CatFree3D: Category-Agnostic 3D Object Detection With Diffusion

Wenjing Bian[1] Zirui Wang[1] Andrea Vedaldi[2]
[1]Active Vision Lab [2]Visual Geometry Group
University of Oxford
{wenjing, ryan, vedaldi}@robots.ox.ac.uk
https://bianwenjing.github.io/CatFree3D

Abstract

Image-based 3D object detection is widely employed in applications such as autonomous vehicles and robotics, yet current systems struggle with generalisation due to complex problem setup and limited training data. We introduce a novel pipeline that decouples 3D detection from 2D detection and depth prediction, using a diffusion-based approach to improve accuracy and support category-agnostic detection. Additionally, we introduce the Normalised Hungarian Distance (NHD) metric for an accurate evaluation of 3D detection results, addressing the limitations of traditional IoU and GIoU metrics. Experimental results demonstrate that our method achieves state-of-the-art accuracy and strong generalisation across various object categories and datasets.

1. Introduction

Image-based 3D object detection systems are designed to identify and localise objects in three-dimensional space from input images. These systems play a critical role in applications like autonomous vehicles and robotics.

Recent developments in deep learning have significantly advanced 3D object detection [6, 11, 32, 39, 40, 50, 51, 69]. While these methods have been well engineered, they remain largely domain-specific and are limited in the number of detectable object categories, especially when compared to state-of-the-art 2D detection systems [8, 60], which can detect hundreds of categories across various domains.

The performance gap between 2D and 3D detection, particularly in generalising to more detectable categories, is mostly due to i) the complex problem setup and ii) insufficient training data. The 3D detection task is closely related to various other tasks such as 2D detection, depth estimation and object pose estimation, each of which is a challenging research area. Additionally, labelling 3D data is more labour-intensive, requiring specifying nine degrees of freedom instead of four in 2D. This combination of complexity and limited data restricts current 3D detection methods to fewer categories, often with reduced accuracy.

To overcome these limitations, we propose a pipeline that decouples the 3D detection task from 2D detection and depth prediction. This decoupling enhances training efficiency, and, most importantly, allows for a category-agnostic approach with improved accuracy.

Our key idea lies in recovering a 3D bounding box from a random noise, conditioned on several visual prompts, using a denoising network inspired by diffusion models [20, 53] in a generative fashion. Specifically, the random noise is sampled from a normal distribution, with visual prompts consisting of the image of the target object, a 2D detection bounding box, and the depth of the target. During training, we take advantage of ground truth labels of 2D bounding boxes and object depths. During inference, the model can be integrated with 2D detectors and depth estimation models or take prompts from various sources, *e.g.* human annotation.

In addition to simplifying existing pipelines and achieving category-agnostic detection, our diffusion-based approach allows us to generate an *arbitrary* number of predictions for a *single* target due to their stochastic nature. We leverage this by estimating multiple 3D bounding boxes for one target, assigning a confidence score to each, and selecting the most confident one, further improving detection accuracy.

While developing our novel 3D detection method, we discovered that conventional metrics, such as Intersection over Union (IoU) and Generalised IoU (GIoU), often struggle to evaluate 3D detection results accurately, particularly for non-overlapping or enclosing cases, which are common for thin and small objects. To address these limitations, we propose a new metric called *Normalised Hungarian Distance* (NHD), which seeks a one-to-one assignment between the corners of the ground truth and predicted 3D bounding boxes, then calculates the Euclidean distance between corresponding corners, offering a more detailed and precise evaluation of 3D object detection.

In summary, we make three key contributions: First, we introduce a novel diffusion-based pipeline for 3D object detection that decouples the 3D detection task from 2D detection and depth prediction, enabling category-agnostic 3D detection. Second, we enhance the 3D detection accuracy by leveraging generative capabilities in our diffusion-based pipeline to predict multiple bounding boxes with confidence scores. Third, we propose the *Normalised Hungarian Distance* (NHD), a new evaluation metric that provides a more precise assessment of 3D detection results.

As a result, our method achieves state-of-the-art accuracy in 3D object detection in a category-agnostic manner and demonstrates strong generalisation to unseen datasets.

2. Related Work

2D Object Detection. 2D object detectors include two-stage detectors [15, 46] that use a coarse-to-fine approach and single-stage detectors [8, 34, 46] that directly estimate the location of objects from the extracted visual features. DiffusionDet [9] was the first to apply diffusion to detection tasks, progressively refining noisy 2D boxes towards the target objects. Category-agnostic 2D detection models are conceptually similar to our approach in that they avoid using category information. These models learn to differentiate generic objects from the image background using low-level visual cues [41, 72] or using supervision from bounding box labels [25, 26, 38]. However, instead of learning objectness like these methods, we focus on mapping 2D boxes to 3D boxes by leveraging visual cues from input images.

Monocular 3D Object Detection Monocular 3D object detectors predict 3D cuboids from a single input image. Depending on the dataset domain, certain models are tailored for outdoor self-driving scenes [11, 23, 35, 37, 39, 39, 44, 45, 61], while others are specifically designed for indoor environments [24, 30, 40, 59]. Additionally, some studies [6, 32, 50] have explored integrating both indoor and outdoor datasets during training. These methods often use category labels for supervision [6, 23, 32, 35], require category information as a prior for initialisation or as input [6, 40, 44], or focus on specific scenes and object categories with strong assumptions about the predictions' dimensions or orientation [11, 39, 51, 64]. This reliance on category and scenario-specific knowledge limits their generalisation to in-the-wild scenes and novel categories. In contrast, our approach does not use category information during training or inference, focusing solely on predicting 3D bounding boxes. This allows the model to be used for novel objects that were not present during training.

Diffusion Models for Visual Perception Diffusion models [20, 52, 55, 56] have demonstrated remarkable results in computer vision [18, 21, 66], natural language processing [3, 17, 31, 67], and multimodal data generation [4, 42, 43, 49, 68]. In visual perception tasks, DiffusionDet [9] was the first to apply box diffusion for 2D object detection from a single RGB image. Additionally, diffusion has been utilised for tasks like image segmentation [2, 10] and human pose estimation [16, 22]. For 3D object detection, 3DiffTection [63] leverages image features from a 3D-aware diffusion model to enhance the detection performance, while directly predicting box parameters. Another line of approaches applies the diffusion process to the box parameters. Zhou *et al.* [70] introduced diffusion of 5 DoF Bird's Eye View boxes as proposals for detection from point clouds. Diffusion-SS3D [19] employs diffusion for semi-supervised object detection in point clouds, denoising object size and class labels. DiffRef3D [27] and DiffuBox [12] apply diffusion to refine proposals/coarse boxes for 3D object detection from point clouds. MonoDiff [44] uses Gaussian Mixture Models to initialise the dimensions and poses of 3D bounding boxes, recovering the target boxes through diffusion conditioned on an image. Unlike these approaches, which assume an initial distribution or proposals for the diffusing box parameters [12, 44], limited DoFs for box orientation and dimensions [12, 27], or diffuse only partial parameters [19, 44], our model initialise all box parameters with random noise for diffusion. By conditioning the model on an image and a 2D box, our model recovers using diffusion the in-plane translations, three sizes, and three DoF for the rotation of the 3D box.

3. Method

Given an image I, a 2D bounding box of an object B, the object depth z, and the camera intrinsics K, our objective is to estimate the centre, 3D size and orientation of a 3D box that tightly encloses the object. We formulate this task as a conditional diffusion process [20], progressively recovering a target box from a noise sampled from normal distributions, conditioned on multiple prompts.

Let us consider a general diffusion setup first. In the forward diffusion process, Gaussian noise is incrementally added to a variable $\mathbf{x}_0$ over T time steps until it follows a normal distribution. During backward denoising, an estimate $\hat{\mathbf{x}}_0$ can be recovered from its noisy version $\mathbf{x}_t$ using a neural network f_θ:

$$f_\theta(\mathbf{x}_t, t) \rightarrow \hat{\mathbf{x}}_0, \tag{1}$$

where $t \in [1, T]$ denotes a diffusion step. *In our 3D object detection task, we consider* $\mathbf{x}_0$ *the parameters of a 3D box.*

To adapt the general diffusion process to this vision-based 3D detection setup, we consider a conditional denoising network f_θ:

$$f_\theta(\mathbf{x}_t, t, \mathbf{c}) \rightarrow \hat{\mathbf{x}}_0, \tag{2}$$

where $\mathbf{c}$ denotes the conditional signal that includes information from the input image I, the 2D bounding box B,

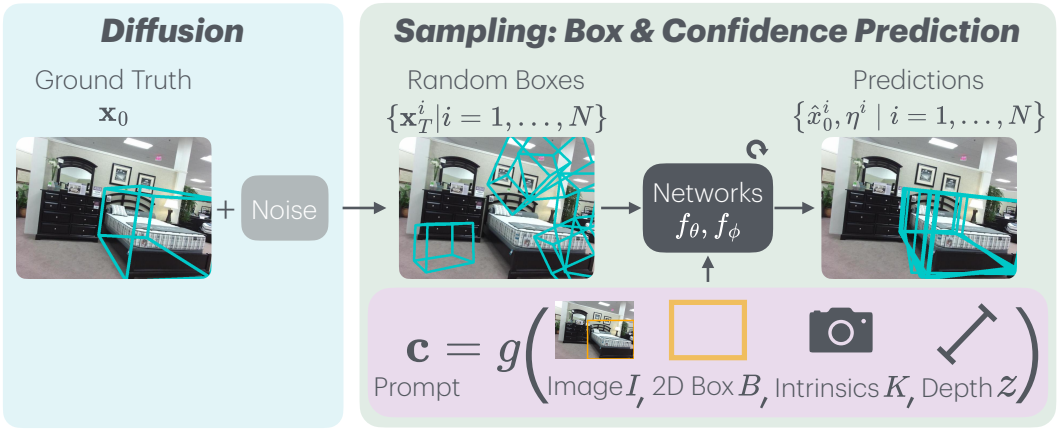

Figure 1. **Method Overview.** During forward diffusion, we add N independent Gaussian noises to a ground truth box $\mathbf{x}_0$ to obtain a number of noisy boxes. We then train a denoising network f_θ to recover the target box parameters $\hat{\mathbf{x}}_0$ from noisy boxes, conditioned on a vision-related signal $\mathbf{c}$. Additionally, we train another network f_ϕ to estimate a confidence score η for each predicted box. The final output is the box with the highest confidence score.

the camera intrinsics K, and the object depth z. This conditional diffusion process is similar to diffusion-based text-to-image generation tasks [49, 68], where the conditional signal typically consists of text descriptions.

In Sec. 3.1 we begin by introducing the parametrisation of the 3D bounding boxes and the prompt encoding. We then explain the forward diffusion and reverse sampling in Sec. 3.2 and Sec. 3.3 respectively, detailing how we use the generative properties of diffusion networks to predict multiple proposals and their associated confidence scores. Finally, we outline our training processes and losses in Sec. 3.4.

3.1. Preparation

Box Parameterisation We consider the position, orientation and size of a 3D bounding box. Specifically, each 3D bounding box $\mathbf{x}_0$ is represented using 11 parameters:

$$\mathbf{x} := [u, v, w, h, l, \mathbf{p}]. \tag{3}$$

The position of the 3D box is defined by the 2D projected coordinates u and v, which represent the object centre on the image plane of the observing camera. This position parametrisation decouples the image plane position components and the depth component, allowing us to leverage object depth from various sources. The orientation of the object relative to the camera of the input image is represented by the continuous 6D allocentric rotation $\mathbf{p} \in \mathbb{R}^6$ [71]. The size of the 3D bounding box is captured by w, h, and l. Overall, this parametrisation follows [6], but with the depth component excluded.

Prompts Encoding Our conditioning signal $\mathbf{c}$ is derived from the image I, a 2D bounding box B, camera intrinsics K, and the object depth z:

$$\mathbf{c} = g(I, B, K, z), \tag{4}$$

where function $g(\cdot)$ denotes a prompt encoding function. This function includes an image encoding backbone, positional encoding functions, and a shallow MLP that summaries all prompt information in preparation for the box prediction network. Further details on the prompt encoding can be found in the supplementary material.

3.2. Diffusion: Adding Noise to a Box

In the forward diffusion process, we start with a noise-free 3D bounding box, denoted as $\mathbf{x}_0$, and iteratively add Gaussian noise over T steps to generate a fully noisy box $\mathbf{x}_T$, which follows a normal distribution. This process follows the standard DDPM schedule [20].

Preprocessing Before applying noise to the original 3D bounding box $\mathbf{x}_0$, we perform additional normalisation and scaling to ensure that $\mathbf{x}_0$ lies within the range $[-s, s]$, where s corresponds to the signal-to-noise ratio of the diffusion process. During the normalisation step, the projected coordinates u and v are normalised relative to the image dimensions. The dimensions of the 3D bounding box—width w, height h, and length l—are normalised against a predefined maximum box size. The orientation of the box $\mathbf{p}$, expressed in the allocentric representation [71], is inherently normalised and does not require further adjustment. In the scaling step, all box parameters are further scaled by a scalar s. As shown in Sec. 5.5, this scaling step, adopted from [9, 10], improves box prediction accuracy.

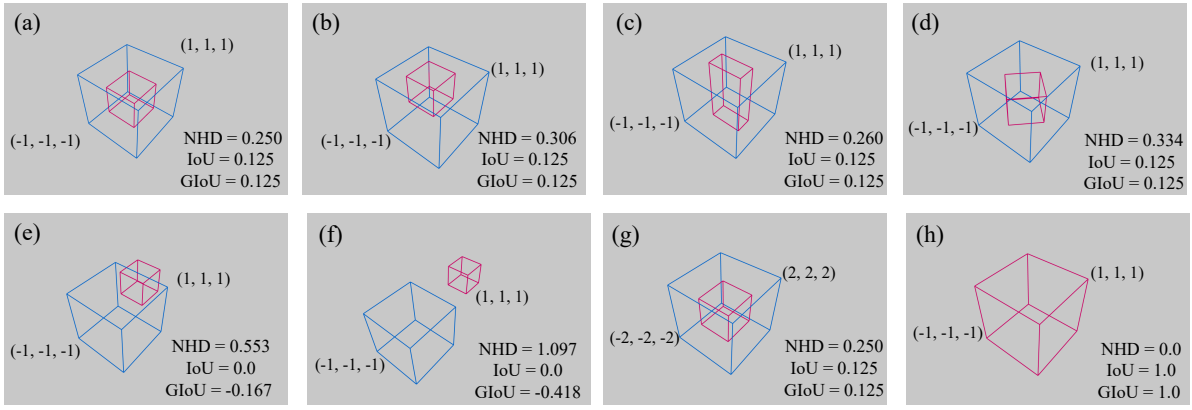

Figure 2. **Comparison Between NHD, IoU and GIoU**. Comparing block (a) with blocks (b, c, d, e, f), we show that NHD provides a more accurate measurement of errors compared to IoU and GIoU, particularly under translation, scaling, and rotation transformations. Block (g) demonstrates that all three metrics are scale-invariant. Block (h) presents metric values when the two boxes are perfectly aligned.

3.3. Sampling: Predicting a Box

Our method predicts a 3D bounding box for a target object using a denoising network f_θ, conditioned on a vision-related prompt $\mathbf{c}$. To enhance detection performance, we introduce a confidence score for each predicted box, generating multiple candidate boxes and selecting the one with the highest confidence.

Single Box Prediction To predict a single 3D box $\hat{\mathbf{x}}_0$, we begin with a randomly sampled noise $\mathbf{x}_T$ and iteratively refine the 3D box $\hat{\mathbf{x}}_t$ with our denoising network f_θ, conditioned on the encoded prompt $\mathbf{c}$. This process continues until the final denoising step $t = 0$, following the standard sampling procedure introduced in DDIM [53].

Multi-Box Prediction and Selection Leveraging the generative capabilities of diffusion networks, our method allows for the prediction of *multiple* 3D boxes for a *single* target object. Specifically, for each target object, we sample N 3D box parameters $\{\mathbf{x}_T^i | i = 1, \dots, N\}$ from a normal distribution, producing N predictions $\{\hat{\mathbf{x}}_0^i | i = 1, \dots, N\}$. We introduce a learnable confidence score η^i for each prediction, where we select the box with the highest confidence score as the final prediction during inference.

Confidence Prediction To estimate the uncertainty $\mu \in (0, \infty)$ for each box prediction, we employ an additional network branch f_ϕ, which takes the current box estimation and the vision conditioning signal $\mathbf{c}$ as inputs:

$$f_\phi(\mathbf{c}, \mathbf{x}_t) \rightarrow \mu. \tag{5}$$

The confidence score $\eta \in (0, 1)$ is derived from the uncertainty through an exponential mapping $\eta = e^{-\mu}$. This score reflects the agreement between the current box estimation and the provided vision prompt.

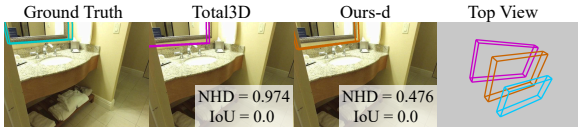

Figure 3. **IoU and NHD in a Practical Example.** For thin objects like mirrors, even a small translational offset can lead to an IoU of 0. In contrast, NHD effectively captures and reflects the box estimation error in these cases.

3.4. Training

We follow the standard DDPM [20] training process and utilise a training loss $\mathcal{L}$ consisting of two loss terms:

$$\mathcal{L} = \mathcal{L}_{\text{3D}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \tag{6}$$

with λ_{reg} being a hyperparameter that balances the regularisation term $\mathcal{L}_{\text{reg}}$ and the reconstruction term $\mathcal{L}_{\text{3D}}$.

The reconstruction loss $\mathcal{L}_{\text{3D}}$ is designed to encourage our denoising network to predict accurate 3D boxes. This is achieved by penalising the Chamfer distance between the corners of predicted boxes and the ground truth, weighted by the confidence η:

$$\mathcal{L}_{\text{3D}} = \frac{1}{N} \sum_{i=1}^{N} \eta^i \mathcal{L}_{\text{chamfer}}(\hat{\mathbf{x}}_0^i, \mathbf{x}_0), \tag{7}$$

whereas the regularisation term $\mathcal{L}_{\text{reg}}$ prevents the uncertainty prediction μ being excessively large:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \mu^i; \tag{8}$$

Details regarding the Chamfer distance $\mathcal{L}_{\text{chamfer}}$ between two 3D boxes are provided in the supplementary material.

Table 1. **Detection Performance: Comparing Our Method with Cube R-CNN [6] and Total3D [40] on Omni3D Dataset.** The top three rows use GT 2D boxes along with predicted depths. The depths of our predictions are set to the same as [6] for fair comparison. The bottom three rows use GT 2D boxes and GT depths for all methods.

Methods	SUN RGB-D		Omni3D Indoor		Omni3D Indoor & Outdoor	
	IoU (%) ↑	NHD ↓	IoU (%) ↑	NHD ↓	IoU (%) ↑	NHD ↓
Total3D	24.8	0.376	-	-	-	-
Cube R-CNN	36.2	0.236	19.5	0.667	23.0	0.593
Ours-d	**40.2**	**0.231**	**20.8**	**0.648**	**23.3**	**0.591**
Total3D*	46.6	0.184	-	-	-	-
Cube R-CNN*	54.5	0.137	41.0	0.189	45.8	0.167
Ours*	**61.4**	**0.114**	**49.7**	**0.143**	**51.4**	**0.142**

Table 2. **Generalisation Performance: Novel Categories.** We train the model on 31 object categories from the SUN RGB-D training set and evaluate its IoU on 7 unseen categories. To show the importance of estimating in-plane offsets and box orientation, we use a Unprojection baseline that converts GT 2D boxes to 3D with GT depth and dimensions, setting 3D rotation to zero degrees.

Methods	Trained on	sofa	table	cabinet	toilet	bathtub	door	oven	avg.
Unprojection	N/A	28.2	27.1	28.6	25.6	23.9	23.6	37.2	27.7
Ours	SUN RGB-D	**56.4**	**56.8**	**53.0**	**61.3**	**46.7**	**27.7**	**62.1**	**52.0**

4. Metric: Normalised Hungarian Distance

Intersection-over-Union (IoU) is a common metric for evaluating 3D object detection [6, 40, 50]. While IoU is scale-invariant, it fails to measure the closeness of predictions when two boxes do not overlap, which is problematic for thin and small objects like mirrors or televisions. An example is shown in Fig. 3. Later, Generalised IoU (GIoU) [47] has been proposed to address this, but it still does not fully capture alignment in terms of centre, scale, and orientation, as seen in Fig. 2.

We propose a new metric, *Normalised Hungarian Distance* (NHD), to provide a more precise evaluation for 3D object detection. NHD is calculated as

$$\text{NHD}(\mathcal{M}_{\text{pred}}, \mathcal{M}_{\text{gt}}) = \frac{1}{d_{\text{gt}}} \sum_{i} \|a_i - b_j\|_2, \qquad (9)$$

where P represents the optimal 1-to-1 mapping between predicted box corners $\mathcal{M}_{\text{pred}}$ and ground truth box corners $\mathcal{M}_{\text{gt}}$. The mapping P is obtained through a linear assignment algorithm by minimising the Euclidean distance between corresponding corners, ensuring corner $a_i \in \mathcal{M}_{\text{pred}}$ in the prediction matches corner $b_j \in \mathcal{M}_{\text{gt}}$ in the ground truth. To make NHD scale-invariant, we normalise it with the maximum diagonal length d_{gt} of the ground truth box.

5. Experiments

We outline our experimental setup in Sec. 5.1. In Secs. 5.2 and 5.3, we compare our approach to baselines, demonstrating its accuracy and generalisation. We also present its application in 3D dataset labelling (Sec. 5.4) and discuss the impact of hyperparameter choices (Sec. 5.5).

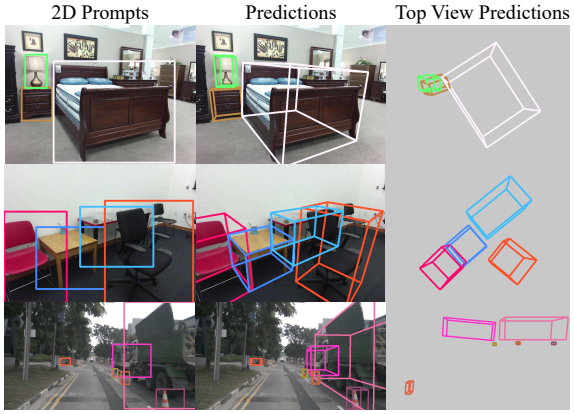

Figure 4. **Detection Performance: Results on Omni3D Test Set**. Estimating 3D box from GT 2D boxes and GT object depths.

5.1. Experimental Setup

Datasets For training and evaluation, we use the Omni3D dataset [6], which is composed of six datasets[54], ARKitScenes [5], Hypersim [48], Objectron [1], KITTI [14], and nuScenes [7], encompassing both indoor and outdoor environments. For the quantitative tests in Sec. 5.4, we use the COCO [33] and nerfstudio [58] datasets.

Baselines To accurately assess the performance of our 3D detector, it is essential to eliminate errors from sources such as 2D detection and depth estimation. In Sec. 5.2, we assume ground truth 2D boxes and depth are provided for all methods during evaluation. We select Cube R-CNN [6] and Total3DUnderstanding [40] as our primary baselines, as their 3D detection heads can function independently. Ad-

Figure 5. **Generalisation Performance: Results for In-the-Wild Objects on COCO Dataset.** We show predictions made by our method without knowing object depths or camera intrinsics. By using constant values for depths and camera intrinsics, our approach accurately predicts 3D boxes with well-aligned projections on the image.

ditionally, since the baseline methods require category information, we provide ground truth category labels to their models.

Metrics In Sec. 5.2, we use IoU and NHD as two metrics to evaluate the performance of 3D detectors. Unlike previous methods [6, 40], we do not compute average precision (AP), as we assume that all 2D boxes provided to each model are true positives.

Model Our model architecture is built on Detectron2 [62]. We use the Swin Transformer [36] pre-trained on ImageNet-22K [13] as the image feature encoder and freeze all parameters during training. The denoising decoder adopts the iterative architecture in Sparse R-CNN [57] with 6 blocks, where the predictions from the previous stage are used as the input for the next stage. Predictions from each stage are used to compute loss against the ground truth.

Training & Inference We train the model for 270k iterations with a batch size of 16 on 2 A5000 GPUs. In comparison, Cube R-CNN is trained on 48 V100 GPUs with a batch size of 192. We use AdamW optimiser [28] with a learning rate of 2.5×10^{-5} which decays by a factor of 10 at 150k and 200k iterations. During training, we use image augmentation including random horizontal flipping and resizing. During evaluation, as the box parameters are initialised with random noise, the result can vary with different random seeds. To obtain a stable result, experiment results reported in Sec. 5.2 and Sec. 5.3 are computed by averaging the results of 10 different random seeds. We also analyse the stability of model performance with random seeds in the supplementary material.

5.2. Detection Performance

We compare our model with other 3D object detection approaches on Omni3D dataset [6] and its subsets. For baseline models [6, 40], we assume ground truth 2D boxes as the oracle 2D detection results and provide category information. In contrast, our method requires 2D boxes only and does not use the category information. As our model requires object depth as input, we use the depths estimated by [6] for a fair comparison. We introduce a variant of each method that substitutes estimated object depths with ground truth depths. As shown in Tab. 1, our approach outperforms the baselines, even with their use of additional category information and dimension priors. The comparison between using predicted depths (upper half of table) and ground truth depths (lower half) highlights that inaccurate monocular depth estimation is a key source of error in monocular 3D detection models. Fig. 4 shows the qualita-

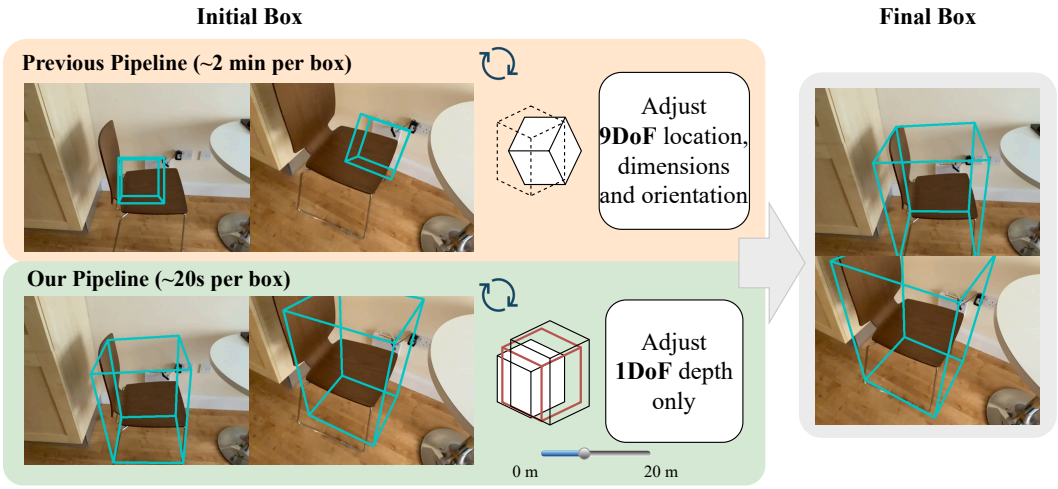

Figure 6. **Application: 3D Detection Dataset Annotation.** In a conventional 3D box annotation pipeline, annotators typically need to adjust a randomly initialised 3D box across nine degrees of freedom (rotation, translation, and size) until it appears correct in every view. This process is time-consuming and labour-intensive. Our model streamlines this workflow by reducing the task to a single degree of freedom, depth, significantly accelerating the dataset labelling process.

tive performance of our model on the Omni3D test set.

5.3. Generalisation Performance

Cross-category Generalisation To verify our model's generalisation ability across different objects, we trained it on the SUN RGB-D dataset using 31 out of 38 categories and evaluated it on the remaining categories. Since there are no previous baselines for comparison, we created a simple baseline using axis-aligned 3D boxes to demonstrate the importance of estimating in-plane offsets and box orientation, as well as the effectiveness of our approach. This baseline method takes the 2D box prompt and back-projects it to 3D using the provided depth and focal length. The bounding box orientation is set to match the camera orientation, and the dimension in the z-direction is set to the maximum length of the ground truth bounding box in that direction. Results in Tab. 2 and Fig. 7 show that our approach significantly outperforms this baseline and demonstrates good generalisation ability.

Cross-dataset Generalisation To assess the model's generalisation across different datasets and input types, we perform a qualitative analysis with Omni3D-trained models on test images from COCO and nerfstudio datasets.

Figure 5 shows the predictions on COCO, using only 2D prompts and uniform depth and camera intrinsics for all objects. Despite inaccuracies in depth and intrinsics, our model accurately estimates up-to-scale 3D boxes where the projections align with the objects in the image.

Figure 8 demonstrates the use of a monocular depth estimation network [65] and a segmentation network [29], combined with 2D prompts, to estimate 3D boxes with accurate top-view projections. The depth for each object is calculated by averaging the depths within the object's mask.

When 3D data, such as point clouds or CAD models, is available, it can also be utilised for depth inference. Figure 9 illustrates an example from nerfstudio dataset where 3D point clouds and a manually annotated 2D prompt are used to estimate the 3D box.

5.4. Application

As discussed in Sec. 1, the availability of 3D detection datasets is significantly more limited compared to 2D detection, primarily due to the high costs associated with annotating 3D bounding boxes. Figure 6 illustrates a typical 3D box annotation process, where annotators must draw a 3D box and fine-tune its projections across multiple views. This manual adjustment of the box's location, dimensions, and orientation can take several minutes per box.

Our model streamlines this process by reducing the complexity from adjusting nine degrees of freedom (DoF) to just one DoF – depth. Starting with a 2D prompt from an annotator, our model generates a 3D box prediction with depth ambiguity. The human annotator only needs to adjust the predicted box depth and verify it across other views. Incorporating this approach into the 3D box annotation pipeline has the potential to greatly improve annotation efficiency.

5.5. Discussion

We conducted a model analysis on the 10 common categories in the SUN RGB-D test set to study its performance. Unless otherwise stated, all inferences are run with 10 predictions for each object at a single sampling step.

Signal-to-noise Ratio In Tab. 3, we show the influence

Table 3. **Discussion: Model Analysis on SUN RGB-D Test Set.**

(a) IoU and NHD under various SNR (s).

SNR	IoU (%) ↑	NHD ↓
1.0	57.2	0.195
2.0	**59.8**	**0.177**
3.0	59.7	0.178

(b) IoU and NHD under various DDIM steps.

Iter steps	IoU (%) ↑	NHD ↓
1	61.19	0.1153
3	61.23	0.1150
5	**61.28**	**0.1147**

(c) IoU under various No. of sampled boxes.

$N_{\text{train}} \backslash N_{\text{eval}}$	1	10	100
1	58.1	60.2	60.2
10	60.9	61.2	61.2
100	61.2	**61.4**	**61.4**

of the diffusion process's signal-to-noise ratio (SNR). Setting SNR to 2 achieves the highest performance in IoU and NHD, which is consistent with the observations in other diffusion models for detection [9, 27].

Inference Iteration Steps During inference, we observed that increasing the number of DDIM sampling steps enhances model performance, albeit at the cost of longer inference times. As shown in Tab. 3b, raising the iteration steps from 1 to 5 results in performance gains, which plateau with additional steps.

Number of Sampled Boxes. Since we use random noise to initialise box parameters, the number of boxes sampled is flexible and can vary between training (N_{train}) and inference (N_{eval}). Table 3c presents the results for different combinations of N_{train} and N_{eval}. We observed that increasing N_{train} enhances model performance, with a cost of larger memory consumption. However, increasing N_{eval} beyond 10 does not yield additional improvement. Consequently, we select $N_{\text{train}} = 100$ and $N_{\text{eval}} = 10$ for other experiments.

6. Conclusion

Our diffusion-based pipeline significantly improves 3D object detection by decoupling it from 2D detection and depth prediction, enabling category-agnostic detection. Furthermore, introducing the Normalised Hungarian Distance metric addresses the limitations of existing evaluation methods, providing a more accurate assessment of 3D detection outcomes, especially for complex scenarios involving small or thin objects. Experimental results confirm that our method achieves state-of-the-art accuracy and strong generalisation across various object categories and datasets.

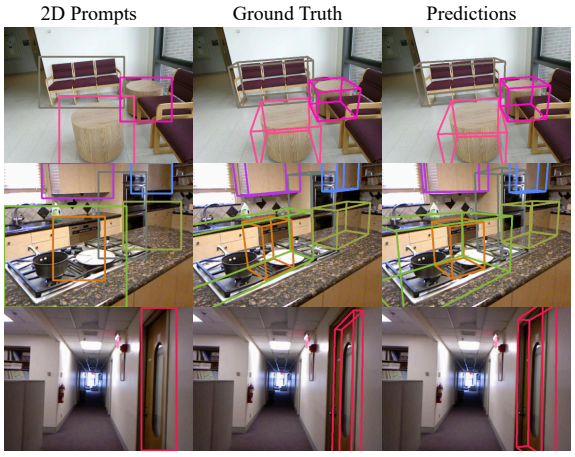

Figure 7. **Generalisation Performance: Predictions on Novel Categories**. We show that our model generalises well to unseen object categories.

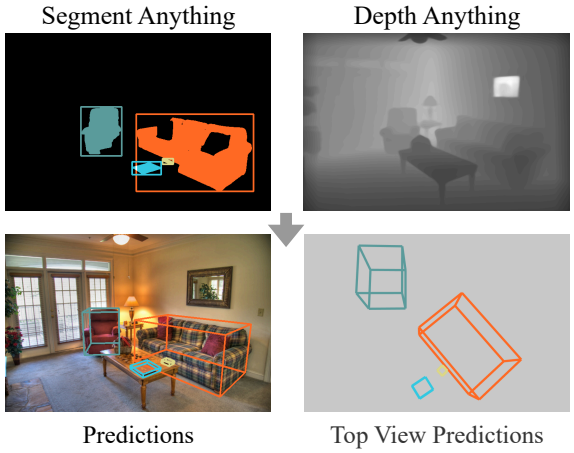

Figure 8. **Generalisation Performance: Predict with Prompts from 2D Detectors and Monocular Depth Estimators.** We infer object depths using DepthAnything [65] and SegmentAnything [29] for COCO Dataset.

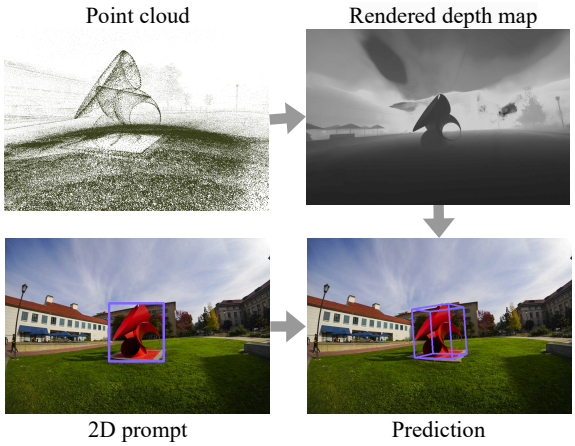

Figure 9. **Generalisation Performance: Predict with Prompts from human annotated 2D bounding boxes and a point cloud.**

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *CVPR*, 2021. 5

[2] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 2

[3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *NeurIPS*, 2021. 2

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 2

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS*, 2021. 5

[6] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *CVPR*, 2023. 1, 2, 3, 5, 6

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 5

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2

[9] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, 2023. 2, 3, 8

[10] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *ICCV*, 2023. 2, 3

[11] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016. 1, 2

[12] Xiangyu Chen, Zhenzhen Liu, Katie Z Luo, Siddhartha Datta, Adhitya Polavaram, Yan Wang, Yurong You, Boyi Li, Marco Pavone, Wei-Lun Chao, et al. Diffubox: Refining 3d object detection with point diffusion. *arXiv preprint arXiv:2405.16034*, 2024. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 5

[15] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[16] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, 2023. 2

[17] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. In *ICLR*, 2023. 2

[18] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *NeurIPS*, 2022. 2

[19] Cheng-Ju Ho, Chen-Hsuan Tai, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Diffusion-ss3d: Diffusion model for semi-supervised 3d object detection. *NeurIPS*, 2024. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1, 2, 3, 4

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 2

[22] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *CVPR*, 2023. 2

[23] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 2

[24] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. *NeurIPs*, 2018. 2

[25] Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic object detection. In *WACV*, 2021. 2

[26] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 2022. 2

[27] Se-Ho Kim, Inyong Koo, Inyoung Lee, Byeongjun Park, and Changick Kim. Diffref3d: A diffusion-based proposal refinement framework for 3d object detection. *arXiv preprint arXiv:2310.16349*, 2023. 2, 8

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 7, 8

[30] Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, and Abhinav Gupta. 3d-relnet: Joint object and relational network for 3d prediction. In *ICCV*, 2019. 2

[31] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *NeurIPS*, 2022. 2

[32] Zhuoling Li, Xiaogang Xu, SerNam Lim, and Hengshuang Zhao. Unimode: Unified monocular 3d object detection. In *CVPR*, 2024. 1, 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

[35] Yuxuan Liu, Yuan Yixuan, and Ming Liu. Ground-aware monocular 3d object detection for autonomous driving. *IEEE Robotics and Automation Letters*, 2021. 2

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6

[37] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *ICCV*, 2019. 2

[38] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *ECCV*, 2022. 2

[39] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *CVPR*, 2017. 1, 2

[40] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *CVPR*, 2020. 1, 2, 5, 6

[41] Pedro O O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. *NeurIPS*, 2015. 2

[42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2022. 2

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[44] Yasiru Ranasinghe, Deepti Hegde, and Vishal M Patel. Monodiff: Monocular 3d object detection and pose estimation with diffusion models. In *CVPR*, 2024. 2

[45] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, 2021. 2

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 2

[47] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[48] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 5

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVPR*, 2021. 2, 3

[50] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *WACV*, 2022. 1, 2, 5

[51] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, 2019. 1, 2

[52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 1, 4

[54] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019. 2

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 2

[57] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 6

[58] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH 2023*, 2023. 5

[59] Shubham Tulsiani, Saurabh Gupta, David F Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *CVPR*, 2018. 2

[60] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6. IEEE, 2024. 1

[61] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *CoRL*, 2022. 2

[62] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[63] Chenfeng Xu, Huan Ling, Sanja Fidler, and Or Litany. 3difftection: 3d object detection with geometry-aware diffusion features. In *CVPR*, 2024. 2

[64] Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. Monocd: Monocular 3d object detection with complementary depths. In *CVPR*, 2024. 2

[65] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 7, 8

[66] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 2023. 2

[67] P Yu, S Xie, X Ma, B Jia, B Pang, R Gao, Y Zhu, S-C Zhu, and YN Wu. Latent diffusion energy-based model for interpretable text modeling. In *ICML 2022*, 2022. 2

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3

[69] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021. 1

[70] Xin Zhou, Jinghua Hou, Tingting Yao, Dingkang Liang, Zhe Liu, Zhikang Zou, Xiaoqing Ye, Jianwei Cheng, and Xiang Bai. Diffusion-based 3d object detection with random boxes. In *PRCV*, 2023. 2

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3

[72] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 2