Semantic View Induction for Object-Guided Scene Structuring

Opeyemi Bamigbade[1], Mark Scanlon[2], and John Sheppard[1]

[1] Forensics and Security Research Group, South East Technological University, Waterford, X91 K0EK, Ireland.
opeyemi.bamigbade@postgrad.setu.ie, John.Sheppard@setu.ie,
[2] Forensics and Security Research Group, School of Computer Science, University College Dublin, D04 V1W8, Ireland.
mark.scanlon@ucd.ie

Abstract. Scene datasets often exhibit substantial intra-class variability, where images belonging to the same semantic category contain multiple latent scene configurations characterised by different object compositions, spatial layouts, and functional contexts. This variability degrades the underlying semantic structure and increases representation ambiguity. This work introduces a two-stage scene structuring framework composed of Semantic View Induction (`SVI`) and `SceneStruct`. `SVI` discovers latent semantic views directly from image representations without predefined subclasses or manual annotations. `SceneStruct` then assigns images to the induced views using open-set object evidence while explicitly preserving ambiguous samples. Experimental evaluation on Hotel-ID and SUN397 demonstrates that the proposed framework induces three and two semantic views, respectively, while confidently structuring 94.08% and 99.38% of samples, leaving 5.92% and 0.61% as ambiguous. On Hotel-ID, restructuring increases retrieval P@1 from 38.60% to 48.28% and mAP@5 from 49.80% to 57.17%, while classification accuracy increases from 36.56% to 40.96%. On SUN397, classification accuracy increases from 63.41% to 69.74%. These results show that semantic view induction can uncover latent scene structure while preserving uncertain samples and is associated with improved downstream performance.

Keywords: Scene Understanding, Data-Centric Learning, Open-Vocabulary Object Detection, Semantic Induction

1 Introduction

Large-scale scene datasets are the foundation for many computer vision applications, including image retrieval, scene recognition, place recognition, robotic navigation, and visual reasoning [23, 3]. One application is multimedia geolocation for human trafficking investigations, where visual evidence is examined to derive location information [6]. Despite their widespread use, scene categories often exhibit substantial intra-class variability arising from differences in object

composition, spatial arrangement, viewpoint, illumination, and functional context [20]. Images assigned to the same category label may therefore correspond to multiple distinct semantic configurations despite sharing a common scene identity.

Conventional visual learning pipelines typically optimise representations using the existing category labels. A coarse category label does not describe the internal configurations of a scene. A model trained on such labels must therefore accommodate variation that is not explicitly represented within the dataset organisation.

Approaches to improving feature discrimination include changes to model architecture, metric learning objectives, and supervision strategies [18, 8, 21]. These approaches focus on the representation or learning objective, while visually distinct scene configurations can remain grouped under the same category during training.

This work investigates a dataset-level approach to representing recurring configurations within existing scene categories. The objective is to discover candidate semantic views and use object evidence to organise images around these views, with an explicit option for leaving uncertain assignments unresolved.

To this end, this work introduces *Semantic View Induction* (`SVI`), a framework that discovers views by clustering image representations within existing scene categories. Each induced view is intended to represent a recurring scene configuration. Building upon these views, `SceneStruct` uses open-set object evidence to assign images to views and retains uncertain samples in a separate ambiguity set. The resulting organisation preserves the original category labels for downstream learning and makes the induced views and assignment decisions available for inspection.

The evaluation examines both dataset properties and downstream task performance. Coverage and detection confidence describe the assignments and their object evidence, while intra-class variability and inter-view similarity describe the geometry of the CLIP representation space used for view induction. These geometric measurements do not independently establish semantic correctness.

Under the reported downstream model and optimisation settings, the Structured condition achieves higher retrieval and classification scores than the Original condition. The ambiguity comparison also shows higher classification accuracy for Structured than for Forced Assignment on both datasets. However, the conditions differ in ambiguity filtering and the resulting training and validation populations. The separate contributions of semantic assignment and sample selection therefore remain unresolved.

The main contributions of this work are summarised as follows:

- *Semantic View Induction* (`SVI`), a framework for discovering candidate semantic views within existing scene categories without predefined subclasses or additional per-image view annotations.
- *SceneStruct*, an object-guided assignment framework that organises images according to `SVI`-induced views and retains uncertain assignments in an explicit ambiguity set.

– A set of descriptive dataset metrics for quantifying view coverage, object detection confidence, within-view dispersion, and inter-view similarity.
– An evaluation on Hotel-ID and SUN397 reporting changes in CLIP representation geometry and higher downstream scores under the reported protocol.

2 Related Work

2.1 Scene Understanding and Visual Representation

Scene understanding plays a central role in computer vision applications such as image retrieval, scene recognition, place recognition, and visual reasoning. Large-scale datasets including SUN397 [23], Hotels-50K [20], Hotel-ID [19], and Google Landmarks [22] have supported the development and evaluation of visual representations across scene recognition, hotel recognition, and instance-level recognition and retrieval.

Scene categories may nevertheless exhibit substantial intra-class variability arising from differences in viewpoint, object composition, illumination, and functional context. Such variability has been considered in fine-grained visual recognition [4] and remains relevant to scene and place recognition, where images associated with the same identity or location can exhibit substantial visual differences [17]. This motivates representations and dataset organisations that can account for internal variation within coarse semantic categories.

2.2 Representation Learning under High Intra-Class Variability

Metric learning provides one direction for handling variability within visual representations. Triplet-based learning [18], angular-margin objectives such as ArcFace [8], and proxy-based methods such as ProxyNCA++ [21] learn embedding spaces that encourage samples with related identities or classes to form more discriminative representations.

For indoor geolocation, Bamigbade et al. [5] combined dominant colours and colour histograms with learned image embeddings, evaluating the fused representations through classification and metric learning. This approach adds explicit colour information to the learned representation.

Other approaches exploit local or structural visual information. RoomNet models room layout for indoor scene understanding [12], while LocalViT investigates the role of locality within vision transformer representations [14]. These approaches modify the representation, architecture, or learning objective while operating on an existing dataset organisation. This work considers a complementary data-centric direction by identifying latent semantic configurations within coarse scene categories and explicitly representing them as semantic views before downstream learning.

2.3 Object-Centric Scene Understanding and Open-Vocabulary Detection

Object-level information provides an interpretable source of evidence for scene understanding. Object ensembles have been explored for hotel recognition [7], while object and convolutional feature representations have also been used to improve indoor scene recognition [10]. These studies demonstrate the value of local and object-related information when global scene appearance alone may be insufficient.

Aftab et al. [1] used detected electrical sockets and their predicted types to constrain candidate countries for indoor images. Their method uses a predefined set of socket types to derive geographical constraints from object evidence.

Recent vision-language methods extend object-level reasoning beyond fixed detection vocabularies. GLIP [13], Grounding DINO [15], OWL-ViT [16], and ViLD [9] enable language-conditioned or open-vocabulary object detection. These methods are primarily designed for detection, grounding, and open-vocabulary recognition. In this work, open-vocabulary detection serves a different purpose: object evidence is used to associate images with latent semantic views discovered through representation-space analysis.

2.4 Dataset Organisation and Semantic Structuring

The organisation of visual collections has been considered in retrieval, indexing, and localisation. NetVLAD learns global image representations for place recognition [3], while data-centric approaches have been explored for image scene localisation [2]. Scene-graph representations provide another form of explicit structure by supporting image retrieval through learned similarities between scene graphs [24]. Earlier work has also investigated subdividing landmark image classes according to spatio-temporal and visual characteristics for large-scale indexing [11].

These approaches demonstrate different ways of exploiting structure within visual collections. The present work instead focuses on discovering latent semantic structure within existing scene categories. SVI induces semantic views without predefined subclasses, while SceneStruct uses object-level evidence to organise images according to the induced views and explicitly preserves ambiguous samples. This provides a data-centric mechanism for representing semantic variability within coarse scene categories.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ denote a scene dataset, where x_i is an image and y_i is a coarse scene label. Conventional scene datasets assume that all images sharing the same label belong to a single semantic category. In practice, however, many scene categories contain multiple latent semantic configurations arising from differences in object composition, spatial layout, viewpoint, and functional context.

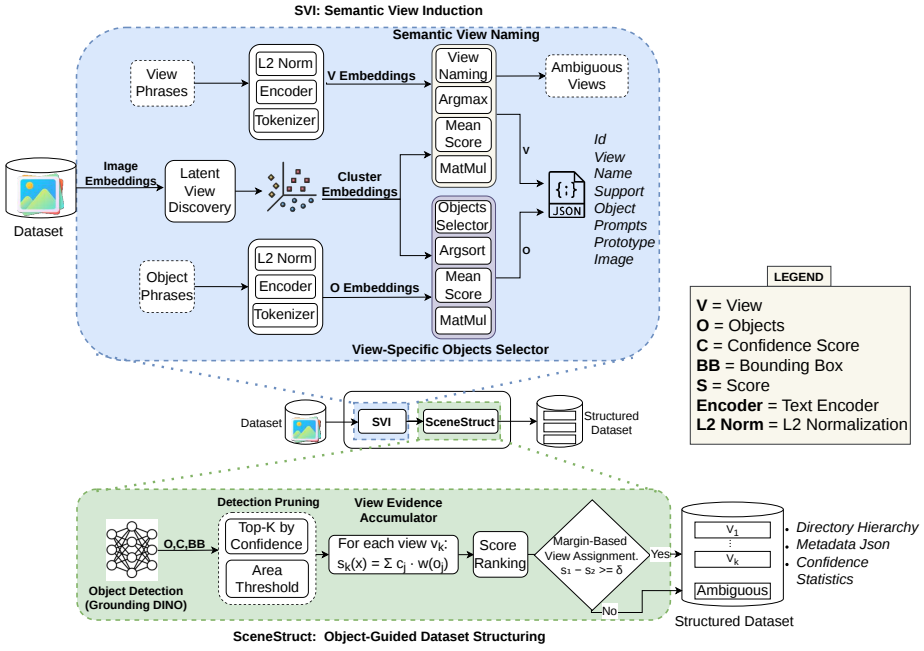


Fig. 1. Overview of the proposed two-stage dataset structuring pipeline. **Center:** High-level pipeline transforming a scene dataset into a structured representation via SVI followed by SceneStruct. **Top:** Internal components of SVI for unsupervised semantic view discovery, naming, and prototype selection. **Bottom:** Internal decision process of SceneStruct.

This work models such variability through a set of latent semantic views

$$\mathcal{V} = v_1, v_2, \ldots, v_K, \tag{1}$$

where each view represents a recurring semantic configuration within a scene category. In addition, some images may not exhibit sufficient evidence to support a reliable assignment to any dominant semantic view. These samples are represented through an ambiguity set

$$\mathcal{A} \subseteq \mathcal{X}. \tag{2}$$

The objective is therefore to recover a mapping

$$g : \mathcal{X} \to \mathcal{V} \cup \{\mathcal{A}\}, \tag{3}$$

where images are assigned either to a semantic view or to an explicit ambiguity state $\mathcal{A}$.

To achieve this objective, this work introduces a two-stage framework consisting of `SVI` and `SceneStruct`. `SVI` discovers latent semantic views directly

Algorithm 1: SVI + SceneStruct

Input: Scene dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, image encoder $f(\cdot)$, open-set object detector
Output: Structured dataset $\mathcal{S}$
Stage 1: Semantic View Induction (SVI)
Extract image embeddings: $e_i = f(x_i)$.
Cluster embeddings to obtain candidate views.
Retain clusters satisfying support threshold τ as semantic views $\mathcal{V}$.
Assign remaining samples to ambiguity set $\mathcal{A}$.
Compute view centroids μ_m and semantic descriptors $\hat{t}_m$.
Stage 2: SceneStruct
Construct view-specific object descriptors $\phi(v)$ for $v \in \mathcal{V}$.
foreach *image* $x_i \in \mathcal{D}$ **do**
  Extract object detections $O(x_i)$.
  foreach *view* $v \in \mathcal{V}$ **do**
    Compute compatibility score $s(v|x_i)$.
  end
  $v^* \leftarrow \arg\max_{v \in \mathcal{V}} s(v|x_i)$
  if $s(v^*|x_i) - s(v'|x_i) \geq \delta$ **then**
    Assign x_i to v^*
  end
  else
    Assign x_i to $\mathcal{A}$
  end
end
Construct structured dataset $\mathcal{S}$ from view assignments.
return $\mathcal{S}$

from image representations, while `SceneStruct` leverages object-level evidence to organise images according to the discovered semantic structure. An overview of the proposed framework is illustrated in Figure 1, and the overall two-stage restructuring pipeline is summarised in Algorithm 1.

3.2 Semantic View Induction

The objective of `SVI` is to discover recurring semantic configurations within a scene category directly from image representations. Given an image x_i, a pre-trained vision-language model is used to obtain a semantic embedding

$$e_i = f(x_i), \tag{4}$$

where $e_i \in \mathbb{R}^d$ denotes the embedding associated with image x_i. All embeddings are ℓ_2 normalised prior to view discovery.

Semantic views are identified through density-based clustering within the embedding space. Let

$$\mathcal{C} = \{c_1, c_2, \ldots, c_M\} \tag{5}$$

denote the set of discovered clusters. Clusters satisfying a minimum support criterion are retained as semantic views,

$$\mathcal{V} = \{c_m \in \mathcal{C} : |c_m| \geq \tau\}, \tag{6}$$

where τ is the minimum cluster size threshold.

Not all images participate in stable semantic structures. Scene datasets frequently contain mixed-content samples, transitional configurations, and atypical viewpoints that exhibit weak affinity to any dominant semantic view. To account for this phenomenon, `SVI` introduces an explicit ambiguity set

$$\mathcal{A} = \{x_i \in \mathcal{X} : x_i \notin v, \forall v \in \mathcal{V}\}. \tag{7}$$

Unlike conventional clustering approaches that enforce a complete partition of the dataset, the ambiguity set preserves uncertain samples and prevents structurally inconsistent assignments.

For each retained semantic view v_m, a centroid representation is computed as

$$\mu_m = \frac{1}{|v_m|} \sum_{e_i \in v_m} e_i. \tag{8}$$

To facilitate interpretation, each semantic view is assigned a semantic descriptor from a predefined vocabulary of scene concepts. The descriptor associated with view v_m is obtained through nearest-neighbour matching in the shared embedding space:

$$\hat{t}_m = \arg\max_{t_j \in \mathcal{T}} \cos(\mu_m, t_j), \tag{9}$$

where $\mathcal{T}$ denotes the descriptor vocabulary and $\cos(\cdot, \cdot)$ represents cosine similarity.

The output of `SVI` therefore consists of a set of semantic views $\mathcal{V}$, an ambiguity set $\mathcal{A}$, and a semantic descriptor associated with each induced view. These components provide the semantic structure used by `SceneStruct` for object-guided dataset organisation.

3.3 SceneStruct

While `SVI` discovers latent semantic views, it does not directly reorganise the dataset. `SceneStruct` operationalises the induced semantic structure through an object-guided assignment process that associates images with semantically coherent views while preserving ambiguity.

Each image x is processed using an open-set object detector to obtain a set of detected objects and associated confidence scores:

$$O(x) = \{(o_j, c_j)\}_{j=1}^{M}, \tag{10}$$

where o_j denotes the detected object label and $c_j \in [0, 1]$ denotes the corresponding confidence score.

For each semantic view $v \in \mathcal{V}$, a view-specific object descriptor is constructed from objects that occur consistently among images belonging to that view:

$$\phi(v) = \{o_1, o_2, \ldots, o_L\}. \tag{11}$$

These descriptors provide an interpretable semantic representation of the dominant object composition associated with each induced view.

The detected object set is filtered using the confidence criterion described in Section 4, with low-confidence object evidence removed prior to compatibility scoring.

Given an image x, a compatibility score is computed between the detected object evidence and each semantic view:

$$s(v|x) = \sum_{(o_j, c_j) \in O(x)} w(o_j) c_j \cdot \mathbb{I}\left(o_j \in \phi(v)\right), \tag{12}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and $w(o_j)$ denotes a predefined object-specific importance weight used to control the contribution of object o_j to the compatibility score. Distinctive objects such as beds and toilets receive higher weights (1.3), while less discriminative objects such as chairs (0.2) and tables (0.3) receive lower weights. Objects without an explicitly specified weight are assigned a default weight of 0.5.

The image is assigned to the view with the highest compatibility score.

$$v^* = \arg\max_{v \in \mathcal{V}} s(v|x), \tag{13}$$

provided sufficient evidence exists to distinguish the selected view from competing alternatives. Let v' denote the second-highest scoring view. Assignment is accepted only when

$$s(v^*|x) - s(v'|x) \geq \delta, \tag{14}$$

where δ is a confidence margin controlling assignment reliability.

Images that fail to satisfy the margin criterion are assigned to the ambiguity set rather than being forcibly incorporated into a semantic view:

$$x \in \mathcal{A} \quad \text{if} \quad s(v^*|x) - s(v'|x) < \delta. \tag{15}$$

This ambiguity-aware assignment strategy prevents semantically inconsistent samples from degrading view coherence and preserves uncertainty as an explicit component of the dataset structure.

The output of `SceneStruct` is a structured dataset representation consisting of semantically coherent views and an associated ambiguity subset. Unlike conventional dataset organisation approaches based solely on global visual similarity, `SceneStruct` combines latent semantic structure with object-level evidence to produce interpretable and ambiguity-aware scene organisation.

3.4 Dataset Structuring Quality Metrics

To evaluate the quality of the induced semantic structure independently of any downstream task, four dataset-centric metrics are introduced. These metrics quantify view coverage, object evidence reliability, within-view compactness, and separation between induced semantic views.

View Coverage Accuracy (VCA). VCA measures the proportion of images confidently assigned to one of the induced semantic views:

$$\mathrm{VCA} = \left(\frac{N_{\mathrm{covered}}}{N_{\mathrm{total}}} \right) \times 100\%, \tag{16}$$

where N_{covered} denotes the number of images assigned to any semantic view and N_{total} denotes the total number of images.

Detection Reliability (DR). DR measures the reliability of object evidence used during object-guided assignment. For a semantic view v, DR is computed as

$$\mathrm{DR}_v = \frac{1}{N_v} \sum_{i=1}^{N_v} \left(\frac{1}{M_i} \sum_{j=1}^{M_i} c_{ij} \right), \tag{17}$$

where N_v is the number of images assigned to view v, M_i is the number of detected objects in image i, and c_{ij} denotes the confidence score of the j-th detected object.

Intra-Class Variability (ICV). ICV measures the compactness of image representations within each semantic view. Lower ICV indicates that images assigned to the same view are more tightly concentrated around their corresponding view centroid. For a view v containing n_v images, the view centroid is defined as

$$\mu_v = \frac{1}{n_v} \sum_{i=1}^{n_v} f(x_i). \tag{18}$$

The ICV of view v is then computed as

$$\mathrm{ICV}(v) = \frac{1}{n_v} \sum_{i=1}^{n_v} \left\| f(x_i) - \mu_v \right\|^2. \tag{19}$$

The dataset-level ICV is obtained by averaging $\mathrm{ICV}(v)$ across all induced semantic views.

Inter-View Similarity (IVS). IVS measures representational overlap between different semantic views. For the original dataset, the original scene categories are treated as coarse views for comparison. Following restructuring, images assigned to the same induced semantic view are pooled across the dataset to form a view-level representation. For two distinct views v and v', the cosine similarity between their centroids is defined as

$$\mathrm{IVS}(v, v') = \frac{\mu_v \cdot \mu_{v'}}{\|\mu_v\| \, \|\mu_{v'}\|}. \tag{20}$$

The average IVS across all pairs of views is computed as

$$\text{Avg. IVS} = \frac{2}{K(K-1)} \sum_{v=1}^{K} \sum_{v'=v+1}^{K} \mathrm{IVS}(v, v'), \tag{21}$$

where K denotes the number of views under the corresponding dataset organisation.

Higher VCA and DR indicate improved coverage and more reliable object evidence, while lower ICV and Avg. IVS indicate improved compactness and reduced overlap between induced semantic views.

ICV and IVS are computed in the same CLIP representation space used for semantic view induction. They are therefore interpreted as descriptive measures of the representational geometry produced by restructuring rather than as independent evidence of semantic validity. Semantic coherence is instead considered jointly through object-level evidence, qualitative analysis, and downstream task performance.

4 Experimental Setup

4.1 Datasets

The proposed framework is evaluated on Hotel-ID [19] and SUN397 [23]. These datasets were selected because they exhibit substantial intra-class variability arising from differences in viewpoint, object composition, scene layout, and functional context.

Hotel-ID is a large-scale hotel recognition dataset containing images collected from diverse hotel environments. Images belonging to the same hotel frequently correspond to distinct semantic configurations such as bedrooms, bathrooms, corridors, and lobbies. SUN397 provides a complementary evaluation setting consisting of diverse indoor and outdoor scene categories and is used to assess the generality of the proposed framework beyond hotel-centric environments.

No additional annotations, manually defined subclasses, object labels, or semantic view labels are introduced during semantic view induction or scene structuring.

4.2 Implementation Details

Image embeddings are extracted using a pretrained CLIP ViT-B/16 model and projected into a shared semantic embedding space. All embeddings are ℓ_2 normalised prior to semantic view induction.

Semantic views are discovered independently within each scene category using HDBSCAN with minimum cluster size $\tau = 10$. The semantic descriptor

vocabulary $\mathcal{T}$ comprises predefined functional scene phrases tailored to each evaluation domain, including sleeping, bathroom, corridor, living, workspace, classroom, library, and restaurant concepts. The object vocabulary used by `SceneStruct` is constructed from the object prompts associated with the semantic views generated by `SVI`. These prompts are combined to form the open-vocabulary detection queries used for object-guided assignment.

Clusters satisfying the support criterion are retained as semantic views and assigned semantic descriptors from the predefined vocabulary of scene concepts in the shared CLIP embedding space, while remaining samples are assigned to the ambiguity set.

Open-set object evidence is extracted using the Grounding DINO-Tiny model, with box and text confidence thresholds of 0.35 and 0.25, respectively. The eight highest-scoring detections are retained per image. Detections occupying less than 2% of the image area are excluded before compatibility scoring. For each semantic view, a view-specific object descriptor is constructed from the five highest-ranked object concepts associated with the corresponding cluster.

Downstream retrieval experiments use a frozen CLIP ViT-B/16 backbone, followed by a trainable projection MLP and ℓ_2 normalisation to obtain the retrieval representation. The model architecture, optimisation procedure, learning schedule, and identity supervision remain consistent across experimental conditions.

For Hotel-ID retrieval, one image per eligible hotel identity is selected from the original held-out test set to form the query population. These queries remain unchanged across experimental conditions and are excluded from training and semantic restructuring. The eligible images outside the held-out query population form the retrieval gallery, against which the queries are evaluated. The training populations reflect the corresponding dataset organisation.

Classification experiments use a frozen CLIP ViT-B/16 backbone followed by a trainable classification MLP and linear classification head. The original class labels remain the prediction targets, and the induced semantic views are not introduced as additional class labels. Across the experimental conditions, the downstream model configuration and optimisation settings are held fixed, while the training and validation populations may differ according to the dataset organisation. The experimental intervention therefore consists of semantic organisation and ambiguity-based sample selection at the dataset level rather than changes to the downstream model. Figure 2 summarises the experimental conditions and common downstream processing pipeline.

The implementation configuration used for `SVI` and `SceneStruct` is summarised in Table 1.

4.3 Evaluation Protocol

Evaluation is conducted from both a dataset-centric and task-centric perspective. For task-centric evaluation, the original dataset is initially partitioned into training, validation, and test sets. The test set is retained unchanged across all experimental conditions. For the Original condition, the initial training and

Table 1. Implementation configuration used for SVI and SceneStruct.

Component	Configuration
Image encoder	CLIP ViT-B/16
Clustering	HDBSCAN
Minimum cluster size (τ)	10
Object detector	Grounding DINO-Tiny
Box confidence threshold	0.35
Text confidence threshold	0.25
Top-K detections per image	8
Objects per view descriptor	5
Assignment margin (δ)	0.15

validation partitions are used without semantic restructuring. For the remaining conditions, the initial training and validation partitions are combined before applying `SVI` and `SceneStruct`. The resulting datasets are subsequently partitioned into training and validation sets for downstream learning.

Three experimental conditions are considered. The *Original* condition preserves the existing dataset organisation. The *Forced Assignment* condition applies semantic structuring but assigns every sample to its highest-scoring semantic view without applying the ambiguity margin. The *Structured* condition retains confident semantic-view assignments while preserving uncertain samples in the ambiguity set. Ambiguous samples are excluded before constructing the downstream training and validation partitions. All conditions use the same downstream architecture and are evaluated on the original held-out test set, as illustrated in Figure 2.

For dataset-centric evaluation, restructuring quality is assessed using four metrics: View Coverage Accuracy (VCA), Detection Reliability (DR), Intra-Class Variability (ICV), and Inter-View Similarity (IVS). VCA measures the proportion of images assigned to stable semantic views, DR evaluates the consistency of object evidence, ICV measures within-view compactness, and IVS quantifies separation between induced semantic views.

For task-centric evaluation, the resulting structured datasets are assessed on retrieval and classification tasks. Retrieval performance is measured using Precision@1 (P@1) and mean Average Precision at rank 5 (mAP@5). Classification performance is evaluated using top-1 accuracy. In all experiments, the downstream model and optimisation settings are held fixed to evaluate the dataset-level intervention.

5 Results

5.1 Dataset Structuring Evaluation

The primary objective of the proposed framework is to uncover latent semantic structure within scene datasets and reorganise images into semantically coherent

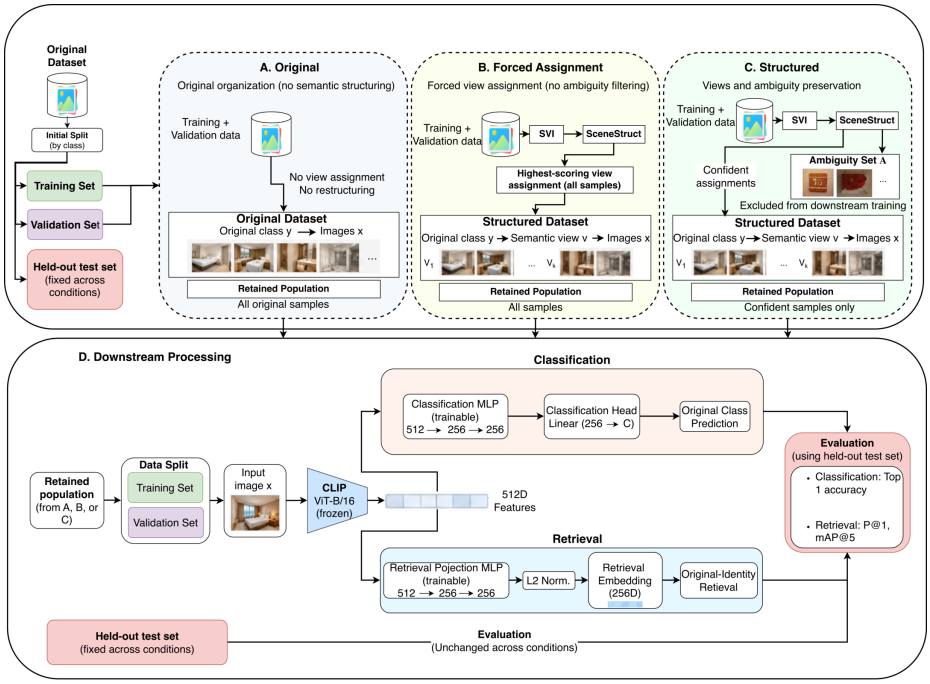


Fig. 2. Experimental conditions and downstream processing pipeline. Original, Forced Assignment, and Structured populations are evaluated using the same classification and retrieval architectures.

views. Evaluation, therefore, begins by analysing the structural properties of the resulting datasets before considering downstream task performance.

Table 2 summarises the dataset structuring characteristics obtained after applying `SVI` and `SceneStruct`. The table reports the number of retained semantic views, view coverage, assignment reliability, ambiguity statistics, and the resulting changes in dataset geometry.

For Hotel-ID, `SVI` discovers an average of three semantic views per category while assigning 94.08% of images to stable views. The resulting dataset exhibits substantially lower intra-class variability and reduced inter-view similarity compared with the original organisation.

For SUN397, `SVI` discovers an average of two semantic views per category and achieves 99.38% coverage. Despite the greater semantic diversity of the dataset, restructuring continues to improve separation between semantic views while maintaining high assignment reliability.

Compared with the original dataset organisation, restructuring reduces ICV from 0.0730 to 0.0390 and Avg. IVS from 0.9409 to 0.8779 on Hotel-ID. Similar behaviour is observed on SUN397, where Avg. IVS decreases from 0.9344 to 0.9195 following semantic view induction and object-guided restructuring.

Table 2. Dataset structuring evaluation. K denotes the number of retained semantic views per category (macro-averaged for SUN397). VCA measures view coverage, DR measures object evidence reliability, and Unassigned denotes the fraction of images assigned to the ambiguity set. Lower ICV and Avg. IVS indicate improved compactness and separation between semantic views.

Dataset	Representation	Views (K)	VCA (%)	Avg. DR	Unassigned (%)	ICV ↓	Avg. IVS ↓
Hotel-ID	Original Dataset	–	–	–	–	0.0730	0.9409
	Structured Dataset	3	94.08	0.4229	5.92	**0.0390**	**0.8779**
SUN397	Original Dataset	–	–	–	–	0.0460	0.9344
	Structured Dataset	2	99.38	0.4935	0.61	**0.0454**	**0.9195**

Table 3. Downstream task evaluation. Retrieval experiments are conducted on Hotel-ID using the same model architecture and optimisation settings across experimental conditions. Classification experiments use a frozen CLIP ViT-B/16 backbone. Improvements are obtained without modifying model architectures, loss functions, optimisation procedures, or supervision strategies.

Task	Dataset	Original	Structured	Improvement
Retrieval (P@1)	Hotel-ID	38.60	**48.28**	+9.68
Retrieval (mAP@5)	Hotel-ID	49.80	**57.17**	+7.37
Classification Accuracy	Hotel-ID	36.56	**40.96**	+4.40
Classification Accuracy	SUN397	63.41	**69.74**	+6.33

Across both datasets, the proposed framework produces a more compact and separated organisation in the CLIP representation space, as reflected by the reductions in ICV and Avg. IVS. Since these metrics are measured in the same representation space used for semantic view induction, they characterise the geometry produced by restructuring rather than providing independent evidence of semantic validity. Their interpretation is therefore considered alongside the object-level evidence, qualitative analysis, and downstream evaluation reported in the following sections.

5.2 Downstream Task Evaluation

The central hypothesis of this work is that improved dataset organisation can improve downstream visual learning. To evaluate this hypothesis, the restructured datasets are assessed on both retrieval and classification tasks under matched downstream model and optimisation settings.

Table 3 summarises the downstream evaluation results. Across all tasks and datasets, models trained on the restructured datasets consistently outperform those trained on the original dataset organisation.

For retrieval, restructuring improves Precision@1 from 38.60% to 48.28% and increases mAP@5 from 49.80% to 57.17%. These gains are consistent with the improved dataset organisation produced by `SVI` and `SceneStruct`.

Table 4. Effect of ambiguity modelling on classification performance. The Forced Assignment condition assigns all samples to their highest-scoring semantic views without applying the ambiguity margin.

Dataset	Structured	Forced Assignment
Hotel-ID	**40.96**	38.22
SUN397	**69.74**	66.51

A similar trend is observed for classification. On Hotel-ID, accuracy improves from 36.56% to 40.96%, while SUN397 achieves an increase from 63.41% to 69.74%. The consistency of these improvements across both fine-grained hotel recognition and general scene recognition suggests that the benefits of semantic view induction extend beyond a single application domain.

Importantly, all experiments use identical model architectures, training schedules, and optimisation procedures, while the training populations reflect the corresponding dataset organisation. Under matched downstream model and optimisation settings and a common held-out test set, the observed gains are associated with the dataset-level intervention introduced by `SVI` and `SceneStruct`, which comprises semantic organisation and ambiguity-based sample selection.

Taken together, these results indicate that dataset geometry constitutes an important factor in visual learning. The proposed framework demonstrates that meaningful performance improvements can be obtained through semantic dataset restructuring, providing a complementary direction to conventional approaches that focus primarily on model design and optimisation.

5.3 Ambiguity Ablation

A central component of the proposed framework is the explicit modelling of ambiguity through the ambiguity set introduced during semantic view induction and preserved during object-guided restructuring. To evaluate the contribution of this mechanism, the Forced Assignment condition removes the ambiguity option and assigns all samples to their highest-scoring semantic views, regardless of the assignment margin.

As reported in Table 4, the Structured condition achieves classification accuracies of 40.96% on Hotel-ID and 69.74% on SUN397, exceeding Forced Assignment by 2.74 and 3.23 percentage points, respectively. Both conditions use the same held-out test set, but differ in ambiguity filtering and the resulting training and validation populations. The comparison therefore does not isolate the contribution of semantic assignment from sample selection.

5.4 Qualitative Analysis

Figure 3 presents selected images assigned to the induced views and to the ambiguity set. The assigned examples show recurring visual and object-level patterns despite changes in appearance and viewpoint. The object detections provide a

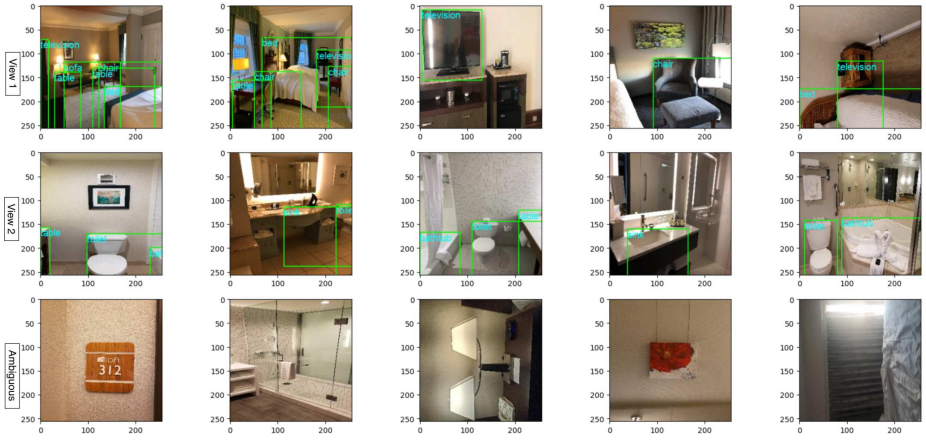


Fig. 3. Selected examples of assignments to two induced views (View 1 and View 2) and to the ambiguity set (Ambiguous). Detected objects are shown where available.

means of inspecting the evidence available to `SceneStruct` when associating images with the views discovered by `SVI`.

The ambiguity row illustrates images for which the pipeline did not produce an accepted view assignment. Such outcomes can arise from limited detection evidence or competing object matches. The ambiguity designation records uncertainty under the configured assignment rule and does not imply that a human observer would find the scene semantically ambiguous.

These examples illustrate the resulting groupings and assignment decisions, but do not independently establish semantic correctness or explain the observed downstream gains. As discussed in Section 5.3, the performance comparison involves differences in sample selection and the resulting training and validation populations, leaving the individual contributions of semantic assignment and sample selection unresolved.

6 Discussion

Across Hotel-ID and SUN397, the structured datasets have lower ICV and Avg. IVS than the original organisation in the CLIP representation space. These measurements describe reduced within-view dispersion and lower similarity between view centroids. Since the same representation space is used to induce the views, these changes do not independently establish their semantic correctness.

The Structured condition also achieves higher retrieval and classification scores than the Original condition under the reported downstream model and optimisation settings. These gains are associated with a dataset-level intervention that combines semantic organisation with ambiguity-based sample selection.

As reported in Section 5.3, the Structured condition achieves higher classification accuracy than Forced Assignment on both datasets. Both conditions use

the same held-out test set, but differ in ambiguity filtering and the resulting training and validation populations. The comparison therefore leaves the individual contributions of semantic assignment and sample selection unresolved.

The proposed approach also has several limitations. `SVI` depends on the quality of the pretrained representation space and predefined semantic vocabulary, while SceneStruct relies on the reliability of open-vocabulary object detections and the available object concepts. Consequently, visually similar scene configurations, incomplete object evidence, or insufficient vocabulary coverage may produce uncertain assignments, particularly for categories with weakly distinguishable semantic structure.

The two-stage restructuring process also introduces additional preprocessing overhead due to representation extraction, semantic view induction, and open-vocabulary object detection. This cost is incurred during dataset restructuring rather than downstream model inference.

The induced views also represent a dataset-dependent organisation rather than a fixed semantic taxonomy, and their number and interpretation may vary across datasets. These limitations motivate future investigation of alternative representation spaces, adaptive view discovery, and broader evaluation across scene domains.

7 Conclusion

This work introduced Semantic View Induction (`SVI`), an unsupervised framework for discovering latent semantic views within scene datasets, and `SceneStruct`, an object-guided scene structuring framework that organises images according to the discovered semantic structure while explicitly preserving ambiguity.

The framework operates at the dataset level without modifying downstream learning architectures. On Hotel-ID and SUN397, the structured datasets exhibit lower within-view dispersion and inter-view similarity in the CLIP representation space used for view induction. These geometric changes do not independently establish semantic correctness.

Under the reported evaluation protocol, the Structured condition achieves higher downstream scores than the Original condition and higher classification accuracy than Forced Assignment on both datasets. These findings support further investigation of semantic organisation and ambiguity-based sample selection as a combined dataset-level intervention. Their individual contributions remain unresolved because the conditions differ in sample selection and the resulting training and validation populations.

Future work will investigate hierarchical semantic view induction, adaptive ambiguity modelling, and the application of semantic dataset structuring to larger-scale vision and multimodal datasets.

References

1. Aftab, K., Adams, G., Scanlon, M.: Plug to Place: Indoor multimedia geolocation from electrical sockets for digital investigation. Forensic Science International: Digital Investigation **56**, 302056 (2026). https://doi.org/10.1016/j.fsidi.2026.302056
2. Alfarrarjeh, A., Kim, S.H., Rajan, S., Deshmukh, A., Shahabi, C.: A data-centric approach for image scene localization. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 594–603 (2018). https://doi.org/10.1109/BigData.2018.8621912
3. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1437–1451 (2018). https://doi.org/10.1109/TPAMI.2017.2711011
4. Bai, Y., Gao, F., Lou, Y., Wang, S., Huang, T., Duan, L.Y.: Incorporating intra-class variance to fine-grained visual recognition. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 1452–1457 (2017). https://doi.org/10.1109/ICME.2017.8019371
5. Bamigbade, O., Scanlon, M., Sheppard, J.: Improving image embeddings with colour features in indoor scene geolocation. IEEE Access **13**, 79860–79870 (2025). https://doi.org/10.1109/ACCESS.2025.3564496
6. Bamigbade, O., Sheppard, J., Scanlon, M.: Computer vision for multimedia geolocation in human trafficking investigation: A systematic literature review (2024), https://arxiv.org/abs/2402.15448
7. Bhavanasi, S.S., Stylianou, A.: Hotel recognition using object ensembles. In: 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). pp. 1–8 (2023). https://doi.org/10.1109/AIPR60534.2023.10440661
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694 (2019). https://doi.org/10.1109/CVPR.2019.00482
9. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=lL3lnMbR4WU
10. Khan, S.H., Hayat, M., Bennamoun, M., Togneri, R., Sohel, F.A.: A discriminative representation of convolutional features for indoor scene recognition. IEEE Transactions on Image Processing **25**(7), 3372–3383 (Jul 2016). https://doi.org/10.1109/tip.2016.2567076
11. Kim, D., Rho, S., Jun, S., Hwang, E.: Classification and indexing scheme of large-scale image repository for spatio-temporal landmark recognition. Integrated Computer-Aided Engineering **22**(2), 201–213 (2015). https://doi.org/10.3233/ICA-140478
12. Lee, C.Y., Badrinarayanan, V., Malisiewicz, T., Rabinovich, A.: RoomNet: End-to-end room layout estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 4875–4884 (2017). https://doi.org/10.1109/ICCV.2017.521
13. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded Language-Image Pre-training. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10955–10965 (2022). https://doi.org/10.1109/CVPR52688.2022.01069

14. Li, Y., Zhang, K., Cao, J., Timofte, R., Magno, M., Benini, L., Van Gool, L.: LocalViT: Analyzing locality in vision transformers. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9598–9605 (2023). https://doi.org/10.1109/IROS55552.2023.10342025
15. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. Lecture Notes in Computer Science, vol. 15105, pp. 38–55. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72970-6_3
16. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. Lecture Notes in Computer Science, vol. 13670, pp. 728–755. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20080-9_42
17. Müller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII. Lecture Notes in Computer Science, vol. 11216, pp. 575–592. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_35
18. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682
19. Stylianou, A., Dane, S.: Hotel-ID to Combat Human Trafficking 2022 - FGVC9. https://kaggle.com/competitions/hotel-id-to-combat-human-trafficking-2022-fgvc9 (2022), Kaggle
20. Stylianou, A., Xuan, H., Shende, M., Brandt, J., Souvenir, R., Pless, R.: Hotels-50K: A Global Hotel Recognition Dataset. Proceedings of the AAAI Conference on Artificial Intelligence **33**(1), 726–733 (07 2019). https://doi.org/10.1609/aaai.v33i01.3301726
21. Teh, E.W., DeVries, T., Taylor, G.W.: ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol. 12369, pp. 448–464. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_27
22. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2572–2581 (2020). https://doi.org/10.1109/CVPR42600.2020.00265
23. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3485–3492 (June 2010). https://doi.org/10.1109/CVPR.2010.5539970
24. Yoon, S., Kang, W.Y., Jeon, S., Lee, S., Han, C., Park, J., Kim, E.S.: Image-to-image retrieval by learning similarity between scene graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10718–10726 (2021). https://doi.org/10.1609/aaai.v35i12.17281