DOI: 10.1111/cgf.15006
EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2

HaLo-NeRF: Learning Geometry-Guided Semantics for Exploring Unconstrained Photo Collections

Chen Dudai*[1], Morris Alper*[1], Hana Bezalel[1], Rana Hanocka[2], Itai Lang[2], Hadar Averbuch-Elor[1]

[1]Tel Aviv University [2]University of Chicago

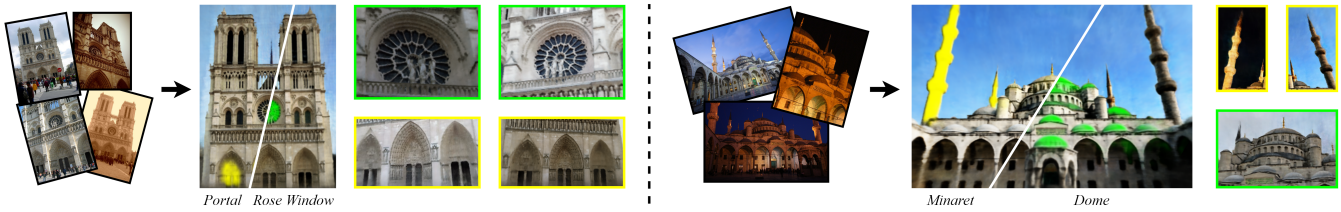


Figure 1: *Given a collection of images in-the-wild depicting a large-scale scene, such as the Notre-Dame Cathedral or the Blue Mosque above, we learn a semantic localization field for each textual description (shown with green and yellow overlay). Our approach enables generating novel views with controlled appearances of these semantic regions of interest (as shown in the boxes of corresponding colors).*

Abstract

Internet image collections containing photos captured by crowds of photographers show promise for enabling digital exploration of large-scale tourist landmarks. However, prior works focus primarily on geometric reconstruction and visualization, neglecting the key role of language in providing a semantic interface for navigation and fine-grained understanding. In more constrained 3D domains, recent methods have leveraged modern vision-and-language models as a strong prior of 2D visual semantics. While these models display an excellent understanding of broad visual semantics, they struggle with unconstrained photo collections depicting such tourist landmarks, as they lack expert knowledge of the architectural domain and fail to exploit the geometric consistency of images capturing multiple views of such scenes. In this work, we present a localization system that connects neural representations of scenes depicting large-scale landmarks with text describing a semantic region within the scene, by harnessing the power of SOTA vision-and-language models with adaptations for understanding landmark scene semantics. To bolster such models with fine-grained knowledge, we leverage large-scale Internet data containing images of similar landmarks along with weakly-related textual information. Our approach is built upon the premise that images physically grounded in space can provide a powerful supervision signal for localizing new concepts, whose semantics may be unlocked from Internet textual metadata with large language models. We use correspondences between views of scenes to bootstrap spatial understanding of these semantics, providing guidance for 3D-compatible segmentation that ultimately lifts to a volumetric scene representation. To evaluate our method, we present a new benchmark dataset containing large-scale scenes with ground-truth segmentations for multiple semantic concepts. Our results show that HaLo-NeRF can accurately localize a variety of semantic concepts related to architectural landmarks, surpassing the results of other 3D models as well as strong 2D segmentation baselines. Our code and data are publicly available at https://tau-vailab.github.io/HaLo-NeRF/.

CCS Concepts

• ***Computing methodologies*** → ***3D imaging; Rendering; Image segmentation;***

1. Introduction

Our world is filled with incredible buildings and monuments that contain a rich variety of architectural details. Such intricately-designed human structures have attracted the interest of tourists and scholars alike. Consider, for instance, the Notre-Dame Cathedral pictured above. This monument is visited annually by over 10 million people from all around the world. While Notre-Dame's

* Denotes equal contribution

facade is impressive at a glance, its complex architecture and history contain details which the untrained eye may miss. Its structure includes features such as portals, towers, and columns, as well as more esoteric items like *rose window* and *tympanum*. Tourists often avail themselves of guidebooks or knowledgeable tour guides in order to fully appreciate the grand architecture and history of such landmarks. But what if it were possible to explore and understand such sites without needing to hire a tour guide or even to physically travel to the location?

The emergence of neural radiance fields presents new possibilities for creating and exploring virtual worlds that contain such large-scale monuments, without the (potential burden) of traveling. Prior work, including NeRF-W [MBRS*21] and Ha-NeRF [CZL*22], has demonstrated that photo-realistic images with independent control of viewpoint and illumination can be readily rendered from unstructured imagery for sites such as the Notre-Dame Cathedral. However, these neural techniques lack the high-level semantics embodied within the scene—such semantic understanding is crucial for exploration of a new place, similarly to the travelling tourist.

Recent progress in language-driven 3D scene understanding has leveraged strong two-dimensional priors provided by modern vision-and-language (V&L) representations [HCJW22, CLW*22, CGT*22, KMS22, KKG*23]. However, while existing pretrained vision-and-language models (VLMs) show broad semantic understanding, architectural images use a specialized vocabulary of terms (such as the *minaret* and *rose window* depicted in Figure 1) that is not well encapsulated by these models out of the box. Therefore, we propose an approach for performing semantic adaptation of VLMs by leveraging Internet collections of landmark images and textual metadata. Inter-view coverage of a scene provides richer information than collections of unrelated imagery, as observed in prior work utilizing collections capturing physically grounded in-the-wild images [WZHS20, IMK20, WAESS21]. Our key insight is that modern foundation models allow for extracting a powerful supervision signal from *multi-modal* data depicting large-scale tourist scenes.

To unlock the relevant semantic categories from noisy Internet textual metadata accompanying images, we leverage the rich knowledge of large language models (LLMs). We then localize this *image-level* semantic understanding to *pixel-level* probabilities by leveraging the 3D-consistent nature of our image data. In particular, by bootstrapping with inter-view image correspondences, we fine-tune an image segmentation model to both learn these specific concepts and to localize them reliably within scenes, providing a 3D-compatible segmentation.

We demonstrate the applicability of our approach for connecting low-level neural representations depicting such real-world tourist landmarks with higher-level semantic understanding. Specifically, we present a text-driven localization technique that is supervised on our image segmentation maps, which augments the recently proposed Ha-NeRF neural representation [CZL*22] with a localization head that predicts volumetric probabilities for a target text prompt. By presenting the user with a visual *halo* marking the region of interest, our approach provides an intuitive interface for interacting with virtual 3D environments depicting architectural landmarks. HaLo-NeRF (Ha-NeRF + **Lo**calization halo) allows the user to "zoom in" to the region containing the text prompt and view it from various viewpoints and across different appearances, yielding a substantially more engaging experience compared to today's common practice of browsing thumbnails returned by an image search.

To quantitatively evaluate our method, we introduce *HolyScenes*, a new benchmark dataset composed of six places of worship annotated with ground-truth segmentations for multiple semantic concepts. We evaluate our approach qualitatively and quantitatively, including comparisons to existing 2D and 3D techniques. Our results show that HaLo-NeRF allows for localizing a wide array of elements belonging to structures reconstructed in the wild, capturing the unique semantics of our use case and significantly surpassing the performance of alternative methods.

Explicitly stated, our key contributions are:

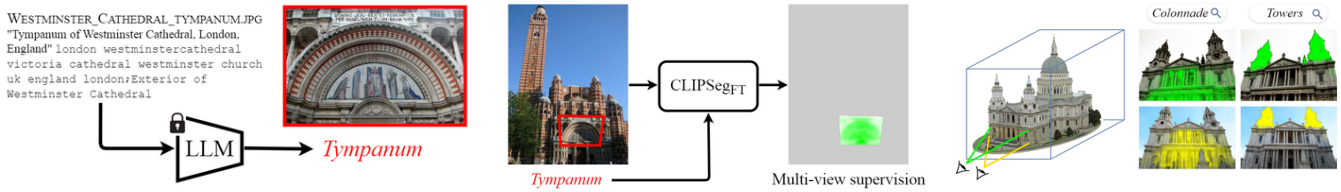
- A novel approach for performing semantic adaptation of VLMs which leverages inter-view coverage of scenes in multiple modalities (namely textual metadata and geometric correspondences between views) to bootstrap spatial understanding of domain-specific semantics;
- A system enabling text-driven 3D localization of large-scale scenes captured in-the-wild;
- Results over diverse scenes and semantic regions, and a benchmark dataset for rigorously evaluating the performance of our system as well as facilitating future work linking Internet collections with a semantic understanding.

2. Related Work

Text-guided semantic segmentation. The emergence of powerful large-scale vision-language models [JYX*21, RKH*21] has propelled a surge of interest in pixel-level semantic segmentation using text prompts [XZW*21, LWB*22, LE22, DXXD22, XDML*22, GGCL22, ZLD22, LWD*23]. A number these works leverage the rich semantic understanding of CLIP [RKH*21], stemming from large-scale contrastive training on text-image pairs.

LSeg [LWB*22] trains an image encoder to align a dense pixel representation with CLIP's embedding for the text description of the corresponding semantic class. OpenSeg [GGCL22] optimizes a class-agnostic region segmentation module to matched extracted words from image captions. CLIPSeg [LE22] leverages the activations of CLIP's dual encoders, training a decoder to convert them into a binary segmentation mask. CLIP's zero-shot understanding on the image level has also been leveraged for localization by Decatur *et al.* [DLH22], who lift CLIP-guided segmentation in 2D views to open-vocabulary localization over 3D meshes.

These methods aim for general open-vocabulary image segmentation and can achieve impressive performance over a broad set of visual concepts. However, they lack expert knowledge specific to culturally significant architecture (as we show in our comparisons). In this work, we incorporate domain-specific knowledge to adapt an image segmentation model conditioned on free text to our setting; we do this by leveraging weak image-level text supervision and pixel-level supervision obtained from multi-view correspon-



(a) LLM-Based Semantic Concept Distillation (b) Semantic Adaptation of V&L Models (c) Text-Driven 3D Localization

Figure 2: ***System overview of our approach.*** *(a) We extract semantic pseudo-labels from noisy Internet image metadata using a large language model (LLM). (b) We use these pseudo-labels and correspondences between scene views to learn image-level and pixel-level semantics. In particular, we fine-tune an image segmentation model (CLIPSeg$_{FT}$) using multi-view supervision—where zoomed-in views and their associated pseudo-labels (such as image on the left associated with the term "tympanum") provide a supervision signal for zoomed-out views. (c) We then lift this semantic understanding to learn volumetric probabilities over new, unseen landmarks (such as the St. Paul's Cathedral depicted on the right), allowing for rendering views of the segmented scene with controlled viewpoints and illumination settings. See below for the definitions of the concepts shown*.*

dences. Additionally, we later lift this semantic understanding to volumetric probabilities over a neural representation of the scene.

Language-grounded scene understanding and exploration. The problem of 3D visual grounding aims at localizing objects in a 3D scene, which is usually represented as a point cloud [HCJW22, CLW*22, CGT*22, LXW*23]. Several works have exploited free-form language for object localization [CCN20, CWNC22] or semantic segmentation [RLD22] of a 3D scene provided as an RGB-D scan. Peng *et al*. [PGJ*22] have leveraged input images in addition to a 3D model, represented as a mesh or a point cloud, to co-embed dense 3D point features with image pixels and natural language.

These works generally assume strong supervision from existing semantically annotated 3D data, consisting of common standalone objects. By contrast, we tackle the challenging real-world scenario of a photo collection in the wild, aiming to localizing semantic regions in large-scale scenes and lacking annotated ground-truth 3D segmentation data for training. To overcome this lack of strong ground-truth data, our method distills *both* semantic and spatial information from large-scale Internet image collections with textual metadata, and fuses this knowledge together into a neural volumetric field.

The problem of visualizing and exploring large-scale 3D scenes depicting tourist landmarks captured *in-the-wild* has been explored by several prior works predating the current deep learning dominated era [SSS06, SGSS08, RMBB*13]. Exactly a decade ago, Russell *et al*. [RMBB*13] proposed 3D Wikipedia for annotating isolated 3D reconstructions of famous tourist sites using reference text via image–text co-occurrence statistics. Our work, in contrast, does not assume access to text describing the landmarks of interest and instead leverages weakly-related textual information of similar landmarks. More recently, Wu *et al*. [WAESS21] also addressed the problem of connecting 3D-augmented Internet image collections to semantics. However, like most prior work, they focused on learning a small set of predefined semantic categories, associated with isolated points in space. By contrast, we operate in the more challenging setting of open-vocabulary semantic understanding, aiming to associate these semantics with volumetric probabilities.

NeRF-based semantic representations. Recent research efforts have aimed to augment neural radiance fields (NeRF) [MST*20] with semantic information for segmentation and editing [TZFR23]. One approach is to add a classification branch to assign each pixel with a semantic label, complementing the color branch of a vanilla NeRF [ZLLD21, KGY*22, SPB*22, FZC*22]. A general drawback of these categorical methods is the confinement of the segmentation to a pre-determined set of classes.

To enable open-vocabulary segmentation, an alternative approach predicts an *entire feature vector* for each 3D point [TLLV22, KMS22, FWJ*22, KKG*23]; these feature vectors can then be probed with the embedding of a semantic query such as free text or an image patch. While these techniques allow for more flexibility than categorical methods, they perform an ambitious task—regressing high-dimensional feature vectors in 3D space—and are usually demonstrated in controlled capture settings (e.g. with images of constant illumination).

To reduce the complexity of 3D localization for unconstrained large-scale scenes captured in the wild, we adopt a hybrid approach. Specifically, our semantic neural field is optimized over a single text prompt at a time, rather than learning general semantic features which could match arbitrary queries. This enables open-vocabulary segmentation, significantly outperforming alternative methods in our setting.

3. Method

An overview of the proposed system is presented in Figure 2. Our goal is to perform text-driven neural 3D localization for landmark scenes captured by collections of Internet photos. In other words, given this collection of images and a text prompt describing a semantic concept in the scene (for example, *windows* or *spires*), we would like to know where it is located in 3D space. These images are *in the wild*, meaning that they may be taken in different seasons,

* *Colonnade* refers to a row of columns separated from each other by an equal distance. A *tympanum* is the semi-circular or triangular decorative wall surface over an entrance, door or window, which is bounded by a lintel and an arch.

Input: ARCHED-WALKWAYS-AT RAJON-KI-BAOLI.JPG "This is a photo of ASI monument number." `Rajon ki Baoli.`
Output: Archways
Input: CATEDRAL-DE-PALMA-DE-MALLORCA,-FACHADA-SUR,-DESDE-EL-PASEO-DE-LA-MURALLA.JPG "Catedral de Palma de Mallorca, fachada sur, desde el Paseo de la Muralla." `mallorca catedral cathedral palma spain mallorca majorca;Exterior of Cathedral of Palma de Mallorca;Cathedral of Palma de Mallorca - Full.`
Output: Facade
Input: SUNDIAL-YENI CAMII2-ISTANBUL.JPG "sundial outside Yeni Camii. On top of the lines the arabic word Asr (afternoon daily prayer) is given. The ten lines (often they are only 9) indicate the times from 20min to 3h before the prayer. Time is read off at the tip of the shadow. The clock was made around 1669 (1074 H)." `New Mosque (Istanbul).`
Output: Sundial

Figure 3: ***LLM-based distillation of semantic concepts****. The full image metadata (**Input**), including* FILENAME*, "*caption*" and* `WikiCategories` *(depicted similarly above) are used for extracting distilled semantic pseudo-labels (**Output**) with an LLM. Note that the associated images on top (depicted with corresponding colors) are not used as inputs for the computation of their pseudo-labels.*

time of day, viewpoints, and distances from the landmark, and may include transient occlusions.

In order to localize unique architectural features landmarks in 3D space, we leverage the power of modern foundation models for visual and textual understanding. Despite progress in general multimodal understanding, modern VLMs struggle to localize fine-grained semantic concepts on architectural landmarks, as we show extensively in our results. The architectural domain uses a specialized vocabulary, with terms such as *pediment* and *tympanum* being rare in general usage; furthermore, terms such as *portal* may have a particular domain-specific meaning in architecture (referring primarily to doors) in contrast to its general usage (meaning any kind of opening).

To address these challenges, we design a three-stage system: the *offline* stages of LLM-based semantic concept distillation (Section 3.1) and semantic adaptation of VLMs (Section 3.2), followed by the *online* stage of 3D localization (Section 3.3). In the offline stages of our method, we learn relevant semantic concepts using textual metadata as guidance by distilling it via an LLM, and subsequently locate these concepts in space by leveraging inter-view correspondences. The resulting fine-tuned image segmentation model is then used in the online stage to supervise the learning of volumetric probabilities—associating regions in 3D space with the probability of depicting the target text prompt.

Training Data The training data for learning the unique semantics of such landmarks is provided by the WikiScenes dataset [WAESS21], consisting of images capturing nearly one hundred *Cathedrals*. We augment these with images capturing 734 *Mosques*, using their data scraping procedure[*]. We also remove all landmarks used in our HolyScenes benchmark (described in Section 4) from this training data to prevent data leakage. The rich data captured in both textual and visual modalities in this dataset, along with large-scale coverage of a diverse set of scenes, provides the needed supervision for our system.

3.1. LLM-Based Semantic Concept Distillation

In order to associate images with relevant semantic categories for training, we use their accompanying textual metadata as weak supervision. As seen in Figure 3, this metadata is highly informative but also noisy, often containing many irrelevant details as well as having diverse formatting and multilingual contents. Prior work has shown that such data can be distilled into categorical labels that provide a supervision signal [WAESS21]; however, this loses the long tail of uncommon and esoteric categories which we are interested in capturing. Therefore, we leverage the power of instruction-tuned large language models (LLMs) for distilling concise, open-ended semantic *pseudo-labels* from image metadata using an instruction alone (i.e. zero-shot, with no ground-truth supervision). In particular, we use the encoder-decoder LLM Flan-T5 [CHL*22], which performs well on tasks requiring short answers and is publicly available (allowing for reproducibility of our results). To construct a prompt for this model, we concatenate together the image's filename, caption, and WikiCategories (*i.e.*, a hierarchy of named categories provided in Wikimedia Commons) into a single description string; we prepend this description with the instruction: "*What architectural feature of* ⟨BUILDING⟩ *is described in the following image? Write "unknown" if it is not specified.*" In this prompt template, the building's name is inserted in ⟨BUILDING⟩ (e.g. *Cologne Cathedral*). We then generate a pseudo-label using beam search decoding, and lightly process these outputs with standard textual cleanup techniques. Out of $\sim 101K$ images with metadata in our train split of WikiScenes, this produces $\sim 58K$ items with non-empty pseudo-labels (those passing filtering heuristics), consisting of 4,031 unique values. Details on text generation settings, textual cleanup heuristics, and further statistics on the distribution of pseudo-labels are provided in the supplementary material.

Qualitatively, we observe that these pseudo-labels succeed in producing concise English pseudo-labels for inputs regardless of distractor details and multilingual data. This matches the excellent performance of LLMs such as Flan-T5 on similar tasks such as text summarization and translation. Several examples of the metadata and our generated pseudo-labels are provided in Figure 3, and a quantitative analysis of pseudo-label quality is given in our ablation study (Section 5.4).

3.2. Semantic Adaptation of V&L Models

After assigning textual pseudo-labels to training images as described in Section 3.1, we use them as supervision for cross-modal

[*] Unlike [WAESS21] that only use images of more common landmarks that can also be reconstructed using *structure-from-motion* techniques, we also include landmarks that are captured by several images only.

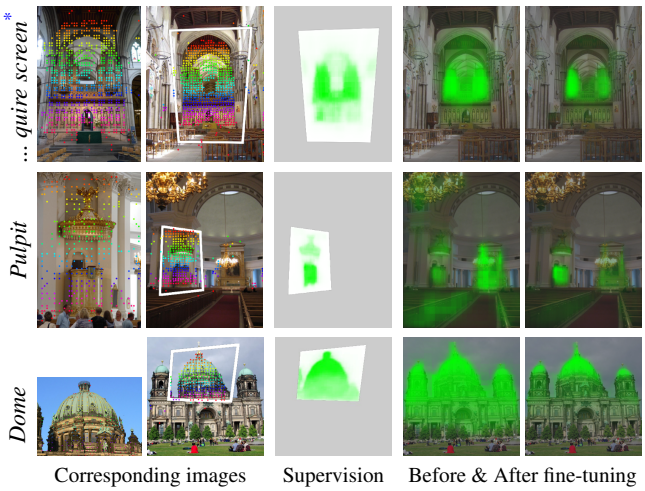


Figure 4: ***Adapting a text-based image segmentation model to architectural landmarks****. We utilize image correspondences (such as the pairs depicted on the left) and pseudo-labels to fine-tune CLIPSeg. We propogate the pseudo-label and pseudo-label of the zoomed-in image to serve as the supervision target, as shown in the central column; we supervise predictions on the zoomed-out image only over the corresponding region (other regions are colored in grayed out for illustration purposes). This supervision (together with using random crops further described in the text) refines the model's ability to recognize and localize architectural concepts, as seen by the improved performance shown on the right.*

understanding, learning image-level and pixel-level semantics. As we show below in Section 5, existing V&L models lack the requisite domain knowledge out of the box, struggling to understand architectural terms or to localize them in images depicting large portions of buildings. We therefore adapt pretrained models to our setting, using image-pseudolabel pairs to learn *image-level* semantics and weak supervision from pairwise image correspondences to bootstrap *pixel-level* semantic understanding. We outline the training procedures of these models here; see the supplementary material for further details.

To learn image-level semantics of unique architectural concepts in our images, we fine-tune the popular foundation model CLIP [RKH*21], a dual encoder model pretrained with a contrastive text-image matching objective. This model encodes images and texts in a shared semantic space, with cross-modal similarity reflected by cosine distance between embeddings. Although CLIP has impressive zero-shot performance on many classification and retrieval tasks, it may be fine-tuned on text-image pairs to adapt it to particular semantic domains. We fine-tune with the standard contrastive learning objective using our pairs of pseudo-labels and images, and denote the resulting refined model by CLIP_{FT}. In addition to being used for further stages in our VLM adaptation pipeline, CLIP_{FT} serves to retrieve relevant terminology for the user who may not be familiar with architectural terms, as we show in our evaluations (Section 5.3).

To apply our textual pseudo-labels and image-level semantics to concept localization, we build on the recent segmentation model CLIPSeg [LE22], which allows for zero-shot text-conditioned image segmentation. CLIPSeg uses image and text features from a CLIP backbone along with additional fusion layers in an added decoder component; trained on text-supervised segmentation data, this shows impressive open-vocabulary understanding on general text prompts. While pretrained CLIPSeg fails to adequately understand architectural concepts or to localize them (as we show in Section 5.4), it shows a basic understanding of some concepts along with a tendency to attend to salient objects (as we further illustrate in the supplementary material), which we exploit to bootstrap understanding in our setting.

Our key observation is that large and complex images are composed of subregions with different semantics (e.g. the region around a window or portal of a building), and pretrained CLIPSeg predictions on these zoomed-in regions are closer to the ground truth than predictions on the entire building facade. To find such pairs of zoomed-out and zoomed-in images, we use two types of geometric connections: multi-view geometric correspondences (i.e. *between images*) and image crops (i.e. *within images*). Using these paired images and our pseudo-label data, we use predictions on zoomed-in views as supervision to refine segmentation on zoomed-out views.

For training across multiple images, we use a feature matching model [SSW*21] to find robust geometric correspondences between image pairs and CLIP_{FT} to select pairs where the semantic concept (given by a pseudo-label) is more salient in the zoomed-in view relative to the zoomed-out view; for training within the same image, we use CLIP_{FT} to select relevant crops. We use pretrained CLIPSeg to segment the salient region in the zoomed-in or cropped image, and then fine-tune CLIPSeg to produce this result in the relevant image when zoomed out; we denote the resulting trained model by $\text{CLIPSeg}_{\text{FT}}$. During training we freeze CLIPSeg's encoders, training its decoder module alone with loss functions optimizing for the following:

Geometric correspondence supervision losses. As described above, we use predictions on zoomed-in images to supervise segmentation of zoomed-out views. We thus define loss terms $\mathcal{L}_{corresp}$ and $\mathcal{L}_{crop}$, the cross-entropy loss of these predictions calculated on the region with supervision targets, for correspondence-based and crop-based data respectively. In other words, $\mathcal{L}_{corresp}$ encourages predictions on zoomed-out images to match predictions on corresponding zoomed-in views as seen in Figure 4; $\mathcal{L}_{crop}$ is similar but uses predictions on a crop of the zoomed-out view rather than finding a distinct image with a corresponding zoomed-in view.

Multi-resolution consistency. To encourage consistent predictions across resolutions and to encourage our model to attend to relevant details in all areas of the image, we use a multi-resolution consistency loss $\mathcal{L}_{consistency}$ calculated as follows. Selecting a random crop of the image from the correspondence-based dataset, we calculate cross-entropy loss between our model's prediction cropped to this region, and CLIPSeg (pretrained, without fine-tuning) applied within this cropped region. To attend to more relevant crops,

* Full pseudo-label text: *Neo-gothic quire screen*. This refers to a screen that partitions the choir (or quire) and the aisles in a cathedral or a church.

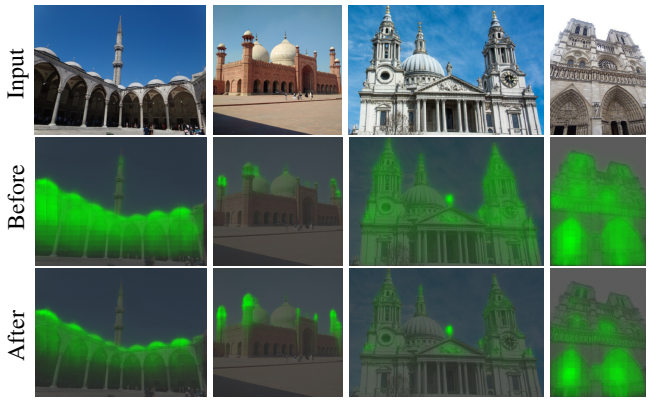


Figure 5: ***Text-based segmentation before and after fine-tuning.*** *Above we show 2D segmentation results over images belonging to landmarks from HolyScenes (unseen during training). As illustrated above, our weakly-supervised fine-tuning scheme improves the segmentation of domain-specific semantic concepts.*

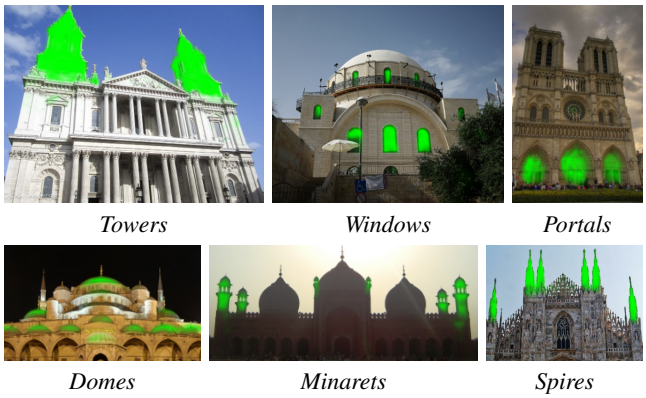


Figure 6: ***Neural 3D Localization Results.*** *We show results from each landmark in our HolyScenes benchmark (clockwise from top: St. Paul's Cathedral, Hurva Synagogue, Notre-Dame Cathedral, Blue Mosque, Badshahi Mosque, Milan Cathedral), visualizing segmentation maps rendered from 3D HaLo-NeRF representations on input scene images. As seen above, HaLo-NeRF succeeds in localizing various semantic concepts across diverse landmarks.*

we pick the random crop by sampling two crops from the given image and using the one with higher CLIP_{FT} similarity to the textual pseudo-label.

Regularization. We add the regularization loss $\mathcal{L}_{reg}$, calculated as the average binary entropy of our model's outputs. This encourages confident outputs (probabilities close to 0 or 1).

These losses are summed together with equal weighting; further training settings, hyperparameters, and data augmentation are detailed in the supplementary material.

We illustrate this fine-tuning process over corresponding image pairs in Figure 4. As illustrated in the figure, the leftmost images (*i.e.*, zoom-ins) determine the supervision signal. Note that while we only supervise learning in the corresponding region in each training sample, the refined model (denoted as $\text{CLIPSeg}_{\text{FT}}$) correctly extrapolates this knowledge to the rest of the zoomed-out image. Figure 5 illustrates the effect of this fine-tuning on segmentation of new landmarks (unseen during training); we see that our fine-tuning gives $\text{CLIPSeg}_{\text{FT}}$ knowledge of various semantic categories that the original pretrained CLIPSeg struggles to localize; we proceed to use this model to produce 2D segmentations that may be lifted to a 3D representation.

3.3. Text-Driven Neural 3D Localization

In this section, we describe our approach for performing 3D localization over a neural representation of the scene, using the semantic understanding obtained in the previous offline training stages. The input to our 3D localization framework is an Internet image collection of a new (unseen) landmark and a target text prompt.

First, we optimize a Ha-NeRF [CZL*22] representation to learn volumetric densities and colors from the unstructured image collection. We then extend this neural representation by adding a semantic output channel. Inspired by previous work connecting neural radiance fields with semantics [ZLLD21], we augment Ha-NeRF with a segmentation MLP head, added on top of a shared backbone (see the supplementary material for additional details). To learn the volumetric probabilities of given target text prompt, we freeze the shared backbone and optimize only the segmentation MLP head.

To provide supervision for semantic predictions, we use the 2D segmentation map predictions of $\text{CLIPSeg}_{\text{FT}}$ (described in Section 3.2) on each input view. While these semantically adapted 2D segmentation maps are calculated for each view separately, HaLo-NeRF learns a 3D model which aggregates these predictions while enforcing 3D consistency. We use a binary cross-entropy loss to optimize the semantic volumetric probabilities, comparing them to the 2D segmentation maps over sampled rays [ZLLD21]. This yields a a representation of the semantic concept's location in space. Novel rendered views along with estimated probabilities are shown in Figures 1 and 2 and in the accompanying videos.

4. The HolyScenes Benchmark

To evaluate our method, we need Internet photo collections covering scenes, paired with ground truth segmentation maps. As we are not aware of any such existing datasets, we introduce the *HolyScenes* benchmark, assembled from multiple datasets (WikiScenes [WAESS21], IMC-PT 2020 [Yi20] MegaDepth [LS18]) along with additional data collected using the data scraping procedure of Wu *et al*. We enrich these scene images with ground-truth segmentation annotations. Our dataset includes 6,305 images associated with 3D structure-from-motion reconstructions and ground-truth segmentations for multiple semantic categories.

We select six landmarks, exemplified in Figure 6: *Notre-Dame Cathedral* (Paris), *Milan Cathedral* (Milan), *St. Paul's Cathedral* (London), *Badshahi Mosque* (Lahore), *Blue Mosque* (Istanbul) and *Hurva Synagogue* (Jerusalem). These landmarks span different geographical regions, religions and characteristics, and can readily be associated with accurate 3D reconstructions due to the large

number of publicly-available Internet images. We associate these landmarks with the following semantic categories: *portal*, *window*, *spire*, *tower*, *dome*, and *minaret*. Each landmark is associated with a subset of these categories, according to its architectural structure (*e.g.*, *minaret* is only associated with the two mosques in our benchmark).

We produce ground-truth segmentation maps to evaluate our method using manual labelling combined with correspondence-guided propagation. For each semantic concept, we first manually segment several images from different landmarks. We then propagate these segmentation maps to overlapping images, and manually filter these propagated masks (removing, for instance, occluded images). Additional details about our benchmark are provided in the supplementary material.

5. Results and Evaluation

In this section, we evaluate the performance of HaLo-NeRF on the HolyScenes benchmark, and compare our method to recent works on text-guided semantic segmentation and neural localization techniques. We also validate each component of our system with ablation studies – namely, our LLM-based concept distillation, VLM semantic adaptation, and 3D localization. Finally, we discuss limitations of our approach. In the supplementary material, we provide experimental details as well as additional experiments, such as an evaluation of the effect of CLIPSeg fine-tuning on general and architectural term understanding evaluated on external datasets.

5.1. Baselines

We compare our method to text-driven image segmentation methods, as well as 3D NeRF segmentation techniques. As HolyScenes consists of paired images and view-consistent segmentation maps, it can be used to evaluate both 2D and 3D segmentation methods; in the former case, by directly segmenting images and evaluating on their ground-truth (GT) annotations; in the latter case, by rendering 2D segmentation masks from views corresponding to each GT annotation.

For text-based 2D segmentation baseline methods, we consider CLIPSeg [LE22] and LSeg [LWB*22]. We also compare to the ToB model proposed by Wu *et al.* [WAESS21] that learns image segmentation over the WikiScenes dataset using cross-view correspondences as weak supervision. As their model is categorical, operating over only ten categories, we report the performance of ToB only over the semantic concepts included in their model.

For 3D NeRF-based segmentation methods, we consider DFF [KMS22] and LERF [KKG*23]. Both of these recent methods utilize text for NeRF-based 3D semantic segmentation. DFF [KMS22] performs semantic scene decomposition using text prompts, distilling text-aligned image features into a volumetric 3D representation and segmenting 3D regions by probing these with the feature representation of a given text query. Similarly, LERF optimizes a 3D language field from multi-scale CLIP embeddings with volume rendering.

The publicly available implementations of DFF and LERF cannot operate on our *in-the-wild* problem setting, as it does not have images with constant illumination or a single camera model. To provide a fair comparison, we replace the NeRF backbones used by DFF and LERF (vanilla NeRF and Nerfacto respectively) with Ha-NeRF, as used in our model, keeping the remaining architecture of these models unchanged. In the supplementary material, we also report results over the unmodified DFF and LERF implementations using constant illumination images rendered from Google Earth.

In addition to these existing 3D methods, we compare to the baseline approach of lifting 2D CLIPSeg (pretrained, not fine-tuned) predictions to a 3D representation with Ha-NeRF augmented with a localization head (as detailed in Section 3.3). This baseline, denoted as HaLo-NeRF-, provides a reference point for evaluating the relative contribution of our optimization-based approach rather than learning a feature field which may be probed for various textual inputs (as used by competing methods), and of our 2D segmentation fine-tuning.

Method	mAP	portal	window	spire	tower	dome	minaret
2D Seg.							
LSeg	0.09	0.05	0.13	0.06	0.19	0.05	0.06
ToB	0.23	0.15	0.04	×	0.49	×	×
CLIPSeg	0.56	0.29	0.44	0.46	0.87	0.69	0.63
CLIPSeg$_{FT}$	**0.66**	**0.49**	**0.51**	**0.50**	**0.87**	**0.77**	**0.81**
3D Loc.							
DFF*	0.11	0.06	0.04	0.09	0.17	0.11	0.17
LERF*	0.14	0.16	0.15	0.18	0.13	0.10	0.09
HaLo-NeRF-	0.62	0.28	0.61	0.55	**0.90**	0.72	0.69
HaLo-NeRF	**0.68**	**0.45**	**0.64**	**0.61**	0.87	**0.74**	**0.80**

*Using a Ha-NeRF backbone -Using CLIPSeg without fine-tuning

Table 1: ***Quantitative Evaluation***. *We report mean average precision (mAP; averaged per category) and per category average precision over the HolyScenes benchmark, comparing our results (highlighted in the table) to 2D segmentation and 3D localization techniques. Note that ToB uses a categorical model, and hence we only report performance over concepts it was trained on. Best results are highlighted in* ***bold****.*

5.2. Quantitative Evaluation

As stated in Section 5.1, our benchmark allows us to evaluate segmentation quality for both both 2D and 3D segmentation methods, in the latter case by projecting 3D predictions onto 2D views with ground-truth segmentation maps. We perform our evaluation using pixel-wise metrics relative to ground-truth segmentations. Since we are interested in the quality of the model's soft probability predictions, we use average precision (AP) as our selected metric as it is threshold-independent.

In Table 1 we report the AP per semantic category (averaged over landmarks), as well as the overall mean AP (mAP) across categories. We report results for 2D image segmentation models on top, and 3D segmentation methods underneath. In addition to reporting 3D localization results for our full proposed system, we also report the results of our intermediate 2D segmentation component (CLIPSeg$_{FT}$).

As seen in the table, $\text{CLIPSeg}_{\text{FT}}$ (our fine-tuned segmentation model, as defined in Section 3.2) outperforms other 2D methods, showing better knowledge of architectural concepts and their localization. In addition to free-text guided methods (LSeg and CLIPSeg), we also outperform the ToB model (which was trained on WikiScenes), consistent with the low recall scores reported by Wu *et al*. [WAESS21]. LSeg also struggles in our free-text setting where semantic categories strongly deviate from its training data; CLIPSeg shows better zero-shot understanding of our concepts out of the box, but still has a significance performance gap relative to $\text{CLIPSeg}_{\text{FT}}$.

In the 3D localization setting, we also see that our method strongly outperforms prior methods over all landmarks and semantic categories. HaLo-NeRF adds 3D-consistency over $\text{CLIPSeg}_{\text{FT}}$ image segmentations, further boosting performance by fusing predictions from multi-view inputs into a 3D representation which enforces consistency across observations. We also find an overall performance boost relative to the baseline approach using HaLo-NeRF without CLIPSeg fine-tuning. This gap is particularly evident in unique architectural terms such as *portal* and *minaret*.

Regarding the gap between our performance and the competing 3D methods (DFF, LERF), we consider multiple contributing factors. In addition to our enhanced understanding of domain-specific semantic categories and their positioning, the designs of these models differ from HaLo-NeRF in ways which may impact performance. DFF is built upon LSeg as its 2D backbone; hence, its performance gap on our benchmark follows logically from the poor performance of LSeg in this setting (as seen in the reported 2D results for LSeg), consistent with the observation of Kobayashi *et al*. [KMS22] that DFF inherits bias towards in-distribution semantic categories from LSeg (e.g. for traffic scenes). LERF, like DFF, regresses a full semantic 3D feature field which may then be probed for arbitrary text prompts. By contrast, HaLo-NeRF optimizes for the more modest task of localizing a particular concept in space, likely more feasible in this challenging setting. The significant improvement provided by performing per-concept optimization is also supported by the relatively stronger performance of the baseline model shown in Table 1, which performs this optimization using pretrained (not fine-tuned) CLIPSeg segmentation maps as inputs.

5.3. Qualitative Results

Sample results of our method are provided in Figures 6–11. As seen in Figure 6, HaLo-NeRF segments regions across various landmarks and succeeds in differentiating between fine-grained architectural concepts. Figure 7 compares these results to alternate 3D localization methods. As seen there, alternative methods fail to reliably distinguish between the different semantic concepts, tending to segment the entire building facade rather than identifying the areas of interest. With LERF, this tendency is often accompanied by higher probabilities in coarsely accurate regions, as seen by the roughly highlighted windows in the middle row. Figure 8 shows a qualitative comparison of HaLo-NeRF with and without CLIPSeg fine-tuning over additional semantic concepts beyond those from our benchmark. As is seen there, our fine-tuning procedure is needed to learn reliable localization of such concepts which may be lifted to 3D.

We include demonstrations of the generality of our method. Besides noting that our test set includes the *synagogue* category which was not seen in training (see the results for the Hurva Synagogue shown in Figure 6), we test our model in the more general case of (non-religious) architectural landmarks. Figure 11 shows results on various famous landmarks captured in the IMC-PT 2020 dataset [Yi20] (namely, Brandenburg Gate, Palace of Westminster, The Louvre Museum, Park Güell, The Statue of Liberty, Las Vegas, The Trevi Fountain, The Pantheon, and The Buckingham Palace). As seen there, HaLo-NeRF localizes unique scene elements such as the *quadriga* in the Brandenburg Gate, the Statue of Liberty's torch, and the Eiffel tower, The Statue of Liberty, and Las Vegas, respectively. In addition, HaLo-NeRF localizes common semantic concepts, such as *clock*, *glass*, and *text* in the Palace of Westminster, The Louvre Museum, and The Pantheon, respectively. Furthermore, while we focus mostly on outdoor scenes, Figure 9 shows that our method can also localize semantic concepts over reconstructions capturing indoor scenes.

Understanding that users may not be familiar with fine-grained or esoteric architectural terminology, we anticipate the use of CLIP_{FT} (our fine-tuned CLIP model, as defined in Section 3.2) for retrieving relevant terminology. In particular, CLIP_{FT} may be applied to any selected view to retrieve relevant terms to which the user may then apply HaLo-NeRF. We demonstrate this qualitatively in Figure 10, which shows the top terms retrieved by CLIP_{FT} on test images. In the supplementary material, we also report a quantitative evaluation over all architectural terms found at least 10 times in the training data. This evaluation further demonstrates that CLIP_{FT} can retrieve relevant terms over these Internet images (significantly outperforming pretrained CLIP at this task).

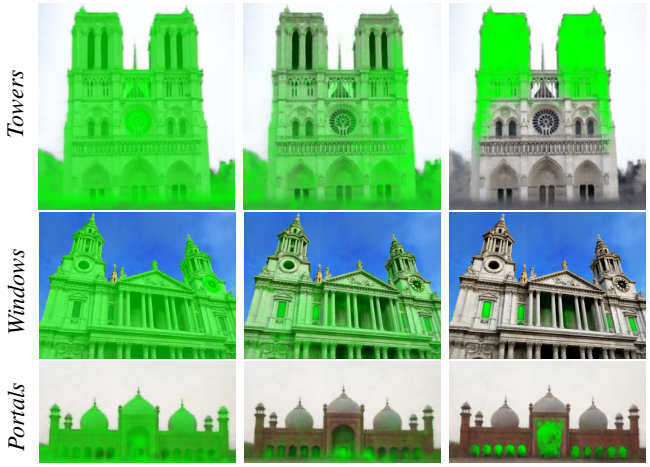
Figure 11 further illustrates the utility of our method for intuitive exploration of scenes. By retrieving scene images having maximal overlap with localization predictions, the user may focus automatically on the text-specified region of interest, allowing for exploration of the relevant semantic regions of the scene in question. This is complementary to exploration over the optimized neural representation, as illustrated in Figures 1-2, and in the accompanying videos.

5.4. Ablation Studies

We proceed to evaluate the contribution of multiple components of our system—LLM-based concept distillation and VLM semantic adaptation—to provide motivation for the design of our full system.

LLM-based Concept Distillation. In order to evaluate the quality of our LLM-generated pseudo-labels and their necessity, we manually review a random subset of 100 items (with non-empty pseudo-labels), evaluating their factual correctness and comparing them to

* *Colonnade* refers to a row of columns separated from each other by an equal distance. *Pediment* refers to a triangular part at the top of the front of a building that supports the roof and is often decorated.
* A *roundel* is an circular shield or figure; here it refers to round panels bearing calligraphic emblems.



*Using a Ha-NeRF backbone

Figure 7: ***Localizing semantic regions in architectural landmarks compared to prior work.*** *We show probability maps for DFF and LERF models on* Milan Cathedral*, along with our results. As seen above, DFF and LERF struggle to distinguishing between different semantic regions on the landmark, while our method accurately localizes the semantic concepts.*

Figure 8: ***3D localization results on additional concepts****, comparing HaLo-NeRF to the baseline HaLo-NeRF- model (using CLIPSeg without fine-tuning as input to HaLo-NeRF) over semantic regions appearing on the* Hurva Synagogue *(left) and* St. Paul's Cathedral *(right). Our model can localize these concepts, while the baseline model fails to reliably distinguish between relevant and irrelevant regions. See below for the definitions of the concepts shown* [*].

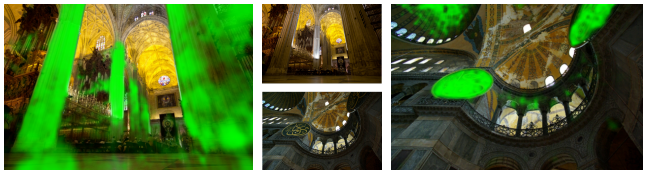


Figure 9: ***Results over indoor scenes.*** *HaLo-NeRF is capable of localizing unique semantic regions within building interiors (shown above over the* Seville Cathedral *and* Blue Mosque *landmarks). The definition of* roundel *is given below* [*].

CLIP_{FT} Img→Text Results

construction	*muqarnas*
scaffolding	*ornate*
bell tower	*stucco decoration*
crossing tower	*dome chamber*
church tower	*tile work*
clock tower	*mihrab*
sail tower	*ceiling tile work*
flying buttresses	*winter prayer hall*

Figure 10: ***Examples of terminology retrieval.*** *By applying CLIP_{FT} to a given view, the user may retrieve relevant architectural terminology which can then be localized with HaLo-NeRF. Above, we display the top eight retrieval results for two test images, using the CLIP_{FT} retrieval methodology described in Section 5.3. As is seen above, CLIP_{FT} returns relevant items such as* scaffolding, church tower, muqarnas, ceiling tile work *which may aid the user in selecting relevant architectural terms.*

Method	mAP	portal	window	spire	tower	dome	minaret
$\text{CLIPSeg}_{\text{FT}}$	**0.66**	**0.49**	**0.51**	**0.50**	0.87	**0.77**	**0.81**
$-\mathcal{L}_{crop}$	0.65	**0.49**	0.50	0.47	**0.88**	**0.77**	0.78
$-\mathcal{L}_{reg}$	0.63	0.45	0.46	0.45	0.87	**0.77**	0.78
−corr. data*	0.45	0.09	0.27	0.25	0.76	0.65	0.71
2D Baseline	0.56	0.29	0.44	0.46	0.87	0.69	0.63

* Refers to removing correspondence supervision losses, namely $\mathcal{L}_{corresp}$ and $\mathcal{L}_{consistency}$.

Table 2: ***Ablation Studies****, evaluating the effect of design choices on the fine-tuning process of CLIPSeg_{FT}. "Baseline" denotes using the CLIPSeg segmentation model without fine-tuning. We report AP and mAP metrics over the HolyScenes benchmark as in Table 1. Best results are highlighted in* ***bold****.*

two metadata-based baselines – whether the correct architectural feature is present in the image's caption, and whether it could be inferred from the last WikiCategory listed in the metadata for the corresponding image (see Section 3.1 for an explanation of this metadata). These baselines serve as upper bounds for architectural feature inference using the most informative metadata fields by themselves (and assuming the ability to extract useful labels from them). We find 89% of pseudo-labels to be factually correct, while only 43% of captions contain information implying the correct architectural feature, and 81% of the last WikiCategories to describe said features. We conclude that our pseudo-labels are more informative than the baseline of using the last WikiCategory, and significantly more so than inferring the architectural feature from the image caption. Furthermore, using either of the latter alone would still require summarizing the text to extract a usable label, along with translating a large number of results into English.

To further study the effect of our LLM component on pseudo-labels, we provide ablations on LLM sizes and prompts in the supplementary material, finding that smaller models underperform ours while the best-performing prompts show similar results. There we also provide statistics on the distribution of our pseudo-labels,

Figure 11: ***Localization for general architectural scenes.*** *HaLo-NeRF can localize various semantic concepts in a variety of scenes in the wild, not limited to the religious domain of HolyScenes. Our localization, marked in green in the first image for each concept, enables focusing automatically on the text-specified region of interest, as shown by the following zoomed-in images in each row.*

showing that they cover a diverse set of categories with a long tail of esoteric items.

VLM Semantic Adaptation Evaluation. To strengthen the motivation behind our design choices of $\text{CLIPSeg}_{\text{FT}}$, we provide an ablation study of the segmentation fine-tuning in Table 2. We see that each element of our training design provides a boost in overall performance, together significantly outperforming the 2D baseline segmentation model. In particular, we see the key role of our correspondence-based data augmentation, without which the fine-tuning procedure significantly degrades due to lack of grounding in the precise geometry of our scenes (both relative to full fine-tuning, and relative to the original segmentation model). These results complement Figure 5, which show a qualitative comparison of the CLIPSeg baseline and $\text{CLIPSeg}_{\text{FT}}$. We also note that we have provided a downstream evaluation of the effect of fine-tuning CLIPSeg on 3D localization in Table 1, showing that it provides a significant performance boost and is particularly crucial for less common concepts.

5.5. Limitations

As our method uses an optimization-based pipeline applied for each textual query, it is limited by the runtime required to fit each term's segmentation field. In particular, a typical run takes roughly two hours on our hardware setup, described in the supplementary material. We foresee future work building upon our findings to accelerate these results, possibly using architectural modifications such as encoder-based distillation of model predictions.

Furthermore, if the user inputs a query which does not appear in the given scene, our model may segment semantically- or geometrically-related regions – behavior inherited from the base segmentation model. For example, the spires of Milan Cathedral are segmented when the system is prompted with the term *minarets*, which are not present in the view but bear visual similarity to spires. Nevertheless, CLIP_{FT} may provide the user with a vocabulary of relevant terms (as discussed in Section 5.3), mitigating this issue (e.g. *minarets* does not appear among the top terms for images depicting Milan Cathedral). We further discuss this tendency to segment salient, weakly-related regions in the supplementary material.

Additionally, since we rely on semantic concepts that appear across landmarks in our training set, concepts require sufficient coverage in this training data in order to be learned. While our method is not limited to common concepts and shows understanding of concepts in the long tail of the distribution of pseudo-labels (as analyzed in the supplementary material), those that are extremely rare or never occur in our training data may not be properly identified. This is seen in Figure 12, where the localization of the

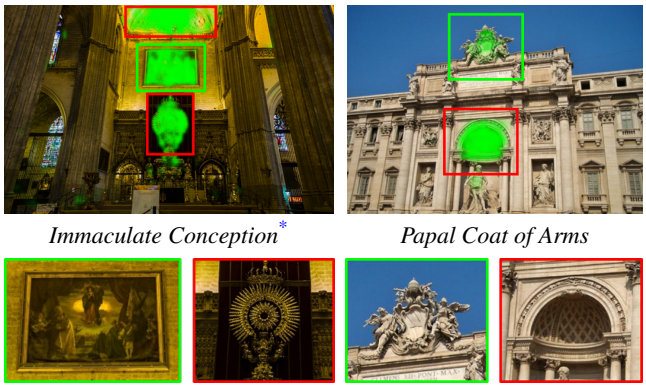


Figure 12: ***Limitation examples.*** *Correct results are marked in green boxes and incorrect ones in red. Our method may fail to properly identify terms that never appear in our training data, such as the* Immaculate Conception* *as on the left and the* Papal Coat of Arms *as on the right.*

scene-specific concepts *Immaculate Conception* and *Papal Coat of Arms* (terms which never occur in our training data; for example, the similar term *coat of arms* appears only seven times) incorrectly include other regions.

6. Conclusions

We have presented a technique for connecting unique architectural elements across different modalities of text, images, and 3D volumetric representations of a scene. To understand and localize domain-specific semantics, we leverage inter-view coverage of a scene in multiple modalities, distilling concepts with an LLM and using view correspondences to bootstrap spatial understanding of these concepts. We use this knowledge as guidance for a neural 3D representation which is view-consistent by construction, and demonstrate its performance on a new benchmark for concept localization in large-scale scenes of tourist landmarks.

Our work represents a step towards the goal of modeling historic and culturally significant sites as explorable 3D models from photos and metadata captured in the wild. We envision a future where these compelling sites are available to all in virtual form, making them accessible and offering educational opportunities that would not otherwise be possible. Several potential research avenues include making our approach interactive, localizing multiple prompts simultaneously and extending our technique to additional mediums with esoteric concepts, such as motifs or elements in artwork.

Acknowledgements

This work was supported by research grants from ISF (application number 2510/23) and BSF (application number 2022363). [Correction added on 29 April 2024, after first online publication: Acknowledgment section has been added.]

* The *Immaculate Conception* is the event depicted in the painting, a work by Alfonso Grosso Sánchez situated in the Seville Cathedral.

References

[CCN20] CHEN D. Z., CHANG A. X., NIESSNER M.: ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2020), pp. 202–221. 3

[CGT*22] CHEN S., GUHUR P.-L., TAPASWI M., SCHMID C., LAPTEV I.: Language conditioned spatial relation reasoning for 3d object grounding. *arXiv preprint arXiv:2211.09646* (2022). 2, 3

[CHL*22] CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI E., WANG X., DEHGHANI M., BRAHMA S., ET AL.: Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416* (2022). 4

[CLW*22] CHEN J., LUO W., WEI X., MA L., ZHANG W.: Ham: Hierarchical attention model with high performance for 3d visual grounding. *arXiv preprint arXiv:2210.12513* (2022). 2, 3

[CWNC22] CHEN D. Z., WU Q., NIESSNER M., CHANG A. X.: D3Net: A Speaker-Listener Architecture for Semi-supervised Dense Captioning and Visual Grounding in RGB-D Scans. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2022). 3

[CZL*22] CHEN X., ZHANG Q., LI X., CHEN Y., FENG Y., WANG X., WANG J.: Hallucinated Neural Radiance Fields in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 12943–12952. 2, 6

[DLH22] DECATUR D., LANG I., HANOCKA R.: 3d highlighter: Localizing regions on 3d shapes via text descriptions. *arXiv preprint arXiv:2212.11263* (2022). 2

[DXXD22] DING J., XUE N., XIA G.-S., DAI D.: Decoupling Zero-Shot Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 11583–11592. 2

[FWJ*22] FAN Z., WANG P., JIANG Y., GONG X., XU D., WANG Z.: Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776* (2022). 3

[FZC*22] FU X., ZHANG S., CHEN T., LU Y., ZHU L., ZHOU X., GEIGER A., LIAO Y.: Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *International Conference on 3D Vision (3DV)* (2022). 3

[GGCL22] GHIASI G., GU X., CUI Y., LIN T.-Y.: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2022), pp. 540–557. 2

[HCJW22] HUANG S., CHEN Y., JIA J., WANG L.: Multi-view transformer for 3d visual grounding. In *CVPR* (2022). 2, 3

[IMK20] IQBAL U., MOLCHANOV P., KAUTZ J.: Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 5243–5252. 2

[JYX*21] JIA C., YANG Y., XIA Y., CHEN Y.-T., PAREKH Z., PHAM H., LE Q., SUNG Y.-H., LI Z., DUERIG T.: Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)* (2021), pp. 4904–4916. 2

[KGY*22] KUNDU A., GENOVA K., YIN X., FATHI A., PANTOFARU C., GUIBAS L. J., TAGLIASACCHI A., DELLAERT F., FUNKHOUSER T.: Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 12871–12881. 3

[KKG*23] KERR J., KIM C. M., GOLDBERG K., KANAZAWA A., TANCIK M.: Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 19729–19739. 2, 3, 7

[KMS22] KOBAYASHI S., MATSUMOTO E., SITZMANN V.: Decomposing NeRF for Editing via Feature Field Distillation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 2, 3, 7, 8

[LE22] LÜDDECKE T., ECKER A.: Image Segmentation Using Text and Image Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 7086–7096. 2, 5, 7

[LS18] LI Z., SNAVELY N.: Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2041–2050. 6

[LWB*22] LI B., WEINBERGER K. Q., BELONGIE S., KOLTUN V., RANFTL R.: Language-Driven Semantic Segmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2022). 2, 7

[LWD*23] LIANG F., WU B., DAI X., LI K., ZHAO Y., ZHANG H., ZHANG P., VAJDA P., MARCULESCU D.: Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 7061–7070. 2

[LXW*23] LU Y., XU C., WEI X., XIE X., TOMIZUKA M., KEUTZER K., ZHANG S.: Open-vocabulary point-cloud object detection without 3d annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 1190–1199. 3

[MBRS*21] MARTIN-BRUALLA R., RADWAN N., SAJJADI M. S., BARRON J. T., DOSOVITSKIY A., DUCKWORTH D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7210–7219. 2

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2020), pp. 405–421. 3

[PGJ*22] PENG S., GENOVA K., JIANG C., TAGLIASACCHI A., POLLEFEYS M., FUNKHOUSER T., ET AL.: OpenScene: 3D Scene Understanding with Open Vocabularies. *arXiv preprint arXiv:2211.15654* (2022). 3

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G., SUTSKEVER I.: Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)* (2021), pp. 8748–8763. 2, 5

[RLD22] ROZENBERSZKI D., LITANY O., DAI A.: Language-Grounded Indoor 3D Semantic Segmentation in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2022), pp. 125–141. 3

[RMBB*13] RUSSELL B. C., MARTIN-BRUALLA R., BUTLER D. J., SEITZ S. M., ZETTLEMOYER L.: 3d wikipedia: Using online text to automatically label and navigate reconstructed geometry. *ACM Transactions on Graphics (TOG) 32*, 6 (2013), 1–10. 3

[SGSS08] SNAVELY N., GARG R., SEITZ S. M., SZELISKI R.: Finding paths through the world's photos. *ACM Transactions on Graphics (TOG) 27*, 3 (2008), 1–11. 3

[SPB*22] SIDDIQUI Y., PORZI L., BULÓ S. R., MÜLLER N., NIESSNER M., DAI A., KONTSCHIEDER P.: Panoptic Lifting for 3D Scene Understanding with Neural Fields. *arXiv preprint arXiv:2212.09802* (2022). 3

[SSS06] SNAVELY N., SEITZ S. M., SZELISKI R.: Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers* (2006), pp. 835–846. 3

[SSW*21] SUN J., SHEN Z., WANG Y., BAO H., ZHOU X.: Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 8922–8931. 5

[TLLV22] TSCHERNEZKI V., LAINA I., LARLUS D., VEDALDI A.: Neural Feature Fusion Fields: 3D Distillation of Self-Supervised 2D Image Representations. In *Proceedings of the International Conference on 3D Vision (3DV)* (2022). 3

[TZFR23] TURKI H., ZHANG J. Y., FERRONI F., RAMANAN D.: Suds: Scalable urban dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12375–12385. 3

[WAESS21] WU X., AVERBUCH-ELOR H., SUN J., SNAVELY N.: Towers of Babel: Combining Images, Language, and 3D Geometry for Learning Multimodal Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 428–437. 2, 3, 4, 6, 7, 8

[WZHS20] WANG Q., ZHOU X., HARIHARAN B., SNAVELY N.: Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (2020), Springer, pp. 757–774. 2

[XDML*22] XU J., DE MELLO S., LIU S., BYEON W., BREUEL T., KAUTZ J., WANG X.: GroupViT: Semantic Segmentation Emerges from Text Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 18134–18144. 2

[XZW*21] XU M., ZHANG Z., WEI F., LIN Y., CAO Y., HU H., BAI X.: A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-language Model. *arXiv preprint arXiv:2112.14757* (2021). 2

[Yi20] YI K. M.: Image matching: Local features & beyond 2020. https://www.cs.ubc.ca/ kmyi/imw2020/data.html, 2020. https://www.cs.ubc.ca/ kmyi/imw2020/data.html. URL: https://www.cs.ubc.ca/~kmyi/imw2020/data.html, arXiv:https://www.cs.ubc.ca/~kmyi/imw2020/data.html. 6, 8

[ZLD22] ZHOU C., LOY C. C., DAI B.: Extract Free Dense Labels from CLIP. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2022), pp. 696–712. 2

[ZLLD21] ZHI S., LAIDLOW T., LEUTENEGGER S., DAVISON A. J.: In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)* (2021), pp. 15838–15847. 3, 6