, underlined the second best.

	HertzianaDP		SemArt+				WikiArt+				
Method	Artist	Period	Artist	Date	Genre	Style	Artist	Genre	Style	Period	Material
ArtSAGENet	**0.808**	0.791	0.121	0.563	0.520	0.374	0.430	0.023	0.563	0.218	0.089
CLIP	0.010	0.015	0.695	0.237	0.350	0.259	0.010	0.021	0.695	0.174	0.140
CLIP-ft	0.464	0.702	0.212	0.480	0.623	0.418	0.343	0.035	0.516	0.546	0.145
ColPali	0.014	0.082	0.128	0.109	0.166	0.102	0.014	0.008	0.128	0.084	0.005
ColQwen	0.020	0.028	0.645	0.036	0.025	0.014	0.020	0.010	0.645	0.062	0.011
GraphCLIP	0.588	0.719	0.074	0.381	**0.738**	0.494	0.109	0.025	0.207	**0.522**	0.112
MSC	0.175	0.062	0.050	0.312	0.123	0.237	0.175	0.000	0.050	0.028	0.007
RCML	0.250	0.548	0.171	0.343	0.580	0.232	0.264	0.023	0.441	0.402	0.177
SigLIP	0.470	0.088	0.702	0.598	0.168	0.091	**0.470**	0.030	0.702	0.257	0.283
SigLIP-ft	0.650	0.776	0.620	0.611	0.585	0.459	0.460	0.060	0.644	0.339	**0.291**
CANVAS	0.472	**0.794**	**0.808**	**0.634**	0.628	**0.637**	0.462	**0.043**	**0.790**	0.485	0.226

5.3 Qualitative analysis

Figure 5 illustrates how **CANVAS** modulates retrieval depending on the relation on WikiArt+. For Matisse's *Sheaf*, the textual queries retrieve different aspects related to the painting and artist: its gestural nature of abstract expressionism is retrieved through Rosenberg's quote; a title query surfaces Cubist compositions; the artist query returns Fauvist-adjacent works; the date yields contemporaneous 50's pieces. Image-to-Image retrieval exhibits analogous behavior, confirming that each subspace captures distinct art-historical dimensions.

Figure 6 provides geometric evidence via a t-SNE projection on WikiArt+. Around Matisse's *Sheaf*, different relational lenses yield different neighborhoods: 50's works for date, Abstract Expressionist/Art Nouveau neighborhood of style, artworks at the boundary between figurative and abstract for genre, and partly figurative flat-color-plane compositions for key characteristics. These neighborhoods are neither identical nor redundant, confirming that **CANVAS** maintains complementary subspaces, which substantiates the results in Tabs. 1 and 3.

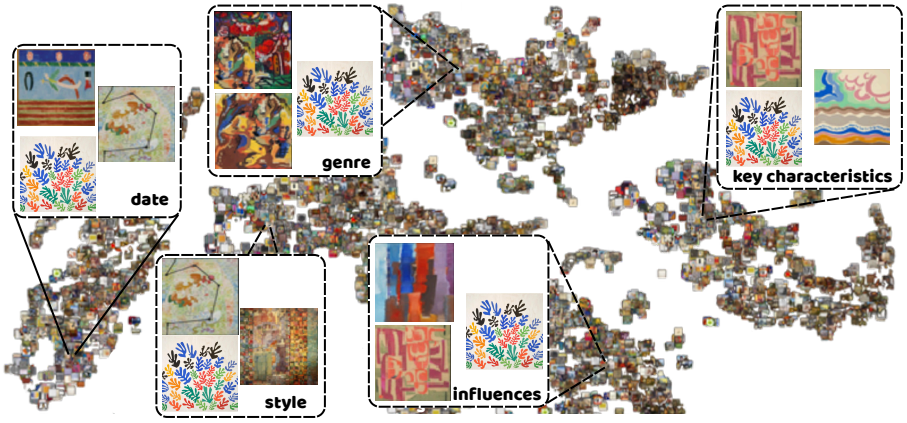

Fig. 6: Visualization of relation-aware embedding space learned by CANVAS on the WikiArt+ dataset. The plot shows a 2D t-SNE projection of the learned multimodal embeddings, with points representing image embeddings. The image shows how **CANVAS** organizes artworks into semantically meaningful clusters based on different relations. The inset boxes highlight specific relational neighborhoods around Matisse's "Sheaf": artworks clustered by temporal information show examples of art from the 50's, stylistic neighbors are at the boundary between abstract expressionism and Art Nouveau, the genre highlights a tension between figurative and abstract art, while influences are primarily from abstract expressionism, with key characteristics mirroring the same tension between recognizable figures and color planes.

6 Conclusion

We presented **CANVAS**, a sheaf-inspired framework learning relation-aware multimodal representations for art understanding. Through aspect-conditioned embeddings and a graph-informed contrastive loss, it captures multi-relational structure while remaining fully inductive at inference. Together with three new benchmarks, WikiArt+, SemArt+ and HertzianaDP, experiments across datasets demonstrate improvements over VLM and graph-based baselines.

Beyond art understanding, **CANVAS** is domain-agnostic and suits settings such as medical imaging and cultural heritage, where images and texts share heterogeneous semantic relations.

Acknowledgements

The authors acknowledge Pietro Liuzzo and the Photographic Collection team of the Bibliotheca Hertziana for the support in providing the HertzianaDP dataset. Antonio Purificato and Fabrizio Silvestri acknowledge project EVOLVE: A Fluid Framework for Dynamic Agentic AI Systems, Progetto Ateneo Sapienza. Noa Garcia acknowledges JSPS KAKENHI No. 23H00497 and No. 22K12091.

References

1. Bai, Z., Nakashima, Y., Garcia, N.: Explain me the painting: Multi-topic knowledgeable art description generation. In: CVPR. pp. 5422–5432 (2021)
2. Barbero, F., Bodnar, C., de Ocáriz Borde, H.S., Bronstein, M., Veličković, P., Liò, P.: Sheaf neural networks with connection laplacians. In: Topological, Algebraic and Geometric Learning Workshops 2022. pp. 28–36. PMLR (2022)
3. Bodnar, C., Di Giovanni, F., Chamberlain, B., Lio, P., Bronstein, M.: Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. Advances in Neural Information Processing Systems **35**, 18527–18541 (2022)
4. Castellano, G., Vessio, G.: Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. Neural Computing and Applications **33**(19), 12263–12282 (2021)
5. Crowley, E., Zisserman, A.: The state of the art: Object retrieval in paintings using discriminative regions. In: Proceedings of the British Machine Vision Conference 2014. British Machine Vision Association (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
7. Efthymiou, A., Rudinac, S., Kackovic, M., Worring, M., Wijnberg, N.: Graph neural networks for knowledge enhanced visual representation of paintings. In: ACM MM. pp. 3710–3719 (2021)
8. El Vaigh, C.B., Garcia, N., Renoust, B., Chu, C., Nakashima, Y., Qian, Y., Nagahara, H.: Gnnboost: boosting artwork classification with graph embeddings. Multimedia Tools and Applications pp. 1–21 (2025)
9. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. In: ICLR (2025)
10. Gao, Y., Liu, J., Xu, Z., Wu, T., Zhang, E., Li, K., Yang, J., Liu, W., Sun, X.: Softclip: Softer cross-modal alignment makes clip stronger. In: AAAI. vol. 38, pp. 1860–1868 (2024)
11. Garcia, N., Renoust, B., Nakashima, Y.: Context-aware embeddings for automatic art analysis. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 25–33 (2019)
12. Garcia, N., Renoust, B., Nakashima, Y.: Contextnet: representation and exploration for painting classification and retrieval in context. International Journal of Multimedia Information Retrieval **9**(1), 17–30 (2020)
13. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Eur. Conf. Comput. Vis. Worksh. pp. 0–0 (2018)
14. Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., Grover, A.: Cyclip: Cyclic contrastive language-image pretraining. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS. vol. 35, pp. 6704–6719 (2022)
15. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. NeurIPS **30** (2017)
16. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: Lightgcn: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 639–648 (2020)
17. Hoffman, A.J.: On the line graph of the complete bipartite graph. The Annals of Mathematical Statistics **35**(2), 883–885 (1964)
18. Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H.: Recognizing image style. arXiv preprint arXiv:1311.3715 (2013)

19. Liuzzo, P.M.: Paintings Gemma-Enriched Dataset. Fotothek - Bibliotheca Hertziana (2025). https://doi.org/10.17617/3.Z8W2JR, https://doi.org/10.17617/3.Z8W2JR
20. Lyu, S., Rockmore, D., Farid, H.: A digital technique for art authentication. Proceedings of the National Academy of Sciences **101**(49), 17006–17010 (2004). https://doi.org/10.1073/pnas.0406398101, https://www.pnas.org/doi/abs/10.1073/pnas.0406398101
21. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
22. Pan, X., Ye, T., Han, D., Song, S., Huang, G.: Contrastive language-image pre-training with knowledge graphs. NeurIPS **35**, 22895–22910 (2022)
23. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI. vol. 32 (2018)
24. Qiao, Y., Hu, Y., Zhao, L.: Multimodal representation learning conditioned on semantic relations (2025), https://arxiv.org/abs/2508.17497
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
26. Ramos, P., Ramos, R., Garcia, N.: Data leakage in visual datasets. In: CVPR. pp. 6309–6319 (2025)
27. Ruthardt, J., Gaur, M., Ramanan, D., Tapaswi, M., Asano, Y.M.: Steerable visual representations. In: ECCV (2026)
28. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855 (2015)
29. Scaringi, R., Fiameni, G., Vessio, G., Castellano, G.: Graphclip: Image-graph contrastive learning for multimodal artwork classification. Knowledge-Based Systems **310**, 112857 (2025)
30. Strezoski, G., Worring, M.: Omniart: multi-task deep learning for artistic data analysis. arXiv preprint arXiv:1708.00684 (2017)
31. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE TIP **28**(1), 394–409 (2019). https://doi.org/10.1109/TIP.2018.2866698, https://doi.org/10.1109/TIP.2018.2866698
32. Teru, K., Denis, E., Hamilton, W.: Inductive relation prediction by subgraph reasoning. In: ICML. pp. 9448–9457. PMLR (2020)
33. Tutte, W.T.: Graph theory, vol. 21. Cambridge university press (2001)
34. Van Noord, N., Postma, E.: Learning scale-variant and scale-invariant features for deep image classification. Pattern Recognition **61**, 583–592 (2017)
35. Xie, Y., Wang, Y., Xie, Y., Tan, X., Li, J., Li, X., Peng, W., Tang, M., Fang, M.: Image-text retrieval with main semantics consistency. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. p. 2629–2638. CIKM '24, Association for Computing Machinery, New York, NY, USA (2024)
36. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: CVPR. pp. 11975–11986 (2023)
37. Zhu, Y., Zhu, Z., Lin, B., Liang, X., Zhao, F., Liu, J.: Relclip: Adapting language-image pretraining for visual relationship detection via relational contrastive learning. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 4800–4810 (2022)
38. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE **111**(3), 257–276 (2023)