UNCOVERING HIERARCHICAL SCENE STRUCTURE FROM HUMAN PERCEPTUAL GROUPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the hierarchical organization of natural scenes remains a persistent challenge in computer vision. In this work, we present a data-driven framework for recovering multi-scale scene hierarchies directly from human visual grouping judgments without requiring semantic labels, demonstrating that universal grouping principles can be learned independently of semantic category knowledge. We introduce Triple-Seg, a benchmark dataset of 210 natural images paired with segmentations and dense human triplet comparison judgments of the image regions. Using a neural embedding model trained on these judgments, segment-level visual features are mapped into a embedding space that enables the reconstruction of structural scene dendrograms via agglomerative clustering. By relying on perceptual rather than semantic cues, the model generalizes to unseen categories from only a few hundred training images. Code and datasets will be made publicly available upon acceptance.

1 INTRODUCTION

Natural images are composed of visual elements that can be organized at multiple levels of abstraction. Small image regions combine to form object parts, object parts combine into objects, and objects interact to form larger scene structures. Humans appear to organize visual information in this hierarchical manner effortlessly (Wagemans et al., 2012), allowing complex scenes to be understood as coherent arrangements of nested visual entities rather than collections of independent pixels or regions (Võ et al., 2019; Vo, 2021). This organization does not necessarily depend on explicit semantic knowledge; rather, perceptual structures can emerge from relations between visual elements and their spatial organization (Palmer & Rock, 1994; Feldman, 2003; Biederman, 1987).

This organization of visual elements is not arbitrary, it emerges from fundamental principles of perceptual grouping. Gestalt psychologists proposed that visual perception relies on a set of organizational principles that determine which elements are perceived as belonging together, including proximity, similarity, good continuation, closure, and common region (Wagemans et al., 2012; Wertheimer, 1938). These principles allow local image elements to be integrated into larger perceptual units, enabling observers to structure complex visual inputs without requiring explicit semantic interpretation. Moreover, grouping mechanisms operate across multiple spatial scales, supporting hierarchical organizations in which smaller elements become integrated into progressively larger structures (Palmer & Rock, 1994; Wagemans, 2018). Such organization is thought to play a fundamental role in object formation, where visual objects emerge from the arrangement and relationships between their constituent parts rather than being directly available from low-level image information (Feldman, 2003).

Recovering hierarchical scene organizations directly from visual information remains a challenge in computer vision. Although recent advances in deep learning have led to substantial improvements in object recognition, detection, and semantic scene understanding(Krizhevsky et al., 2012; Dosovitskiy et al., 2020), most representations are not explicitly organized into hierarchies. Classical approaches sought to address this limitation through hierarchical segmentation and compositional representations (Felzenszwalb & Huttenlocher, 2004; Zhu & Mumford, 2007), while more recent methods increasingly rely on semantic object representations and learned category-level knowledge (Xu et al., 2017; Johnson et al., 2015). Because these semantic methods group pixels based on learned object identity, they inherently require massive datasets of dense, category-specific anno-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tations and struggle to generalize to unseen object classes. Consequently, standard segmentation benchmarks evaluate semantic categorization (Lin et al., 2014; Cordts et al., 2016), missing the underlying perceptual organization that groups visually coherent regions regardless of their semantic meaning.

Existing approaches leave a fundamental gap in visual representation learning. Standard deep networks largely fail to learn explicit part-whole structural representations, a limitation Hinton (2023) identified as a primary bottleneck in current machine learning architectures. Current hierarchical representations are typically driven either by hand-crafted low-level image cues (e.g., region merging) or by supervised high-level semantic labels (e.g., scene graphs and category-centric models). In this paper, we ask a fundamental question: is it possible to computationally recover multi-scale scene hierarchies without relying on semantics, purely based on low-level perceptual cues? We hypothesize that because perceptual grouping relies on universal visual regularities rather than category-specific knowledge, it does not require millions of densely annotated images. We demonstrate that by learning from a sparse set of human judgments across just a few hundred images, our model successfully captures fundamental grouping principles that generalize to unseen scenes and categories—circumventing the need for the expensive, exhaustive annotations typical of semantic segmentation.

In this work, we bridge this gap by demonstrating that human perceptual grouping judgments can serve as a direct supervisory signal for learning structured visual feature spaces. Our framework successfully maps segment-level visual features to human judgments, enabling the automatic extraction of multi-scale scene trees via agglomerative clustering and demonstrating predictable linear scaling with data volume. We summarize our contributions as follows:

- **The Triple-Seg Dataset**: We introduce Triple-Seg, a benchmark dataset comprising 210 diverse natural images paired with fine-grained segmentations and 33,257 human visual grouping judgments collected across dense triplet comparisons. This type of supervision allows the model to learn to capture the core statistical regularities of visual grouping, enabling effective perceptual learning without requiring massive training sets.
- **Predictive Perceptual Embedding Framework**: We propose an end-to-end framework that maps segment-level visual features to a normalized embedding space, successfully predicting human grouping behavior.
- **Hierarchical Scene Extraction**: We present a data-driven pipeline that reconstructs multi-scale scene hierarchies via agglomerative clustering on learned embeddings, accompanied by extensive ablation and scaling studies characterizing how visual features and dataset volume drive perceptual organization.

2 Related Work

A longstanding goal in computer vision has been to move beyond flat image representations and capture the hierarchical organization of visual scenes. Early approaches addressed this problem through hierarchical segmentation, where images are recursively partitioned into regions based on low-level visual cues. Graph-based segmentation methods demonstrated that local relationships between image elements could be used to construct multiscale image representations (Felzenszwalb & Huttenlocher, 2004), while contour-based approaches further explored hierarchical organization through boundary and region information (Arbelaez et al., 2010). Beyond segmentation, compositional models proposed representing images as recursive combinations of simpler visual elements, using frameworks such as stochastic grammars and AND-OR graphs (Zhu & Mumford, 2007). However, these approaches typically rely on predefined grouping criteria or manually designed representations, limiting their ability to capture flexible visual organization.

Recent advances in deep learning have enabled the learning of structured scene representations directly from visual data. Scene graph approaches represent images as collections of objects connected by semantic relationships, providing an explicit description of scene structure (Johnson et al., 2015; Xu et al., 2017). Similarly, object-centric learning methods aim to decompose scenes into discrete object representations that can support downstream reasoning and generalization (Greff et al., 2019; Locatello et al., 2020). More broadly, large-scale representation learning has produced increasingly powerful visual features that capture object and category-level information (Radford et al., 2021).

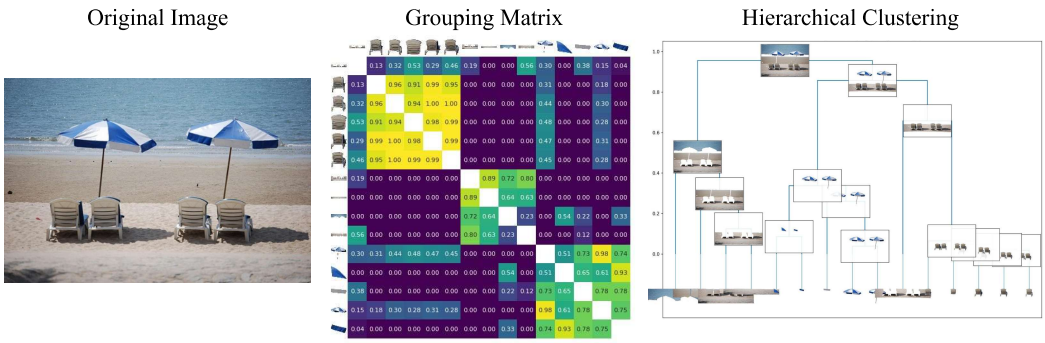


Figure 1: From left to right: Original image. Predicted grouping matrix expressing the perceptual grouping relationships between all pairs of segments in the image, computed with MLP-SPoSE's embeddings. Hierarchical grouping tree generated from the predicted grouping matrix. Leaf nodes show the individual image segments, while internal nodes show the progressively merged groups according to the hierarchical clustering. The y-axis represents the linkage distance, where lower merge heights indicate more strongly grouped segments and higher merge heights indicate weaker grouping relationships.

While these approaches provide rich semantic descriptions of visual scenes, they typically derive structure from learned object representations and category-level knowledge rather than from perceptual grouping relationships between visual elements.

Inspired by Gestalt theory, a complementary line of research has investigated how visual elements can be grouped into coherent perceptual units based on relationships such as similarity, proximity, connectedness, and contour continuity. Early computational models incorporated these principles through hand-crafted grouping cues and features (Feldman, 2024; Froyen et al., 2015), demonstrating that perceptual organization could emerge from local interactions between image elements (Palmer & Rock, 1994; Elder & Goldberg, 2002). While these approaches provide valuable insights into the mechanisms underlying perceptual organization, they typically focus on specific grouping cues or local grouping decisions, rather than learning perceptual grouping relationships that can be used to construct hierarchical scene representations.

Human similarity judgments have long been used to investigate the structure of perceptual representations and infer latent dimensions underlying visual perception (Shepard, 1987). Building on this tradition, recent work has employed odd-one-out and triplet-comparison tasks to learn embedding spaces directly from behavioral data, enabling perceptual relationships to be recovered without requiring explicit category labels (Hebart et al., 2020). Sparse Positive Similarity Embedding (SPoSE) provides a framework for learning interpretable and sparse representations from human judgments by identifying dimensions that explain perceptual similarities in the data (Hebart et al., 2020). Recent applications have extended this approach beyond object representations, using triplet judgments and SPoSE to uncover the core dimensions underlying human material perception (Schmidt et al., 2025). These studies demonstrate that human behavioral judgments can provide a powerful source of supervision for learning perceptual representations. However, existing applications have primarily focused on similarities between isolated stimuli, leaving largely unexplored how such representations can be used to learn grouping relationships between image regions and construct hierarchical organizations within complex scenes.

3 METHODS

This work presents a pipeline designed to map a natural input image to a tree graph representing its hierarchical organization. The framework operates in four sequential stages: (1) image segmentation, (2) segment-level feature extraction, (3) feature embedding generation via a trained model, and (4) structural representation of embedding relationships using hierarchical clustering. To align the model's representations with human visual perception, the architecture was trained on behavioral data collected from a large-scale online experiment.

To train and evaluate the proposed pipeline, we introduce and utilize the **Triple-Seg** dataset. This dataset comprises 210 natural images along with their corresponding segmentations and human grouping judgments over a dense set of segment triplets per image.

3.1 Image Segmentation

All input images were resized to a maximum dimension of 1024 pixels while maintaining aspect ratio. Image segmentation was performed using the Segment Anything Model (SAM3) Carion et al. (2026) under default hyperparameter settings without prompt engineering. Segments were sorted according to their predicted Intersection Over Union (IoU), and the top 20 masks per image were retained. To eliminate redundant regions, we removed duplicate masks, segments with greater than 95% area overlap, and segments occupying less than 0.5% of the total image area. This filtering step yielded a final dataset with an average of 14.52 segments per image (median: 15, standard deviation: 3.099, range: 6–20). Figures 3 and 4 visualize instances of segmentation from **Triple-Seg**.

3.2 Feature Extraction

To extract features relevant to perceptual grouping, six hand-crafted descriptors were computed for each segment using traditional computer vision algorithms. To enhance model explainability, algorithms were selected to yield mutually independent features (e.g., shape representations invariant to scale and orientation). The extracted features are defined as follows:

- **Size:** Represented as a 1-dimensional scalar computed by dividing the segment area by the total image area.
- **Position:** Encoded as a 3-dimensional vector comprising the segment's centroid coordinates (x, y) normalized to the image dimensions, along with the normalized Euclidean distance from the centroid to the image center.
- **Color:** Formulated as a 6-dimensional vector containing the mean and standard deviation for each channel in the CIELAB color space. The CIELAB space was chosen for its perceptual uniformity, where Euclidean distances correlate closely with human color perception (de l'Eclairage, 1978).
- **Shape:** Described using Zernike Moments to capture mask geometry (Teague, 1980). Moments were calculated with a radius of 21 pixels and a maximum degree of 4. Taking the absolute value of the complex moments yields a compact, rotation-invariant 9-dimensional descriptor that captures global shape characteristics while maintaining robustness to noise.
- **Texture:** Computed using Local Binary Patterns (LBP) (Ojala et al., 2002) with $P = 8$ sampling points, a radius of $R = 1$, and uniform encoding. This configuration captures local intensity transitions by thresholding the eight neighboring pixels, resulting in a 10-dimensional feature vector.
- **Skeleton:** Derived from the topological skeleton of the segment (Zhang & Suen, 1984) and represented as a 3D vector containing the skeleton length, branch ratio, and endpoint ratio.

Concatenating these individual descriptors yields a comprehensive 32-dimensional feature vector per segment. To prevent feature groups with higher dimensionalities from dominating distance calculations during similarity computation, each sub-vector was normalized and scaled by $1/\sqrt{d}$, where d represents the dimensionality of that specific feature set.

3.3 Data Collection

To collect human visual grouping judgments for **Triple-Seg**, we designed a web-based perceptual experiment. Participants were presented with a natural image alongside three sequentially highlighted segment regions blinking at 1s per segment. Tasks required selecting two regions deemed to belong together most strongly; participants were permitted to replay the highlight sequence as needed. To avoid biasing judgments, instructions explicitly avoided mentioning potential grouping criteria (e.g., color, size, proximity). Each participant evaluated 30 images with 5 segment triplets per image, completing a total of 150 trials. The task took an average of 26.5 minutes to complete. A total of 236 participants were recruited, of whom 213 (90.25%) completed all 150 trials.

A triplet is defined as an unordered set of three unique segments extracted from a single image. Across all 210 images in **Triple-Seg**, there are 98,940 unique candidate triplets. Due to data collection constraints, our target sampling rate was to evaluate at least 30% of all possible triplets. In total, we collected human judgments for 33,257 unique triplets (representing 33.61% of all possible triplets), achieving an average coverage of 38.49% per image. This sparse sampling is not merely a pragmatic constraint, it reflects the underlying psychological process of human visual learning, demonstrating that perceptual hierarchies can be effectively recovered without requiring explicit, exhaustive coding of all pairwise grouping relationships. The 210 images from the **Triple-Seg** dataset were split into training (80%), validation (10%), and test (10%).

3.4 Model Training

To learn a perceptual representation of image segments, we employ a neural network inspired by the SPoSE (Hebart et al., 2020) framework. Rather than learning an independent embedding for every segment, the network predicts an embedding directly from handcrafted visual features, allowing the model to generalize to previously unseen segments.

Each segment is represented by a feature vector x describing its color, shape, texture, size, orientation, and position. These features are passed through a backbone and projected into a 32-dimensional embedding space. The resulting embedding is L2-normalized, where $f(\cdot)$ denotes the neural network (Equation 1. Training is performed using human triplet judgments. Given three segments (a, b, c), the network computes embeddings $\mathbf{z}_a$, $\mathbf{z}_b$, and $\mathbf{z}_c$. Pairwise similarities are computed using cosine similarity, $\mathbf{z}_i^\top \mathbf{z}_j$, which is equivalent to the dot product because all embeddings are normalized to unit length. The three possible pairwise similarities $s(a, b)$, $s(a, c)$, $s(b, c)$, are interpreted as logits and passed directly to a softmax cross-entropy loss. The network is therefore trained to assign the highest similarity to the pair selected by human observers.

To evaluate the robustness of this learning framework, we experiment with several network backbones for predicting the segment embeddings while keeping the training objective unchanged. In addition to models based on handcrafted segment features, we evaluate two image-based variants, **CNN-SPoSE** and **ResNet-SPoSE**. Unlike the feature-based models, these architectures receive as input a resized version of the original image together with the three corresponding segment masks, allowing them to learn representations from scratch directly from the image data without relying on handcrafted descriptors. This comparison was included to assess the effectiveness of the proposed handcrafted feature set against representations learned end-to-end from raw visual input. The following are the backbones we tested:

- **Linear-SPoSE**: A linear baseline in which the handcrafted feature vector is projected directly into the embedding space using a single fully connected layer without bias.
- **MLP-SPoSE**: An extension of the linear model that uses a multilayer perceptron consisting of two fully connected hidden layers with batch normalization, ReLU activations, and dropout. This architecture enables the model to learn non-linear relationships between the handcrafted features before projecting them into the normalized embedding space.
- **Attention-SPoSE**: A transformer-inspired variant that first projects the handcrafted features into an embedding space and then applies a multi-head self-attention layer over the three segments in each triplet. A residual connection followed by layer normalization integrates the self-attention output with the original projected embeddings. This enables each segment to contextually incorporate relational information from the other two segments while preserving its intrinsic feature identity.
- **CNN-SPoSE**: An image-based variant that uses a lightweight convolutional neural network (CNN) (Krizhevsky et al., 2012) to learn segment embeddings directly from the RGB image and corresponding segment mask, concatenated as a four-channel input. The shared convolutional encoder consists of three convolutional blocks followed by global average pooling and a linear projection, producing normalized embeddings.
- **ResNet-SPoSE**: An image-based variant that replaces the custom convolutional encoder with a shared ResNet-18 backbone (He et al., 2016). The first convolutional layer is modified to a four-channel input consisting of the RGB image and the corresponding segment mask, while the final classification layer is replaced with an embedding projection head.

3.5 HIERARCHY EXTRACTION

To construct an image's underlying hierarchy, we first construct a pairwise **grouping matrix** $\mathbf{G} \in \mathbb{R}^{n \times n}$ for an image containing n segments, where entry G_{ij} quantifies the probability that segments i and j are perceptually grouped. For non-contextual backbones, $\mathbf{G}$ can be obtained by computing the cosine similarity between the embeddings of every segment pair, $\max(0, \mathbf{z}_i^\top \mathbf{z}_j)$, where negative values are truncated to zero. Alternatively, a grouping matrix can be empirically constructed via triplet-based model inference. For every candidate triplet (i, j, k), the model performs a forward pass to predict the most strongly grouped pair via an. We tally the frequency with which each pair (i, j) is selected across all evaluated triplets containing that pair, resulting in a co-selection count matrix (Equation 2). The final grouping matrix G is then obtained by min-max normalizing M by its maximum value ($G = M / \max(M)$). Notably, this inference-based aggregation is required for **Attention-SPOSE**, as its self-attention mechanism dynamically conditions a segment's embedding on the contextual presence of the other two segments in the input triplet.

$$\mathbf{z} = \frac{f(\mathbf{x})}{\|f(\mathbf{x})\|_2} \qquad (1)$$

$$M_{ij} = \sum_{k \neq i,j} \mathbf{1}\left[\operatorname{argmax} f(i, j, k) = (i, j)\right] \qquad (2)$$

Once the grouping matrix has been computed average-linkage hierarchical agglomerative clustering is then applied to iteratively merge the most similar segment groups, producing a hierarchical tree. To facilitate interpretation, each leaf node is visualized using the corresponding segmented object, while each internal node displays the union of all segments contained in that cluster. The resulting dendrogram represents the hierarchical organization of the scene, with merge height indicating the linkage distance between the merged groups. Figure 1 visualizes an instance of this process.

3.6 TRIPLE-SEG DATASET

The Triple-Seg dataset comprises 210 images curated from several benchmark datasets, including EVA (Kang et al., 2020), FCDB (Chen et al., 2017), COCO (Lin et al., 2014), Unsplash (Unsplash, n.d.), and MemCat (Goetschalckx & Wagemans, 2019). The collection captures a broad diversity of indoor and outdoor scenes, natural and manmade objects, and varying levels of visual complexity. In addition to raw images, the dataset provides segmented masks averaging 14.52 ± 3.10 segments per image (median $= 15$, range: 6–20). It also incorporates 33,257 human perceptual grouping judgments spanning 33.61% of all candidate triplets, yielding a mean image coverage of 38.49%. Additionally, 929 triplets were evaluated by multiple participants to assess inter-subject agreement.

4 RESULTS

4.1 BACKBONE

The 210 images from the **Triple-Seg** dataset were split into training (80%), validation (10%), and test (10%) sets to train SPoSE across various backbone architectures. All models were optimized for 100 epochs using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and cross-entropy loss. To establish a upper performance bound, human agreement was computed across multi-annotated triplets, yielding a benchmark ceiling of 0.7034. For comparative evaluation, two baseline models were included: a feature-space baseline based on Euclidean distances between segment feature vectors, and a random-choice baseline yielding an accuracy of $1/3$ across the three possible triplet pairings. Performance comparisons across backbones are visualized in Figure 2.

All trained models significantly outperform both the random baseline ($1/3$) and the raw feature baseline (0.4036), confirming that learning embedded representations yields stronger alignment with human perceptual grouping. The **MLP** model architecture achieved the highest accuracy at 0.6610, approaching the human agreement upper bound of 0.7034. CNN-based backbones also demonstrated strong predictive performance (0.6493 for standard CNN and 0.6423 for Pretrained-ResNet), while linear and attention-based mechanisms yielded lower accuracy scores (0.5853 and 0.6110, respectively). Based on these findings, the MLP backbone was selected as the core architecture for all subsequent experiments due to its superior performance.

SPoSE Backbone	Accuracy
Human Agreement	0.7034
Linear	0.5853
MLP	**0.6610**
Attention	0.6110
CNN	0.6493
ResNet	0.6173
Pretrained-ResNet	0.6423
Raw Features	0.4036
Random Guess	0.3333

Dataset Size	Accuracy
5000	0.5500
10000	0.5780
15000	0.6107
20000	0.6200
25000	0.6368
30000	0.6500
Full Dataset	**0.6610**

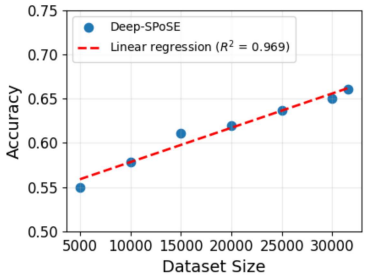


Figure 2: **Left:** Test accuracy across SPoSE backbones and baselines. **Middle:** Test accuracy for MLP-SPoSE across different size datasets. **Right:** Performance visualization of test accuracy (y-axis) against datset size (x-axis).

4.2 Qualitative Results

To qualitatively evaluate the pipeline's output, we present representative hierarchical decompositions generated for images within the held-out test set. Figure 3 illustrates the structural tree output for a complex natural landscape scene. The resulting dendrogram organizes the segment embeddings into three distinct structural clusters based on visual properties: foreground elements consisting of the dry grassy terrain and the standalone tree, distant landforms constituting the background mountains, and a higher-level grouping that merges the visual features of the open water and sky.

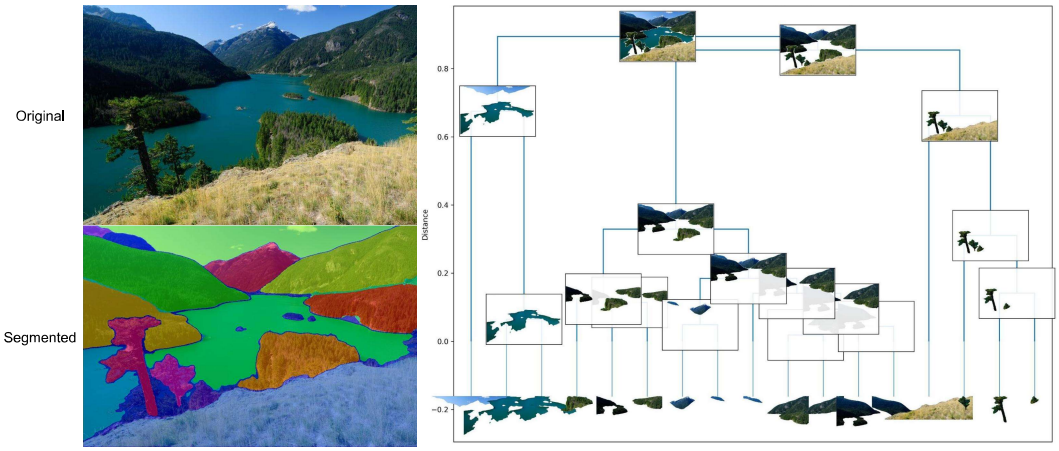


Figure 3: Hierarchical decomposition of a natural landscape (right). Original image on the top left, segmentation visualization on the bottom left.

Figure 4 shows the hierarchical decomposition of an indoor still-life scene featuring ceramic objects on a shelf. The resulting tree structure demonstrates a clear figure-ground segregation at the highest level of the hierarchy, cleanly partitioning the background shelf elements from the foreground objects. Furthermore, within the foreground cluster, the hierarchy exhibits secondary sub-groupings driven primarily by color, successfully separating the white ceramic vessels from the terracotta-colored pottery. Additional qualitative examples of hierarchical image decompositions across diverse scenes are provided in Appendix A.5.

4.3 Dataset Size

To evaluate the impact of dataset size on model performance, we trained the MLP-SPoSE architecture using subsets of 5,000 triplets across all 210 images, incrementally increasing the dataset size by 5,000 triplets up to the full dataset. The results, illustrated in Figure 2, demonstrate a strong linear



Figure 4: Hierarchical decomposition of a natural landscape (right). Original image on the top left, segmentation visualization on the bottom left.

relationship between sample size and predictive accuracy ($R^2 = 0.969$). Test accuracy improved from 0.5500 with 5,000 triplets to a peak of 0.6610 on the full dataset, indicating that model performance scales predictably with additional human grouping annotations and has not yet saturated.

4.4 Feature Ablation Study

To isolate the relative contribution of each visual property to perceptual grouping, we conducted a feature ablation study by training the MLP-SPoSE model independently on each individual feature set. **Color** and **Texture** proved to be the most informative single modalities for predicting human grouping judgments, both achieving test accuracies around 0.60. Geometric features exhibited moderate predictive utility, with **Shape** and **Position** each yielding accuracies around 0.50. Conversely, **Skeleton** and **Size** produced lower performances (around 0.43 and 0.41, respectively), though both still exceeded the random baseline (0.333). Table 5 shows the results for this experiment.

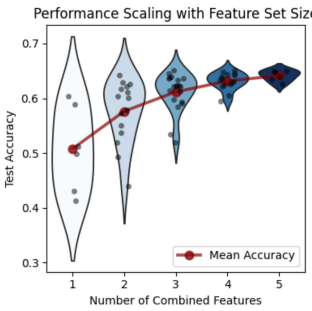
To further evaluate feature synergy, we conducted a combinatorial ablation study evaluating MLP-SPoSE across all $2^6 - 1 = 63$ possible feature subsets. As illustrated in Figure 5, test accuracy exhibits a clear upward trend as a function of feature set cardinality, with mean accuracy progressively rising as more descriptors are integrated. Increasing the number of combined features systematically reduces performance variance across combinations, stabilizing performance at higher dimensionalities. The complete 6-feature model achieves the highest overall test accuracy (0.6610), though high-cardinality subsets yield near-equivalent performance—such as the 5-feature combination excluding **Skeleton** (0.6502). Table 1 in the appendix shows the full ablation study.

4.5 Embedding Space Analysis

To evaluate whether the learned segment representations capture file-specific context or general structural pose features, we examined the distribution of segment embeddings of the test set in the lower-dimensional vector space. High-dimensional segment representations were projected into two-dimensional space using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton, 2008) ($perplexity = 15$) as our primary visualization tool to examine local neighborhood structures. To ensure these observations are not artifacts of t-SNE, a secondary projection using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) was performed as a robustness check (see Appendix A.2).

As illustrated in Figure 5, the t-SNE projection demonstrates significant inter-file mixing throughout the embedding space, where segments originating from the same file (indicated by distinct shape and color combinations) are broadly distributed across multiple distinct regions of the manifold. While small, tight micro-clusters of identical markers do appear, reflecting highly similar local segments within the same image (e.g., repeating objects) other segments from that very same file land far

Feature	# Dim	Accuracy
Size	1	0.407
Position	3	0.499
Skeleton	3	0.425
Shape	9	0.503
Texture	10	0.592
Color	6	0.601



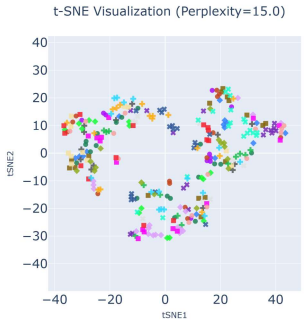


Figure 5: **Left:** Test accuracy of the MLP-SPoSE model trained on individual features, along with their respective dimensionalities. **Center:** Test accuracy across all 63 possible feature combinations evaluated with the MLP-SPoSE architecture. Violin plots illustrate the distribution of test accuracies for each feature set (1 to 6 features), with individual combinations plotted as data points. The red line tracks the mean test accuracy, demonstrating a consistent performance gain and variance reduction as additional feature modalities are integrated. **Right:** 2D t-SNE projection ($perplexity = 15$) of segment embeddings across test. Points are uniquely categorized by distinct markers and color hues corresponding to their source file ID. Rather than forming isolated clusters grouped by image origin, segments from the same image (matching symbols) are distributed across distinct regions.

apart in entirely different neighborhoods. This spatial distribution confirms that the embedding model successfully abstracts visual and structural pose features beyond simple image-level context.

5 Future Work

Building on our findings, several promising avenues exist for future research. First, given the strong linear scaling between dataset size and model performance ($R^2 = 0.969$), expanding the dataset may further bridge the gap toward the human agreement ceiling (0.7034). Second, capturing 100% of all possible triplet judgments for a selected subset of images would allow for a dense comparative analysis between the fully realized human ground-truth and our predicted grouping matrices. Third, extending this evaluation to artworks, ranging from representational art to non-figurative compositions, would test the generalizability of our feature representations across aesthetic styles. Abstract art in particular serves as an ideal domain for studying visual grouping, by removing explicit semantic references and object identities, abstract works isolate purely low- and mid-level visual attributes, offering a clean benchmark to evaluate the pipeline without influence from high-level semantic features.

6 Conclusion

In this work, we presented a data-driven framework and the Triple-Seg dataset to demonstrate that hierarchical scene structures can be recovered purely from human perceptual grouping, entirely independent of high-level semantic labels. While semantic hierarchies require massive, expensive datasets and standard benchmarks to evaluate category recognition, perceptual hierarchies rely on universal visual regularities. By training on a sparse set of human triplet comparisons from few hundred images, our model learns fundamental grouping principles that naturally generalize, unseen scenes. By learning a visual embedding space trained on human triplet comparisons, our model captures perceptual similarity across diverse visual elements and enables the reconstruction of multi-scale scene hierarchies via agglomerative clustering. Our empirical analyses show that model performance scales predictably with dataset size and multi-modal feature integration, offering a principled approach to bridge the gap between low-level visual features and visual scene organization. Ultimately, this semantics-free approach provides both a computational baseline for human perception and a flexible structural backbone for downstream tasks like image aesthetics, image understanding and layout prediction.

AI USE STATEMENT

In this work, we used generative AI tools for drafting parts of the research paper, identifying relevant literature, editing software code, and editing the research paper to improve readability. We have not used generative AI tools to: Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis and interpret results. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness, literature identified by LLMs was also reviewed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 226–234. IEEE, 2017.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Commission Internationale de l'Eclairage. Recommendations on uniform color spaces, color-difference equations, psychometric color therms. CIE, 1978.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

James H Elder and Richard M Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):5–5, 2002.

Jacob Feldman. What is a visual object? *Trends in Cognitive Sciences*, 7(6):252–256, 2003.

Jacob Feldman. Probabilistic origins of compositional mental representations. *Psychological Review*, 131(3):599, 2024.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

Vicky Froyen, Jacob Feldman, and Manish Singh. Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological review*, 122(4):575, 2015.

Lore Goetschalckx and Johan Wagemans. Memcat: a new category-based image set quantified on memorability. *PeerJ*, 7:e8169, 2019.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International conference on machine learning*, pp. 2424–2433. PMLR, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 35(3):413–452, 2023.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.

Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. Eva: An explainable visual aesthetics dataset. In *Joint workshop on aesthetic and technical quality assessment of multimedia and media analytics for societal trends*, pp. 5–13, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

Stephen Palmer and Irvin Rock. Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic bulletin & review*, 1(1):29–55, 1994.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Filipp Schmidt, Martin N Hebart, Alexandra C Schmid, and Roland W Fleming. Core dimensions of human material perception. *Proceedings of the National Academy of Sciences*, 122(10): e2417202122, 2025.

Roger N Shepard. Toward a universal law of generalization for psychological science. *Science*, 237 (4820):1317–1323, 1987.

Michael Reed Teague. Image analysis via the general theory of moments. *Journal of the optical society of America*, 70(8):920–930, 1980.

Unsplash. Unsplash dataset, n.d. URL https://unsplash.com/data. Accessed August 7, 2026.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Melissa Le-Hoa Vo. The meaning and structure of scenes. *Vision Research*, 181:10–20, 2021.

Melissa Le-Hoa Võ, Sage EP Boettcher, and Dejan Draschkow. Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current opinion in psychology*, 29:205–210, 2019.

Johan Wagemans. Perceptual organization. In John T. Wixted and John Serences (eds.), *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, volume 2, pp. 803–872. John Wiley & Sons, Inc., Hoboken, NJ, 2018. doi: 10.1002/9781119170174.epcn218.

Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

Max Wertheimer. Laws of organization in perceptual forms. 1938.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5419, 2017.

Tongjie Y Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and trends in computer graphics and vision*, 2(4):259–362, 2007.