Reconstructing Animals *and* the Wild

Peter Kulits[1] Michael J. Black[1] Silvia Zuffi[2]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]IMATI-CNR, Milan, Italy
{kulits,black}@tue.mpg.de, silvia.zuffi@cnr.it

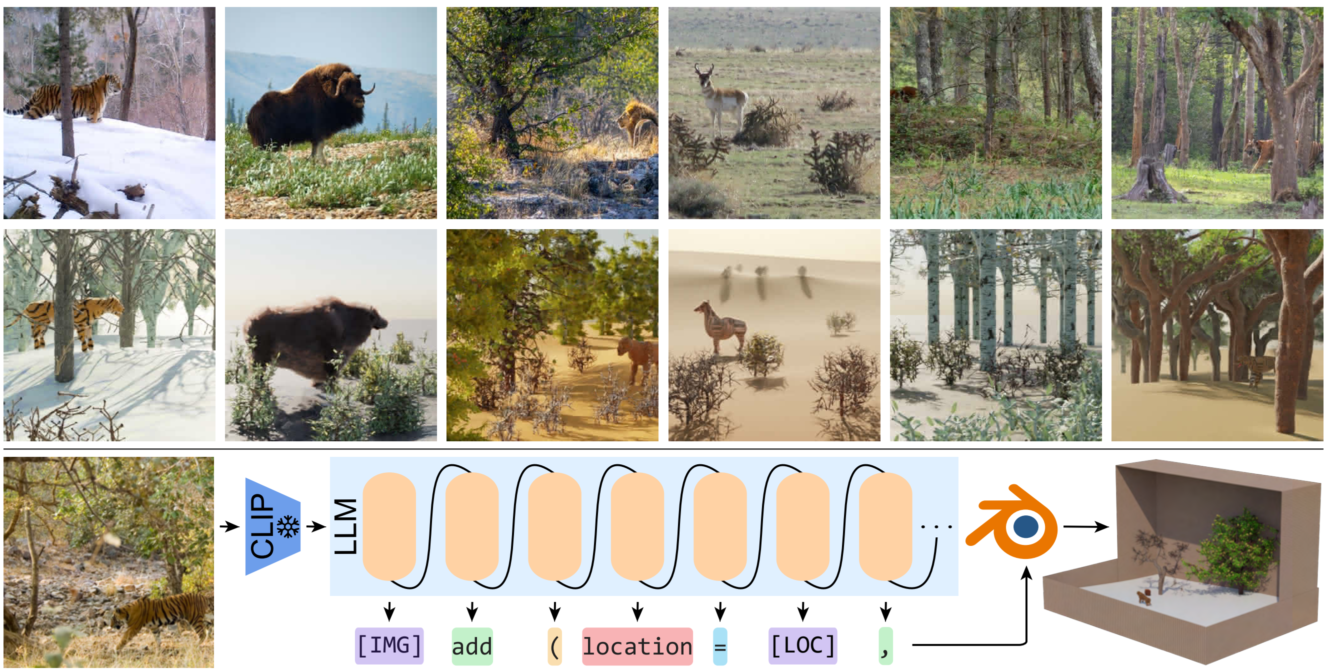


Figure 1. We train an LLM to decode a frozen CLIP embedding of a natural image into a structured, compositional scene representation encompassing explicit representations of both animals and their surrounding environment, reconstructing animals *and* the wild (RAW).

Abstract

The notion of 3D reconstruction as scene understanding is foundational in computer vision. Reconstructing 3D scenes from 2D visual observations necessitates strong priors to disambiguate structure. Much work has been focused on the anthropocentric, which, characterized by smooth surfaces, coherent normals, and regular edges, allows for the integration of strong geometric inductive biases. Here, we consider a more challenging problem where such assumptions do not hold: the reconstruction of natural scenes containing trees, bushes, boulders, and animals. While numerous works have attempted to tackle the problem of reconstructing animals in the wild, they have focused solely on the animal, neglecting environmental context. This limits their usefulness for analysis tasks, as animals exist inherently within the 3D world, and information is lost when environmental factors are disregarded. We propose a method to reconstruct natural scenes from single images. We base our approach on recent advances leveraging the strong world priors ingrained in Large Language Models and train an autoregressive model to decode a CLIP embedding into a structured, compositional scene representation, encompassing both animals and the wild. To enable this, we propose a synthetic dataset comprising one million images and thousands of assets. Through our introduction of a CLIP-projection head, we demonstrate that our approach generalizes to the task of reconstructing animals and their environments in real-world images, despite having been trained solely on synthetic data. We release our dataset and code to encourage future research at https://raw.is.tue.mpg.de/.

1. Introduction

The 3D reconstruction of the physical world from visual observations plays a fundamental role in computer vision, providing the foundation for Marr and Nishihara [56]'s computational model of visual perception. The process culminates in a structured 3D representation of the environment. Compositional 3D reconstructions of scenes, where objects are distinguished into semantic classes, are particularly amenable to analysis, enabling editing and simulation. When such reconstructions are represented in a compact form that encapsulates and can be used to reproduce the scene, they become expressive models of 3D reality, supporting applications in modeling of physical behavior [47].

Recent work has built on developments in Large Language Models (LLMs) to reconstruct simple scenes composed of a few objects from a monocular image by inferring graphics code [41], or to reproduce architectural layouts of indoor scenes represented in an ad-hoc vector format [4].

We consider a more challenging problem, the reconstruction of outdoor natural scenes containing diverse vegetation and animals. These open settings present distinct difficulties: unlike man-made scenes, natural environments are harder to interpret, as animals often blend into their surroundings with camouflaging colors and patterns; objects may be positioned at varying distances, some very close and others very far, or under a range of lighting conditions; and natural scenes can feature complex interactions between elements, such as trees, animals, and other natural objects. Unlike Avetisyan et al. [4] and in line with Kulits et al. [41], we reconstruct scenes in compositional graphics code from a single image, producing interpretable, editable, and animatable scenes that integrate with existing graphics assets.

While reconstructing natural scenes is itself an unsolved computer-vision problem, we are motivated by the goal of enabling a next-generation computational ethology [2]. Early vision-aided animal-behavior analysis methods relied on 2D pose observations [65]. However, 2D pose provides only limited information and, given that the solution is view-dependent, it is typically only applicable in controlled environments for problems like animal-gait analysis from a fixed camera [19]. The transition to 3D reconstruction of animals represents a natural progression [13, 38, 102], offering a more comprehensive picture. However, reconstructing animals in isolation presents limitations for analysis; for example, studying animal behavior in an empty volume cannot account for occlusions, physical boundaries, or natural interactions. Precise environmental context is useful for understanding animal behavior, yet natural environments pose challenges for both representation and reconstruction. To date, no work has attempted to concurrently reconstruct both 3D animals and their 3D environment. Instead, research in recent years has largely focused on creating increasingly detailed 3D representations of isolated animals. We take a step back, opting instead to work in a complementary direction: rather than pursuing ever-finer models of animals, we prioritize estimating precise layout of the greater scene context with relatively coarse shape representations, to capture the overall environment. In this work, we are the first to tackle the problem of compositionally reconstructing natural scenes from monocular images, presenting the first approach to Reconstruct Animals *and* the Wild (RAW); see Fig. 1.

Figure 2. **Dataset Samples.** Training images from our synthesized dataset. See the Supp. Mat. for additional sample visualizations.

Modeling Natural Environments in 3D. The 3D reconstruction of natural environments from monocular images presents challenges, stemming both from the fundamental ill-posedness of inverting 2D images into their originating 3D scenes and from a lack of adequate graphics models for representing natural environments. The natural world is notably more complex and varied than anthropocentric environments with their geometric regularity. Consequently, modeling natural environments in a manner conducive to analysis is not straightforward. To address this, we propose a compositional approach that represents environments as ordered sets of objects in conjunction with various scene-level attributes. This explicit, object-based representation is interpretable and low-dimensional, abstracting away complexity in a manner facilitating downstream analysis [90].

Data. Teaching a model to decompose single images of animals and their natural environments into structured 3D representations requires it having broad compositional understanding, yet acquiring suitable training data presents its own obstacles. Because the 3D scanning of nature at scale is impractical, here we exploit synthetic-data generation. Building upon tools introduced by the Infinigen project [64], we design a data generator to construct RAW, a million-image dataset comprising both synthetic animals and their environments. Our scenes encompass a range of elements types, including birds, carnivores, herbivores, bushes, boulders, and trees. See Fig. 2 for dataset samples.

Reconstruction. Our goal is to approximately reconstruct 3D animals, scene objects, and layout from a single natural image. To that end, we design a structured graphics-program representation, or language. Akin to [41], we train an LLM to decode CLIP [63] image features into graphics code where objects are represented by their asset identities.

However, when naively training the language model to produce this sequence, we observe that, while it may learn to capture well the scene layout, it fails to generalize to estimating out-of-distribution objects in a semantically meaningful way, preventing real-world generalization. This manifests as reconstructing a tiger with a bush or a bird with a boulder; see Fig. 5a. We hypothesize this inconsistency arises due to limitations in representation/train-time supervision. Built upon a causal LLM, the model operates autoregressively, reconstructing scenes in incremental chunks (Fig. 1). These discretized units, known as tokens, represent bits of text. During training, language models typically are optimized through a cross-entropy next-token objective, whereby they are trained to predict the probabilities of the subsequent token, conditioned on the preceding ones. Although this discrete text-based representation and supervision excel in capturing distinct categorical attributes such as "small," "purple," "shiny," or "cube," difficulties arise when representing quantities that are naturally continuous [41].

Asset names, represented as discrete tokens, lack a meaningful distance metric between one another. This becomes problematic when the objective is not to retrieve the precise object, for which no dataset match may exist, but to provide the best estimate. We hypothesize that, rather than training the LLM to infer exact assets by their IDs, the model can be taught to estimate continuous visual appearance, representing assets by their CLIP encodings and supervising prediction by a loss in semantic CLIP space. We do so by adding a unique token, `[CLIP]`, which signals LLM hidden state should bypass the discretizing tokenization process and be projected as a CLIP embedding (Fig. 3).

We observe that with the incorporation of the CLIP-projection head, the model demonstrates the ability to scale, estimating objects in scenes featuring much-expanded diversity. Our approach successfully reconstructs animals and their environments in real images. We will release our dataset and code to encourage further research in this area.

2. Related Work

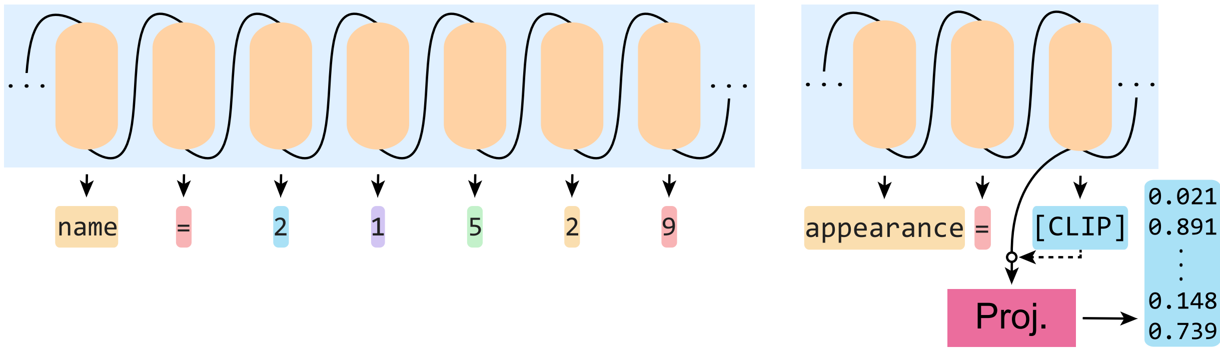
Animal Pose and Shape. Many works have attempted to estimate animal pose and shape from visual observations, evolving from primitive 2D representations to parametric 3D models. Early work by Ramanan et al. [65] focused on recovering 2D articulated models of animals from video. Research later progressed to 3D representations, with Cashman and Fitzgibbon [13] developing a 3D morphable model of a dolphin from images. Kanazawa et al. [38] extended the idea, additionally learning to model articulations and pose-dependent deformations. Zuffi et al. [102] advanced this further by constructing an articulatable multi-species 3D morphable model from scans of toy animals, used to recover 3D shape and pose of quadrupeds [8, 9, 58, 68, 103], while others built models to estimate the shape of birds [5, 89].

Recent approaches have somewhat diverged from a clear progression. Kanazawa et al. [39] learned to recover 3D shape and texture of deformable objects from an image. Sanakoyeu et al. [71] adapted 2D dense pose from humans to animals and Kulkarni et al. [42] developed canonicalized surface mappings across articulated objects. Yang et al. [95] extracted template-free 3D neural models of articulated shape from video, while Yao et al. [97] and Wu et al. [91] learned articulated shape models using DINO [12]-feature-aided part discovery or correspondence. Sharing an asset-based approach, Wu et al. [92] estimated shape and pose from video by retrieving proximal 3D templates from a collection of video-game assets and deforming the templates.

Inverse-Graphics Approaches. The inverse-graphics problem – the task of *inverting* an image into physical variables that, when rendered, enable reproduction of the observed scene – has a long history, dating back to Larry Roberts's Blocks-World thesis [67]. Considerable efforts have focused on tasks such as estimating object pose [46, 49, 53, 61, 79, 85–87, 93] and reconstructing shape from single images [16, 25, 30, 57, 60, 76, 88]. However, works addressing multi-object scenes [21, 28, 74] often neglect object semantics and relationships, limiting deeper reasoning. Holistic 3D-scene understanding aims to reconstruct individual objects along with scene layout. Initial efforts centered on 3D bounding boxes [18, 33, 48, 55, 66], with recent advancements emphasizing finer shape reconstruction [29, 51, 98]. Relatedly, some methods also involve retrieving CAD or mesh models, followed by 6-DoF pose estimation for objects or scenes [3, 6, 24, 31, 36, 37, 43–45, 50, 70, 82]. In contrast, our work, like IG-LLM [41], explores the use of LLMs for the inverse-graphics problem, seeking a possibly simpler and more generalizable solution.

Synthetic Data in Vision. The use of synthetic data for training transferable models has proven successful in recent years. Applications involve learning to detect [81], segment [14], track objects [101], navigate [20, 69], and estimate optical flow [22], depth [64], and human pose and shape [10, 84]. We extend this paradigm by learning to extract compositional 3D scene representations from natural images using procedurally generated Blender [17] scenes, building off tools introduced by the Infinigen project [64].

LLMs and 3D Understanding. Recent applications of LLMs have extended to various 3D-related tasks. These tasks include 3D question answering [23, 34], task planning [34, 52, 99], text-to-3D scene synthesis [35, 78, 96], procedural model editing [40], multi-modal representation learning [34, 94], and 3D scene reconstruction from calibrated RGB-D image sequences [4]. These applications demonstrate the broad applicability of LLMs to tasks not traditionally considered text-based. We continue along the line of IG-LLM [41] and employ an LLM to decode CLIP embeddings into structured 3D scene representations.



(a) **Discrete Name** (b) **Continuous Name**

Figure 3. **CLIP Head.** Rather than teaching the LLM to generate asset names as discrete tokens without a semantic distance metric, we train the model to produce a special token to signal when the hidden state should instead be projected into a continuous CLIP embedding.

3. Method

3.1. Preliminaries

Autoregressive Language Generation. Causal language models generate text in an autoregressive manner, proceeding chunk by chunk. Each generated chunk, known as a token, is conditioned on the preceding sequence of generated chunks. Individual tokens can represent bytes, or one or more characters [72]. The models are typically trained with only a simple next-token prediction objective [7], conditioned on the sequence of previously observed tokens:

$$p(x) = \prod_{i=1}^{n} p\left(s_i | s_1, \ldots s_{i-1}\right) \tag{1}$$

The loss is cross-entropy over predicted token probabilities.
Inverse Graphics With LLMs. LLMs are known for their robust zero-shot generalization capabilities [1, 11, 62], owing to their scale in parameters and the vast amounts of data on which they are trained. Departing from traditional approaches, the success of LLMs to diverse tasks stems from training on large, diverse datasets with a simple objective, followed by fine-tuning on smaller, task-specific sets. This contrasts with previous paradigms that relied heavily on increasing task-specific data for improvements on that task.

Motivated by the remarkable generalization ability of LLMs, IG-LLM [41] treats inverse graphics as LLM-backed inductive program synthesis. It employs an LLM, aligns a CLIP [63] vision encoder to its token space as a visual tokenizer, and finetunes it on primitive demonstrations of images paired with graphics programs, teaching the LLM to decode CLIP embeddings into structured code representations that can be used to reproduce the observed scene using a standard 3D graphics engine. The demonstrations are produced using procedurally generated image–code pairs.
Continuous-Parameter Estimation in LLMs. Tokens are discrete entities, and the cross-entropy loss applied does not impose any particular ordering. In this loss space, a '4' token is equally distant from a '5' as it is from an '8.' The discrete nature of the tokens makes it difficult to enforce metric supervision. IG-LLM [41] addresses this challenge by introducing a numeric module for continuous-parameter estimation. Rather than passing numbers through the text tokenizer, IG-LLM trains the model to produce a special mask token, `[NUM]`, indicating that the token embedding should bypass the gradient-breaking token discretization and be processed by an MLP to produce a continuous, gradient-preserving parameter estimate. By circumventing tokenizer discretization, IG-LLM maintains differentiability, facilitating the use of metric supervision on floats. This adaptation led to stronger parameter-space generalization and smoother training dynamics. We adopt a similar approach of using a special token to signal the re-routing of a token embedding. See Sec. 3.4 for details on our design decisions.

3.2. Base Architecture

We adopt the framework established by IG-LLM and base our architecture on an instruction-tuned version [15][1] of LLaMA-7b [80], incorporating a frozen CLIP [63] visual tokenizer[2] and applying a learnable linear projection to link the vision embeddings with the word-embedding space of the LLM. Following IG-LLM's coarse vision–language alignment strategy, we pre-train the projector using image–caption pairs sourced from the Conceptual Captions dataset (CC3M) [73]. See also IG-LLM [41] for additional details.

3.3. Data-Generation Setting

We design an image–code training-data generator, building upon the tooling of the Infinigen project [64]. Infinigen is a procedural data-generation framework designed to create realistic 3D Blender [17] scenes of the natural world. The

[1] https://huggingface.co/lmsys/vicuna-7b-v1.3
[2] https://huggingface.co/openai/clip-vit-large-patch14-336

framework not only generates diverse terrain but produces a broad range of 3D assets to populate these environments, including various types of plants, trees, rocks, and creatures. These assets are fully parameterized through mathematical rules. The framework boasts 182 unique procedural asset generators and 1,070 interpretable parameter degrees of freedom, in addition to those parameterized by seeds.

The complexity of the scenes is notable, with the authors reporting an average "wall time" of 4.5 hours to create a single image. Although experiments were conducted on 30k generated images, only ten samples were released, none having 3D-object ground truth. To effectively model the natural world, data generation must be made scalable.

In its public iteration, Infinigen renders an image from a single camera location within each scene; the scene is generated around the camera's field of view to mitigate unnecessary rendering complexity. This setup presents challenges for randomly placing multiple cameras within the scene ad hoc, as scenes are designed for a single camera perspective. To improve efficiency, we implement a number of simplifying modifications, which primarily include the following:

- We limit the generated assets to the object types of boulders, bushes, trees, carnivores, herbivores, and birds.
- We pre-generate 1,000 instances of each above asset type.
- We produce two resolutions for each individual asset.
- In each scene, we sample five types of tree, bush, boulder, or creature from those pre-generated. Unlike Infinigen, where assets are individually unique, we instance them.
- We populate the entire scene with assets, rather than solely in the area observed from a single camera location.

Following our modifications, we generate 100 images of each of 10,000 distinct scenes, resulting in 1M total images. Each scene is backed by an explicit Blender representation.

Only the carnivore, herbivore, and bird assets are natively orientable, that is, have a canonical "front." The trees, boulders, and bushes may have very different visual appearances from different angles, such as a left-leaning tree. To be able to incorporate and estimate this information, we assign labels to each object based on its yaw (rotation around the vertical axis) relative to the camera. We divide the yaw into increments of five degrees, resulting in 72 orientations for each object. This increases the total effective number of assets to 432,000. In constructing our ground truth, we zero the yaw of the non-orientable objects local to the camera.

3.4. RAW

Here we define our model objective. We show a partial scene representation in Fig. 4, where `[ROT]` represents a variation of the `[NUM]` token as applied in IG-LLM, signaling the token embedding to instead be put through an MLP to regress a nine-parameter rotation matrix. This choice was motivated pragmatically by reducing code dimensionality to enable the use of up to twenty-five objects per sequence in model context, and through the greater generalization demonstrated in IG-LLM. The semantic token, `[CLIP]`, is used to signal whether the token embedding should be projected to CLIP space with a projection layer. During training, we set the target of the embedding head to be that of the rendered asset image at the given yaw of the scene asset. See Fig. 3 for a visualization of the projection head.

```
set_sun_intensity(0.981)
set_sun_elevation(0.691)
set_sun_size(0.811)
set_camera(88.130)
set_atmospheric_density(0.009)
set_ozone(1.499)
set_sun_rotation(231.110)
set_dust(0.169)
set_sun_strength(0.212)
set_air(0.771)
set_ground([CLIP])
add(pixels=1582, loc=(-0.553, -0.809, -22.591),
↪  height=1.365, rotation=[ROT], appearance=[CLIP])
add(pixels=111, loc=(-1.524, -0.939, -30.159),
↪  height=1.224, rotation=[ROT], appearance=[CLIP])
...
```

Figure 4. **Code Sample.** The figure displays a partial scene code. `height` is estimated as embeddings do not capture object scale.

Scene-level attributes are estimated at the beginning of the sequence prior to defining objects. These include sun parameters (intensity, elevation, size, strength, and rotation relative to the camera) and atmospheric conditions (density, ozone, dust content, and air density). Additionally, a CLIP embedding is estimated to retrieve ground texture, improving the visual realism of the resulting scene reconstruction.

Objects are ordered in the code-sequence objective by the number of pixels they occupy: from the visually largest to the least significant. In this way, the model is taught to first focus on the most-salient objects before attempting to explain bushes in the background. See the project page for a step-wise reconstruction visual, highlighting the ordering.

3.5. RAW-RW

To complement our synthetic test setting and enable quantitative evaluation of image syn-to-real generalization, we curate a set of online, permissively licensed photos of animals in their natural habitats, which we refer to as RAW-RW.

3.6. Supervision

In addition to the next-token prediction objective loss applied to the text of the generated code, our introduction of additional heads for rotation and CLIP-appearance estimation enables – and necessitates – further supervision. Following Geist et al. [26], we apply a mean-squared-error loss on rotation matrices after performing symmetric orthogonalization. The CLIP estimation is supervised by a cosine-similarity loss between the estimated and target embeddings and an additional regularization term because similarity is vector-norm invariant. See Supp. Mat. for further details.

(a) **ID** (b) **RAW-RW**

Figure 5. **Qualitative Comparison.** (a) In-distribution, the discrete-name baseline confuses classes: a tiger and a bush; a bird and a boulder. (b) When tasked with the reconstruction of real scenes, it well-estimates layout, but struggles to generalize in an aligned manner.

4. Evaluations

In this section, we evaluate the ability of our model to reconstruct both held-out synthetic in-distribution (ID) and natural real-world (RAW-RW) scenes. We begin by measuring the effect of representing assets by semantic CLIP embeddings in Sec. 4.2. In Sec. 4.3, we evaluate the effect of conditioning-memorization on model generalization. Later, in Sec. 4.4, we test how introducing additional conditioning into the sequence affects learning. Finally, in Sec. 4.5, we compare alternatives to CLIP for appearance estimation.

4.1. Metrics

To evaluate our method quantitatively, we render the reconstructed scene and evaluate it primarily with perceptual metrics against the source image. Prior to rendering, we warp a ground plane to the estimated object locations using an RBF kernel. Central to our evaluation is LPIPS [100], which measures perceptual similarity between two images – the source and rendered reconstruction – using learned VGG [75] features. LPIPS serves as a spatial metric as it is patch-based. We additionally compute cosine similarity between the input and reconstruction using CLIP [63], BioCLIP [77], and DINOv2 [59]. We however de-emphasize the metrics due to their indirect usage in model ablations, but find they offer a complementary perspective to LPIPS. See Supp. Mat. for object-wise 3D evaluations and a quantitative comparison against a YOLOX-6D-Pose [54] baseline.

	↓LPIPS	↑S_{CLIP}	↑S_{BioCLIP}	↑S_{DINOv2}
IG-LLM	0.720	0.748	0.421	0.833
+ CLIP	0.654	0.806	0.537	**0.858**
+ Fuzz.	0.612	0.807	0.526	0.849
+ Cond.	**0.598**	**0.815**	**0.539**	**0.858**

(a) **ID**

	↓LPIPS	↑S_{CLIP}	↑S_{BioCLIP}	↑S_{DINOv2}
IG-LLM	0.772	0.385	0.254	0.559
+ CLIP	0.762	0.445	0.335	0.561
+ Fuzz.	0.762	0.445	0.349	0.565
+ Cond.	**0.724**	**0.490**	**0.409**	**0.631**

(b) **RAW-RW**

Table 1. **Quantitative Ablation Effects**. Each change impacts performance. S_{CLIP}, S_{BioCLIP}, and S_{DINOv2} represent cosine similarity between embeddings of the input and rendered reconstruction.

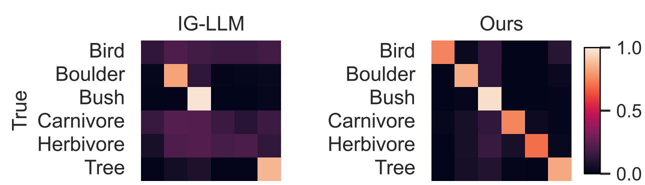


Figure 6. **ID Class-Confusion Matrices.** While the discrete model variant estimates objects with another of the same type on average, it biases heavily toward those with the greatest frequency.

4.2. Discrete Names and Continuous Embeddings

We begin with evaluating the naive discrete-shape-name version of our pipeline (IG-LLM) trained on our synthetic dataset. In this template, rather than `set_ground` and appearance being parametrized by `[CLIP]` tokens and corresponding embeddings, the assets are named by integers, which represent the scene ID in the case of ground texture. See representative scene reconstructions in Figs. 5 and 7.

We observe that both models estimate scene layout fairly consistently both in- and out-of-distribution. However, while the majority of assets estimated by each of the models appear as reasonable approximations in the ID case (Fig. 5a), the discrete variant often confuses classes of similar sizes (Fig. 6). Rather than another instance of the type of object portrayed, the reconstructions mix assets with distinct semantics: a tiger with a bush or a bird with a boulder.

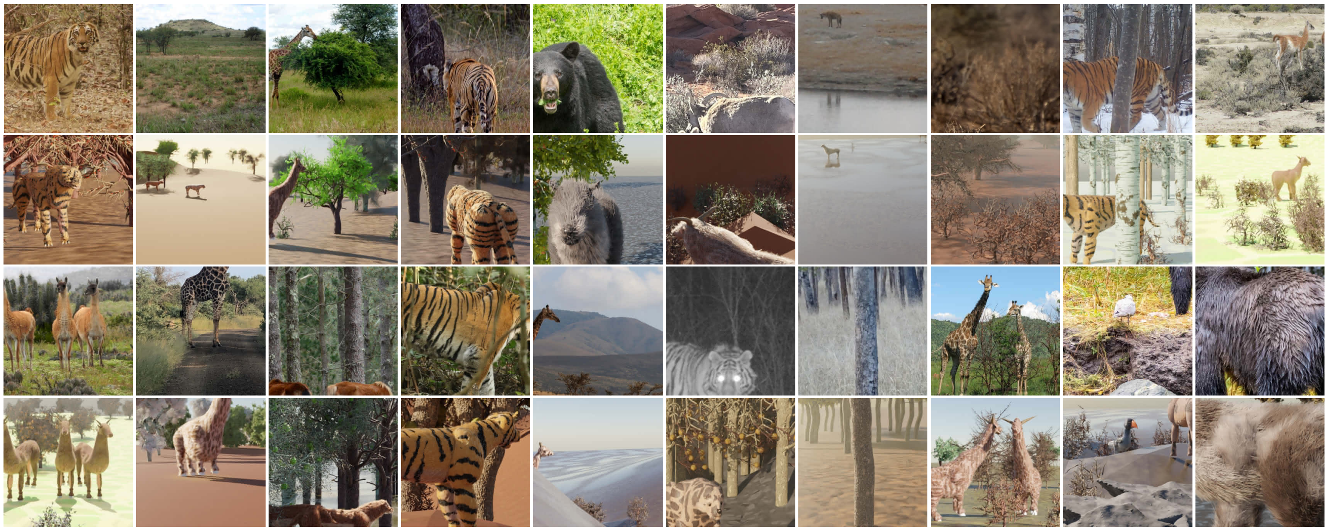


Figure 7. **Additional Reconstructions.** Additional real-world-generalization samples (top: input; bottom: output). Note how our model is able to reconstruct scenes where the animal is very far or close to the camera, under severe occlusion, and with difficult lighting conditions.

Moving to the OOD case, while the baseline continues to somewhat consistently predict layout, it fails to reconstruct the scene with meaningful assets (Fig. 5b). In its uncertainty it will often select a tree for a tree, but not the right tree.

Predicting instead semantic asset appearance in the CLIP-estimation variant yields improved results. Images are explained with similar layout, but by assets with noticeably greater perceptual alignment. This is reflected by a jump across metrics in both evaluation settings (Tab. 1a).

4.3. Value Fuzzing

We next investigate the effect of memorization of conditioning. Producing the graphics-code sequence autoregressively, the model conditions the generation of any token on all preceding it. In causal-language model training, the context tokens the model sees are always ground-truth values.

In estimating the value of `atmospheric_density`, the model should have all the information necessary to produce the quantity based on only the conditioned-on image features. However, in practice, if the model knew that the `sun_intensity` was 0.981, and that the `sun_intensity` is only ever 0.981 in scene 3,389, it may be simpler to memorize a table mapping scene identities to scene attributes. While some token conditioning is necessary – the model must know that it is now estimating `atmospheric_density` and hasn't already done so – conditioning on ground-truth scene-level values might hurt the model's ability to estimate these values in new scenes. If the training scenes were unique, and each of the million images were from distinct scenes, this could not be expected to be such an issue, but as one hundred images are produced in each scene, we suspect it may harm generalization ability.

We hypothesize that adding a small amount of noise, or "fuzzing" to the target scene-level attributes during training will force the model to pay more attention to image features, and learn to better avoid memorization of exact scenes. For each value, we add a uniformly distributed $\pm 0.5\%$ of noise to each of the scene-level attributes. The magnitude of the noise added is small enough to not have a noticeable effect on the values themselves, but results in a measurable improvement in LPIPS (Tab. 1a), supporting our intuition.

4.4. Additional Conditioning

We then evaluate the effect of introducing additional value conditioning to the sequences. The model, when generating, conditions object predictions off all preceding objects in the sequence. In the base version, it is able to leverage positional information (`loc`) and estimated object height (`height`) to determine where in the sequence it is and what should be produced next. As the objects are ordered in the objective by the count of their pixels visible in the source image, the model is expected to reason about – and disentangle – objects by saliency. It must follow the GT sequence during training, and can not learn its own object ordering.

We hypothesize that reasoning about this ordering from only the list of what came before is difficult for the model to learn. Motivated by this, we task the model to additionally estimate the number of pixels visible for each object. In doing so, it must explicitly model visibility, and it can also condition off the information during training to reduce uncertainty. We evaluate adding pixel count to the sequence (`pixels`), and find quantitative improvement (Tab. 1).

	↓LPIPS	↑S_{CLIP}	↑S_{BioCLIP}	↑S_{DINOv2}
CLIP	<u>0.598</u>	<u>0.815</u>	<u>0.539</u>	<u>0.858</u>
BioCLIP	0.676	0.795	0.512	0.833
DINOv2	**0.597**	**0.850**	**0.603**	**0.865**

(a) **ID**

	↓LPIPS	↑S_{CLIP}	↑S_{BioCLIP}	↑S_{DINOv2}
CLIP	**0.724**	**0.490**	**0.409**	**0.631**
BioCLIP	<u>0.730</u>	0.443	0.339	0.575
DINOv2	0.743	<u>0.473</u>	<u>0.387</u>	<u>0.611</u>

(b) **RAW-RW**

Table 2. **Embedding Ablation.** While DINOv2 [59] features are most discriminative, we suspect they are more difficult to interpret.

4.5. Choice of Embedding

Finally, we explore the use of alternate embeddings for appearance estimation, namely DINOv2 [59] and BioCLIP [77], to determine both their efficacy and the LLM's ability to estimate them effectively. For clarity, this evaluation does not examine the differences in behavior of alternatives to the CLIP visual tokenizer, which remains fixed throughout. DINOv2 was trained with a self-supervised learning objective roughly based on masked-image modeling [32], and BioCLIP is a fine-tuned version of CLIP trained on image–label data of taxonomic species from the iNaturalist [83] and BIOSCAN-1M [27] image databases.

We quantitatively evaluate the variants in terms of the earlier evaluation setting. Results can be seen in Tab. 2. Contrary to our initial speculation, we find that BioCLIP performs worst as the target embedding across metrics. We suspect that the finetuning applied to the model hinders its generalization ability. On the other hand, DINOv2, performed best in-distribution (Tab. 2a), but did not effectively generalize (Tab. 2b). We suggest that, while the features may increase the separability of the assets, the space is less interpretable, and more difficult to learn a general mapping.

5. Limitations and Future Work

While assets produced using Infinigen tooling are parametrically defined, reconstructing them solely as code presents challenges, as they are frequently the result of complex non-invertible physical processes, including seeded noise operations. Consequently, while the assets may have compact low-dimensional representations, the parameters are not smooth: small seed changes might result in an object with dramatically different shape or appearance. Future work could explore a compromise between retrieval and asset generation (predicting some parameters while retrieving others), rely on differentiable proxies for the non-invertible steps, or employ models with fully interpretable parameters.

Figure 8. **Limitations Samples.** Our model struggles reconstructing images with highly out-of-distribution layout or camera pose.

The creature-articulation system in Infinigen is non-functional (as pictured throughout Raistrick et al. [64], all creatures are of static pose, many with feet visibly off the ground). This restricts the expressivity of the data framework. Future work may involve inferring articulated pose.

While we observe fairly consistent semantically aligned reconstruction of real-world environments, our model can struggle to reconstruct images of scenes with highly out-of-distribution configurations, such as those in which the pose of the camera is outside the distribution seen during training. We suspect that some such generalization issues might be abated with better layout sampling during data-generation, but without a more-diverse pool of assets, the model will not be able to sufficiently explain all relevant aspects of the scene, such as a vehicle on the road (Fig. 8).

While our goal was to compositionally reconstruct natural scenes from monocular images, we expect our findings may generalize to other domains too. Future work should expand to additionally model the anthropocentric world.

6. Conclusion

Our investigation represents the first compositional reconstruction of natural scenes from a monocular image that captures both animals and their environments, bridging the gap between reconstructing animals *and* the wild (RAW). In summary, we make the following key contributions:

First, we identify and address a limitation in scaling LLM-backed inverse graphics [41] to real-world scenes. By teaching an LLM to "name" objects in terms of semantic CLIP appearance, rather than as a sequence of discrete tokens, we overcame limitations inherent in pure token-based supervision, making the problem effectively continuous.

Second, we introduce a million-image synthetic dataset built on tooling introduced in the Infinigen project [64], addressing data limitations. Despite training exclusively on synthetic data, our approach demonstrated successful generalization to reconstructing natural scenes from an image.

Finally, by enabling comprehensive scene reconstruction that includes both animals and their environmental context, we lay the groundwork for a next generation of computational ethology. This opens new possibilities for automated interpretation of animal behavior grounded in their habitat.

7. Acknowledgements

We thank Tomasz Niewiadomski, Taylor McConnell, and Tsvetelina Alexiadis for study assistance, Nikos Athanasiou for discussions, and Rick Akkerman for proofreading.

Disclosure. While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society. SZ is supported by PNRR FAIR Future AI Research (PE00000013), Spoke 8 Pervasive AI (CUP H97G22000210007) and NBFC National Biodiversity Future Center (CN00000033), Spoke 4 (CUP B83C22002930006) under the NRRP MUR program by NextGenerationEU.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736. Curran Associates, Inc., 2022. 4

[2] David J. Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014. Publisher: Elsevier. 2

[3] Mathieu Aubry, Daniel Maturana, Alexei A. Efros, Bryan C. Russell, and Josef Sivic. Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models. In *CVPR*, pages 3762–3769, Los Alamitos, CA, USA, 2014. IEEE Computer Society. 3

[4] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. SceneScript: Reconstructing scenes with an autoregressive structured language model. In *ECCV*, 2024. 2, 3

[5] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G. Pfrommer, Marc F. Schmidt, and Kostas Daniilidis. 3D bird reconstruction: A dataset, model, and shape recovery from a single view. In *ECCV*, pages 1–17, 2020. 3

[6] Aayush Bansal, Bryan C. Russell, and Abhinav Gupta. Marr revisited: 2D-3D alignment via surface normal prediction. In *CVPR*, pages 5965–5974, Los Alamitos, CA, USA, 2016. IEEE Computer Society. 3

[7] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In *NeurIPS*. MIT Press, 2000. 4

[8] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *ACCV*, pages 3–19, Cham, 2019. Springer International Publishing. 3

[9] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, pages 195–211. Springer, 2020. 3

[10] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 3

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901. Curran Associates, Inc., 2020. 4

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3

[13] Thomas J. Cashman and Andrew W. Fitzgibbon. What shape are dolphins? Building 3D morphable models from 2D images. *TPAMI*, 35(1):232–244, 2013. 2, 3

[14] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 3

[15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, 2023. 4

[16] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, pages 628–644, Cham, 2016. Springer International Publishing. 3

[17] Blender Online Community. *Blender – A 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 3, 4

[18] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. DeLay: Robust spatial layout estimation for cluttered indoor scenes. In *CVPR*, Los Alamitos, CA, USA, 2016. IEEE Computer Society. 3

[19] Charles E. DeCamp, Robert W. Soutas-Little, Joe Hauptman, Bari Olivier, Terrance Braden, and Aaron Walton. Kinematic gait analysis of the trot in healthy Greyhounds. *American Journal of Veterinary Research*, 54(4):627–634, 1993. 2

[20] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *NeurIPS*, pages 5982–5994. Curran Associates, Inc., 2022. 3

[21] Maximilian Denninger and Rudolph Triebel. 3D scene reconstruction from a single viewport. In *ECCV*, pages 51–67. Springer, 2020. 3

[22] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 3

[23] Mohammed Munzer Dwedari, Matthias Niessner, and Zhenyu Chen. Generating context-aware natural answers for questions in 3D scenes. In *BMVC*. BMVA, 2023. 3

[24] Francis Engelmann, Konstantinos Rematas, Bastian Leibe, and Vittorio Ferrari. From points to multi-object 3D reconstruction. In *CVPR*, pages 4588–4597, 2021. 3

[25] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 3

[26] Andreas René Geist, Jonas Frey, Mikel Zhobro, Anna Levina, and Georg Martius. Learning with 3D rotations, a hitchhiker's guide to SO(3). In *ICML*, 2024. 5

[27] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, Iuliia Zarubiieva, Joakim Bruslund Haurum, Scott Lowe, Jaclyn McKeown, Chris Ho, Joschka McLeod, Yi-Yun Wei, Jireh Agda, Sujeevan Ratnasingham, Dirk Steinke, Angel Chang, Graham W. Taylor, and Paul Fieguth. A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset. In *NeurIPS*, pages 43593–43619. Curran Associates, Inc., 2023. 8

[28] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. In *ICCV*, 2019. 3

[29] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. Learning 3D object shape and layout without 3D supervision. In *CVPR*, pages 1695–1704, 2022. 3

[30] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. A papier-mâché approach to learning 3D surface generation. In *CVPR*, pages 216–224, 2018. 3

[31] Can Gümeli, Angela Dai, and Matthias Nießner. ROCA: Robust CAD model retrieval and alignment from a single image. In *CVPR*, pages 4022–4031, 2022. 3

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 8

[33] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, pages 1849–1856, 2009. 3

[34] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. *arXiv preprint arXiv:2307.12981*, 2023. 3

[35] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A. Ross, Cordelia Schmid, and Alireza Fathi. SceneCraft: An LLM agent for synthesizing 3D scenes as blender code. In *ICML*, 2024. 3

[36] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3D scene parsing and reconstruction from a single RGB image. In *ECCV*, 2018. 3

[37] Hamid Izadinia, Qi Shan, and Steven M. Seitz. IM2CAD. In *CVPR*, 2017. 3

[38] Angjoo Kanazawa, Shahar Kovalsky, Ronen Basri, and David Jacobs. Learning 3D deformation of animals from 2D images. In *Eurographics*, page 365–374, Goslar, DEU, 2016. Eurographics Association. 2, 3

[39] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 3

[40] Milin Kodnongbua, Benjamin T. Jones, Maaz Bin Safeer Ahmad, Vladimir G. Kim, and Adriana Schulz. ReparamCAD: Zero-shot CAD re-parameterization for interactive manipulation. *SIGGRAPH Asia (Conference track)*, 2023. 3

[41] Peter Kulits, Haiwen Feng, Weiyang Liu, Victoria Fernandez Abrevaya, and Michael J. Black. Re-thinking inverse graphics with large language models. *TMLR*, 2024. 2, 3, 4, 8

[42] Nilesh Kulkarni, Abhinav Gupta, David F. Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, 2020. 3

[43] Abhijit Kundu, Yin Li, and James M. Rehg. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare. In *CVPR*, 2018. 3

[44] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2CAD: 3D shape prediction by learning to segment and retrieve. In *ECCV*. Springer International Publishing, 2020.

[45] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Patch2CAD: Patchwise embedding learning for in-the-wild shape retrieval from a single image. In *ICCV*, pages 12589–12599, 2021. 3

[46] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. CosyPose: Consistent multi-view multi-object 6D pose estimation. In *ECCV*, pages 574–591. Springer, 2020. 3

[47] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. 2

[48] David C. Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, pages 2136–2143. IEEE, 2009. 3

[49] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate O(n) solution to the PnP problem. *IJCV*, 81(2):155–166, 2009. 3

[50] Joseph J. Lim, Aditya Khosla, and Antonio Torralba. FPM: Fine pose parts-based model with 3D CAD models. In *ECCV*, pages 478–493. Springer, 2014. 3

[51] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *ECCV*, pages 429–446. Springer Nature Switzerland, 2022. 3

[52] Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. GPT4Motion: Scripting physical motions in text-to-video generation via Blender-oriented GPT planning. In *CVPRW*, pages 1430–1440, 2024. 3

[53] Wufei Ma, Angtian Wang, Alan Yuille, and Adam Kortylewski. Robust category-level 6D pose estimation with coarse-to-fine rendering of neural features. In *ECCV*, pages 492–508. Springer, 2022. 3

[54] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. YOLO-6D-Pose: Enhancing YOLO for single-stage monocular multi-object 6D pose estimation. In *3DV*, pages 1616–1625, Los Alamitos, CA, USA, 2024. IEEE, IEEE Computer Society. 6

[55] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *ICCV*, pages 936–944, 2015. 3

[56] David Marr and H. Keith Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 200(1140):269–294, 1978. 2

[57] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 3

[58] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J. Black, Silvia Zuffi, and Peter Kulits. Generative zoo, 2024. 3

[59] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 6, 8

[60] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 3

[61] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G. Derpanis, and Kostas Daniilidis. 6-DOF object pose from semantic keypoints. In *ICRA*, pages 2011–2018. IEEE, 2017. 3

[62] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. 4

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 4, 6

[64] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *CVPR*, pages 12630–12641, 2023. 2, 3, 4, 8

[65] D. Ramanan, D.A. Forsyth, and K. Barnard. Building models of animals from video. *TPAMI*, 28(8):1319–1334, 2006. 2, 3

[66] Yuzhuo Ren, Shangwen Li, Chen Chen, and C-C Jay Kuo. A coarse-to-fine indoor layout estimation (CFILE) method. In *ACCV*, pages 36–51. Springer, 2017. 3

[67] Lawrence G. Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 3

[68] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *CVPR*, pages 3876–3884, 2022. 3

[69] Fereshteh Sadeghi and Sergey Levine. CAD2RL: Real single-image flight without a single real image. In *Robotics: Science and Systems*, 2017. 3

[70] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly, and Andrew J. Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. In *CVPR*, pages 1352–1359, 2013. 3

[71] Artsiom Sanakoyeu, Vasil Khalidov, Maureen S. McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *CVPR*, 2020. 3

[72] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics. 4

[73] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565. Association for Computational Linguistics, 2018. 4

[74] Daeyun Shin, Zhile Ren, Erik B. Sudderth, and Charless C. Fowlkes. 3D scene reconstruction with multi-layer depth and epipolar transformers. In *ICCV*, pages 2172–2182, 2019. 3

[75] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 6

[76] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. *NeurIPS*, 32, 2019. 3

[77] Samuel Stevens, Jiaman Wu, Matthew J. Thompson, Elizabeth G. Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M. Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *CVPR*, pages 19412–19424, 2024. 6, 8

[78] Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3D-GPT: Procedural 3D modeling with large language models. *arXiv preprint arXiv:2310.12945*, 2023. 3

[79] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class Hough forests for 3D object detection and pose estimation. In *ECCV*, pages 462–477. Springer International Publishing, 2014. 3

[80] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. 4

[81] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *CVPRW*, 2018. 3

[82] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *CVPR*, pages 1510–1519, 2015. 3

[83] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, 2018. 8

[84] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[85] Angtian Wang, Adam Kortylewski, and Alan L. Yuille. NeMo: Neural mesh models of contrastive features for robust 3D pose estimation. In *ICLR*, 2021. 3

[86] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. In *CVPR*, pages 16611–16621, 2021.

[87] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *CVPR*, pages 2642–2651, 2019. 3

[88] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single rgb images. In *ECCV*, 2018. 3

[89] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *CVPR*, pages 14739–14749, 2021. 3

[90] Jiajun Wu, Joshua B. Tenenbaum, and Pushmeet Kohli. Neural scene de-rendering. In *CVPR*, pages 699–707, 2017. 2

[91] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3D animals in the wild. In *CVPR*, pages 8792–8802, 2023. 3

[92] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. CASA: Category-agnostic skeletal animal reconstruction. In *NeurIPS*, pages 28559–28574. Curran Associates, Inc., 2022. 3

[93] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes, 2018. 3

[94] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In *CVPR*, pages 1179–1189, 2023. 3

[95] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. BANMo: Building animatable 3D neural models from many casual videos. In *CVPR*, pages 2863–2873, 2022. 3

[96] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3D embodied AI environments. In *CVPR*, pages 16227–16237, 2024. 3

[97] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. LASSIE: Learning articulated shapes from sparse image ensemble via 3D part discovery. In *NeurIPS*, pages 15296–15308. Curran Associates, Inc., 2022. 3

[98] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3D scene understanding from a single image with implicit representation. In *CVPR*, pages 8833–8842, 2021. 3

[99] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. In *ICLR*, 2024. 3

[100] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[101] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, pages 19855–19865, 2023. 3

[102] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, 2017. 2, 3

[103] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*, 2019. 3