AWOL: Analysis WithOut synthesis using Language

Silvia Zuffi[1] and Michael J. Black[2]

[1] IMATI-CNR, Milan, Italy
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany
silvia.zuffi@cnr.it, black@tuebingen.mpg.de
http://awol.is.tue.mpg.de

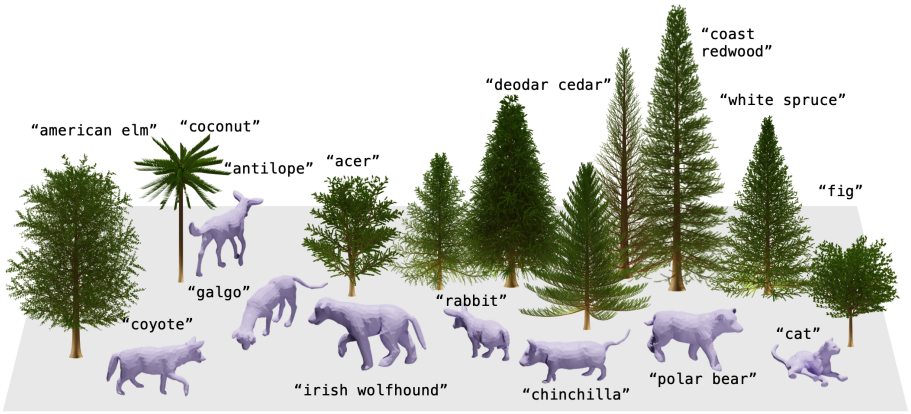

Fig. 1: Generated trees and animals. AWOL learns to generate animals and trees from text and images. We show examples of tree and animal species not seen during training (except for the Cat).

Abstract. Many classical parametric 3D shape models exist, but creating novel shapes with such models requires expert knowledge of their parameters. For example, imagine creating a specific type of tree using procedural graphics or a new kind of animal from a statistical shape model. Our key idea is to leverage language to control such existing models to produce novel shapes. This involves learning a mapping between the latent space of a vision-language model and the parameter space of the 3D model, which we do using a small set of shape and text pairs. Our hypothesis is that mapping from language to parameters allows us to generate parameters for objects that were never seen during training. If the mapping between language and parameters is sufficiently smooth, then interpolation or generalization in language should translate appropriately into novel 3D shapes. We test our approach with two very different types of parametric shape models (quadrupeds and arboreal trees). We use a learned statistical shape model of quadrupeds and

show that we can use text to generate new animals not present during training. In particular, we demonstrate state-of-the-art shape estimation of 3D dogs. This work also constitutes the first language-driven method for generating 3D trees. Finally, embedding images in the CLIP latent space enables us to generate animals and trees directly from images.

1 Introduction

We address the problem of generating new, realistic samples from various 3D shape models using language. The key idea is to relate language (e.g. names of dog breeds or types of trees) to the model's parameters and then leverage language to generate shapes that were never seen during training. To make this possible, we leverage the shared latent space of large vision-language foundation models (VLM), like CLIP (Contrastive Language-Image Pretraining) [43]. Such models relate how objects appear in images to how we describe them with language. Since how objects appear is related to their 3D shape, we can assume that CLIP implicitly also relates object shape with language. Since models like CLIP are learned from large data corpora, VLM latent spaces are rich and dense; in other words, they know a lot about objects and their shape, but not explicitly. Given a small training set, we learn a mapping between the CLIP space and the shape parameters of various models. Finally, our central hypothesis is that the CLIP space is well-behaved such that interpolation or extrapolation in this space produce appropriate interpolation or extrapolation of the associated shape parameters. This allows us to exploit the general knowledge of a VLM to control the parameters of the shape model to produce, within the shape space of the parametric 3D model, novel shapes outside its training set. We show in our experiments that this ability extends to fine-grain control (i.e. generating different dog breeds) and 3D generation with attributes. We test our hypothesis using two very diverse object classes, animals and trees, that use two very different generation processes. For animals, we use an analytic, statistical, parametric shape model, named SMAL$^+$, that we introduce here as a new, extended version of previous models [27, 47, 76]. For trees, we use a procedural, non-differentiable, tree generator implemented as a Blender add-on [17]; this is very different from SMAL$^+$. Trees are an interesting case because they are composed by thin structures (branches) and thin surfaces (leaves) that cannot easily be fit with the 3D implicit representations used in many current text-to-3D solutions (Fig. 2). With our method, named AWOL, we generate trees and animals that are unseen during training and that are expressed as triangular meshes, thus supporting easy rendering and animation in graphics engines; see Fig. 1.

There is growing interest in generating 3D content with easy-to-use tools. An abundance of methods have been proposed to create 3D assets from simple text prompts [5, 6, 6, 18, 20, 20, 30, 34, 35, 41, 59, 60, 62, 66, 68], or single images [20, 32, 51]. Such methods are able to generate compelling rigid objects, with realistic appearance. Such models do not, however, produce articulated objects that are rigged for animation. With AWOL, we obtain animal models that share

Fig. 2: Comparison with existing methods. From left: AWOL from text, Genie (LumaAI) [1] from text, image prompt and two views of Instant Mesh [67].

the same skeleton and mesh topology. This is important: a standardized 3D generation would allow easy motion transfer and facilitate analysis, promoting the application of 3D computer vision methods (i.e. 3D model-based articulated motion estimation) to the animal research and conservation fields. Existing 3D parametric shape models for articulated subjects, like SMPL [33], for humans, or SMAL [74], for animals, are generative models for body shape, and consequently they are widely used to create 3D avatars, either by sampling the generative model, or by aligning the model to data [4, 7, 11, 12, 15, 22, 23, 37, 39, 40, 54, 55, 57, 74, 75]. Alignment is made possible by the differentiable nature of the models, which support reconstruction through the analysis-by-synthesis paradigm. While the SMPL model can arguably represent a large portion of the world population, given its large training set and uniqueness of the human species, the SMAL model has been trained on a small set of quadrupeds to represent animals from 5 different families (canine, equine, bovine, hippopotamids, and feline). As such, naively sampling the model shape space can produce non-existing animals that are often a mixture of more species. Sampling with family-specific shape priors (i.e. Gaussian distributions centered at the family mean shape variables) allows generating instances with realistic shape. However, as illustrated in the paper [76], when aligned to data, the SMAL model can broadly represent species that are not present in its training set, for example representing a boar with a mane borrowed by lions, a long mouth from hippos, and bulky body from cows. The question then is: how can we generate species that are not in one of the five SMAL families without using analysis-by-synthesis? The question is of broader application, as it regards the possibility of generalizing the generation of 3D assets given parametric models defined on a small set of samples. Identifying the manifold of realistic samples may be difficult: some regions of the space can correspond to shapes that, although not seen during training, are realistic, while other regions can correspond to non-existing class instances. Therefore, there is a problem of realistic interpolation for data generation. In addition, shape models based on continuous latent spaces do not offer extrapolation capabilities, as their dimensions generally do not correspond to semantic deformations. While space transformations can be applied to identify axes with semantic meaning, this does not address the generalization principle, as how to move along these axes to generate new, realistic samples remains undefined. In both the animals and trees models, the set of training samples is scarce. This limits the application of highly flexible generative models that are popular today, such as diffusion models. We employ Real-NVP [10], a generative model characterized by a set of explicit

transformations defined through a cascade of layers that selectively couple the different dimensions of the input data using fixed binary masks. While Real-NVP has been used for text-to-3D generation before [50], here we show that learning the binary masks improves performance, and adds realistic relative scaling to the predicted shapes. In summary, our contributions are: a 3D parametric shape model for animals that includes more species than previous models; a method to generate 3D rigged animals from text or images; and a method to generate 3D trees from text or images, which can output a triangular mesh with branches and leaves details.

2 Related Work

Text-to-3D Our work is related to text-driven model-based 3D content creation systems. An early example is BodyTalk [53], which correlates textual shape attributes with transformed dimensions of the SMPL shape space. Semantify [16] also addresses the problem of controlling the SMPL body model with shape attributes, but it exploits CLIP [43]. Recent work uses text to control 3D face generation [64]. In the past few years, numerous methods have addressed the text-driven generation of images [14, 44–46, 48, 49], and more recently 3D objects [6, 18, 20, 34, 59, 66, 68]. Training is often based on the similarity between textual queries and rendered 3D shapes when encoded in a joint latent space (i.e., CLIP), with the gradient back propagated through a differentiable renderer. Many methods are thus based on differentiable 3D neural representations, often Neural Radiance Field (NeRF) [36], with a few mesh-based exceptions [30, 58]. Directly regressing a 3D triplane representation speeds up the text-to-3D generation [28]. The scarcity of 3D data is overcome by exploiting 2D losses. DreamFields [20] generates open-set 3D objects by optimization. The output is a NeRF that is trained by optimizing for rendered views to have high semantic similarity, given the text prompt. The method uses CLIP in synergy with geometric priors. DreamFusion [41] leverages powerful text-to-image diffusion models (here Imagen [49]) and introduces Score Distillation Sampling (SDS) to exploit diffusion priors as losses for 3D object optimization, an approach also adopted in [5,30,35,59,60,62]. Recent methods [28,52] benefit from 3D supervision thanks to the availability of large 3D datasets [8,9]. Our work is related to CLIP-Forge [50], which trains a normalizing flow network to learn the mapping between the CLIP and the latent space of a 3D shape model, learned over a collection of 3D rigid objects.

3D Animal Models Three-dimensional differentiable articulated shape models have been defined for a few common species. SMAL [76] is a multi-species model that can represent a wide range of quadrupeds. SMALR [75] extends SMAL to capture 3D shapes of animals from a set of images. SMALST [74] learns a 3D model for the Gravy's zebra from images. AVES [61] learns the 3D shape of birds from images, starting from a reference template. hSMAL [27] and D-SMAL [47] are 3D parametric shape models for horses and dogs, respectively. Many recent methods do not assume an existing reference template. Lassie [70] and Hi-Lassie

Fig. 3: Training set for the tree network. From left: Poplar, Maple, Palm, Silver Birch, English Oak, European Larch, Weeping Willow, Balsam Fir, Black Tupelo, Sphere Tree, Black Oak, Hill Cherry, Sassafras, Douglas Fir, Apple, Willow, Cypress, Magnolia, Pine, Fan Palm, Quaking Aspen.

[71] create 3D models from a small collection of images. Like SMALR [75], they require different images with a clear, non-occluded view of the animal. Artic3D [72] supports noisy images. Lepard [31] reconstructs 3D animals from images using a part-based neural representation. While applicable to animals with a different number of body parts, these methods do not reconstruct realistic fine-grained details, as the synthesis losses are based on matching silhouettes or image features. Moreover, they only reconstruct single animal instances. Methods exist to learn category-specific shape priors from images: MagicPony [65] learns models for horses, and 3D-Fauna [29] extends the approach to arbitrary quadrupeds. RAC [69] learns category-level 3D models from video. GART [25] learns a subject specific model from monocular video.

3D Arboreal Trees Generation The modeling of trees and vegetation has a long history. Early approaches focused on modeling the branching structure using fractals [2,38], grammars and particle systems [21], and L-systems [42], with the latter proving effective for modeling a large variety of realistic trees given a set of production rules. Weber and Penn [63] define a procedural model that, instead of accurately modeling how trees grow, focuses on the tree's global geometry. Using such systems is complicated and requires extensive knowledge to define a non-intuitive set of parameters. Recent methods exploit learning systems to simplify parameter definition and automate the synthetic tree generation process. The recent DeepTree [73] learns rules from traditional procedural methods and define a network that can automatically grow trees while taking into account environmental constraints. Lee et al. [24] train a neural network to generate parameters for procedural tree generation. None of these methods allow for obtaining parameters from text, as we do. Li et al. [26] grow tree branches using a multi-cylindrical shape, estimated from an image mask, the as surface limit.

3 Method

3.1 Animal Model

The SMAL$^+$ parametric animal model is an extension of SMAL [76]. SMAL is defined by a triangular mesh template $\mathbf{v}_t$ with n_V vertices, a matrix B of shape $3n_V \times n_B$ containing the n_B basis vectors of a linear shape deformation space, a joint regressor J_r that maps model vertices to a set of n_J joint locations, and a

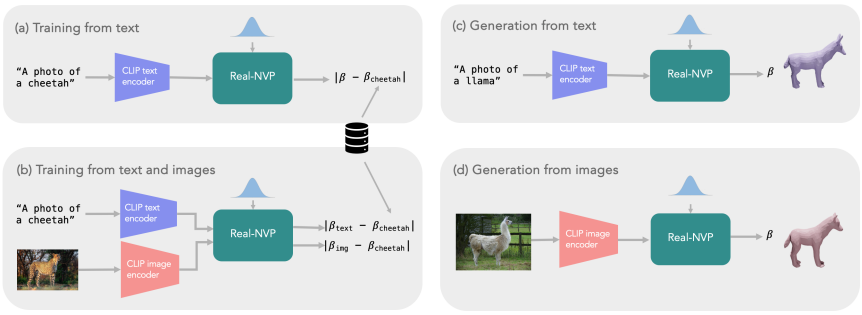

Fig. 4: Network architecture. At training, we can consider only text as input (a), or also provide reference images (b), with about $3-10$ examples for each breed/species. At inference, we can query the text-only network with text (c), or the text-and-image network with images (d).

skinning weight matrix W. An animal is generated, given shape parameters β and pose parameters θ, by first deforming the template into an intrinsic shape $\mathbf{v}_s$, then applying Linear Blend Skinning (LBS) to rotate the body parts according to the given pose:

$$\begin{aligned} \mathbf{v}_s &= \mathbf{v}_t + B\beta^T \\ \mathbf{v} &= LBS(\mathbf{v}_s, \theta; W, J_r). \end{aligned} \tag{1}$$

The linear shape space is learned using Principal Component Analysis (PCA) on a set of 41 quadruped toy scans. The SMAL$^+$ we introduce here is obtained by leveraging the training samples of SMAL [74], D-SMAL [47] and hSMAL [27]. We register the training horses from the hSMAL model, along with additional horse toy scans to the SMAL topology, obtaining a set of 60 registrations. We also add new species: Giraffe, Bear, Mouse, and Rat, learning an animal model from a total of 145 animals. Note that D-SMAL defines dog breeds for the training samples, while in hSMAL the breed of the training horses is undefined. After learning, we collect the set of shape variables for all the training samples, along with their associated species or, in the case of dogs, breed name. This constitutes the training set for the AWOL animal shape prediction.

3.2 Tree Model

The tree model corresponds to the Tree-Gen add-on for Blender [17]. Tree-Gen procedurally generates realistic 3D models of trees based on the method proposed by Weber and Penn [63] and thw Blender's Bézier curve system. The add-on supports saving the generated tree as a triangular mesh. The model generation is controlled by a set of parameters. Some parameters are categorical, referring to a set of defined tree or leaf shapes, while others are numerical, controlling the density of branches and leaves. Additionally, ranges of variation for the numerical parameters are defined, allowing the add-on to generate diverse results from the

same set of parameters. Tree-Gen provides reference parameters labeled with the species name for a set of representative tree shapes. We added the Italian Cypress and Magnolia to the reference trees. This extended set of names and parameters constitutes the training set for the tree shape prediction (Fig. 3).

3.3 Text-to-Shape Model

We base our approach on the real-valued non-volume preserving (Real-NVP) model [10]. Real-NVP is a generative probabilistic model specifically designed for high-dimensional and highly structured data. Formulated with a set of stably invertible transformations and allowing exact and efficient reconstruction, Real-NVP is particularly suited for our task of latent space mapping with limited training data. We summarize Real-NVP here. Let $x \in X$ be an observed, high-dimensional variable, and $z \in Z$ a latent variable, with an associated simple prior distribution p_Z. Let f be a bijection $f : X \to Z$, with $f^{-1} = g : Z \to X$. Using the change of variable formula, a model on x can be defined as:

$$p_X(x) = p_Z(f(x)) \Big| det \left(\frac{\partial f(x)}{\partial x^T} \right) \Big|, \tag{2}$$

where the determinant is computed over the Jacobian of f. In order to generate samples from $p_X(x)$, one would first sample a latent variable z from p_Z, then compute $x = g(z)$. Obtaining the density at x requires computing the Jacobian (Eq. 2). Dinh et al. [10] introduce a convenient construction of f using a set of bijective functions that are easy to invert. They formulate f in a way that its Jacobian is a triangular matrix, allowing for the determinant computation as the product of the diagonal terms. Specifically, f is obtained by stacking a set of *Affine Coupling Layers*. Each coupling layer computes a transformation from the input $x \in \mathbb{R}^D$ to the output $y \in \mathbb{R}^D$ as follows:

$$\begin{aligned} y_{1:d} &= x_{1:d} \\ y_{d+1:D} &= x_{d+1:D} \odot \texttt{exp}(s(x_{1:d})) + t(x_{1:d}), \end{aligned} \tag{3}$$

where $d < D$ and $s()$ and $t()$ are scale and translation functions that convert the input into a vector of dimension $D - d$. These transformations are easy to invert, and obtaining the Jacobian does not require computing derivatives for the scale and translation functions [10]. The partitioning of the input vectors can be modeled with a binary mask. In [10], two strategies are considered: checkerboard masking and dimension-wise masking. In AWOL, we employ Real-NVP to model the conditional distribution of the shape parameters (either shape variables β in SMAL$^+$ or the parameters of the tree Blender add-on), given the CLIP encoding of the textual or visual input. Following [50], we define the input variable x in Eq. 3 as the concatenation between the CLIP encoding and the shape parameters. The output variable z follows a unit Gaussian distribution. We adopt the Real-NVP model with important differences. First, instead of using a fix masking like in previous work [10, 50], we use trainable masks. Second,

Fig. 5: Dog breeds. We verify that CLIP can discriminate the dog breeds in the D-SMAL training set by running a zero-shot classification test on the images above, which achieved 100% accuracy.

unlike the original formulation [10] and ClipForge [50], we aim for data reconstruction during training, employing a reconstruction loss rather than a density estimation loss. This approach has been proved effective for training generative diffusion models [56]. We compare different training losses in our ablation studies. For the reconstruction loss, we use the $L1$ norm between predicted and ground truth shape parameters (See Fig. 4). Note that we also considered a $L2$ loss during development, but it yielded poor results. Finally, we follow previous work in defining simple small networks to implement the scale and translation functions, specifically two Multi Layer Perceptron (MLP) networks. Differently from previous work, we add two additional fully-connected layers that compress the hidden space of those functions. We found that this compression layer is necessary when learning the binary masks, although it hurts performance when the traditional masking approaches are considered. We demonstrate the advantages of our design choices in our experiments.

4 Experiments

We first verify that CLIP can understand and discriminate between different dog breeds and tree species. We consider an image for each of the dog breeds in the D-SMAL model (see Fig. 5), and perform zero-shot classification using the prompt "A photo of a <breed> dog". We found that CLIP can recognize all our training

Fig. 6: Horse breeds. We found with a zero-shot classification test that among the horse breeds above, CLIP can correctly recognize only for the Tinker/Shire horses (violet box) and the Icelandic/Welsh ponies (blue box).

Fig. 7: Tree prediction from text. First row: AWOL; second row: Genie (LumaAI) [1]. The generated tree species are, from left: Ginkgo, Coconut, Cedar of Lebanon, Fig, Cocoa, Bigleaf Maple, Deodar Cedar, Eucalyptus, Tulip, Oak, Banyan, American Elm, Acer, Coast Redwood, Sequoia, Western Red Cedar, White Spruce. None of these species is in the AWOL training set.

breeds. Interestingly, the Chevalier King Charles Spaniel is correctly detected only if indicated as King Charles Spaniel. We perform a similar experiment for our training tree species (Fig. 3) and a set of representative horse breeds (see Fig. 6). We found that the most distinctive trees are correctly recognized, while the majority of the horse breeds cannot be identified, except for ponies and big horses. Therefore, we identify such cases in our animal training set and assign corresponding labels, while the remaining horses are generically labeled as "Horse".

4.1 Implementation

We implement the AWOL network in Pytorch. We define a single network for both animal and tree data, with similar training parameters, the main difference being the dimension of the shape space. The latent shape space for the animal network is the 145-dimensional space of the SMAL$^+$ model. The shape variables are Gaussian distributed with zero mean and identity variance by construction. The latent space for the tree network corresponds to the parameters of the Blender add-on for tree generation. We set the parameters that define the degree of randomness to zero, and we consider only parameters that vary across the reference species, resulting in a latent space with 60 parameters out of the 105 defined by Tree-Gen. We center and normalize the variables by subtracting the mean and dividing by the standard deviation, so that the animal and tree parameters are defined within similar ranges. We do not apply centering and normalization to the categorical variables, which we represent with a one-hot encoding. We use 5 affine coupling layers, and the hidden space for the scale and translation networks has dimension 1024, which we compress to 512 with an additional layer. We encode the text of the sentence "A photo of a <animal>" and "A photo of a <species> tree" for the animal and tree networks, respectively. We train the networks on the text and shape data for 6000 epochs, until the loss stabilizes. We then train the same networks on text and images (Fig. 4b). To do this, we download a set of images from the Web[3], between 3 to 10 for each

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

tree/animal species or breed, and create training tuples composed by the CLIP image encoding and parameters. This training data is larger than previously, and we train the networks for 3000 epochs, until the loss stabilizes. The batch size is 16, and we use the Adam optimizer with a learning rate that varies from $1e-4$ to $1e-6$. We use CLIP ViT-B/32-LAION-2B [19].

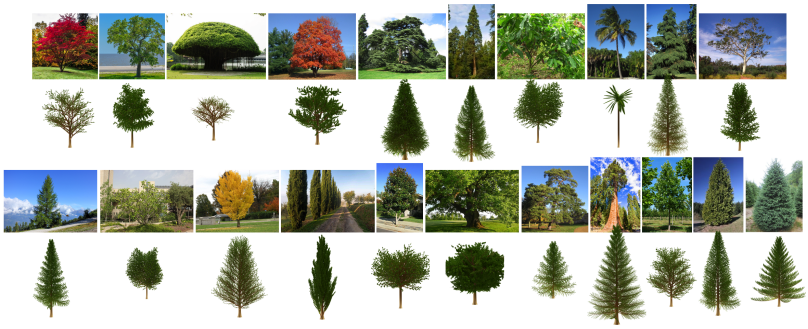

Fig. 8: Tree prediction from images. For each row, the input image is shown at the top and the generated tree is displayed below. Note that we do not predict the tree colors; instead, we show all the trees in a random green color.

4.2 Evaluation

We evaluate our AWOL method in two settings: interpolation and generalization. **Interpolation**. We consider the prediction of new breeds for the dog class as an interpolation task. In nature, dogs of different breeds can mix, and many breeds have been created by mixing existing ones [13]. We argue that, given the large number of breeds included in the model, it is likely that new breed shapes can be generated by interpolation in the space of dog shapes, even though there may be unseen breeds with specific shape features not seen during training. We qualitatively demonstrate interpolation by generating dog breeds and comparing them with BITE [47] (Fig. 14). We also show interpolation for age and size. We query for "Giant Schnauzer", "Standard Schnauzer", "Miniature Schnauzer" and "Toy Schnauzer", and similarly for the Poodle. Note in Figure 9 how the network correctly predicts the scale of the different varieties of the breeds (it is worth noting that for the Schnauzer, the breed varieties are only Giant, Standard and Miniature). We then investigate if AWOL can interpolate shapes and age-dependent features by querying for "Baby", "Young", 'Adult" and 'Old" animals. Figure 9 shows the results for seen and unseen species. Figure 10 presents an analogous analysis for trees. We quantitatively compare the dog breed predictions from textual input with BITE [47] using a perceptual study. For each breed in the StanfordExtra test set [3], we generate a 3D dog, and compare it with the dog reconstructed by BITE from a randomly selected image of the same breed. We

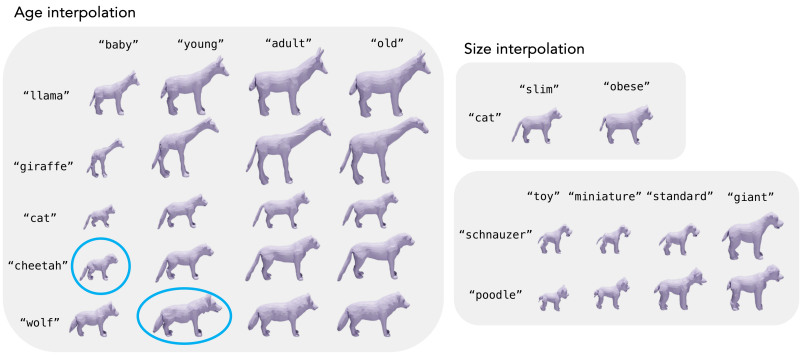

Fig. 9: Age interpolation (left) and size interpolation (right). The circles indicate the animals that are present in the training set as "Baby Cheetah" and "Young Wolf". Giraffe, Cat, and Wolf are in the training set without attributes, while the Llama is not. The small and large Poodles are present in the training as 3D shapes, but their text attribute is "Poodle". Only one shape example for the Schnauzer is present, named "Schnauzer". Note how we can recover the different Poodle breed size variations. For the Schnauzer, the actual breed variations are Miniature, Standard, and Giant.

asked Amazon Mechanical Turk workers to judge which method better represents the dog breed in the picture. Overall, BITE outperforms AWOL with 971 votes versus 884 votes, as confirmed by a binomial test with a p-value of 0.02 (BITE better than AWOL). We noticed that the task favors BITE when the subject in the image is a puppy, as AWOL used without age input generates an adult subject. By removing from the evaluation the images with baby dogs from the evaluation, we obtain votes of 830 for BITE versus 850 for (AWOL) (with a p-value of 0.3; AWOL better than BITE), indicating the ability of AWOL to faithfully generate a large variety of breeds.

Generalization. To test generalization, we prompt the model to create new quadruped species. Figure 11 shows examples of generations from text. We also show examples of reconstructed novel trees from textual and image input in Figure 7 and Figure 8, respectively. Finally, Figure 12 presents examples of animal generation from images, many of which are taken from [75] for comparison. The unseen animals include the Llama, Thylacine, Panda, Pig, Rhino, and Cougar. Figure 13 provides a comparison with DeepTree [73].

4.3 Ablation Studies

We perform our ablation studies on the animal model using CLIP for evaluation, as we found that CLIP can successfully classify animals and dog breeds, enabling quantitative testing on a larger set of cases. The ablation studies evaluate: the effect of learning the binary masks in Real-NVP; the effect of training with a density loss; and the effect of adding the compression layer in the scale and translation functions. We also compare results when reducing the shape

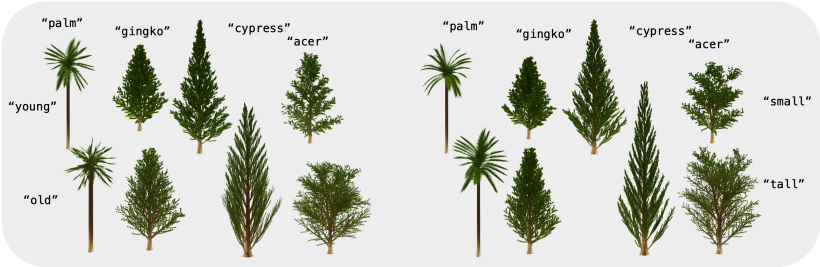

Fig. 10: Age and size interpolation for trees. Palm and Cypress are in the AWOL training set, while Ginkgo and Acer are unseen species. Here the query is "A photo of a <age> <species> tree".

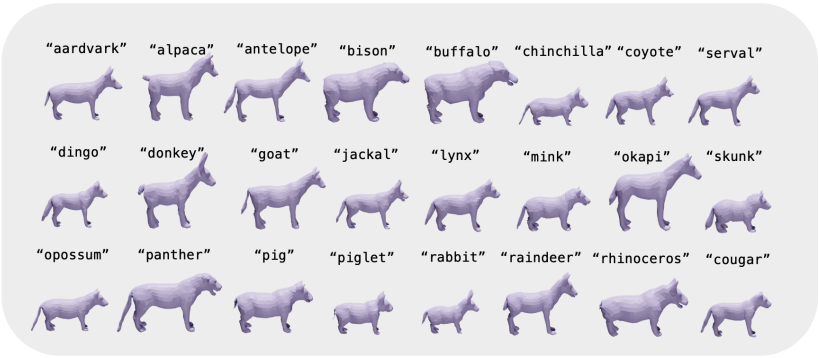

Fig. 11: Animal prediction from text. We generate species that are not present in the SMAL$^+$ and AWOL training sets. The image shows the actual model size. Note that AWOL generates animals within the SMAL+ shape space, and therefore, it cannot create specific details such as the horn on the rhino.

space dimension from 145 (the space of the SMAL$^+$ model) to 40, approximately matching the dimension of the single SMAL models for dogs and horses [27, 47]. We generate a set of 122 animals, none of which are present in the SMAL$^+$ model training set. This selection covers most common quadrupeds, and several unseen dog breeds. We query the network with the sentence "A photo of a <animal name>", where <animal name> is either a quadruped species or a dog breed. We then render the predicted 3D models in grayscale to prevent any color bias. Since the networks can predict different animal sizes, we consider bounding boxes and render the animals to maximize visibility of their profile. We found that the lateral view is the most informative, while adding further views gave inconsistent results. We perform paired comparisons between different networks by testing, for each animal, which of the two network predictions, encoded in CLIP, is closest to the CLIP encoding of the animal name. This corresponds to a CLIP "vote". Note that, even if we base our method on CLIP, we believe it is

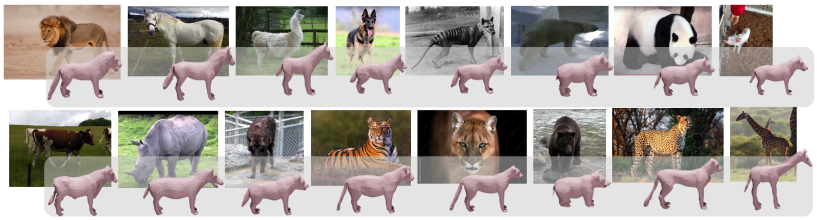

Fig. 12: Animals prediction from images. The images of the Horse, Dog, Thylacine, Polar Bear, Panda, Pig, Cow, Rhino, and Bear are taken from [75]. We replace the green screen images in [75] with natural images for the Lion, Tiger, Cougar, and Cheetah.

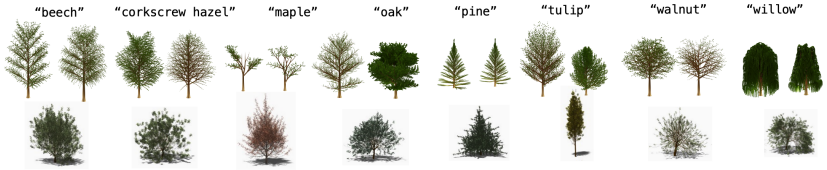

Fig. 13: Comparison with DeepTree. We show predicted trees from text (top), compared with DeepTree (bottom, images taken from [73]). For each predicted pair: (left) network trained only on text, (right) network trained on text and images.

appropriate to use CLIP for the ablation studies, as we are comparing different architectures, under the same conditions. Results are reported in Table 1. Our ablation studies confirm that the network with learned masks and compression of the hidden space for the scale and translation networks provides the best performance on the whole test set.

5 Conclusion

We have addressed the problem of generating 3D objects from text and images using parametric 3D models. Inspired by recent work on learning multimodal latent spaces, we use language to control the selection of the 3D model parameters. We make the hypothesis that using language, we can achieve interpolation and generalization in parametric shape spaces. We demonstrate our hypothesis on two different 3D generative models: on a novel differentiable 3D parametric shape model for animals, which extends previous models with new training samples and species, and on a non-differentiable model for trees, represented by a Blender add-on. Our qualitative and quantitative experiments confirm our hypothesis. The proposed AWOL is the first system that allows generating rigged 3D animals and trees with a simple text prompt. **Acknowledgements**. We thank Tsvetelina Alexiadis, Taylor McConnell and Tomasz Niewiadomski for their help in running the Amazon Mechanical Turk evaluation. We also thank Charlie Hewitt for making his tree generation method available and the authors of [50] for sharing their code. SZ is funded NRRP, Miss. 4

	CLIP-based Comparison: % of votes			
	All	p-value	Dogs	Other Species
A. Check vs. Dims	61:39	0.19	68:32	43:57
B. Dims vs. Dims + Comp.	52:48	0.47	51:49	54:46
C. Check vs. Check. + Comp.	63:37	0.20	64:36	60:40
D. Learn + Comp. vs. Learn	61:39	0.19	59:41	69:31
E. Learn + Comp., 145 vs. 40	50:50	0.58	48:52	54:46
F. Learn + Comp. vs. Dims	62:38	0.13	68:32	46:54
G. Learn + Comp vs. Check	54:46	0.38	53:47	57:43
H. Learn + Comp, density loss	86:14	1.24e-7	86:14	86:14

Table 1: Ablation results. Comparison between different networks. "Check" refers to checkerboard masking, "Dims" to dimension-wise masking, "Comp" to hidden space compression, "Learn" to learned masks. (E) compares the Learn + Comp network with 145 (default) versus 40 shape parameters. The table shows that the best performance on the whole test set is achieved by the network with learned mask and compression (D, F, G). When training also includes a density loss [10], performance degrades significantly.

Comp. 2 Inv. 1.4 - Call No. 3138 16/12/21, rect. by Decree n.3175 18/12/21 of MUR funded by NextGenerationEU; Award N.: Proj. code CN00000033, Conc. Decree No. 1034 17/06/22 CUP B83C22002930006, title: National Biodiversity Future Center - NBFC. SZ is also supported by PNRR FAIR Future AI Research (PE00000013), Spoke 8 Pervasive AI (CUP H97G22000210007) under the NRRP MUR program by NextGenerationEU. **Disclosure**: `https://files.is.tue.mpg.de/black/CoI_ECCV_2024.txt`.

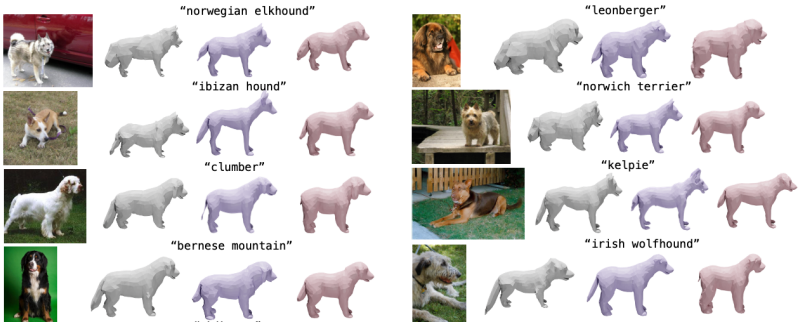

Fig. 14: Comparison with BITE [47]. Randomly chosen images from the StanfordExtra test set. From left: input image, BITE in natural pose (gray), AWOL with textual input (purple), AWOL with image input (red). For both BITE and AWOL with text input, we use the breed label to rotate the ears. For AWOL from images, the ears are down by default. None of these breeds are in the AWOL training set.

References

1. https://lumalabs.ai/genie
2. Aono, M., Kunii, T.L.: Botanical tree image generation. IEEE Computer Graphics and Applications **4**(5), 10–34 (1984). https://doi.org/10.1109/MCG.1984.276141
3. Biggs, B., Boyne, O., Charles, J., Fitzgibbon, A., Cipolla, R.: Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In: ECCV. pp. 195–211. Lecture Notes in Computer Science, Springer International Publishing (Aug 2020)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV. p. 561–578. Lecture Notes in Computer Science, Springer International Publishing (Sep 2016)
5. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: ICCV. pp. 22246–22256 (Oct 2023)
6. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A., Gui, L.: SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In: CVPR. pp. 4456–4465 (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00433
7. Choutas, V., Müller, L., Huang, C.H.P., Tang, S., Tzionas, D., Black, M.J.: Accurate 3d body shape regression using metric and semantic attributes. In: CVPR. pp. 2708–2718. IEEE, Piscataway, NJ (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.00274
8. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)
9. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR. pp. 13142–13153. IEEE (Jun 2023), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2023.html#DeitkeSSWMVSEKF23
10. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. In: ICLR (Apr 2017)
11. Dwivedi, S.K., Athanasiou, N., Kocabas, M., Black, M.J.: Learning to regress bodies from images using differentiable semantic rendering. In: ICCV. pp. 11230–11239. IEEE, Piscataway, NJ (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01106
12. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: International Conference on 3D Vision (3DV). pp. 792–804. IEEE, Piscataway, NJ (Dec 2021). https://doi.org/10.1109/3DV53792.2021.00088
13. G., P.H., L., D.D., M., R., W., D.B., B., M.A., G., C.R., A., O.E.: Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. Cell Reports **4**(19), 697–708 (Apr 2017)
14. Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text. In: ICCV. pp. 7545–7556 (Oct 2023)
15. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (Oct 2023)

16. Gralnik, O., Gafni, G., Shamir, A.: Semantify: Simplifying the control of 3d morphable models using clip. In: ICCV. pp. 14554–14564 (Oct 2023)
17. Hewitt, C.: Procedural generation of tree models for use in computer graphics (2017), https://github.com/friggog/tree-gen
18. Hu, J., Hui, K.H., Liu, Z., Zhang, H., Fu, C.W.: Clipxplore: Coupled clip and shape spaces for 3d shape exploration. In: SIGGRAPH Asia 2023 Conference Papers. SA '23, Association for Computing Machinery, New York, NY, USA (Dec 2023). https://doi.org/10.1145/3610548.3618144
19. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773
20. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: CVPR (Jun 2022)
21. Jain, A., Sunkara, J., Shah, I., Sharma, A., Rajan, K.S.: Automated tree generation using grammar & particle system. In: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing. pp. 1–9. ICVGIP '21, Association for Computing Machinery, New York, NY, USA (Dec 2021). https://doi.org/10.1145/3490035.3490285
22. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: ICCV. pp. 11107–11117. IEEE, Piscataway, NJ (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01094
23. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: SPEC: Seeing people in the wild with an estimated camera. In: ICCV. pp. 11015–11025. IEEE, Piscataway, NJ (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01085
24. Lee, J.J., Li, B., Benes, B.: Latent l-systems: Transformer-based tree generator. ACM Trans. Graph. **43**(1), 1–16 (Nov 2023). https://doi.org/10.1145/3627101
25. Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: Gart: Gaussian articulated template models. In: CVPR (Jun 2024)
26. Li, B., Kałużny, J., Klein, J., Michels, D.L., Pałubicki, W., Benes, B., Pirk, S.: Learning to reconstruct botanical trees from single images. ACM Trans. Graph. **40**(6), 1–15 (Dec 2021). https://doi.org/10.1145/3478513.3480525
27. Li, C., Ghorbani, N., Broomé, S., Rashid, M., Black, M.J., Hernlund, E., Kjellström, H., Zuffi, S.: hsmal: Detailed horse shape and pose reconstruction for motion pattern recognition. arXiv preprint arXiv:2106.10102 (2021). https://doi.org/10.48550/arXiv.2106.10102
28. Li, M., Zhou, P., Liu, J.W., Keppo, J., Lin, M., Yan, S., Xu, X.: Instant3d: Instant text-to-3d generation. arXiv:2311.08403 (2024), https://arxiv.org/abs/2311.08403
29. Li, Z., Litvak, D., Li, R., Zhang, Y., Jakab, T., Rupprecht, C., Wu, S., Vedaldi, A., Wu, J.: Learning the 3d fauna of the web. In: CVPR (Jun 2024)
30. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (Jun 2023)
31. Liu, D., Zhangli, Q., Gao, Y., Metaxas, D.N.: Lepard: learning explicit part discovery for 3d articulated shape reconstruction. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (Dec 2024). https://doi.org/10.5555/3666122.3668481
32. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (Oct 2023)

33. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Trans. Graph. **34**(6) (Oct 2015). https://doi.org/10.1145/2816795.2818013
34. Lorraine, J., Xie, K., Zeng, X., Lin, C.H., Takikawa, T., Sharp, N., Lin, T.Y., Liu, M.Y., Fidler, S., Lucas, J.: Att3d: Amortized text-to-3d object synthesis. In: ICCV. pp. 17946–17956 (Oct 2023)
35. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: CVPR. pp. 12663–12673. IEEE (Jun 2023), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2023.html#MetzerRPGC23
36. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (Dec 2021). https://doi.org/10.1145/3503250
37. Müller, L., Ye, V., Pavlakos, G., Black, M.J., Kanazawa, A.: Generative proxemics: A prior for 3D social interaction from images. In: CVPR (Jun 2024)
38. Oppenheimer, P.E.: Real time design and animation of fractal plants and trees. In: Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques. p. 55–64. SIGGRAPH '86, Association for Computing Machinery, New York, NY, USA (Aug 1986). https://doi.org/10.1145/15922.15892
39. Pavlakos, G., Malik, J., Kanazawa, A.: Human mesh recovery from multiple shots. In: CVPR (Jun 2022)
40. Pavlakos, G., Weber, E., Tancik, M., Kanazawa, A.: The one where they reconstructed 3d humans and environments in tv shows. In: ECCV. pp. 732–749. Lecture Notes in Computer Science, Springer International Publishing (Oct 2022)
41. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: DreamFusion: Text-to-3d using 2d diffusion. In: ICLR (May 2023)
42. Prusinkiewicz, P., Lindenmayer, A.: Graphical modeling using L-systems, pp. 1–50. Springer New York, New York, NY (1990). https://doi.org/10.1007/978-1-4613-8476-2_1
43. 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: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
44. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022). https://doi.org/10.48550/arXiv.2204.06125
45. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8821–8831. PMLR (Jul 2021), https://proceedings.mlr.press/v139/ramesh21a.html
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685 (Jun 2022)
47. Rüegg, N., Tripathi, S., Schindler, K., Black, M.J., Zuffi, S.: BITE: Beyond priors for improved three-D dog pose estimation. In: CVPR. pp. 8867–8876 (Jun 2023)
48. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (Jun 2023)

49. Saharia, C., Chan, W., Saxena, S., Lit, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (Nov-Dec 2022)
50. Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.Y., Fumero, M.: Clip-forge: Towards zero-shot text-to-shape generation. In: CVPR (Jun 2022)
51. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
52. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. In: ICLR (May 2024)
53. Streuber, S., Quiros-Ramirez, M.A., Hill, M.Q., Hahn, C.A., Zuffi, S., O'Toole, A., Black, M.J.: Body Talk: Crowdshaping realistic 3D avatars with words. ACM Trans. Graph. (Proc. SIGGRAPH) **35**(4), 54:1–54:14 (Jul 2016). https://doi.org/10.1145/2897824.2925981
54. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3D people. In: ICCV. pp. 11159–11168. IEEE, Piscataway, NJ (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01099
55. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments. In: CVPR. pp. 8856–8866 (Jun 2023)
56. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (May 2023), https://openreview.net/forum?id=SJ1kSyO2jwu
57. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: CVPR. pp. 4713–4725 (Jun 2023), https://ipman.is.tue.mpg.de
58. Tsalicoglou, C., Manhardt, F., Tonioni, A., Niemeyer, M., Tombari, F.: Textmesh: Generation of realistic 3d meshes from text prompts. In: International Conference on 3D Vision (3DV). pp. 1554–1563. IEEE Computer Society, Los Alamitos, CA, USA (Mar 2024). https://doi.org/10.1109/3DV62453.2024.00154
59. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: CVPR (Jun 2023)
60. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: CVPR (2023)
61. Wang, Y., Kolotouros, N., Daniilidis, K., Badger, M.: Birds of a feather: Capturing avian shape models from images. In: CVPR (Jun 2021)
62. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In: NeurIPS (Dec 2023)
63. Weber, J., Penn, J.: Creation and rendering of realistic trees. ACM TOG pp. 119–128 (Dec 1995)
64. Wu, M., Zhu, H., Huang, L., Zhuang, Y., Lu, Y., Cao, X.: High-fidelity 3d face generation from natural language descriptions. In: CVPR (Jun 2023)
65. Wu, S., Li, R., Jakab, T., Rupprecht, C., Vedaldi, A.: Magicpony: Learning articulated 3d animals in the wild. In: CVPR. pp. 8792–8802 (June 2023)
66. Xie, K., Lorraine, J., Cao, T., Gao, J., Lucas, J., Torralba, A., Fidler, S., Zeng, X.: Latte3d: Large-scale amortized text-to-enhanced3d synthesis. arXiv preprint arXiv:2403.15385 (2024), https://arxiv.org/abs/2403.15385

67. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024), https://arxiv.org/abs/2404.07191
68. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3D: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. arXiv preprint arXiv:2212.14704 (2023), https://arxiv.org/abs/2212.14704
69. Yang, G., Wang, C., Reddy, N.D., Ramanan, D.: Reconstructing animatable categories from videos. In: CVPR. pp. 16995–17005 (Jun 2023)
70. Yao, C.H., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Lassie: Learning articulated shape from sparse image ensemble via 3d part discovery. In: NeurIPS (Dec 2022)
71. Yao, C.H., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In: CVPR (Jun 2023)
72. Yao, C.H., Raj, A., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Artic3d: Learning robust articulated 3d shapes from noisy web image collections. In: NeurIPS (Dec 2023)
73. Zhou, X., Li, B., Benes, B., Fei, S., Pirk, S.: Deeptree: Modeling trees with situated latents. Transactions on Visualization & Computer Graphics **1**, 1–14 (2023). https://doi.org/10.1109/TVCG.2023.3307887
74. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.J.: Three-D safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In: ICCV. pp. 5359–5368 (Nov 2019)
75. Zuffi, S., Kanazawa, A., Black, M.J.: Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In: CVPR. pp. 3955–3963. IEEE Computer Society (Jun 2018)
76. Zuffi, S., Kanazawa, A., Jacobs, D., Black, M.J.: 3D menagerie: Modeling the 3D shape and pose of animals. In: CVPR. pp. 5524–5532 (Jul 2017)