Towers of Babel: Combining Images, Language, and 3D Geometry for Learning Multimodal Vision

Xiaoshi Wu[1]* Hadar Averbuch-Elor[2]* Jin Sun[2] Noah Snavely[2]
[1]Tsinghua University [2]Cornell Tech, Cornell University

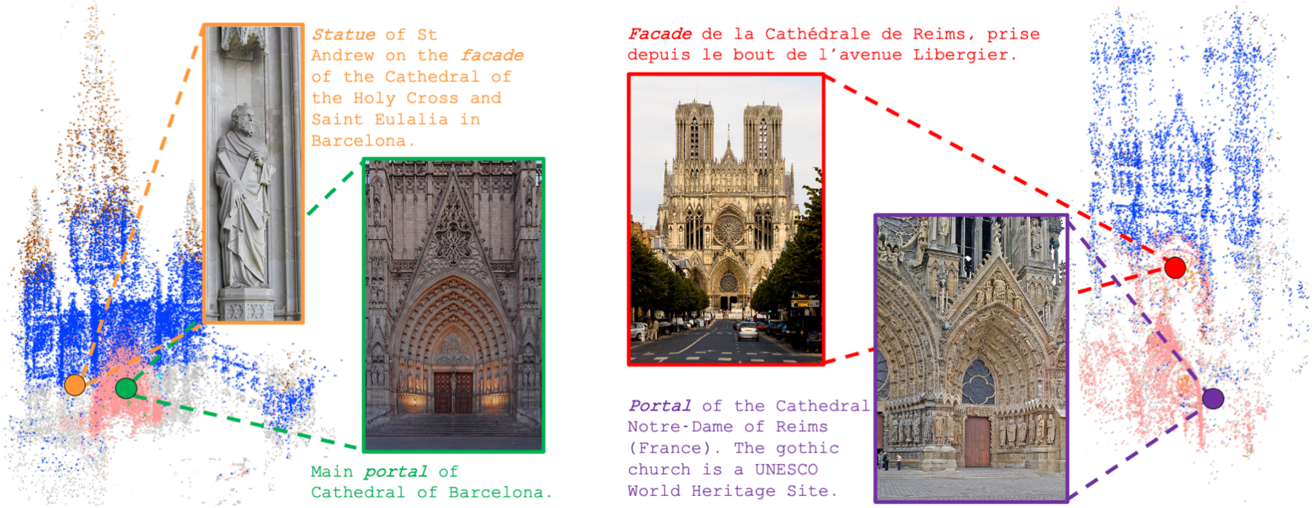

Statue of St Andrew on the *facade* of the Cathedral of the Holy Cross and Saint Eulalia in Barcelona.
Main *portal* of Cathedral of Barcelona.
Facade de la Cathédrale de Reims, prise depuis le bout de l'avenue Libergier.
Portal of the Cathedral Notre-Dame of Reims (France). The gothic church is a UNESCO World Heritage Site.

Figure 1: Our *WikiScenes* dataset combines 3D reconstructions, images, and language descriptions for dozens of landmarks, like the Barcelona and Reims Cathedrals pictured above. WikiScenes enables new tasks that combine different modalities, such as associating semantic concepts like "portal", "facade", and "tower" (colored in pink, blue and brown, respectively) with 3D structure across all cathedrals.

Abstract

The abundance and richness of Internet photos of landmarks and cities has led to significant progress in 3D vision over the past two decades, including automated 3D reconstructions of the world's landmarks from tourist photos. However, a major source of information available for these 3D-augmented collections—namely language, e.g., from image captions—has been virtually untapped. In this work, we present WikiScenes, a new, large-scale dataset of landmark photo collections that contains descriptive text in the form of captions and hierarchical category names. WikiScenes forms a new testbed for multimodal reasoning involving images, text, and 3D geometry. We demonstrate the utility of WikiScenes for learning semantic concepts over images and 3D models. Our weakly-supervised framework connects images, 3D structure, and semantics—utilizing the strong constraints provided by 3D geometry—to associate semantic concepts to image pixels and 3D points.[1]

1. Introduction

Internet photos capturing tourist landmarks around the world have driven research in 3D computer vision for over a decade [43, 18, 16, 2, 40, 30]. Diverse photo collections of landmarks are unified by the underlying 3D scene geometry, despite the fact that a scene can look dramatically different from one image to the next due to varying illumination, alternating seasons, or special events. This geometric anchoring can be exploited when learning a range of geometry-related vision tasks, such as novel view synthesis [31, 26], single-view depth prediction [25], and relighting [51, 50], that require large amounts of diverse training data. However, prior work on tourist photos of landmarks has focused almost exclusively on lower-level reconstruction tasks, and not on

*: indicates equal contribution.

[1] `https://www.cs.cornell.edu/projects/babel/`

higher-level scene understanding or recognition tasks.

We seek to connect such 3D-augmented image collections to a new domain: *language*. Natural language is an effective way to describe the complexities of the 3D world; 3D scenes exhibit features such as compositionality and physical and functional relationships that are easily captured by language. For instance, consider the images of the Barcelona and Reims Cathedrals in Fig. 1. Cathedrals like these have common elements, such as facades, columns, arches, portals, domes, etc., that tend to be physically assembled in consistent ways across all cathedrals (and related buildings like basilicas). Using modern structure from motion methods, we can reconstruct 3D models of the world's cathedrals, but we cannot directly infer such rich semantic connections that exist *across* all cathedrals. Such reasoning calls for methods that jointly consider language, images, and 3D geometry.

However, despite impressive progress connecting images to natural language descriptions across tasks such as image captioning [48, 28, 4] and visual grounding [47, 21, 19], little attention has been given to joint analysis of 3D vision and language. In this work, we facilitate such multimodal analysis with a new framework for creating 3D-augmented datasets from Wikimedia Commons, a diverse, crowdsourced and freely-licensed large-scale data source. We use this framework to create **WikiScenes**, a new dataset that contains 63K paired images and textual descriptions capturing 99 cathedrals, along with their associated 3D reconstructions, illustrated in Fig. 1. WikiScenes enables a range of new explorations at the intersection of language, vision, and 3D.

We demonstrate the utility of WikiScenes for the specific task of mining and learning semantic concepts over collections of images and 3D models. Our key insight is that while raw textual descriptions represent a weak, noisy form of supervision for semantic concepts, the underlying 3D structure of scenes yields powerful physical constraints that grants robustness to data noise and can ground models. In particular, we devise a novel 3D contrastive loss that leverages scene geometry to regularize learning of semantic representations. We also show that 3D scene geometry leads to improved vision-language models in a caption-based image retrieval task, where geometry helps in augmenting the training data with semantically-related samples.

In summary, our key contributions are:

- **WikiScenes**, a large-scale dataset combining language, images, and 3D models, which can facilitate research that jointly considers these modalities.
- A contrastive learning method for learning semantic image representations leveraging 3D models.
- Results that demonstrate that our proposed model can associate semantic concepts with images and 3D models, even for never-before-seen locations.

2. Related Work

Joint analysis of 3D and language. We have recently seen pioneering efforts to jointly analyze 3D and language. Chen *et al*. [11] learn a joint embedding of text and 3D shapes belonging to the ShapeNet dataset [9], and demonstrate these embeddings on text-to-shape retrieval and text-to-shape generation. Achlioptas *et al*. [1] learn language for differentiating between shapes. To do so, they generate a dataset consisting of triplets of ShapeNet chairs with utterances distinguishing one chair from the other two. In contrast to these object-centric works, Chen *et al*. [10] consider full 3D scenes. They construct a multimodal dataset for indoor scenes and localize 3D objects in the scene using natural language. We also consider 3D scenes, but in our case, the 3D scenes capture complex architectural landmarks, and their images and textual descriptions are gathered from Wikimedia Commons.

Vision and language. Many recent works connect images to natural language descriptions. Popular tasks include instruction following [5, 32, 8], visual question answering [6, 15, 22, 4], and phrase localization [29, 49, 45]. However, prior work has shown that models combining vision and language often rely on simple signals or fail to jointly consider both modalities. For instance, visual question answering techniques often ignore the image content [3], and visually-grounded syntax acquisition methods essentially learn a simple noun classifier [23]. We assemble Internet collections that are grounded to a 3D model, providing physical constraints that can better connect language and vision.

Distilling information from Internet collections. Several works mine Internet collections capturing famous landmarks for objects [17, 36], events [38], or named parts [46] using image clustering techniques. Other work analyzes camera viewpoints in large-scale tourist imagery to automatically summarize a scene [42] or segment it into components [41].

Other prior work analyzes image content together with textual tags, geotags, and other metadata to organize image collections. Crandall *et al*. use image features and user tags from geotagged Flickr images to discover and classify world landmarks [13]. 3D Wikipedia analyzes textual descriptions of tourist landmarks, leveraging photo co-occurrences to annotate specific 3D models like the Pantheon [39]. In contrast to the above methods, which operate on each location in isolation, our work aims to discover semantic concepts spanning a whole category of locations, such as all the world's cathedrals. We further use a contrastive learning framework for detecting these concepts in unseen landmarks.

3. The WikiScenes Dataset

Our WikiScenes dataset consists of paired images and language descriptions capturing world landmarks and cultural sites, with associated 3D models and camera poses.

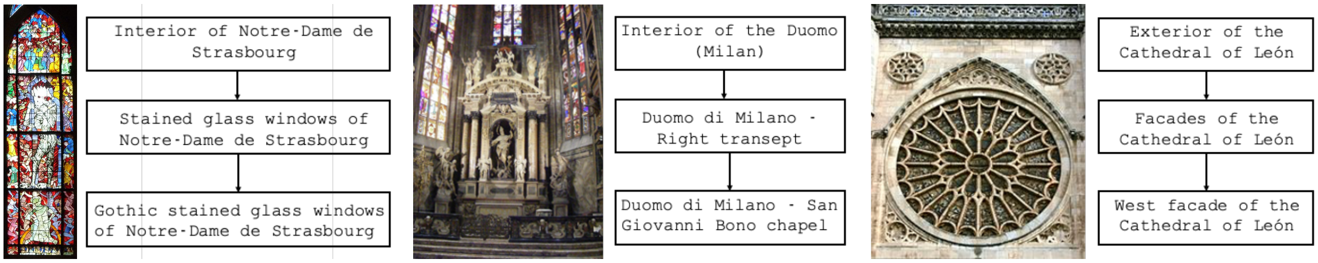

Interior of Notre-Dame de Strasbourg
Stained glass windows of Notre-Dame de Strasbourg
Gothic stained glass windows of Notre-Dame de Strasbourg
Interior of the Duomo (Milan)
Duomo di Milano - Right transept
Duomo di Milano - San Giovanni Bono chapel
Exterior of the Cathedral of León
Facades of the Cathedral of León
West facade of the Cathedral of León

Figure 2: Images paired with hierarchical WikiCategories from the root (top) to the leaf (bottom).

WikiScenes is derived from the massive public catalog of freely-licensed crowdsourced data available in Wikimedia Commons,[2] which contains a large variety of images with captions and other metadata. Within Wikimedia Commons, landmarks are organized into a hierarchy of semantic categories. In this work, we focus on cathedrals as a showcase of our framework, although our methodology is general and can be applied to other types of landmarks. We will also release companion datasets featuring mosques and synagogues.

To create WikiScenes, we first assembled a list of cathedrals using prior work on mining landmarks from geotagged photos [13]. Each cathedral corresponds to a specific *category* on Wikimedia Commons, at which is rooted a hierarchy of sub-categories that each contain photos and other relevant information. We refer to a Wikimedia Commons category as a *WikiCategory*. For example, "Cathédrale Notre-Dame de Paris"[3] is the name of a WikiCategory corresponding to the Notre Dame Cathedral in Paris. It has a descendent WikiCategory called "Nave of Notre-Dame de Paris"[4] that features photos of the nave (a specific region of a cathedral interior), as well as yet more detailed WikiCategories. Each landmark's root WikiCategory node contains "Exterior", "Interior" and "Views" subcategories. We download all images and associated descriptions under these subcategories. We extract two forms of textual descriptions for each image:

- **Captions** associated with images, describing the image using free-form language (Figure 1).
- The **WikiCategory hierarchy** associated with each image. Example hierarchies are shown in Figure 2.

Because data stored in Wikimedia Commons is not specific to any single language edition of Wikipedia, our dataset contains text in numerous languages, allowing for future multilingual tasks like learning of cross-lingual representations [44]. However, one can also train with text from a single language, such as English. Overall, WikiScenes contains 63K images of cathedrals with textual descriptions.

We integrate these Wikimedia Commons–sourced images with 3D reconstructions of landmarks built using COLMAP [40], a state-of-the-art SfM system that reconstructs camera poses and sparse point clouds. For each 3D point in the reconstructed scene, we track all its associated images and corresponding pixel locations. In total, 26K images of cathedrals were successfully registered in 3D. Example 3D reconstructions are shown in Figure 1.

Dataset statistics. WikiScenes is assembled from 99 cathedrals spanning five continents and 23 countries. The languages most common in the captions are English (45.8%), French (11.1%) and Spanish (10.9%). The Notre Dame Cathedral in Paris represents the largest subset, with 5,700 images-description pairs. The median number of words in a caption is seven; the average is significantly higher as some captions contain detailed excerpts about their associated landmark. 8.39% of all captions contain at least one spatial connector,[5] suggesting that our captions describe rich relationships between different parts of a structure. Please see the supplemental material for detailed distributions over attributes including language and collection size.

4. Mining WikiScenes for Semantic Concepts

To demonstrate the semantic knowledge encoded in our dataset, we mine WikiScenes for semantic concepts associated with the Cathedral landmark category. While the raw textual descriptions are noisy, we show that we can distill a clean set of concepts by exploiting within-scene 3D constraints (Sec. 4.1). We then associate these concepts to images (Sec. 4.2), and show that these concepts can be used to train neural networks to visually recognize these concepts.

4.1. Distilling semantic concepts

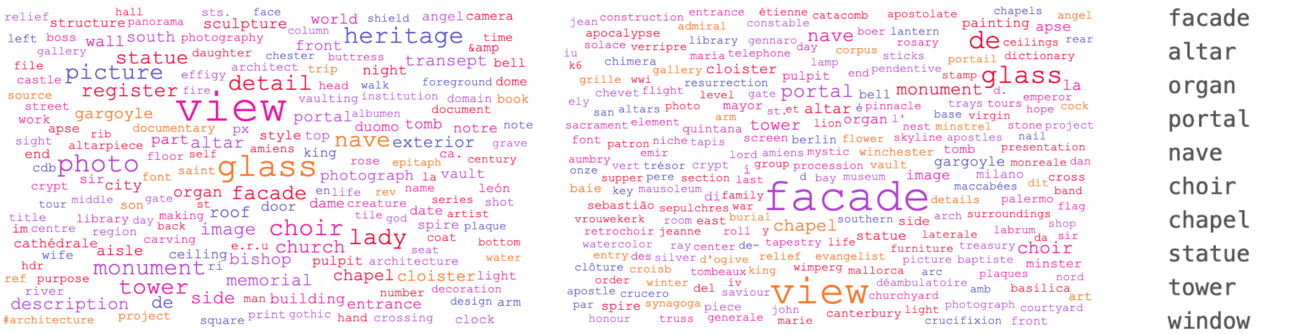
To determine a set of candidate concepts, we first assemble a list of all nouns found in the leaf nodes of the WikiCategories, hereby denoted as the *leaf categories*, as empirically we found that the leaf categories are most representative of the image content. Since we are interested in a

[2] https://commons.wikimedia.org

[3] https://commons.wikimedia.org/wiki/Category:Cath%C3%A9drale_Notre-Dame_de_Paris

[4] https://commons.wikimedia.org/wiki/Category:Nave_of_Notre-Dame_de_Paris

[5] The spatial connectors we consider are: above, over, below, under, beside, behind, from, towards, left, right, east and west.



Candidates *from captions* Candidates *from the leaf categories* Distilled concepts

Figure 3: We visualize the raw text captured in WikiScenes captions (left) and leaf tags (center). Larger words are more frequent in the dataset. Our distilled concepts, obtained according to the algorithm described in Sec. 4.1, are listed on the right.

list of *abstract* concepts and not in detecting specific places and objects, we filter out nouns detected as entities using an off-the-shelf Named Entity Recognition (NER) tagger [37]. Figure 3 (middle) visualizes the initial candidate list as a word cloud (more frequent words appear larger). As the figure illustrates, this list contains nouns that indeed describe semantic regions in the "Cathedral" category, but also contains many outliers, or nouns not specifically related to the "Cathedral" category, such as "view" or "photograph".

As an alternative, we can also extract nouns directly from the captions (Figure 3, left). This results in a noisier list, as the captions are generally longer with more diverse and detailed descriptions. In addition, leveraging category names leads to more images with noun descriptions—over 56K images have at least one noun in their leaf category, whereas only 22K images have an English caption with a noun.

To distill a clean set of semantic concepts from the initial list, we identify and select concepts that pass two tests: they are (1) *well-supported* in the collection (i.e., they occur frequently in the textual descriptions) and (2) *coherent*, in the sense that they consistently reference identical or visually similar elements. While well-supported concepts can be determined by simple frequency measurements, coherence is more difficult to assess from noisy Internet images and their descriptions. However, because these images are physically grounded via a 3D model, we can measure coherence in 3D.

For each candidate concept, e.g., "facade", we construct multiple visual adjacency graphs (one per landmark) over the images associated with that concept. Note that an image can be associated with multiple concepts, according to the nouns detected in its leaf category. For each graph, nodes $v \in V$ correspond to images and two images are connected by an edge $e \in E$ if they share at least K common keypoints in the 3D model (where K is empirically set to 10). We are interested in measuring the degree to which the images of the candidate concept are clustered together in 3D. Therefore, for each landmark ℓ, we compute the graph density:

$$\rho^{\ell} = \frac{2|E|}{|V|(|V|-1)}. \quad (1)$$

The coherence of the candidate concept is measured as the average graph density ρ, obtained by taking the average over all corresponding landmark graphs with at least 10 nodes.

Finally, candidate concepts that appear in at least 25 landmarks (roughly a quarter of the "Cathedral" category) and have a coherency score $\rho \geq 0.08$ are added to our distilled set (Figure 3, right).

4.2. Associating images with distilled concepts

Although the distilled set of semantic concepts is constructed only from text appearing in the leaf categories, we utilize both the image captions and leaf categories when generating labels: an image is associated with a concept if the concept is present either in the caption or in its leaf categories. An image can be associated with multiple concepts.

One exception is that text often includes concepts that are spatially related to the main concept present in an image using *spatial* connectors such as "beside", "next", "from", "towards". For example, an image associated with the text "nave looking towards portal" should be associated with "nave", but not necessarily with "portal". Hence, we do not associate concepts with images if the concept appears anywhere after a spatial connector.

5. Learning Semantic Representations

WikiScenes can be used to study a range of different problems. Here, we focus on semantic reasoning over 2D images and 3D models. In the previous section, we proposed a technique for discovering semantic concepts and associating these with images in WikiScenes. Now, we show how these image-level pseudo-labels can provide a supervision signal for learning semantic feature representations over an entire category of landmarks.

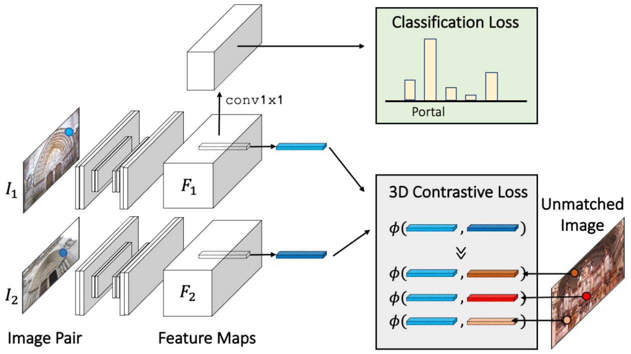


Figure 4: Overview of our contrastive learning framework. Given an image pair with shared keypoints (left), we jointly train a model to classify the images into one of the C concepts from the learned score maps and to output a higher similarity for pixels mapping to the same point in 3D (in blue). Negative pairs are constructed by sampling non-corresponding points from other images in the batch.

We seek to learn pixel-wise representations (in contrast to whole-image representations), because we wish to easily map knowledge from 2D to 3D and vice versa. We would also like our learned representations to be semantically meaningful. In other words, our distilled concepts should be identifiable from these pixel-wise representations. To this end, we devise a contrastive learning framework that computes a feature descriptor for every pixel in the image. We also show how our trained model can be directly utilized to estimate feature descriptors for 3D points through their associated images.

5.1. Training objectives

Our training data consists of image pairs (I_1, I_2) with shared keypoints, obtained from the corresponding SfM model. We use convolutional networks with shared weights to extract dense feature maps F_1 and F_2 whose width and height match those of the original images. For simplicity of notation, we assume both images have dimensions $w \times h$. To train a feature descriptor model with such data, we propose to use two complimentary loss terms: a novel 3D contrastive loss that utilizes within-scene physical constraints and a classification loss (Figure 4).

3D contrastive loss. We design a new 3D contrastive loss to encourage within-scene consistency, such that pixels from different images corresponding to the same 3D point should have similar features. This is unlike prior works on contrastive learning that use handcrafted data augmentations [12, 20] or synthetic images [35] to generate positive pairs—in our case the positive pairs are 2D pixels that are projections of the same point in 3D. This loss relates images with different characteristics, such as lighting and scale, allowing to better focus on semantics and providing higher robustness against such nuisance factors.

Our learning method works as follows: For each point p in I_1 corresponding to point p^+ in I_2 (i.e., they are both projections of the same 3D point P), we formulate a contrastive loss to maximize the mutual information between their descriptors $F_1(p)$ and $F_2(p^+)$. We consider a noise contrastive estimation framework [34], consisting of the positive pair (p, p^+) and m negative pairs $\{(p, p_i^-)\}$:

$$\mathcal{L}_{\text{3D}} = -\log\left[\frac{e^{\phi(p,p^+)}}{e^{\phi(p,p^+)} + \sum_{i=1}^{m} e^{\phi(p,p_i^-)}}\right], \tag{2}$$

where the similarity $\phi(p, p^*)$ is computed as the dot product of feature descriptors scaled by a temperature τ:

$$\phi(p, p^*) = F_1(p) \cdot F_2(p^*) \,/\, \tau. \tag{3}$$

This loss can be interpreted as the log loss of a $(m+1)$-way softmax classifier that learns to classify p as p^+. The points p_i^- are sampled uniformly from other images in the same batch. To avoid collapsing the feature space, we normalize all feature descriptors to unit length.

Semantic classification loss. For each image we also compute a semantic classification loss. Given C unique semantic concepts, we obtain unnormalized score maps from the feature descriptors using a simple `conv1x1` layer. That is, we map the $[K \times h \times w]$ feature descriptor tensor to a $[C \times h \times w]$ score map tensor, where each slice corresponds to one of the semantic concepts.

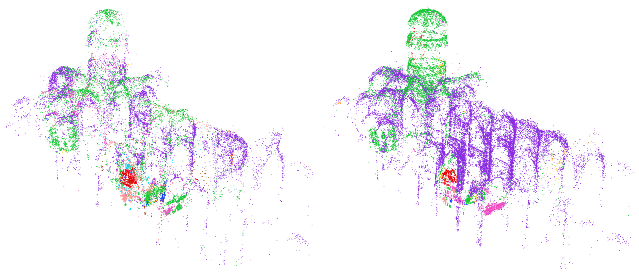
Following the design proposed by Araslanov *et al*. [7], we add a background channel and compute a pixel-wise `softmax` to obtain normalized score maps $y_{\text{pix}} \in \mathbb{R}^{C+1 \times h \times w}$ and image-level classification scores $y \in \mathbb{R}^C$, derived from the score maps using the method of Araslanov *et al*. Our semantic classification loss is defined as

$$\mathcal{L}_{\text{cls}} = \mathcal{L}^{\text{im}}_{\text{cls}} + \mathcal{L}^{\text{pix}}_{\text{cls}}, \tag{4}$$

where $\mathcal{L}^{\text{cls}}_{\text{im}}$ is a classification loss on image-level scores y and $\mathcal{L}^{\text{cls}}_{\text{pix}}$ is a self-supervised semantic segmentation loss over pixel-wise predictions (where high-confidence pixel predictions serve as self-supervised labels). For both training and evaluation, we only consider images labeled with a single concept and the one-hot class label is set according to our pseudo image label. We minimize a cross-entropy loss for both image-level and pixel-level predictions.

5.2. Inference

At inference time, we can feed an image from a never-before-seen location into our model (Figure 5). The model outputs pixel-wise feature descriptors and probability scores over the semantic concepts for each pixel (and also for the full image, if that is desired). We follow the procedure described in [7] to extract 2D segmentations. To output probability scores for a 3D point in the scene, we process all the



Baseline (w/o $\mathcal{L}_{3D}$) Ours

Figure 5: Segmenting an unseen 3D model of the interior of the Aachen Cathedral in Germany. Color legend: *nave*, *chapel*, *organ*, *altar*, *choir*, *statue*, *portal*, *facade*.

images associated with this 3D point. The feature descriptors of all its 2D projections are averaged, and we process this average descriptor to output its associated probability scores. We associate a 3D point with one of the semantic concepts if its corresponding confidence score is greater than $\varphi = 0.5$.

6. Evaluation

In this section, we demonstrate our ability to learn semantic concepts shared across multiple landmarks. Specifically, we seek to answer the following questions:

- Is WikiScenes suitable for learning these concepts?
- How important is the 3D contrastive loss?
- How well does our model generalize to Internet photos from never-before-seen locations?

We perform a variety of experiments to evaluate performance across multiple tasks, including classification, segmentation, and a caption-based image retrieval task that operates on the raw captions directly. These experiments are complemented with a visual analysis that highlights the unique characteristics and challenges of our data.

6.1. Implementation details

Data. Out of the 99 WikiScenes landmarks, 70 landmarks contain sufficient labeled data that can serve for training and evaluating our models (images are labeled using the approach described in Section 4.2). We create a 9:1 split at the landmarks level, forming a test set for landmarks *unseen* during training (WS-U). For the 63 landmarks in the training set, we create a 9:1 split at the images level, forming a test set for *known* landmarks (WS-K) to evaluate how well our model can classify unseen images in familiar locations. Overall, we use almost 9K labeled images for training, with balanced class frequencies across the ten semantic concepts.

Training. We use a batch size of 32, corresponding to 16 image pairs. Only half of these are *real* pairs with shared keypoints, as we also want to consider labeled images that are not associated with any 3D reconstruction, possibly due of a sparse sampling of views in these regions. Please refer to the supplementary for additional implementation details.

6.2. Label quality

We assess the accuracy of our pseudo-labels by manually inspecting 50 randomly sampled training images for each concept, and identifying images with incorrect labels (i.e., the image does not picture all or part of the semantic concept). We found an accuracy greater than 98%, suggesting that our pseudo-labels are highly accurate. We found that most errors are due to images that contain schematic diagrams or scans of the concept (and not natural images capturing it). Please refer to the supplementary material for visualizations of our training samples.

6.3. 3D-consistency guided classification

Next we evaluate to what extent semantic concepts can be learned across a multitude of landmarks, and the effect of the 3D consistency regularization allowed by our dataset on classification results. We perform an image classification evaluation using our pseudo-labels, which we consider ground-truth for evaluation purposes. We compare our model to a model with the same architecture, trained using the semantic classification loss but without our 3D contrastive loss, hereby denoted as the *baseline* model—adapted from the one proposed in Araslanov *et al.* [7].

For each model, we report the overall mean average precision (mAP), as well as a breakdown of AP per concept, in Table 1. Results are reported for test images from known locations (WS-K) and unseen locations (WS-U). As the table illustrates, our model outperforms the baseline model in most of the concepts and yields significant gains in mAP, boosting overall performance by 4.5% and 3.7%, when evaluating on WS-K and WS-U, respectively (and an improvement of 3.3% when averaging across images, which is less affected by class frequencies). We provide additional experiments and an analysis of errors in the supplementary material.

6.4. 2D and 3D segmentation

Our framework learns pixel-wise features that are useful beyond classification, e.g., for producing segmentation maps for 2D images and 3D reconstructions. We show segmentation results for 2D images in Figure 6 and for 3D reconstructions in Figures 1 and 5.

We manually label a random subset of test images (from unseen landmarks) for evaluating 2D segmentation performance and report standard segmentation metrics in Table 2. Specifically, we labeled 237 images spanning six concepts that have definite boundaries (*facade*, *portal*, *window*, *organ*, *tower* and *statue*). The distributions across these classes are roughly uniform (with 24-50 images per class).

Test Set	Model	mAP	mAP*	facade	window	chapel	organ	nave	tower	choir	portal	altar	statue
WS-K	Baseline (w/o $\mathcal{L}_{3D}$)	70.8	77.7	87.2	**89.2**	60.2	89.7	**85.8**	**64.1**	61.5	68.0	50.0	52.0
	Ours	**75.3**	**81.0**	**90.0**	88.5	**68.7**	**90.7**	85.7	61.1	**77.2**	**76.5**	**54.4**	**59.9**
WS-U	Baseline (w/o $\mathcal{L}_{3D}$)	48.3	64.0	71.0	92.2	10.7	**57.3**	71.0	**53.4**	43.6	31.1	25.8	27.1
	Ours	**52.0**	**67.3**	**77.7**	**93.4**	**16.5**	49.4	**77.3**	46.1	**44.1**	**35.2**	**39.9**	**40.0**

Table 1: Classification Performance. We report mean average precision (mAP, * indicates averaging over all images, and not per class), and per distilled concept average precision (AP). Results of our model are compared against a model trained without our 3D contrastive loss. Performance is reported on images from known landmarks (WS-K) and unseen landmarks (WS-U). The best results are highlighted in bold.

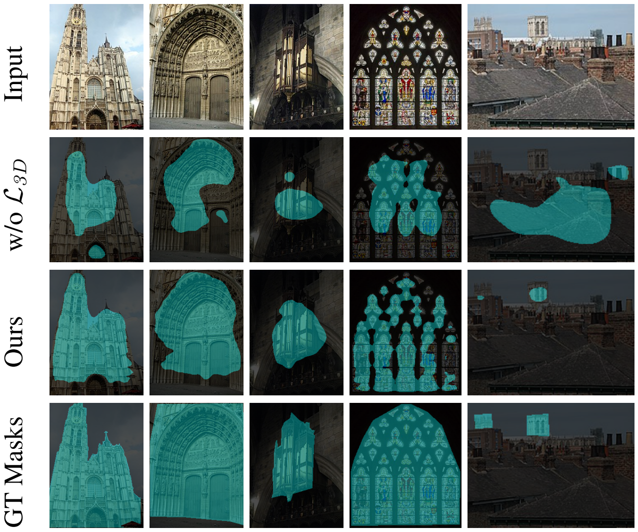


Figure 6: Segmenting images of unseen landmarks. Pixels are labeled *facade*, *portal*, *organ*, *window*, *tower* from left to right.

Model	IoU	Precision	Recall
Baseline (w/o $\mathcal{L}_{3D}$)	25.4	68.6	28.4
Ours	**27.2**	**80.8**	**29.6**

Table 2: Image segmentation performance on manually labeled set.

	WS-K				WS-U			
Method	$\theta_{0.5}$	$\theta_{0.75}$	$\Delta_{0.5}$	$\Delta_{0.75}$	$\theta_{0.5}$	$\theta_{0.75}$	$\Delta_{0.5}$	$\Delta_{0.75}$
Baseline	0.50	0.78	**0.10**	0.09	0.56	0.83	0.13	0.10
Ours	**0.43**	**0.70**	**0.10**	**0.06**	**0.40**	**0.69**	**0.11**	**0.06**

Table 3: 3D Segmentation Evaluation. Proxy metrics θ and Δ are described in detail in Section 6.4. For both metrics, lower is better.

Table 2 shows the average intersection-over-union (IoU), precision and recall on the manually labeled set. These results show that our 3D-contrastive loss boosts performance over all metrics. Precision is significantly higher (81% vs. 69%), with a modest increase in IoU and recall.

To evaluate 3D segmentation performance, as it is difficult to obtain ground-truth 3D segmentations for large-scale landmarks whose reconstructions span thousands of points, we design two proxy metrics to assess both *completeness* and *accuracy* of the 3D results. These metrics are (i) the fraction of ambiguous points θ_φ, and (ii) the interior-exterior error Δ_φ (both dependent on the confidence scores φ).

The fraction of ambiguous points θ_φ quantifies the extent to which the model associates concepts to 3D points with high confidence. To compute θ_φ, we measure the fraction of points that are not associated with a concept, averaging over all landmarks. For example, $\theta_\varphi = 0$ means that for all points, the model's predictions across all images was consistent in 3D space, and thus the points were successfully associated with concepts, and $\theta_\varphi = 1$ means that all points are ambiguous in their semantic association.

Due to limited visual connectivity, 3D reconstructions of landmarks typically are broken into one or more exterior reconstructions and one or more interior reconstructions. Thus, we devise the interior-exterior error Δ_φ to quantify to what extent concepts that should be exclusively found in either an exterior reconstruction or an interior reconstruction are mixed into a single reconstruction. For example, for the interior 3D reconstruction shown in Figure 5, we do not expect to see points labeled as "facade" or "tower", since those concepts appear outdoors. *Interior* concepts include "organ", "nave", "altar", and "choir", and *exterior* concepts include "portal", "facade", and "tower". For each 3D reconstruction m, the error Δ_φ is defined as

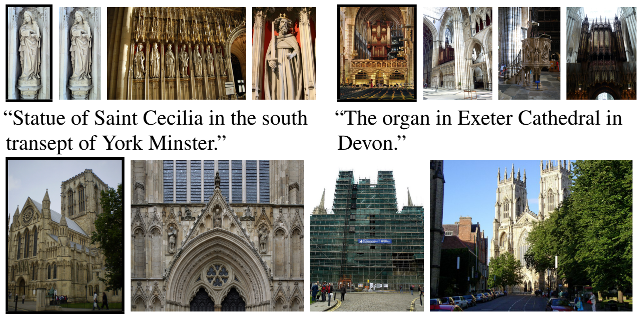
$$\Delta_\varphi^m = \min\left(p_{\text{ext}}, 1 - p_{\text{ext}}\right), \tag{5}$$

where p_{ext} is the probability of an exterior concept in the 3D reconstruction (normalized over the sum of exterior and interior concepts in the reconstruction). We perform a weighted averaging over all the reconstructions, such that larger 3D reconstructions affect the average accordingly.

We report results for both $\theta_\varphi = 0.5$ and $\theta_\varphi = 0.75$ in Table 3 (note that all our qualitative results are generated using $\theta_\varphi = 0.5$). As illustrated in the table, our model surpasses the baseline model (trained without the 3D contrastive loss) on both metrics, demonstrating that more points are consistently associated with concepts, and that each point cloud is more consistently segmented into exterior or interior concepts. Note that some structural parts are inherently more ambiguous (for example, a "statue" is often placed on a "facade"), hence many 3D points are not associated with concepts (also for our model). We explore this further in

Model	R1	R5	R10	S1	S5	S10	S1*	S5*	S10*
Pretrained	1.2	4.3	6.6	22.9	51.0	67.2	44.2	73.9	85.8
Baseline	3.2	11.9	19.2	51.9	80.6	88.0	69.2	89.3	94.6
Ours	**4.0**	**13.9**	**22.5**	**64.0**	**81.9**	**91.2**	**76.0**	**91.2**	**96.3**

Table 4: Caption-Based Image Retrieval Performance. We report performance using a standard retrieval metric and our proposed semantic metric (* indicates averaging over all images, and not per class). Results of our model are compared against a model trained without our 3D augmentations (baseline) and on the pretrained model [27]. Performance is reported on images from unseen landmarks (WS-U). The best results are highlighted in bold.



"Statue of Saint Cecilia in the south transept of York Minster."

"The organ in Exeter Cathedral in Devon."

"York Minster as seen from across the street, York, England."

Figure 7: Retrieving images from captions of unseen landmarks. Above we show the top three retrievals next to the target image (left), corresponding to the caption below.

the supplementary material, showing a confusion matrix for our image classification model as well as the ancestor labels associated with each concept.

6.5. Learning semantics from raw captions

To explore the utility of the raw captions without first distilling concepts, we train a joint vision-language model on images and their raw captions and evaluate it on a caption-based image retrieval task. As with other tasks like classification, we explore the benefit of having 3D geometry in this experiment, showing that geometry can be used to perform data augmentation and boost retrieval performance.

We finetune a state-of-the-art multi-task joint visual and textual representation model [27] using the same landmarks-level splits as above, training on landmarks from WS-K and testing on unseen landmarks in WS-U. We compare models finetuned on two different subsets: (1) a *baseline* subset, provided with pairs of English-only captions and their corresponding images, and (2) a *3D-augmented* subset, where, in addition to the real image-caption pairs, we create new image-caption pairs by associating images with captions from other images with a large visual overlap (measured by thresholding on an IoU ratio of 3D keypoints, set empirically to 0.3). Performing such 3D-aware augmentation enables use of additional images—for which a caption may be unavailable—but whose content is similar to the original image (while appearance and viewpoint may vary). Our 3D-augmentation strategy yields a training dataset with roughly 1.5K more images and 9K more image-caption pairs (the original training set contains nearly 20K pairs).

Table 4 shows caption-based image retrieval performance using Recall@K (R1,R5,R10 in the table), which is a standard metric that measures the percentage of successful retrievals for which the target image is among the top-K retrievals. Additionally, to quantify how *semantically* accurate these retrievals are, we use our semantic labels (obtained according to the method described in Section 4.2) as a proxy and propose a semantic measure S that measures the percentage of retrievals containing at least one image labeled correctly. All metrics are reported for the two models and also for the pretrained model [27] (without finetuning). For our semantic metric, we report an average per class and average over all the images in the test set.

Using 3D augmentations gives a boost in performance across all metrics. Figure 7 illustrates several retrieval results from our model. As illustrated in the bottom row, the model can also align general concepts to our images, such as what a cathedral should look like "from across the street". We show additional qualitative results in the supplementary material.

7. Conclusion

We have presented a new large-scale dataset at the intersection of vision, language, and 3D. We demonstrated the use of our dataset for mining semantic concepts and for learning to associate these concepts with images and 3D models from never-before-seen locations. We show that these tasks benefit from having access to 3D geometry, allowing robust distillation of semantics from noisy Internet collections.

Future applications. We believe our dataset could spark research into many new problems. Automatic captioning of images capturing tourist attractions is one interesting avenue for future research. The rich textual descriptions in our dataset could allow users to virtually explore any tourist attraction, serving as a virtual "tour guide". Our dataset could also enable automatic generation of new 3D scenes and language-guided scene editing. While text-based 2D image generation is a very active research area [14, 33, 24], the problem of generating and modifying 3D scenes using language is largely unexplored. Finally, our focus was on discovery of well-supported concepts, but our dataset can also benefit zero- or few-shot settings via the detailed descriptions present in image captions, enabling rich conceptualization of general visual concepts.

Acknowledgments. This work was supported by the National Science Foundation (IIS-2008313), by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program, by the Zuckerman STEM leadership program, and by an AWS ML Research Award.

References

[1] Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. ShapeGlot: Learning language for shape differentiation. In *ICCV*, 2019.

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building Rome in a day. *Communications of the ACM*, 54(10), 2011.

[3] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[5] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[7] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020.

[8] Valts Blukis, Nataly Brukhim, Andrew Bennett, Ross A. Knepper, and Yoav Artzi. Following high-level navigation instructions on a simulated quadcopter with imitation learning. In *Proceedings of the Robotics: Science and Systems Conference*, 2018.

[9] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[10] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *arXiv preprint arXiv:1912.08830*, 2019.

[11] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *ACCV*, 2018.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. on Machine Learning*. PMLR, 2020.

[13] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proc. Int. Conf. on World Wide Web*, 2009.

[14] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *ICCV*, 2017.

[15] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016.

[16] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010.

[17] Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. I know what you did last summer: object-level auto-annotation of holiday snaps. In *ICCV*, pages 614–621, 2009.

[18] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007.

[19] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. *arXiv preprint arXiv:2006.09920*, 2020.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[21] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *PAMI*, 2019.

[22] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, pages 804–813, 2017.

[23] Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. What is learned in visually grounded neural syntax acquisition. In *ACL*, 2020.

[24] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In *NeurIPS*, 2019.

[25] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

[26] Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. In *ECCV*, 2020.

[27] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.

[28] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.

[29] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

[30] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021.

[31] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *CVPR*, pages 6878–6887, 2019.

[32] Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, 2017.

[33] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *NeurIPS*, 2018.

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[35] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*. Springer, 2020.

[36] James Philbin and Andrew Zisserman. Object mining using a matching graph on very large image collections. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 738–745. IEEE, 2008.

[37] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.

[38] Till Quack, Bastian Leibe, and Luc Van Gool. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 47–56, 2008.

[39] Bryan C Russell, Ricardo Martin-Brualla, Daniel J Butler, Steven M Seitz, and Luke Zettlemoyer. 3D Wikipedia: Using online text to automatically label and navigate reconstructed geometry. In *SIGGRAPH*, 2013.

[40] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[41] Ian Simon and Steven M. Seitz. Scene segmentation using the wisdom of crowds. In *ECCV*, pages 541–553, 2008.

[42] Ian Simon, Noah Snavely, and Steven M. Seitz. Scene summarization for online image collections. In *ICCV*, 2007.

[43] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. In *SIGGRAPH*, 2006.

[44] Dídac Surís, Dave Epstein, and Carl Vondrick. Globetrotter: Unsupervised multilingual translation from visual alignment. *arXiv preprint arXiv:2012.04631*, 2020.

[45] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *ECCV*, pages 696–711, 2016.

[46] Tobias Weyand and Bastian Leibe. Discovering details and scene structure with hierarchical iconoid shift. In *ICCV*, 2013.

[47] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2017.

[48] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[49] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016.

[50] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William A. P. Smith. Self-supervised outdoor scene relighting. In *ECCV*, 2020.

[51] Ye Yu and William A. P. Smith. InverseRenderNet: Learning single image inverse rendering. In *CVPR*, pages 3155–3164, 2019.