ChatGPT

Hierarchical Visual Representations (World Model)

Early vision research emphasized **compositional grammars** for images. For example, Zhu & Mumford (2006) propose a *stochastic image grammar* embodied as an And–Or graph, where scenes decompose into objects, parts, and primitives in a hierarchy [1]. A **parse graph** for a given image is then a subtree of this global grammar, explaining the image via selected parts. This matches the idea of a *world model* as a graph of visual concepts (objects and parts) whose subtrees generate each image. Modern deep models echo this: Bear et al. (2020) introduce **Physical Scene Graphs (PSGs)** that explicitly represent scenes as hierarchical graphs of object parts at different scales, with learned latent vectors at each node [2]. PSGNet uses graph pooling and feedback to extract such PSGs from an image, improving unsupervised segmentation and generalization. Similarly, Deng et al. (2019) propose RICH, a *deep generative model* that learns a latent **compositional hierarchy** for scenes: its latent "scene graph" organizes entities into a tree of parts→objects, and higher-level codes guide lower-level parsing [3]. These models show how to formalize a hierarchical representation: they assume a tree-structured world of parts and use top-down inference to instantiate an image's sub-tree.

Importantly, real-world "world models" can leverage **external knowledge graphs** or ontologies. For instance, Zareian et al. (2020) show that linking a scene graph to common-sense knowledge (WordNet, ConceptNet) helps vision tasks: knowing "Person–Bike→riding" simplifies scene interpretation [4]. Likewise, Toro et al. (2017) demonstrate that injecting a commonsense ontology (ConceptNet) can improve vision tasks like image retrieval [5]. In other words, structuring the world model via known part-of and semantic relationships can guide image parsing and reduce ambiguity. These approaches suggest building our hierarchical feature graph (the "world model") by combining bottom-up statistics from images with top-down prior knowledge (e.g. from WordNet, ConceptNet, or learned from data).

Object- and Part-Centric Models

A crucial step is identifying and localizing "features" or parts in each image. Recent **object-centric** models learn to segment images into parts or objects without labels. The Slot Attention module (Locatello et al. 2020) uses iterative attention to produce a set of **slot vectors**, each binding to an object in a scene [6]. In effect, each slot is a learned embedding of an image segment (e.g. an object or part), found unsupervised. Advances like Slot-VAE incorporate hierarchical VAEs over slots to capture scene-level and object-level representations [7]. More recently, Kucuksozen & Yemez (CVPR 2025) introduce **COCA-Net**: a multi-scale clustering-attention network that extracts object-centric masks and embeddings in a bottom-up hierarchy [8]. COCA explicitly clusters pixels into object-centroid-based segments in a cascading hierarchy, yielding interpretable part/object masks without fixed slot counts.

These models provide practical techniques: one can use Slot Attention or COCA as a preprocessing stage to obtain masks or features corresponding to image parts. Each mask (e.g. for "eye", "arm", "wheel") can then be embedded (via CNN or transformer encoders) to produce semantic feature vectors. Critically, the hierarchy of slots or masks (e.g. eyes + nose → face, wheels + body → car) can be learned by stacking layers

or by top-down composition (as in RICH). Thus, object-centric segmentation networks give a domain-agnostic way to discover fine-grained parts, which can then map to our predetermined labels.

Self-Supervised Embeddings for Semantics

To encode each part or object as an embedding, we can leverage **self-supervised learning (SSL)**. SSL methods like SimCLR, MoCo, BYOL or MAE learn powerful image features without labels. In particular, recent SSL on transformers (e.g. DINO) yields surprisingly interpretable mid-level features: Caron et al. (2021) show that DINO-trained Vision Transformers produce attention maps that align with semantic segments [9]. In their words, *"self-supervised ViT features contain explicit information about the semantic segmentation of an image"*, even though the model was not trained to segment [9]. This means we can use SSL encoders to get embeddings for patches or objects that carry semantic content.

For example, one could use a DINO or iBOT vision transformer to extract patch embeddings, then cluster those into object-level vectors. The resulting vectors can be associated with labels (e.g. via nearest neighbor to labeled prototypes). More broadly, vision-language models like CLIP learn image embeddings aligned with text: these embeddings generalize across domains (thanks to large image-text pretraining) and can recognize varied concepts zero-shot. Thus CLIP or similar multimodal encoders can provide *domain-agnostic semantic embeddings* for detected parts (e.g. by embedding the cropped image and matching to label text).

Beyond Euclidean embeddings, **hyperbolic representations** have proven effective for hierarchies. Liu et al. (NeurIPS 2024) propose *HypStructure*, a regularizer that embeds known label hierarchies into hyperbolic space during training [10]. They show that enforcing a tree-based hyperbolic loss on features reduces distortion and improves generalization. Likewise, HypDAE (Li et al. 2024) embeds images in a hyperbolic latent space for few-shot generation, because *"images exhibit semantic hierarchies… where high-level attributes define the core identity, while low-level attributes add variation"*, and hyperbolic space naturally encodes such tree-like relationships [11]. These works suggest that encoding our hierarchical labels (and their embeddings) in hyperbolic space would preserve the part-to-whole relations with low distortion.

In practice, one could use any SOTA SSL encoder (e.g. a ViT or CNN) to extract features, then train or regularize those features to reflect our hierarchy. For instance, given a tree label hierarchy, apply a hyperbolic contrastive or tree-loss (as in HypStructure) so that child-part embeddings are close in hyperbolic distance to their parent feature. This could encourage the learned embeddings to respect the predefined hierarchy, making them interpretable.

Domain-Agnostic and Adaptable Methods

To keep the method general, **domain-agnostic architectures** are key. Foundation models exemplify this: e.g. **Segment Anything (SAM)** provides promptable masks for *any* object, enabling label-free segmentation across domains. One could use SAM to propose segments, then map each segment to the nearest concept embedding (perhaps via CLIP). In transfer learning, models like CLIP or large SSL encoders can be used off-the-shelf and fine-tuned on new domains with few labels, because they already capture broad visual concepts. Hierarchical prototypes (from WordNet or other ontologies) can be aligned with the visual features to adapt to a new domain's label set. For example, Radford et al. (CLIP) showed that adding language supervision yields embeddings that generalize to unseen classes and domains.

Moreover, many vision tasks now use open-vocabulary or zero-shot learning (embedding labels as text tokens). We could similarly treat each hierarchy label as a "concept", embed it with a language model (concept-net, WordNet gloss, or a word embedding), and align image segments to it. This bridges our visual hierarchy with existing semantic hierarchies (the "world model").

Integrating Knowledge and Testing Approaches

In summary, key insights from the literature: - **Hierarchical grammars and scene graphs** (Zhu & Mumford; PSGs) provide a blueprint: treat the world as a graph of parts→wholes, and parse images as subgraphs [1] [2].
- **Unsupervised part discovery** (Slot Attention, COCA) can find fine-grained segments that correspond to potential leaf nodes. One can then climb up by grouping slots (either learned or via known rules) to build higher-level embeddings [6] [8].
- **Self-supervised embeddings** (DINO, MAE, CLIP, etc.) give powerful features for each part or region. Techniques like hyperbolic embedding losses ensure those features respect a hierarchy [9] [11] [10].
- **Knowledge infusion** from ontologies (ConceptNet, WordNet) or scene-graph priors can bias the model towards plausible hierarchies (e.g. knowing "wheel" is part of "car"). Past work shows this can improve perception tasks [4] [5].

A practical approach could be: first use an SSL encoder + object-centric module to extract a set of part features from each image; then, use a hyperbolic or structured embedding loss (informed by our world hierarchy) so that each feature is tagged with the correct semantic label and placed appropriately in the hierarchy. Testing should include scene understanding tasks (segmentation, detection with hierarchy) as well as generalization to new domains. Overall, the literature points to combining *unsupervised part discovery* with *structured representation learning*, guided by external knowledge, to build the desired hierarchical embeddings.

Sources: Recent deep learning studies on image grammars, scene graphs, object-centric networks, and self-supervised embeddings [1] [2] [3] [6] [8] [9] [11] [10] [4] [5] provide both conceptual frameworks and concrete techniques for this hierarchical vision model.

---

[1] nowpublishers.com
https://www.nowpublishers.com/article/DownloadSummary/CGV-018

[2] [2006.12373] Learning Physical Graph Representations from Visual Scenes
https://arxiv.org/abs/2006.12373

[3] [1910.09119] Generative Hierarchical Models for Parts, Objects, and Scenes
https://arxiv.org/abs/1910.09119

[4] ecva.net
https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123680596.pdf

[5] [1705.08844] How a General-Purpose Commonsense Ontology can Improve Performance of Learning-Based Image Retrieval
https://arxiv.org/abs/1705.08844

6 Object-Centric Learning with Slot Attention
https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html

7 [PDF] Slot-VAE: Object-Centric Scene Generation with Slot Attention
https://proceedings.mlr.press/v202/wang23r/wang23r.pdf

8 Hierarchical Compact Clustering Attention (COCA) for Unsupervised Object-Centric Learning
https://openaccess.thecvf.com/content/CVPR2025/papers/Kucuksozen_Hierarchical_Compact_Clustering_Attention_COCA_for_Unsupervised_Object-Centric_Learning_CVPR_2025_paper.pdf

9 [2104.14294] Emerging Properties in Self-Supervised Vision Transformers
https://arxiv.org/abs/2104.14294

10 Learning Structured Representations with Hyperbolic Embeddings | OpenReview
https://openreview.net/forum?id=wBtmN8SZ2B&referrer=%5Bthe%20profile%20of%20Han%20Zhao%5D(%2Fprofile%3Fid%3D~Han_Zhao1)

11 HypDAE: Hyperbolic Diffusion Autoencoders for Hierarchical Few-shot Image Generation
https://arxiv.org/html/2411.17784v2