Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI
A. Salatino et al. (Eds.)

doi:10.3233/SSW240008

Stitching Gaps: Fusing Situated Perceptual Knowledge with Vision Transformers for High-Level Image Classification

Delfina Sol MARTINEZ PANDIANI [a,1], Nicolas LAZZARI [b] and Valentina PRESUTTI [b]

[a] *Centrum Wiskunde & Informatica, Amsterdam, Netherlands*
[b] *University of Bologna, Bologna, Italy*

Abstract.
The increasing demand for automatic high-level image understanding, including the detection of abstract concepts (AC) in images, presents a complex challenge both technically and ethically. This demand highlights the need for innovative and more interpretable approaches, that reconcile traditional deep vision methods with the situated, nuanced knowledge that humans use to interpret images at such high semantic levels. To bridge the gap between the deep vision and situated perceptual paradigms, this study aims to leverage situated perceptual knowledge of cultural images to enhance performance and interpretability in AC image classification. We automatically extract perceptual semantic units from images, which we then model and integrate into the ARTstract Knowledge Graph (AKG). This resource captures situated perceptual semantics gleaned from over 14,000 cultural images labeled with ACs. Additionally, we enhance the AKG with high-level linguistic frames. To facilitate downstream tasks such as AC-based image classification, we compute Knowledge Graph Embeddings (KGE). We experiment with relative representations [1] and hybrid approaches that fuse these embeddings with visual transformer embeddings. Finally, for interpretability, we conduct posthoc qualitative analyses by examining model similarities with training instances. The adoption of the relative representation method significantly bolsters KGE-based AC image classification, while our hybrid methods outperform state-of-the-art approaches. The posthoc interpretability analyses reveal the visual transformer's proficiency in capturing pixel-level visual attributes, contrasting with our method's efficacy in representing more abstract and semantic scene elements. Our results demonstrate the synergy and complementarity between KGE embeddings' situated perceptual knowledge and deep visual model's sensory-perceptual understanding for AC image classification. This work suggests a strong potential of neuro-symbolic methods for knowledge integration and robust image representation for use in downstream intricate visual comprehension tasks. All the materials and code are available at `https://github.com/delfimpandiani/Stitching-Gaps`

Keywords. Abstract Concepts, Knowledge Graph Embeddings, Vision Transformers, Image Classification, Interpretability

[1]Corresponding Author. E-mail: dsmp@cwi.nl.

1. Introduction

In the rapidly evolving field of Computer Vision (CV), the enduring challenge is to equip machines with human-like cognitive abilities, surpassing data-driven pattern recognition to bridge the gap between bottom-up signal processing and top-down knowledge retrieval and reasoning [2]. This goal is rooted in the understanding that "humans are not mere appearance-based classifiers; we acquire knowledge from experience and language" [3]. While explicit knowledge has historically been recognized as a way to improve automatic image understanding, modern data-driven techniques are rooted in the deep learning (DL) paradigm and aim to acquire the majority of this knowledge from the training data itself.

Meanwhile, CV endeavors to address increasingly complex tasks have been proposed, including discerning abstract concepts like personality traits, political affiliations, and beauty from visual cues [4,5,6,7]. However, the limitations of the deep learning paradigm become evident in these tasks of abstract concept-based (AC) image classification, where performance remains notably low [8]. Cognitive science suggests that ACs differ from concrete concepts in that they serve as specifiers of relations between entities, relying more on *semantic* and associative relations rather than categorical distinctions [9,10]. Detecting ACs, therefore, often requires integrating and inferring over perceptual information [11,12]. Indeed, abstract and cognitively complex tasks benefit from an explicit understanding of perceptual semantics, such as objects and colors [13], as well as symbolic representations like common-sense associations [14] and high-level linguistic frames [15].

AC image classification emphasizes the need to complement CV models with the capacity to comprehend the relationships within a scene [16,17]. Innovative approaches are needed to bridge the deep learning paradigm with explicit knowledge for complex image interpretation. Key to this endeavor is the integration of resources where semantics is represented explicitly, which plays a crucial role in enhancing interpretability [18]. This integration can be realized by combining knowledge-driven methods with data-driven methods [19]. Promising results have been achieved by leveraging Knowledge Graphs (KGs) to integrate background knowledge in CV models [20,21].

Based on these insights, we introduce the "situated perceptual knowledge" paradigm to abstract concept-based (AC) image classification. This paradigm is centered on the development of a KG integrating automatically detected perceptual semantics of images, commonsense knowledge, and ACs via the SituAnnotate [22] ontology. We inject KG embeddings (KGE), computed on the situated KG, with image representations obtained from visual transformers through varying fusion techniques, including absolute and relative representations [1]. We also conduct qualitative analyses to understand the models' abilities to capture symbolic and embodied aspects of image content by analyzing relevant similarities with training instances.

This work is structured as follows: In Section 2, we review related work. Section 3 outlines our method to construct and embed the situated perceptual ARTstract Knowledge Graph (AKG), while Section 4 presents AC image classification experiments using the embeddings. Section 5 presents our results, and in Section 6, we discuss and perform post-hoc interpretation of the AC image classification results, as well as propose potential future directions. We conclude in Section 7 with a summary of our findings and contributions.

2. Background

High-level Image Understanding The field of CV aims to understand images as data [23,24] and interprets scene content at various levels [25], seeking high-level interpretation from visual data [26,27]. Recent advancements focus on automating recognition of abstract and high-level meanings in images, including situational analysis [28,29,30,31], event recognition [32], and visual persuasion and intent analysis [5,33,34,35], visual sentiment analysis [32,36,37], aesthetic analysis [38,6], social signal processing [39,40,41,42], and visual rhetorical analysis [43,44,45].

Injecting Background Knowledge in CV models Knowledge and reasoning have been used in CV tasks for decades now. Methods based on First Order Logic [46,47] and Description Logic [48] have been proposed to perform AC classification and recognition. Despite their promising results, such methods suffer from the lack of flexibility in the data representation. Images are difficult to encode in a structured form. To overcome this issue, different approaches have been proposed to exploit structured knowledge within neural networks, which are far more flexible and can work directly on the raw image. Most works focus on injecting knowledge from large structured resources, such as Visual Genome [49] and ConceptNet [50]. Such approach enables the creation of multi-modal architectures that are able to learn image representations that are informed by the structured resource [51]. Indeed, successful results have been obtained in various CV tasks, including image classification [52], to visual scene recognition [14], image captioning [53], image understanding [54] and scene generation [55]. Combining different representations (including KG) altogether is an active research field [56]. Approaches that integrate the knowledge representation within the model architecture [52,53] or exploit it to learn better features [55] have been proposed. Those methods, however, require an extensive training data set. Recently, with the advent of large pre-trained models, techniques to merge different approaches have been proposed. This includes techniques that unify latent spaces trained on different modalities [57,1] as well as techniques that integrate different representations [58].

Cognitive Insights into AC Representation Cognitive theories of AC representation explore two paths: distributional models, which infer meaning from word co-occurrence statistics [59], and embodied cognition, which grounds meaning in sensory, perceptual, and motor interactions, emphasizing context [60,61]. The "multiple representations view" reconciles these perspectives, integrating distributional and embodied information as mutually influential [62]. For ACs, this involves sensorimotor systems, linguistic data, emotional experiences, and social interactions [63]. Cognitive substrates of ACs, such as acquired embodiment, relationality, and emotionality, highlight their complex representation in the brain [62]. Acquired embodiment connects abstract words to sensory-motor (S-M) information through associations with concrete words, involving context-dependent activation of S-M features [64]. Relationality suggests ACs specify multiple semantic relations between entities, relying on semantic and associative relations [65,10], implying that detecting ACs may involve identifying relationships between objects in scenes [17]. Emotionality emphasizes that abstract words generally have higher emotional associations than concrete words, supported by imaging studies showing greater activation in emotion-related brain areas during AC processing [66,67]. This integrated approach reflects ACs' grounding in both linguistic and perceptual experiences [68].

3. Method

3.1. Task Definition and Data Selection

In light of recent studies on AC image classification [8], we adopt a single-label multi-class classification approach for our task. This decision was made to prioritize identifying the most prominent AC category associated with each image, rather than using a multi-label multi-class approach. This approach simplifies the complexities in visual representations of ACs, where images may not strictly belong to a single category, and ACs may overlap. By focusing on a single-label approach, we aim to capture the primary association between an image and its most salient AC. Each image I_i in our dataset $X = [I_1, I_2, \ldots, I_m]$ is paired with a ground truth label y_i drawn from a set of K potential AC categories. Our objective is to determine the optimal image representation I_i and model parameters θ that maximize the conditional probability $p(y_i \mid I_i, \theta)$. This formulation assumes ACs as mutually exclusive for classification purposes, despite their potential overlap within real-world visual scenes. We aim to enhance AC image classification by automatically integrating situated perceptual knowledge into image representations. This involves three steps: extracting perceptual semantic (PS) features from images (Section 3.2), integrating them with contextual knowledge into the ARTstract Knowledge Graph (Section 3.3), and embedding the AKG for novel image representations suited for AC classification (Section 3.4).

We experiment on the ARTstract dataset, consisting of 14,795 cultural images labeled with abstract concepts (ACs) [8], amalgamates data from ArtPedia [69], ARTemis [37], the Ads Dataset [45], and the Tate Collection metadata. This curated collection includes seven defined AC labels: *comfort*, *danger*, *death*, *fitness*, *freedom*, *power*, and *safety*.The dataset's images were selected by querying the original datasets for images tagged with the words associated with each of the AC clusters [8]. Importantly, the dataset utilizes evoked clusters initially identified in the Ads Dataset, where these clusters originated from analyzing AC co-occurrences in advertising images. These clusters often reflect symbols and themes primarily observed in Western and Euro-centric contexts [45], potentially introducing biases rooted in specific cultural perspectives.

3.2. Perceptual Semantic Units Extraction

Cognitive neuroscience research highlights that ACs in the human brain are linked to concrete items, activating sensory-motor features associated with objects, actions, and colors [64]. Additionally, emotions play a significant role in AC modeling and perception [70,71], suggesting that ACs are grounded in tangible experiences and sensory perceptions. Building on this insight, we extract perceptual semantics (PS) as cognitive-based intermediary semantic units for image representation. These include actions, age, art style, dominant colors, evoked emotions, human presence, depicted objects, and an automatically generated image caption. Table 1 provides an overview of the selected extractors, including their architectural backbones, the datasets on which they are trained, and a description of their task. They have been selected through manual investigation, focusing on easily available, off-the-shelf models trained or fine-tuned for relevant semantic units and prioritizing popularity and ratings.[2] For each extractor, we manually align the output labels to the respective nodes in ConceptNet [72] (except for the textual caption). By leveraging those detectors, we reflect the interpretative capabilities of CV tools across various semantic levels, facilitating automated processing for new or unseen images without requiring human-annotated ground truths.

Table 1. Perceptual Semantic (PS) units and their associated artificial annotators, their model backbones, pretraining datasets, and other details.

PS Unit	Backbone	Dataset	Description
Action	ViT	HAR dataset [73]	Computes the probabilities for detected actions in an image such as *running*, *eating*, and *sleeping*.
Age Tier	ViT	Fair Face [74]	Categorizes individuals into age groups ranging from *0-2* to *70+*.
Art Style	ViT	ArtBench-10 [75]	Detects artistic styles such as *Art Nouveau*, *Baroque*, and *Expressionism*.
Top Colors	ColorThief	*N/A*	Detects up to 5 dominant colors in an image. We convert each RGB color to the CSS3 web color with the closest Euclidean distance. If a distance ≥ 50 is detected, the color is discarded.
Emotion	Artemis [37]	Artemis [37]	Detects the prominent emotion in an image from nine emotion categories such as *amusement*, *awe*, and *contentment*.
Human Presence	ViT	Deep Fashion v1 [76]	Detects whether a human presence is in an image.
Image Caption	BLIP [77]	COCO [78]	Generate a textual description of an image.
Detected Objects	DETR [79]	COCO [78]	Detects the objects in an image. Only objects whose probability is ≥ 0.4 are retained.

3.3. ARTstract Knowledge Graph Creation

We use the *SituAnnotate* ontology [22], which models the situated assignment of annotation labels to information objects, and includes a module tailored for *image* annotation situations. To reify the PS labels, we represent each as an instance of the *Annotation* class and connect it to its `AnnotationSituation`, associated `Image`, utilized `LexicalEntry`, assigned `AnnotationStrength`, label `AnnotationRole`, and the ConceptNet concept that provided its typification (see Figure 1). To formally represent the annotation contexts, each entry row in Table 1 is reified as an instance of a subclass of `ImageAnnotationSituation`. The resulting triples contain detailed information about these annotation situations, including geographical locations, timestamps, annotators, specific model architectures, datasets, and more. To further enhance the KG, following

[2] The models utilized for each detection are as follows: Action detection: https://huggingface.co/DunnBC22/vit-base-patch16-224-in21k_Human_Activity_Recognition; Age Tier: https://huggingface.co/nateraw/vit-age-classifier; Art Style: https://huggingface.co/oschamp/vit-artworkclassifier); Top Colors: https://github.com/lokesh/color-thief; Emotion detection: https://github.com/optas/artemis/blob/master/artemis/neural_models/image_emotion_clf.py; Human Presence: https://huggingface.co/adhamelarabawy/human_presence_classifier; Image Captioning: https://huggingface.co/Salesforce/blip-image-captioning-largemodel; Object Detection: https://huggingface.co/Salesforce/facebook/detr-resnet-50.

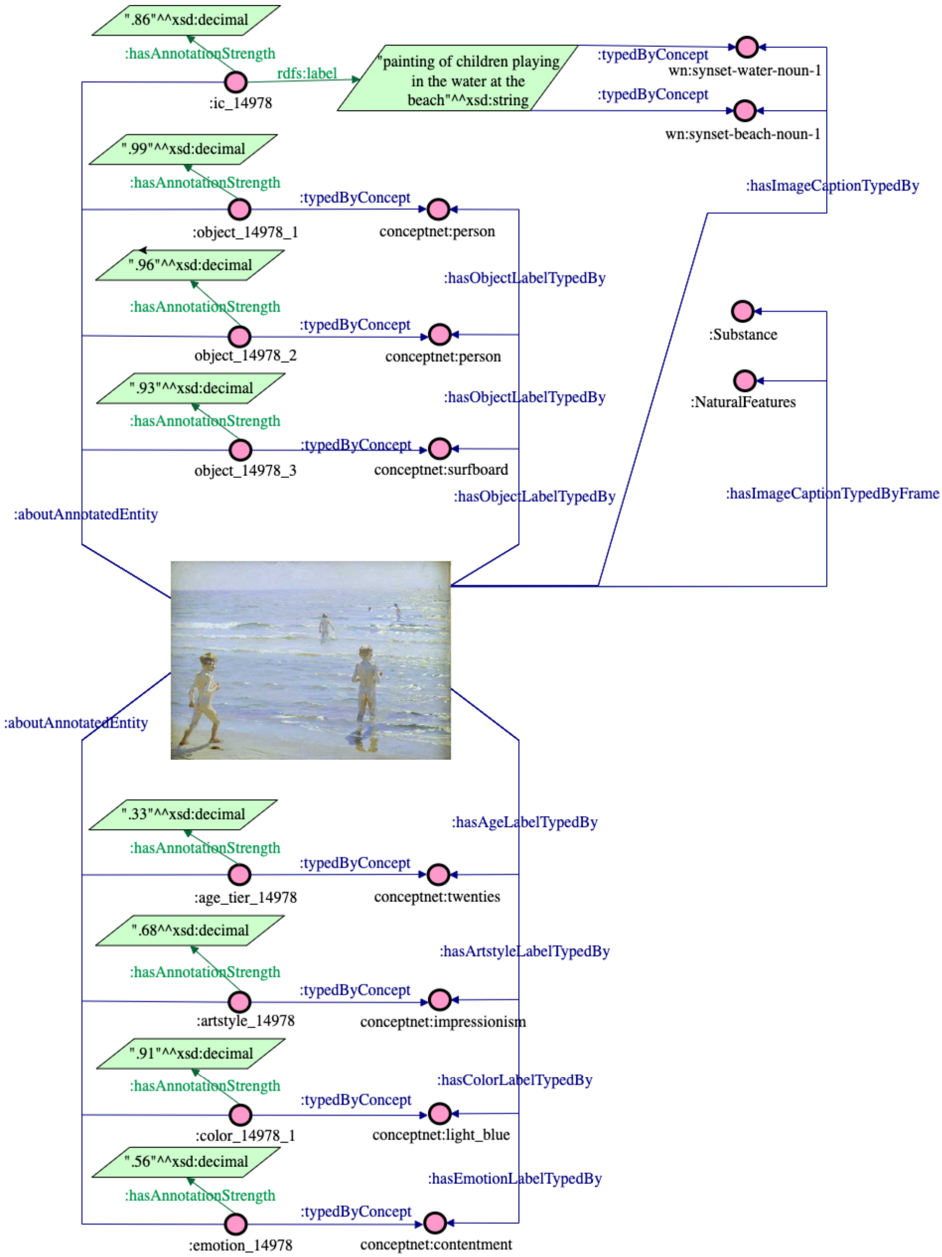

Figure 1. Subset of the A-Box of ARTstract-KG, showing the types of commonsense linguistic knowledge connected to a single image instance. Most annotations are typed by ConceptNet concepts, while the image captions are typed by WordNet concepts as well as by linguistic frames.

[15] we extract WordNet synsets from the captions using FRED [80], and employ these as triggers for the extraction of high-level linguistic frames. The KG was built using RDFlib, which facilitated the mapping of PS from a JSON file to the SituAnnotate ontol-

ogy, which was accessed directly via its permanent IRI[3]. Additionally, we employed the Framester [80] schema to reference ConceptNet and WordNet IRIs.

3.4. ARTstract Knowledge Graph Embedding

In order to exploit the information encoded in the situated AKG, we compute the KGE of AKG by relying on TransE [81]. KGEs transform KG components into continuous vector spaces, so as to simplify the manipulation while preserving the inherent structure of the KG [82]. The representation of a node associated with each image encodes all the selected PS features of an image without taking into account the raw features of the image. Since each PS is aligned to ConceptNet, the representation of two images that share the same PS feature will be similar. Before computing the embeddings, we preprocess the AKG to prevent data leakage: we remove all rows containing AC cluster names in subjects or objects. This filtering maintains the KGEs' separation from the target AC clusters, preserving integrity for the downstream task.

4. Experiments

4.1. Encoding Phase

In the encoding phase, we explore three primary approaches to represent image data: (i) rely only on the image representation produced by the KGE method (I_{KGE}); (ii) rely only on features extracted by a deep CV model (I_{CV}); (iii) combine both the KGE and CV model representations (I_{H}). When relying solely on the KGE method, we use the AKG embeddings generated with TransE. In the case of using only the CV model (I_{CV}), we evaluate three architectures: VGG [83], ResNet [84] and ViT [85]. VGG and ResNet are Convolutional Neural Networks, while ViT is based on the transformer architecture. In the combined approach (I_{H}), we concatenate the KGE embedding with the most effective CV model representation, which in our case is ViT.

While concatenation has demonstrated effectiveness [58], it merges vectors from disparate latent spaces without considering their structural differences. To address this limitation, we adopt the relative representation approach [1]. This method constructs representations where each sample is defined in relation to a subset A of the training data X, which is selected as anchor samples. Each training sample is represented with respect to the embedded anchors $e_{a^{(j)}} = E(a^{(j)})$ with $a^{(j)} \in A$ via a generic similarity function $sim : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. This yields a scalar score r between two absolute representations $r = sim(e_{x^{(i)}}, e_{x^{(j)}})$. Thus, the relative representation of $x^{(i)} \in X$ is defined as:

$$r_{x^{(i)}} = (sim(e_{x^{(i)}}, e_{a^{(1)}}), sim(e_{x^{(i)}}, e_{a^{(2)}}), \ldots, sim(e_{x^{(i)}}, e_{a^{(|A|)}})) \qquad (1)$$

We adopt this approach to construct the relative representations of both the KGE embedding and the most effective CV model representation, ViT. Each embedding in the training distribution is represented in relation to a set of embedded anchor vectors. We randomly select 700 anchors from the training set, with 100 corresponding to each tar-

[3] https://w3id.org/situannotate

get AC class. For each image, we transform the two vector representations, I_{KGE} and I_{ViT}, into their relative versions, $I_{\text{R-KGE}}$ and $I_{\text{R-ViT}}$ respectively. Subsequently, we apply hybrid methods to combine the relative embeddings using either concatenation ($||$) or the Hadamard product (element-wise multiplication, $\odot$). Notably, when using relative representations, the vectors maintain consistent dimensionality, eliminating the need for extension through padding.

4.2. Classification Phase

During the *classification phase*, the output from each encoding approach is inputted into a classifier. We utilized a Multi-Layer Perceptron (MLP) model for final classification, consisting of two linear layers sequentially activated by Rectified Linear Units (ReLU), with a dropout layer (dropout rate of 0.3) for regularization. Training utilized Cross-Entropy Loss with a fixed learning rate ($lr = 0.001$) and 50 epochs per architecture. The MLP outputs the probability of an image belonging to one of the 7 labels in the ARTstract dataset. Data processing efficiency was optimized through multi-threading employing 16 workers. All experiments were conducted on an RTX3080 with 24GB of RAM.

5. Results

5.1. AC Image Classification Performances

Table 2. Comparison of KGE-based models with state-of-the-art DL computer vision models in terms of Macro F1 Score. The top-performing model is highlighted in both bold and italics. The second-best performing models are denoted in bold. SPK: Situated Perceptual Knowledge, DL: Deep Learning.

Input Embedding	Macro F1	Paradigm
Absolute KGE	0.22	SPK
Absolute VGG-16	0.23	DL
Absolute ResNet-50	0.24	DL
Absolute ViT	0.30	DL
Absolute KGE || Absolute ViT	**0.31**	Hybrid
Relative KGE	0.27	SPK
Relative ViT	0.28	DL
Relative KGE $\odot$ Relative ViT	0.29	Hybrid
Relative KGE || Relative ViT	***0.33***	Hybrid

In Table 2, we present the performance metric, specifically the Macro F1 score, of our approaches compared to the state of the art. Among these models, ResNet-50 and VGG-16 achieved Macro F1 scores of 0.24 and 0.23, respectively, while ViT achieved a score of 0.30. Despite lacking access to pixel-level features, our KGE-only model demonstrated competitive results compared to the CNNs, scoring 0.22. Impressively, the Relative KGE version outperformed both CNN methods with a score of 0.27.

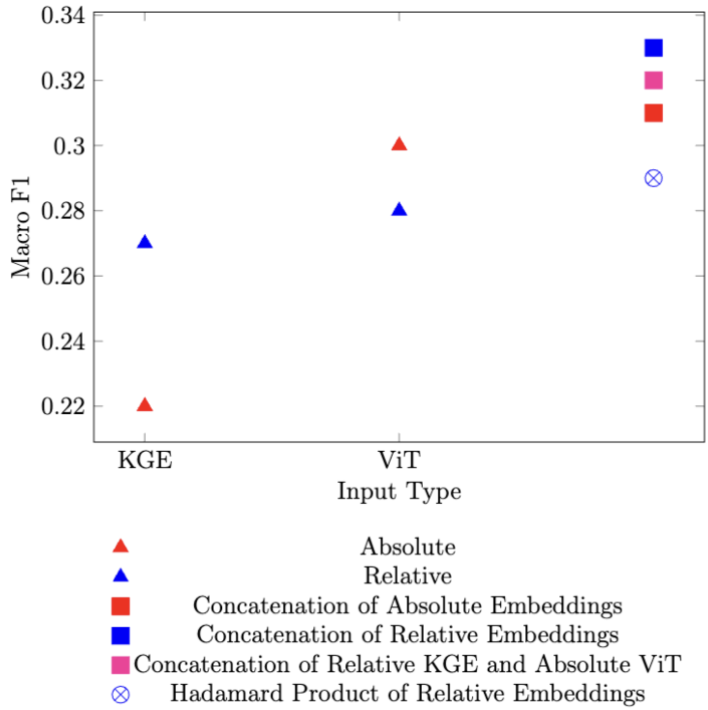

Figure 2. Macro F1 scores on the AC image classification tasks for different input embeddings.

Absolute versus Relative Embeddings RelKGE outperformed absKGE, achieving a higher Macro F1 score of 0.27 compared to 0.22 (see Figure 2). This suggests that using relative representation significantly improves KGE's performance in AC image classification. For ViT embeddings, absViT scored 0.3 in Macro F1, slightly higher than relViT at 0.28. These findings reveal a nuanced difference, indicating relative embeddings may slightly degrade ViT performance.

Hybrid Embeddings Our hybrid approaches exhibited the best overall performance (as depicted in Figure 2), surpassing other methods in our study as well as the existing state of the art. Notably, the Relative KGE || Relative ViT approach, which combines the two relative embeddings, achieved the highest F1 score of 0.33, representing a new benchmark for this task. Additionally, the concatenation of Absolute KGE and Absolute ViT embeddings attained a score of 0.31, further illustrating the effectiveness of hybrid methods. Lastly, the Hadamard product of the two relative embeddings scored slightly lower than Absolute ViT (0.29 versus 0.30), but it remains comparable and offers enhanced interpretability, making it a valuable option for analysis.

6. Discussion

6.1. The ARTstact-KG

The ARTstact-KG is a comprehensive resource containing over 1.9 million triples derived from the ARTstract dataset, encompassing situated annotation data from more than

14,000 unique images. It provides detailed information about perceptual semantics, facilitated by the reification of annotation situations and semantic labels. Annotation situations capture various details such as geographical locations, timestamps, annotators, model architectures, and datasets. Similarly, semantic labels assigned to images are reified as instances of the `Annotation` class, forming connections between annotations, annotation situations, images, lexical entries, annotation strengths, annotation roles, and ConceptNet concepts, while linguistic frames extracted from image captions further enhance its expressiveness, offering a comprehensive linguistic context for each image.

6.2. AC Image Classification

Our results show the effectiveness of various embedding approaches in enhancing AC image classification performance. In our study, we implemented the relative representation method [1], encoding each instance relative to selected anchor points. Our results suggest that the relative representation method improves KGE-based models by providing more meaningful cluster-level semantic information, enhancing semantic resolution. Conversely, Absolute ViT outperformed Relative ViT, indicating that ViT may not benefit from the semantic bias introduced by relative representation, potentially losing fine-grained local differences and spatial resolution critical for pixel-level models like ViT. These findings underscore the relative representation method's potential to boost KGE-based image classification, offering a valuable alternative to ViT. Additionally, our hybrid embedding approaches, particularly the combination of Relative KGE and Relative ViT embeddings, showcase the highest F1 score attained in this task, setting a new benchmark for AC image classification. The competitive performance of hybrid methods underscores the effectiveness of integrating different types of embeddings to leverage their respective strengths.

6.3. Post-Hoc Interpretability

6.3.1. Perceptual Disparities: ViT vs. KGE

To better understand the results, we conducted a post-hoc analysis on randomly selected test images, for which we retrieved the top 5 most similar training images using both ViT and KGE embeddings, and compared them. In Figure 3, we illustrate two examples. In the first example, when using ViT embeddings, 4 out of 5 of the top similar images correctly share the ground truth label, *freedom*. These images prominently feature the United States flag, indicating that ViT's encoding accentuates features reminiscent of the flag's presence. This observation suggests potential geographical and cultural bias in ViT's training data. Contrastingly, all top similar images based on Absolute KGE embeddings share the ground truth *comfort*. These images exhibit a strong visual and semantic connection with the lower portion of the test image, including elements such as grass, fields, trees, and greens. This suggests that KGE embeddings may be biased towards parts of images associated with comfort, a bias possibly inherited from the dataset itself. In the second example, none of the top images based on ViT share the correct ground truth *freedom*, nor do they share evident perceptual semantics. Conversely, the top similar images based on KGE not only share the correct ground truth label but also prominently feature the Statue of Liberty. In this instance, KGE successfully associates the test image with semantically relevant training instances, whereas ViT fails to encode similarity for coherent results.

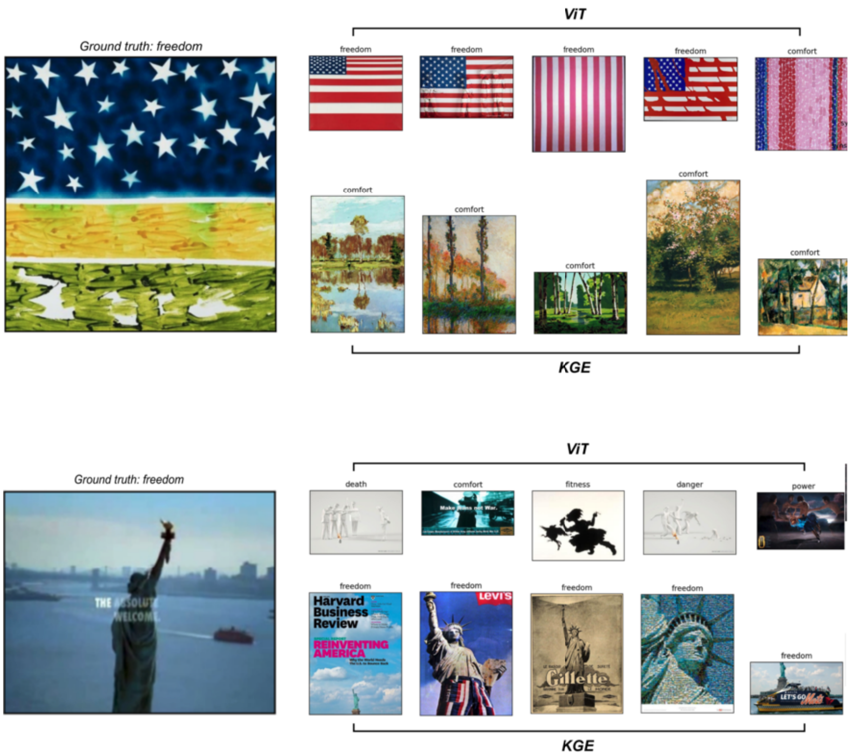

Figure 3. Absolute ViT vs. Absolute KGE embeddings capture different aspects of ARTstract images. Top: Absolute ViT captures aspects that resemble the United States flag while KGE captures more landscape-related features, Bottom: Absolute KGE demonstrates superior semantic performance than ViT by encoding similarities with perceptually diverse visions of the Statue of Liberty

6.3.2. *High-Level Semantic Proficiency of KGE*

Further examples reveal that, even when both embeddings make correct predictions, they exhibit distinct understandings of images. Notably, KGE embeddings appear to encapsulate more “high-level” semantic features compared to ViT embeddings. For instance, in Figure 4, the images identified as most similar by ViT to the test image predominantly share visual characteristics reminiscent of the image’s “aesthetics” or “style,” emphasizing elements like colors, shapes, and artistic composition. However, while some of these images share the correct ground truth label *comfort*, others are labeled *power*. In contrast, the most similar instances identified by KGE embeddings all share the correct ground truth label *comfort*, consistently conveying the same higher-level semantics–in this case, the depiction of a comfortable situation with a woman reading. KGE achieves this by aligning the top 5 similar images with depictions of women reading, whereas ViT fails to correlate the test image with any training images portraying the same scene semantics.

Multiple test instances suggest that the KGE method exhibits superior performance over ViT in capturing higher-level semantics, as illustrated in Figure 5, a test image portraying two individuals in an intimate scenario serves as an exemplary case. While ViT-similar images primarily focus on pixel-level resemblances, such as dark colors and textures, KGE emphasizes the higher level “situation” of individuals engaged in an intimate

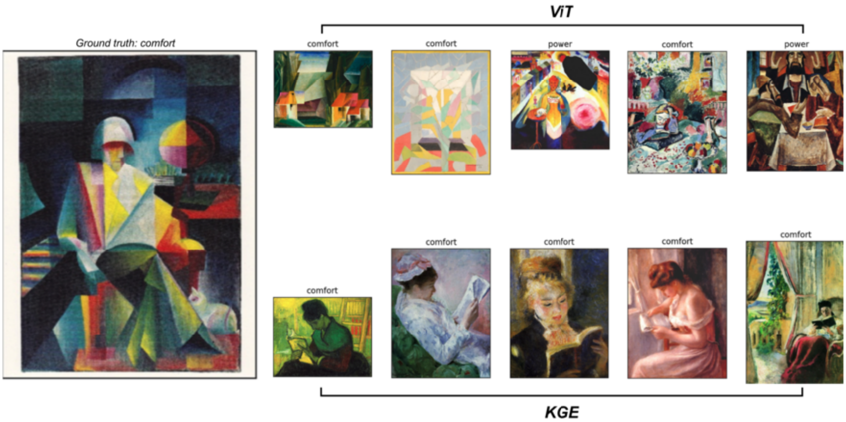

Figure 4. Contrasting semantic proficiency of Absolute KGE vs. Absolute ViT. The top image illustrates ViT's focus on colors and textures (aesthetics), whereas KGE excels in recognizing explicit semantics, particularly women sitting on couches. In the bottom image, KGE effectively encodes the semantics of reading a book in the test artwork.

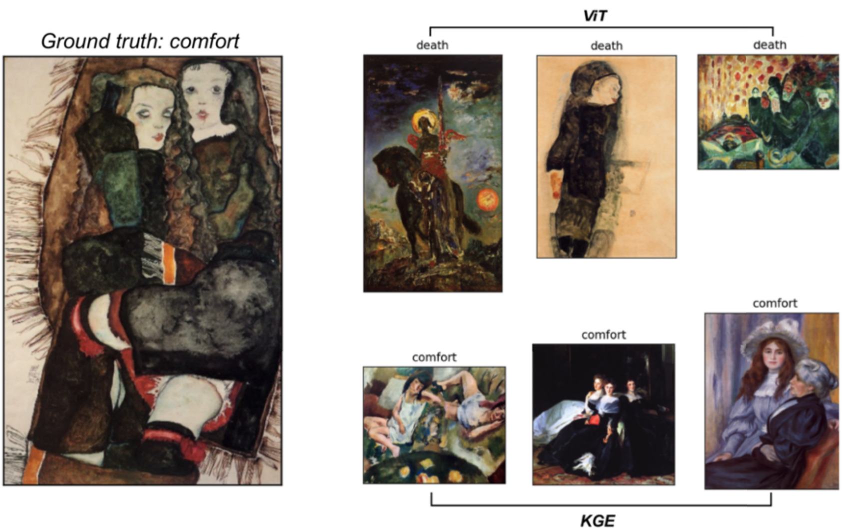

Figure 5. ViT misclassifies as *death*, but KGE successfully associates images with crosses to the concept of *comfort*, indicating ViT's focus on colors and textures.

interaction. Notably, the majority of KGE-generated similar images depict scenes with two or more people in intimate settings, contrasting with the single individuals predominantly shown in ViT-similar images. While ViT may excel in recognizing detailed visual elements, these results suggest the KGE method's potential applicability in tasks requiring the interpretation of social interactions, relationships, or other complex high-level visual cues. A final but compelling example of this trend is seen in Figure 6, in which the KGE method is able to identify a "trigger" of a high-level semantic concept. The test image, categorized under the label *death*, depicts a convoy resembling ambulances,

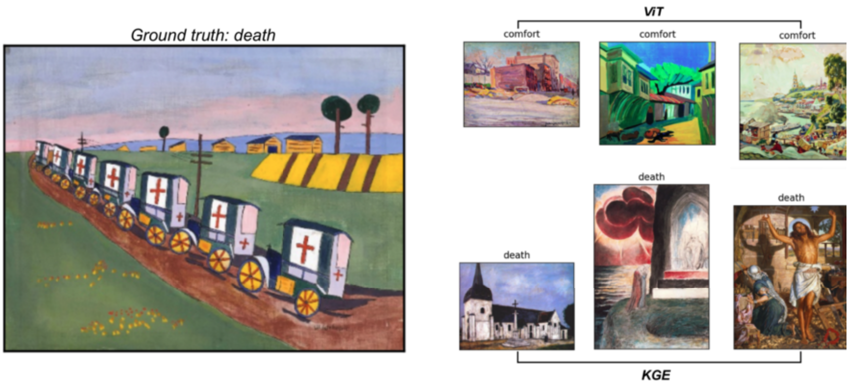

Figure 6. ViT misclassifies as *comfort*, but KGE successfully associates images with crosses to the concept of *death*.

reminiscent of those found in war zones. While ViT retrieves images primarily tagged with *comfort*, likely due to the original image's warm colors, landscape composition and drawing/cartoon-like drawing features, the top similar images as based on vit feature outdoor scenes irrelevant to the ground truth of death. In contrast, the top three similar images based on KGE embeddings share the correct label of *death*, evoked through the presence of crosses and crucifixion imagery. This indicates that the KGE model successfully associates images featuring crosses with the concept of death, prioritizing this connection over visual elements associated with *comfort* (the ViT misclassification). A final but compelling example highlighting this disparity is showcased in Figure 6, where the KGE method discerns the pivotal perceptual semantic unit that acts as a"trigger" for the high-level abstract concept. The test image, categorized under the label *death*, portrays a convoy resembling ambulances, reminiscent of those found in war zones. ViT retrieves images primarily tagged with *comfort*, likely due to the test image's warm colors, landscape composition, and cartoon-like drawing lines; the top similar images based on ViT feature outdoor scenes irrelevant to the ground truth of *death*. Conversely, the top three similar images based on KGE embeddings accurately share the *death* ground truth, and all share the depiction of crosses and crucifixion imagery. This signifies the KGE model's adeptness at associating images featuring crosses with the concept of death, prioritizing this connection over visual elements associated with other ACs like *comfort*.

The proficiency demonstrated by KGE embeddings is particularly remarkable considering the automated pipeline employed in constructing the ARTstract-KG. All perceptual semantic units were annotated using artificial annotators (models), indicating that they represent non-human evaluated perceptual semantics without human or manual semantic coherence checks. Consequently, this pipeline introduces inherent noise, compounded by the complexities of cultural art images, which often lack discrete objects and other detectable categories. Despite this noise, our qualitative analyses underscore the capacity of KGE embeddings to implicitly encode essential high-level semantics, a crucial aspect of our study. We believe that this discrepancy may primarily arise from the prototype selection process, wherein images are represented based on their similarity to these prototypes. Essentially, ViT's latent space heavily relies on the noise accumulated from its extensive training dataset. However, transforming this deep representation into a relative form introduces a strong prior assumption, expecting images that evoke the same AC to exhibit semantic similarity. This transformation does not perfectly align with ViT's latent space; instead, it confines the representation to specific regions within that space. This constraint potentially limits ViT's ability to express semantic relationships, as it can no longer rely solely on pixel-wise perceptual features but must effectively position images within its latent space. Consequently, the images obtained in this process may appear perplexing because the model's internal representation significantly differs from human perception. It primarily depends on subtle pixel differences, which, while effective for simple cognitive tasks, fall short in generalizing to the human internal understanding of the world.

6.3.3. *Hybridity and Complementarity*

One critical finding was that hybrid embeddings yielded the highest classification performance, with fusing two relative embeddings resulting in more significant improvements than fusing two absolute embeddings. This underscores the complementary nature of deep vision and situated perceptual paradigms. Post-hoc interpretability analyses support this result. For instance, Figure 7 illustrates the top 5 most similar anchor images using relative ViT embeddings (top row), relative KGE embeddings (middle rows), and hybrid (Hadamard product) embeddings. Each row includes the top ARTstract-KG nodes shared by the images, extracted via a SPARQL query on the knowledge graph. In this example, both relative ViT and relative KGE embeddings independently encode high similarity with anchor images related to the correct ground truth, *fitness*, and they overlap in selecting certain anchor images. However, each also includes images that do not belong to the correct class in their top selections. Remarkably, the hybrid embedding's top 5 anchors combine correct images from both unimodal embeddings, resulting in all images sharing the correct ground truth *fitness*. Moreover, nodes highly shared in either single embeddings are prevalent in the hybrid, indicating a complementary integration of information between them.

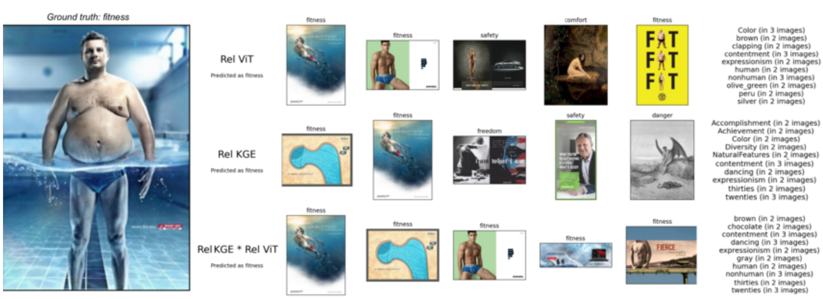

Figure 7. Interpretability results for a test image labeled as *fitness*. Top similar anchors are shown for the test instance using relative ViT embeddings (top row), relative KGE embeddings (middle rows), and hybrid embeddings. Shared ARTstract-KG nodes accompany each row. The hybrid embedding integrates complementary information from both relative embeddings to prioritize anchors tagged as *fitness*.

These findings highlight the potential of utilizing the Hadamard product ($A \circ B$) specifically on relative representations, which emphasizes anchors showing high similarity to a given image from both spatial and semantic perspectives. Through Hadamard multiplication, we assess the agreement between ViT and KGE regarding an image's similarity to prototypes, likely maintaining KGE's semantics while re-ranking images based on perceptual features detected by ViT. This operation aids in identifying anchor images with pronounced similarities to the image of interest across both embedding spaces, facilitating the recognition of anchors with dual-mode significance and unique characteristics captured by each modality. The results in Figure 7 underscore the proficiency of the hybrid embedding in recognizing spatially-semantically similar anchors, attributed to the complementary nature of relative ViT and relative KGE embeddings. By combining them, we capture information sometimes missed when using them individually, offering significant benefits, particularly in situations where both pixel-level and semantic understanding are essential. The hybrid approach shows promise for various applications where understanding the underlying factors contributing to image similarity is critical.

6.4. Limitations and Future Directions

Refining task definitions and evaluation metrics for AC image classification is crucial for future research. The current single-label multi-class approach may not fully capture nuanced relationships between images and ACs, where multiple ACs can co-exist [8]. Alternative ranking-based tasks assess relative AC relevance [5], offering a broader perspective. Future metrics should prioritize reasonability over strict objectivity, considering semantic AC relationships [37,86] to align with human perception. While our method ensures balanced class sampling in training, impacts of imbalance on test performance are unclear. Future evaluations should integrate class support metrics and macro averages for comprehensive model assessment. Variability in AC visual representation introduces technical complexities and ethical concerns. Bias frameworks in CV address labels on human images [87,88,89,90], shaped by social and cultural contexts [91,92], impacting diverse communities [93], perpetuating AI biases and racism. Improved dataset curation and transparency are essential for equitable AI.

7. Conclusion

This study introduces the ARTstract Knowledge Graph (AKG) and its pivotal role in advancing interpretability and performance in AC image classification. AKG is a foundational resource that captures perceptual semantics from over 14,000 cultural images labeled with abstract concepts (ACs), enhancing contextual understanding in visual analysis. By reifying perceptual semantics, encoding annotation context, and integrating with ConceptNet [94] and Framester [80], AKG provides a robust framework for interpretable reasoning in image analysis. The study demonstrates the effectiveness of Knowledge Graph Embeddings (KGE), both absolute and relative, in enhancing AC image classification performance and interpretability. Relative representation significantly strengthens KGE-based models, with hybrid KGE-ViT embeddings emerging as top performers, surpassing state-of-the-art approaches in AC image classification. Post-hoc interpretability analyses illuminate model strengths: ViT excels in capturing detailed pixel-level features, while KGE demonstrates proficiency in interpreting scenes and high-level semantics. The relative approach, by constraining ViT's latent space, raises crucial considerations for interpretability and semantic understanding. These findings provoke critical questions about how models learn representations and their implications for interpreting images. Future research includes exploring hybrid approaches for complex visual tasks.

References

[1] Moschella L, Maiorca V, Fumero M, Norelli A, Locatello F, Rodolà E. Relative representations enable zero-shot latent space communication. In: The Eleventh International Conference on Learning Representations; 2022. .

[2] Aditya S, Yang Y, Baral C. Explicit reasoning over end-to-end neural architectures for visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32; 2018. .

[3] Marino K, Salakhutdinov R, Gupta A. The More You Know: Using Knowledge Graphs for Image Classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society; 2017. p. 20-8.

[4] Segalin C, Cheng DS, Cristani M. Social Profiling through Image Understanding: Personality Inference Using Convolutional Neural Networks. Computer Vision and Image Understanding. 2017 Mar;156:34-50.

[5] Joo J, Li W, Steen FF, Zhu SC. Visual Persuasion: Inferring Communicative Intents of Images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 216-23.

[6] Gray D, Yu K, Xu W, Gong Y. Predicting Facial Beauty without Landmarks. In: Daniilidis K, Petros Maragos, Paragios N, editors. Computer Vision – ECCV 2010. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2010. p. 434-47.

[7] Martinez Pandiani DS, Presutti V. Seeing the Intangible: Survey of Image Classification into High-Level and Abstract Categories. arXiv preprint arXiv:230810562. 2023.

[8] Martinez Pandiani DS, Lazzari N, van Erp M, Presutti V. Hypericons for Interpretability: Decoding Abstract Concepts in Visual Data. International Journal of Digital Humanities (IJDH). 2023.

[9] Crutch SJ, Ridha BH, Warrington EK. The different frameworks underlying abstract and concrete knowledge: Evidence from a bilingual patient with a semantic refractory access dysphasia. Neurocase. 2006;12(3):151-63.

[10] Duñabeitia JA, Avilés A, Afonso O, Scheepers C, Carreiras M. Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. Cognition. 2009;110(2):284-92.

[11] Bruner J. Culture and human development: A new look. Human development. 1990;33(6):344-55.

[12] Firestone C, Scholl BJ. Cognition does not affect perception: Evaluating the evidence for “top-down” effects. Behavioral and brain sciences. 2016;39.

[13] Martinez Pandiani DS, Presutti V. Automatic Modeling of Social Concepts Evoked by Art Images as Multimodal Frames. In: Proceedings of the Workshops and Tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference (LDK 2021). Zaragoza, Spain; 2021. p. arXiv-2110.

[14] Kalanat N, Kovashka A. Symbolic image detection using scene and knowledge graphs. arXiv preprint arXiv:220604863. 2022.

[15] Ciroku F, De Giorgis S, Gangemi A, Martinez Pandiani DS, Presutti V. Automated multimodal sense-making: Ontology-based integration of linguistic frames and visual data. Computers in Human Behavior. 2024;150:107997.

[16] Isola P, Lim JJ, Adelson EH. Discovering states and transformations in image collections. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1383-91.

[17] Sadeghi MA, Farhadi A. Recognition using visual phrases. In: Cvpr 2011. Ieee; 2011. p. 1745-52.

[18] Chen X, Li LJ, Fei-Fei L, Gupta A. Iterative visual reasoning beyond convolutions. In: Proc. of CVPR 2018. IEEE; 2018. p. 7239-48.

[19] van Bekkum M, de Boer M, van Harmelen F, Meyer-Vitali A, Teije At. Modular Design Patterns for Hybrid Learning and Reasoning Systems: a taxonomy, patterns and use cases. arXiv:210211965 [cs]. 2021 Mar;51(9):6528-46.

[20] Aditya S, Yang Y, Baral C. Integrating knowledge and reasoning in image understanding. In: 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. International Joint Conferences on Artificial Intelligence; 2019. p. 6252-9.

[21] Tiddi I, Schlobach S. Knowledge graphs as tools for explainable machine learning: A survey. Artificial Intelligence. 2022;302:103627.

[22] Martinez Pandiani DS, Presutti V. Situated Ground Truths: Enhancing Bias-Aware AI by Situating Data Labels with SituAnnotate. [Under Review] Special Issue on Trustworthy Artificial Intelligence of ACM Transactions on Knowledge Discovery from Data (TKDD). 2024.

[23] Hoiem D, Efros AA, Hebert M. Putting objects in perspective. International Journal of Computer Vision. 2008;80:3-15.

[24] Arnold T, Tilton L. Distant Viewing Toolkit: A Python Package for the Analysis of Visual Culture. Journal of Open Source Software. 2020 Jan;5(45):1800.

[25] Szeliski R. Computer vision: algorithms and applications. Springer Nature; 2022.

[26] Borji A. Negative results in computer vision: A perspective. Image and Vision Computing. 2018;69:1-8.

[27] Hussain Z, Zhang M, Zhang X, Ye K, Thomas C, et al. Automatic understanding of image and video advertisements. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1705-15.

[28] Yatskar M, Zettlemoyer L, Farhadi A. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. p. 5534-42.

[29] Suhail M, Sigal L. Mixture-Kernel Graph Attention Network for Situation Recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE; 2019. p. 10362-71.

[30] Pratt S, Yatskar M, Weihs L, Farhadi A, Kembhavi A. Grounded Situation Recognition. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2020. Lecture Notes in Computer Science. Springer. Cham: Springer International Publishing; 2020. p. 314-32.

[31] Li R, Tapaswi M, Liao R, Jia J, Urtasun R, et al. Situation Recognition with Graph Neural Networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE; 2017. p. 4183-92.

[32] Yao X, She D, Zhao S, Liang J, Lai YK, et al. Attention-Aware Polarity Sensitive Embedding for Affective Image Retrieval. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE; 2019. p. 1140-50.

[33] Jia M, Wu Z, Reiter A, Cardie C, Belongie S, et al. Intentonomy: a Dataset and Study towards Human Intent Understanding. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE; 2021. p. 12981-91.

[34] Huang X, Kovashka A. Inferring Visual Persuasion via Body Language, Setting, and Deep Features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2016. p. 778-84.

[35] Guo M, Hwa R, Kovashka A. Detecting Persuasive Atypicality by Modeling Contextual Compatibility. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021. p. 952-62.

[36] Toisoul A, Kossaifi J, Bulat A, Tzimiropoulos G, Pantic M. Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions. Nature Machine Intelligence. 2021 Jan;3(1):42-50.

[37] Achlioptas P, Ovsjanikov M, Haydarov K, Elhoseiny M, Guibas LJ. ArtEmis: Affective Language for Visual Art. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE; 2021. p. 11569-79.

[38] Workman S, Souvenir R, Jacobs N. Understanding and Mapping Natural Beauty. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE; 2017. p. 5590-9.

[39] Sun Q, Schiele B, Fritz M. A Domain Based Approach to Social Relation Recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE; 2017. p. 435-44.

[40] Li W, Duan Y, Lu J, Feng J, Zhou J. Graph-Based Social Relation Reasoning. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2020. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2020. p. 18-34.

[41] Li J, Wong Y, Zhao Q, Kankanhalli MS. Dual-Glance Model for Deciphering Social Relationships. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE; 2017. p. 2669-78.

[42] Goel A, Ma KT, Tan C. An End-To-End Network for Generating Social Relationship Graphs. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE; 2019. p. 11178-87.

[43] Ye K, Nazari NH, Hahn J, Hussain Z, Zhang M, et al. Interpreting the Rhetoric of Visual Advertisements. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019 Apr;43(4):1308-23.

[44] Ye K, Kovashka A. ADVISE: Symbolism and External Knowledge for Decoding Advertisements. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. vol. 11219 LNCS. Cham: Springer International Publishing; 2018. p. 868-86.

[45] Hussain Z, Zhang M, Zhang X, Ye K, Thomas C, Agha Z, et al. Automatic Understanding of Image and Video Advertisements. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1705-15.

[46] Zhu Y, Fathi A, Fei-Fei L. Reasoning about object affordances in a knowledge base representation. In: European conference on computer vision. Springer; 2014. p. 408-24.
[47] London B, Khamis S, Bach S, Huang B, Getoor L, et al. Collective activity detection using hinge-loss Markov random fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2013. p. 566-71.
[48] Dasiopoulou S, Kompatsiaris I, Strintzis MG. Applying fuzzy DLs in the extraction of image semantics. In: Journal on data semantics XIV. Springer; 2009. p. 105-32.
[49] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. arXiv:160207332 [cs]. 2016 Feb;123(1):32-73.
[50] Havasi C, Speer R, Alonso J. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In: Recent advances in natural language processing. John Benjamins Philadelphia, PA; 2007. p. 27-9.
[51] Ektefaie Y, Dasoulas G, Noori A, Farhat M, Zitnik M. Multimodal learning with graphs. Nat Mac Intell. 2023;5(4):340-50.
[52] Novack Z, McAuley JJ, Lipton ZC, Garg S. CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. vol. 202 of Proceedings of Machine Learning Research. PMLR; 2023. p. 26342-62.
[53] Li X, Lian D, Lu Z, Bai J, Chen Z, Wang X. GraphAdapter: Tuning Vision-Language Models With Dual Knowledge Graph. CoRR. 2023;abs/2309.13625.
[54] Guo W, Wang J, Wang S. Deep multimodal representation learning: A survey. IEEE Access. 2019;7:63373-94.
[55] Buffelli D, Tsamoura E. Scalable Theory-Driven Regularization of Scene Graph Generation Models. In: Williams B, Chen Y, Neville J, editors. Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023. AAAI Press; 2023. p. 6850-9.
[56] Jabeen S, Li X, Shoib AM, Omar B, Li S, Jabbar A. A Review on Methods and Applications in Multimodal Deep Learning. ACM Trans Multim Comput Commun Appl. 2023;19(2s):76:1-76:41.
[57] Norelli A, Fumero M, Maiorca V, Moschella L, Rodolà E, Locatello F. ASIF: Coupled Data Turns Unimodal Models to Multimodal Without Training. CoRR. 2022;abs/2210.01738.
[58] Bollegala D, O'Neill J. A Survey on Word Meta-Embedding Learning. In: Raedt LD, editor. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022. ijcai.org; 2022. p. 5402-9.
[59] FIRTH JR. A synopsis of linguistic theory, 1930-1955. Studies in Linguistic Analysis. 1957.
[60] Barsalou LW. Perceptual symbol systems. Behavioral and Brain Sciences. 1999 Aug;22(4):577-660.
[61] Yee E, Thompson-Schill SL. Putting concepts into context. Psychonomic bulletin and Review. 2016;23:1015-27.
[62] Davis CP, Yee E. Building semantic memory from embodied and distributional language experience. WIREs Cognitive Science. 2021;12(5):e1555.
[63] Andrews M, Frank S, Vigliocco G. Reconciling Embodied and Distributional Accounts of Meaning in Language. Topics in Cognitive Science. 2014;6(3):359-70.
[64] Hoffman P. Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. Psychological Review. 2018;125(3):293.
[65] Crutch SJ, Connell S, Warrington EK. The different representational frameworks underpinning abstract and concrete knowledge: Evidence from odd-one-out judgements. Quarterly Journal of Experimental Psychology. 2009;62(7):1377-90.
[66] Kousta ST, Vigliocco G, Vinson DP, Andrews M, Del Campo E. The representation of abstract words: why emotion matters. Journal of Experimental Psychology: General. 2011;140(1):14.
[67] Vigliocco G, Kousta ST, Della Rosa PA, Vinson DP, Tettamanti M, et al. The Neural Representation of Abstract Words: The Role of Emotion. Cerebral Cortex. 2014 Jul;24(7):1767-77.
[68] Louwerse MM. Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. Topics in Cognitive Science. 2018;10(3):573-89.
[69] Stefanini M, Cornia M, Baraldi L, Corsini M, Cucchiara R. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In: Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20. Springer;

2019. p. 729-40.
[70] Kousta ST, Vigliocco G, Vinson DP, Andrews M, Del Campo E. The representation of abstract words: Why emotion matters. Journal of Experimental Psychology: General. 2011;140(1):14-34.
[71] Vigliocco G, Kousta S, Vinson D, Andrews M, Del Campo E. The representation of abstract words: What matters? Reply to Paivio's (2013) comment on Kousta et al.(2011). 2013.
[72] Speer R, Chin J, Havasi C. Conceptnet 5.5: An open multilingual graph of general knowledge. In: Thirty-first AAAI Conference on Artificial Intelligence; 2017. .
[73] Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL, et al. A public domain dataset for human activity recognition using smartphones. In: Esann. vol. 3; 2013. p. 3.
[74] Karkkainen K, Joo J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2021. p. 1548-58.
[75] Liao P, Li X, Liu X, Keutzer K. The artbench dataset: Benchmarking generative models with artworks. arXiv preprint arXiv:220611404. 2022.
[76] Liu Z, Luo P, Qiu S, Wang X, Tang X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 1096-104.
[77] Li J, Li D, Xiong C, Hoi S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. PMLR; 2022. p. 12888-900.
[78] Lin TY, Maire M, Belongie S, Hays J, Perona P, et al. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer; 2014. p. 740-55.
[79] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European conference on computer vision. Springer; 2020. p. 213-29.
[80] Gangemi A, Alam M, Asprino L, Presutti V, Recupero DR. Framester: A wide coverage linguistic linked data hub. In: Blomqvist E, Ciancarini P, Poggi F, Vitali F, editors. European Knowledge Acquisition Workshop. Lecture Notes in Computer Science. Springer. Cham: Springer International Publishing; 2016. p. 239-54.
[81] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems. 2013;26.
[82] Wang Q, Mao Z, Wang B, Guo L. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering. 2017;29(12):2724-43.
[83] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Bengio Y, LeCun Y, editors. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings; 2015. .
[84] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. p. 770-8.
[85] Zhai X, Kolesnikov A, Houlsby N, Beyer L. Scaling Vision Transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE; 2022. p. 1204-13.
[86] Zhao S, Huang Q, Tang Y, Yao X, Yang J, et al. Computational Emotion Analysis From Images: Recent Advances and Future Directions. arXiv:210310798 [cs]. 2021 Mar:85-113.
[87] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Hal D, et al. Datasheets for datasets. Communications of the ACM. 2021;64(12):86-92.
[88] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency; 2019. p. 220-9.
[89] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR; 2018. p. 77-91.
[90] Buolamwini J. Facing the Coded Gaze with Evocative Audits and Algorithmic Audits. Massachusetts Institute of Technology; 2022.
[91] Mohamed S, Png MT, Isaac W. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology. 2020;33:659-84.
[92] Birhane A. Algorithmic Colonization of Africa. SCRIPTed. 2020;17(2).
[93] Ciston S. A CRITICAL FIELD GUIDE FOR WORKING WITH MACHINE LEARNING DATASETS;

2023. K. Crawford and M. Ananny, Eds., Knowing Machines project. https://knowingmachines.org/critical-field-guide.

[94] Liu H, Singh P. ConceptNet–a practical commonsense reasoning tool-kit. BT technology journal. 2004;22(4):211-26.