# 3DiffTection: 3D Object Detection with Geometry-Aware Diffusion Features

Chenfeng Xu[1,2] Huan Ling[1,3,4] Sanja Fidler[1,3,4] Or Litany[1,5]
[1]NVIDIA [2]UC Berkeley [3]Vector Institute [4]University of Toronto [5]Technion

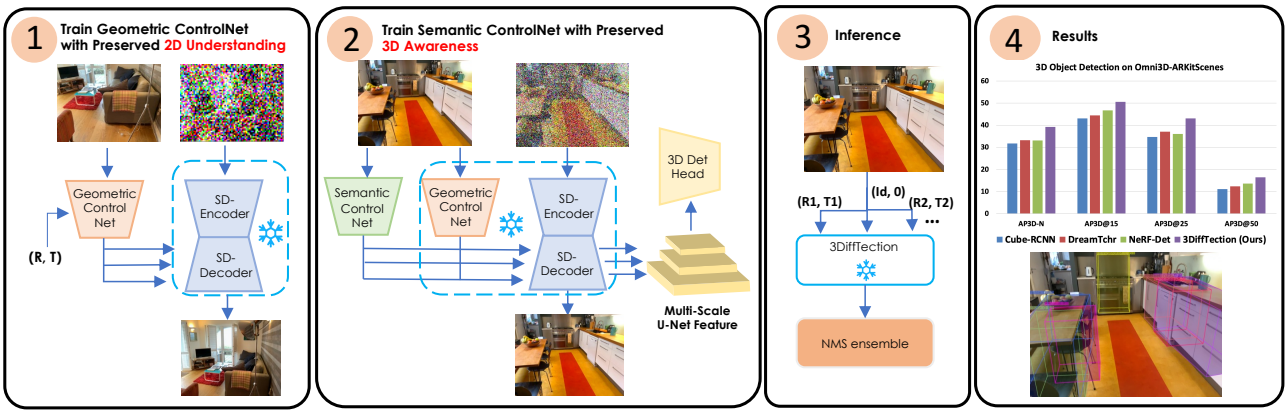


Figure 1. (1) We enhance pre-trained diffusion features with 3D awareness by training a *geometric* ControlNet (Sec. 3.2). (2) We employ a *semantic* ControlNet (Sec. 3.3) to refine generative features for targeted data and downstream tasks, specifically focusing on enhancing features for 3D object detection. (3) During the inference process, we further enhance 3D detection accuracy by ensembling the bounding box predictions from virtual views (Sec. 3.4).

## Abstract

*3DiffTection introduces a novel method for 3D object detection from single images, utilizing a 3D-aware diffusion model for feature extraction. Addressing the resource-intensive nature of annotating large-scale 3D image data, our approach leverages pretrained diffusion models, traditionally used for 2D tasks, and adapts them for 3D detection through geometric and semantic tuning. Geometrically, we enhance the model to perform view synthesis from single images, incorporating an epipolar warp operator. This process utilizes easily accessible posed image data, eliminating the need for manual annotation. Semantically, the model is further refined on target detection data. Both stages utilize ControlNet, ensuring the preservation of original feature capabilities. Through our methodology, we obtain 3D-aware features that excel in identifying cross-view point correspondences. In 3D detection, 3DiffTection substantially surpasses previous benchmarks, e.g., Cube-RCNN, by 9.43% in AP3D on the Omni3D-ARkitscene dataset. Furthermore, 3DiffTection demonstrates robust label efficiency and generalizes well to cross-domain data, nearly matching fully-supervised models in zero-shot scenarios. Project page:* https://research.nvidia.com/labs/toronto-ai/3difftection/.

## 1. Introduction

Detecting objects in 3D from a single image presents a significant challenge in computer vision, involving not only object recognition and localization but also depth and orientation prediction. This task, crucial for applications in robotics and augmented reality, demands advanced 3D reasoning from computational models.

Training a 3D detector from scratch is resource-intensive due to the high labeling costs [5]. Recently, large self-supervised models have emerged as compelling learners for image representation [10, 16, 17]. They acquire robust semantic features that can be fine-tuned on smaller, annotated datasets. Image diffusion models, trained on internet-scale data, have proven to be particularly effective in this context [24, 46, 56]. However, these models often lack 3D awareness and exhibit a domain gap in 3D applications. Recent work have aimed to bridge this gap by lifting 2D image features to 3D and refining them for specific 3D tasks. NeRF-Det [54] trained a view synthesis model alongside a detection head using pretrained image feature extractors. However, this approach is constrained by the need for dense scene views and fully annotated data. Efforts in novel view synthesis using diffusion models have shown promise [7, 58]. Yet, these models are generally

trained from scratch, thereby foregoing the advantages of using pretrained semantic features.

To overcome these limitations, our work, 3DiffTection, introduces a novel framework that repurposes pretrained 2D diffusion models for 3D object detection (see overview Fig. 1). We enhance these models with 3D awareness through a view synthesis task, employing epipolar geometry to warp features from source images to target views. This process utilizes ControlNet [57] to maintain the integrity of the original features (See Fig. 3). Utilizing image pairs from videos, which are abundant and do not require manual annotation, our approach is scalable and efficient. To demonstrate that our approach successfully imparts 3D awareness to the model, we assess the performance of its features in establishing point correspondences across multiple views. Our results indicate that these features outperform those of the base model, both qualitatively and quantitatively. For 3D detection, 3DiffTection trains a standard detection head with 3D box supervision, incorporating a second ControlNet to adapt the features to specific detection tasks and domains, preserving feature quality and view synthesis capabilities. At test time, we capitalize on both geometric and semantic capabilities by generating detection proposals from multiple virtual synthesized views, which are then consolidated through Non-Maximum Suppression (NMS).

Our primary contributions are as follows: (1) We introduce a scalable technique for enhancing pretrained 2D diffusion models with 3D awareness through a novel geometric ControlNet, enhanced with an epipolar warp operator; (2) We adapt these features to a 3D detection task and target domain by introducing a second, semantic ControlNet; and (3) We integrate both view synthesis and 3D detection capabilities to further improve detection performance through ensemble prediction.

3DiffTection emerges as a powerful 3D detector, substantially surpassing previous benchmarks, *e.g.,* CubeRCNN, by 9.43% in AP3D on the Omni3D-ARkitscene dataset. Furthermore, 3DiffTection demonstrates robust label efficiency, achieving a 2.28 AP3D-N improvement over previous methods trained with full supervision while using only 50% of the labels. 3DiffTection also exhibits the ability to generalize to cross-domain data, nearly matching the performance of previously established fully-supervised models without any tuning (zero-shot).

## 2. Related works

**3D Object Detection from Images.** 3D object detection from posed images is widely explored [26, 32, 37, 51, 54]. However, assuming given camera extrinsic is not a common scenario, especially in applications such as AR/VR and mobile devices. The task of 3D detection from single images, relying solely on camera intrinsics, presents a more generalized yet significantly more challenging problem. The model is required to inherently learn 3D structures and harness semantic knowledge. While representative methods [8, 21, 23, 31, 47, 50] endeavor to enforce 3D detectors to learn 3D cues from diverse geometric constraints, the dearth of semantics stemming from the limited availability of 3D datasets still impede the generalizability of 3D detectors. Brazil et al. [5], in an effort to address this issue, embarked on enhancing the dataset landscape by introducing Omni3D dataset. Rather than focusing on advancing generalizable 3D detection by increasing annotated 3D data, we propose a new paradigm, of enhancing semantic-aware diffusion features with 3D awareness.

**Diffusion Models for 2D Perception.** Trained diffusion models [30, 34, 36, 39] have been shown to have internal representations suitable for dense perception tasks, particularly in the realm of image segmentation [6, 14, 45, 56]. These models demonstrate impressive label efficiency [2]. Similarly, we observe strong base performance in both 2D and 3D detection (see Tab. 3); our method also benefits from high label efficiency. Diffusion models have further been trained to perform 2D segmentation tasks [11, 22, 53]. In [1] the model is trained to output a segmentation map using an auxiliary network that outputs residual features. Similarly, we use a ControlNet to refine the diffusion model features to endow them with 3D awareness. We note that several works utilize multiple generations to achieve a more robust prediction [1], we go a step further by using our controllable view generation to ensemble predictions from multiple views. Few works have studied tasks other than segmentation. DreamTeacher [24] proposed to distil the diffusion features to an image backbone and demonstrated excellent performance when tuned on perception tasks[24]. [40] trained a diffusion model for dense depth prediction from a single image. Recently, DiffusionDet [9] proposed an interesting method for using diffusion models for 2D detection by directly denoising the bounding boxes conditioned on the target image. Diffusion features have been analyzed in [49] showing that different UNet layer activations are correlated with different level of image details. We utilize this property when choosing which UNet layer outputs to warp in our geometric conditioning. Remarkably, [46] have shown strong point correspondence ability with good robustness to view change. Here we demonstrate that our 3D-aware features can further boost this robustness.

**Novel View Synthesis with Diffusion Models** Image synthesis has undergone a significant transformation with the advent of 2D diffusion models, as demonstrated by notable works [12, 18, 19, 28, 29, 33, 36, 38, 43, 44]. These models have extended their capabilities to the Novel View Synthesis (NVS) task, where 3DiM [52] and Zero-123 [25] model NVS of objects as a viewpoint-conditioned image-to-image translation task with diffusion models. The models are trained on a synthetic dataset with camera anno-


Figure 2. **Visualization of semantic correspondence prediction using different features** Given a **Red Source Point** in the left most reference image, we predict the corresponding points in the images from different camera views on the right (**Blue Dot**). The ground truth points are marked by **Red Stars**. Our method, 3DiffTection, is able to identify precise correspondences in challenging scenes with repetitive visual patterns.

tation and demonstrate zero-shot generalization to in-the-wild images. NerfDiff [15] distills the knowledge of a 3D-aware conditional diffusion model into a Nerf. RealFusion [27] uses a diffusion model as a conditional prior with designed prompts. NeuralLift [55] uses language-guided priors to guide the novel view synthesis diffusion model. Most recently, inspired by the idea of video diffusion models [4, 20, 42], MVDream [41] adapts the attention layers to model the cross-view 3D dependency. The most relevant work to our approaches is SparseFusion [58], where authors propose to incorporate geometry priors with epipolar geometries. However, while their model is trained from scratch, in our approach, we use NVS merely as an auxiliary task to enhance the pre-trained diffusion features with 3D awareness and design the architecture for tuning a minimal number of parameters by leveraging a ControlNet.

## 3. 3DiffTection

We introduce 3DiffTection, designed to harness diffusion model features for 3D detection. As depicted in Fig. 1, 3DiffTection comprises three core components: 1) Instilling 3D awareness into the diffusion features by training a geometric ControlNet for view synthesis. 2) Bridging the domain and task gaps using a semantic ControlNet, which is concurrently trained with a 3D detection head on the target data distribution. 3) Amplifying 3D box predictions through a virtual view ensembling strategy. We further detail each of these steps in the subsequent sections.

### 3.1. Diffusion Model as a Feature Extractor

Recent works demonstrate that features extracted from text-to-image diffusion models, such as Stable Diffusion [36], capture rich semantics suitable for dense perception tasks, including image segmentation [56] and point correspondences [46]. In this work, our interest lies in 3D object detection. However, since Stable Diffusion is trained on 2D image-text pairs—a pre-training paradigm proficient in aligning textual semantics with 2D visual features—it might lack 3D awareness. We aim to explore this by examining point correspondences between views. We hypothesize that features with 3D awareness should demonstrate the capability to identify correspondences that point to the same 3D locations when provided with multi-view images.

Following [46, 56] we employ a single forward step for feature extraction. However, unlike these works, we only input images without textual captions, given that in real-world scenarios, textual input is typically not provided for object detection. Formally, given an image $\mathbf{x}$, we sample a noise image $\mathbf{x}_t$ at time $t$, and obtain the diffusion features

$$\mathbf{f} = \mathcal{F}(\mathbf{x}_t; \Theta), \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x} + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \epsilon_t \sim \mathbb{N}(0,1), \tag{1}$$

where $\mathbf{f}$ represents the multi-scale features from the decoder module of UNet $\mathcal{F}$ (parameterized by $\Theta$), and $\alpha_t$ represents a pre-defined noise schedule, satisfying $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$.

Interestingly, as illustrated in Fig. 2, the point localization of Stable Diffusion features depends on 2D appearance matching. This can lead to confusion in the presence of repeated visual patterns, indicating a deficiency in 3D spatial understanding. Given this observation, we aim to integrate 3D awareness into the diffusion features.

### 3.2. Injecting 3D Awareness to Diffusion Features

**ControlNet** [57] is a powerful tool that allows the addition of conditioning into a pre-trained, static Stable Diffusion (SD) model. It has been demonstrated to support various types of dense input conditioning, such as depth and semantic images. This is achieved through the injection of conditional image features into trainable copies of the

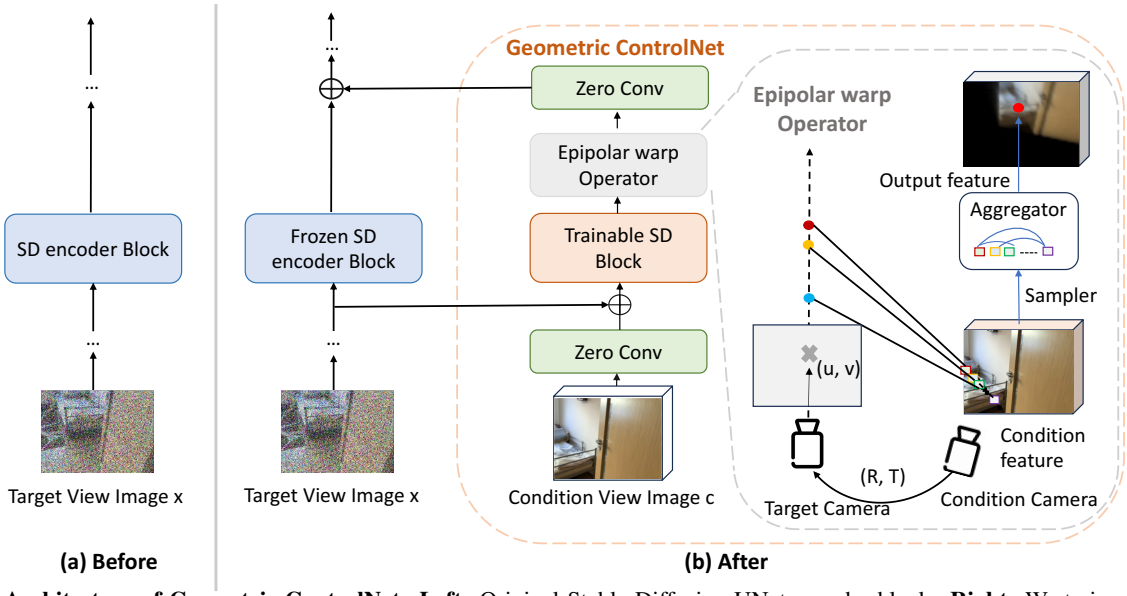


Figure 3. **Architecture of Geometric ControlNet. Left**: Original Stable Diffusion UNet encoder block. **Right**: We train novel view image synthesis by adding a geometric ControlNet to the original Stable Diffusion encoder blocks. The geometric ControlNet receives the conditional view image as an additional input. Using the camera pose, we introduce an epipolar warp operator, which warps intermediate features into the target view. With the geometric ControlNet, we significantly improve the 3D awareness of pre-trained diffusion features.

original SD blocks. A significant attribute of ControlNet is its ability to resist overfitting to the dataset used for tuning while preserving the original model's performance. As a result, ControlNet is well-suited for enhancing diffusion features with 3D awareness without compromising their 2D semantic quality.

Formally, we denote one block of UNet $\mathcal{F}$ as $\mathcal{F}_s(\cdot;\Theta_s)$ parameterized by $\Theta_s$. In particular, the original ControlNet block copies each pre-trained Stable Diffusion module $\mathcal{F}_s(\cdot;\Theta_s)$ denoted as $\mathcal{F}'_s(\cdot;\Theta'_s)$, and accompanying with two zero convolutions $\mathcal{Z}_{s1}$ and $\mathcal{Z}_{s2}$, parameterized by $\Theta_{zs1}$ and $\Theta_{zs2}$, respectively. We slightly abuse the notation of $\mathbf{x} \in \mathcal{R}^{H\times W\times C}$ as the arbitrary middle features of $\mathbf{x}_t$ in $\mathcal{F}$. Then a ControlNet block with the corresponding frozen Stable Diffusion block is given by

$$\mathbf{y}_s = \mathcal{F}_s(\mathbf{x};\Theta_s) + \mathcal{Z}_{s2}(\mathcal{F}'_s(\mathbf{x} + \mathcal{Z}_{s1}(\mathbf{c};\Theta_{zs1});\Theta'_s);\Theta_{zs2}), \tag{2}$$

where $\mathbf{c} \in \mathcal{R}^{H\times W\times C}$ is the condition image feature and $\mathbf{y}_s \in \mathcal{R}^{H\times W\times C}$ is the output.

**Epipolar warp operator.** We utilize ControlNet to enhance the 3D awareness of diffusion features by training it to perform view synthesis. Specifically, we select pairs of images with known relative camera poses and train the ControlNet conditioned on the source view to produce the output view. Since the features induced by the condition in ControlNet are additive, it is a common practice to ensure alignment between these features and the noisy input features. However, the input for our view synthesis task is, by definition, not aligned with the noisy input of the target view. As a solution, we propose to warp the source view features to align with the target using epipolar geometry. We denote the epipolar warp operator as $\mathcal{G}(\cdot, T_n)$, and our *geometric* ControlNet is formulated as:

$$\mathbf{y}_s = \mathcal{F}_s(\mathbf{x};\Theta_s)+\mathcal{Z}_{s2}(\mathcal{G}(\mathcal{F}'_s(\mathbf{x}+\mathcal{Z}_{s1}(\mathbf{c};\Theta_{zs1});\Theta'_s),T_n);\Theta_{zs2}), \tag{3}$$

Formally, to obtain the target novel-view image at position $(u, v)$, we assume that relative camera extrinsic from the source view is described by $T_n = [[R_n, 0]^T, [t_n, 1]^T]$, and the intrinsic parameters are represented as $K$. The equation for the epipolar line is:

$$l_c = K^{-T}([t_n] \times R_n)K^{-1}[u, v, 1]^T, \tag{4}$$

Here, $l_c$ denotes the epipolar line associated with the source conditional image. We sample a set of features along the epipolar line, denoted as $\{\mathbf{c}(p_i)\}$, where the $p_i$ are points on the epipolar line. These features are then aggregated at the target view position $(u, v)$ via a differentiable aggregator function, resulting in the updated features:

$$\mathbf{c}'(u, v) = \text{aggregator}(\{\mathbf{c}(p_i)\}), \quad p_i \sim l_c. \tag{5}$$

The differentiable aggregator can be as straightforward as average/max functions or something more complex like a transformer, as demonstrated in [13, 58], and $\mathbf{c}'$ is the warped condition image features, *i.e.*, the output of epipolar warp operator $\mathcal{G}$. The geometric warping procedure is illustrated in Fig. 3.

Interestingly, we found it beneficial to avoid warping features across all the UNet decoder blocks. As highlighted by

[48], middle-layer features in Stable Diffusion emphasize high-level semantics, while top stages capture appearance and geometry. Given the shared semantic content in novel-view synthesis, even amidst pixel deviations, we warp features only in the final two stages of Stable-Diffusion. This maintains semantic consistency while accommodating geometric warping shifts. Our geometric ControlNet notably enhances the 3D awareness of diffusion features, evident in the 3DiffTection examples in Fig. 2.

### 3.3. Bridging the Task and Domain Gap

We leverage the 3D-enhanced features for 3D detection by training a standard detection head with 3D box supervision. To further verify the efficacy of our approach in adapting diffusion features for 3D tasks, we train a 3D detection head, keeping our fine-tuned features fixed. Notably, we observe a substantial improvement compared to the baseline SD feature. We report details in Tab. 3.

Nevertheless, we acknowledge two potential gaps. Firstly, our view synthesis tuning is conceptualized as a universal 3D feature augmentation method. Hence, it is designed to work with a vast collection of posed image pairs, which can be inexpensively gathered (e.g., from videos) without the need for costly labeling. Consequently, there might be a domain discrepancy when comparing to target data, which could originate from a smaller, fully annotated dataset. Secondly, since the features aren't specifically fine-tuned for detection, there is further potential for optimization towards detection, in tandem with the detection head. As before, we aim to retain the robust feature characteristics already achieved and choose to deploy a second ControlNet.

Specifically, we freeze both the original SD and the geometric ControlNet modules. We then introduce another trainable ControlNet, which we refer to as *semantic* ControlNet. For our model to execute single-image 3D detection, we utilize the input image $x$ in three distinct ways. First, we extract features from it using the pretrained SD, denoted as $\mathcal{F}(x)$, through a single SD denoising forward step. Next, we feed it into our geometric ControlNet, represented as $\mathcal{F}_{geo}(x, T_n)$, with an identity pose ($T_n = [Id, 0]$) to obtain our 3D-aware features. Lastly, we introduce it to the semantic ControlNet, denoted by $\mathcal{F}_{sem}(x)$, to produce trainable features fine-tuned for detection within the target data distribution. We aggregate all the features and pass them to a standard 3D detection head, represented as $\mathcal{D}$ [5]. The semantic ControlNet is trained with 3D detection head.

$$y = \mathcal{D}(\mathcal{F}(x) + \mathcal{F}_{geo}(x, [Id, 0]) + \mathcal{F}_{sem}(x)) \quad (6)$$

The figure overview is in the supplementary material.

### 3.4. Ensemble Prediction

ControlNet is recognized for its ability to retain the capabilities of the pre-tuned model. As a result, our semantically tuned model still possesses view synthesis capabilities. We exploit this characteristic to introduce a test-time prediction ensembling that further enhances detection performance. Specifically, our box prediction $y$ is dependent on the input view. Although our detection model is trained with this pose set to the identity (i.e., no transformation), at test time, we can incorporate other viewing transformations denoted as $\xi_i$,

$$y(\xi) = \mathcal{D}(\mathcal{F}(x) + \mathcal{F}_{geo}(x, \xi) + \mathcal{F}_{sem}(x)). \quad (7)$$

The final prediction is derived through a non-maximum suppression of individual view predictions:

$$y_{final} = NMS(\{y(\xi_i\}). \quad (8)$$

We note that our objective isn't to create a novel view at this stage but to enrich the prediction using views that are close to the original pose. The underlying intuition is that the detection and view synthesis capabilities complement each other. Certain objects might be localized more precisely when observed from a slightly altered view.

## 4. Experiments

In this section, we present a comprehensive experimental evaluation of 3DiffTection and its constituent components. Initially, in Section 4.1, we establish 3DiffTection as a powerful 3D detection framework, particularly when fine-tuned on a specific target dataset. We then validate its capacity for generalization to new datasets, both with and without tuning of the detection head (Section 4.2). Subsequently, we demonstrate its ability to maintain strong performance with limited labels (Section 4.3). Finally, in Section 4.4, we confirm 3DiffTection's enhanced 3D awareness by measuring its feature correspondence accuracy. We also validate the importance of each module in our design and conclude with visualizations of our auxiliary view synthesis ability.

**Datasets and implementation details** For all our experiments, we train the geometric ControlNet on the official ARKitscene datasets [3], which provide around 450K posed low-resolution (256 $\times$ 256) images. We sample around 40K RGB images along with their intrinsics and extrinsics. **Note that in the following experiments, the pre-trained geometric ControlNet is kept frozen.** For training 3D object detection, we use Omni3D-ARkitscenes as our primary in-domain experiment dataset, and Omni3D-SUNRGBD for our cross-dataset experiments. To evaluate the performance, we compute a **mean** AP3D across all categories in Omni3D-ARkitscenes and over a range of IoU3D thresholds in $[0.05, 0.10, ..., 0.50]$, simply denoted as **AP3D**. We also report AP3D at IoU 15, 25, and 50 (AP3D@15, AP3D@25 and AP3D@50) as following [5]. We take the publicly available text-to-image LDM [36], Stable Diffusion as the preliminary backbone. Unlike previous diffusion models which require multiple images for training

| Methods | Resolution | NVS Train Views | Det. Train Views | AP3D↑ | AP3D@15↑ | AP3D@25↑ | AP3D@50↑ |
|---|---|---|---|---|---|---|---|
| CubeRCNN-DLA | 256×256 | - | 1 | 31.75 | 43.10 | 34.68 | 11.07 |
| DreamTchr-Res50 | 256×256 | - | 1 | 33.20 | 44.54 | 37.10 | 12.35 |
| NeRF-Det-R50 | 256×256 | $\geq 10$ | $\geq 10$ | 33.13 | 46.81 | 36.03 | 13.58 |
| ImVoxelNet | 256×256 | - | $\geq 10$ | 32.09 | 46.71 | 35.62 | 11.94 |
| 3DiffTection | 256×256 | 2 | 1 | **39.22** | **50.58** | **43.18** | **16.40** |
| CubeRCNN-DLA | 512×512 | - | 1 | 34.32 | 46.06 | 36.02 | 12.51 |
| DreamTchr-Res50 | 512×512 | - | 1 | 36.14 | 49.82 | 40.51 | 15.48 |
| 3DiffTection | 512×512 | 2 | 1 | **43.75** | **57.13** | **47.32** | **20.30** |
| CubeRCNN-DLA-Aug | 512×512 | - | 1 | 41.72 | 53.09 | 45.42 | 19.26 |

Table 1. **3D Object Detection Results on Omni3D-ARKitScenes testing set**. 3DiffTection significantly outperforms baselines, including CubeRCNN-DLA-Aug, which is trained with 6x more supervision data.

a novel-view synthesis task, we only take *two* views, one as the source view and another one as the target view. Moreover, we only consider two views with an overlap of less than 30%. Regarding novel-view synthesis ensemble, we use pseudo camera rotations, *i.e.*, $\pm 15\,\mathrm{deg}$ and ensemble the predicted bounding boxes via NMS.

**Methods in comparison.** CubeRCNN [5] extends Fast-RCNN [35] to 3D object detection by incorporating a cube head. In our work, we aim to provide a stronger 3D-aware image backbone, and compare it with other image backbones using the Cube-RCNN framework. Specifically, we compare with DreamTeacher [24], which distills knowledge from a Pre-trained Stable Diffusion to a lighter network, *ResNet-50*. We also compare with DIFT [46], which directly employs the frozen Stable Diffusion as the image feature extractor. Additionally, we evaluate methods designed for multi-view 3D detection, such as NeRF-Det [54] and ImVoxelNet [37]. While these methods typically require more images during training, we use them for single-image 3D object detection during testing.

### 4.1. 3D Object Detection on Omni3D-ARKitscenes

In Tab. 1, we analyze the 3D object detection performance of 3DiffTection compared to several baseline methods. Notably, 3DiffTection significantly outperforms CubeRCNN-DLA [5], a prior art in single-view 3D detection on the Omni3D-ARKitScenes dataset, achieving a margin of 7.4% at a resolution of 256×256 and 9.43% at a resolution of 512×512 on the AP3D metric. We further compare our approach to NeRF-Det-R50 [54] and ImVoxelNet [37], both of which utilize multi-view images during training (indicated in Tab. 1 as NVS Train Views and Det. Train Views). In contrast, 3DiffTection which does not rely on multi-view images for training the detection network and uses only view-pairs for geometric network training, surpasses these methods by 6.09% and 7.13% on the AP3D metric, respectively. Additionally, we compare our approach to DreamTeacher-Res50 [24], which distills StableDiffusion feature prediction into a ResNet backbone to make it amenable for perception tasks. 3DiffTection exceeds DreamTeacher by 6.02% and 7.61% at resolutions of 256×256 and 512×512, respectively. Lastly, we evaluate our model against CubeRCNN-DLA-Aug, which denotes the training of CubeRCNN on the complete Omni3D dataset, comprising 234,000 RGB images with a more robust training recipe. Remarkably, our model outperforms CubeRCNN-DLA-Aug by 2.03% on AP3D while using nearly 6x less data, demonstrating its data efficiency.

We also show visualization results in Fig. 4. Compared to CubeRCNN, our proposed 3DiffTection predicts 3D bounding boxes with better pose, localization and significantly fewer false defections. As seen in the middle column, our model can even handle severe occlusion cases, *i.e.,* the sofa in the middle image and the sink in the right image.

### 4.2. Cross-dataset Generalization

To assess the capability of 3DiffTection's geometric ControlNet to carry its 3D awareness to other datasets, we employed a 3DiffTection model with its geometric ControlNet trained on the OMni3D-ARKitscene dataset, and conduct cross-dataset experiments on the Ommni3D-SUNRGBD dataset. We evaluate it with two settings: (1) finetune the parameters on the Omni3D-SUNRBGD dataset and test the performance on Omni3D-SUNRGBD dataset, and (2) train the parameters on the Omni3D-ARKitscenes dataset and directly test the performance on Omni3D-SUNRGBD dataset in a zero-shot setting. The performance is shown in Tab. 2.

In the first setting (shown in the fourth column), as a baseline, we trained the 3D head using DIFT-SD features. 3DiffTection w/o Semantic-ControlNet and w/ Semantic-ControlNet outperform DIFT-SD by 1.21% and 5.99%, respectively. We further compare our approach with CubeRCNN. To ensure a fair comparison, we take CubeRCNN-DLA trained on Omni3D-ARKitscene datasets and fine-tuned its entire model on the Omni3D-SUNRGBD. Without any training of the geometric ControlNet on the Omni-SUNRGBD, 3DiffTection (w/o Semantic-ControlNet) with only tuned a 3D head surpasses the fully fine-tuned CubeRCNN-DLA by 0.39%. Then, we reintegrate the semantic ControlNet and jointly train it with the 3D head. This yield a performance boost of 5.09%. These results indicate that even without training the geometric ControlNet in the target domain, the semantic ControlNet adeptly adapts features for perception tasks.

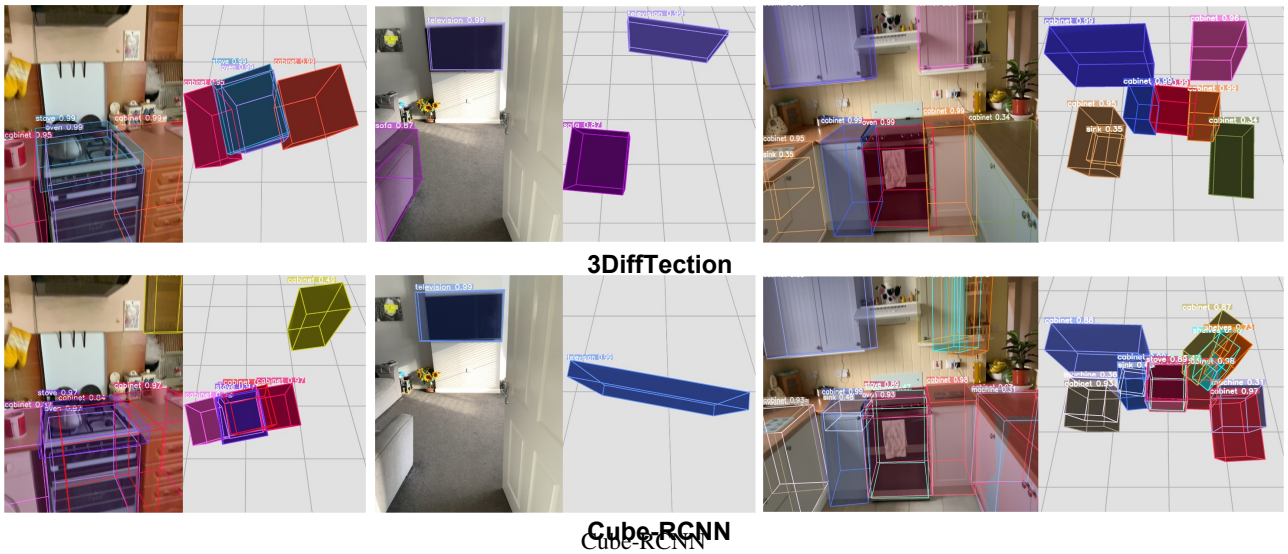


Figure 4. **Qualitative results on Omni3D-ARKitScene 3D Detection.** In contrast to Cube-RCNN (bottom), our approach (top) accurately predicts both the box class and 3D locations. The bird's-eye-view visualization further demonstrates that our predictions surpass the baseline performance of Cube-RCNN.

| Methods | Backbone | Pretrained on ARKit | Tuned on SUNRGBD | Zero-shot(w/o 2D GT) | Zero-shot(w/ 2D GT) |
|---|---|---|---|---|---|
| DIFT-SD | StableDiff | ✗ | 21.92 | 16.74 | 25.31 |
| CubeRCNN | DLA34 | ✓ | 22.72 | 16.81 | 25.05 |
| 3DiffTection | StableDiff+Geo-Ctr | ✓ | 23.11 | 17.37 | 26.94 |
| 3DiffTection | StableDiff+Geo-Ctr+Sem-Ctr | ✓ | 27.81 | 22.64 | 30.14 |

Table 2. **Cross-domain experiment on the Omni3D-SUNRGBD dataset.** The "Pre-trained on ARKit" denotes we pre-train the **backbone** on Omni3D-ARkitscenes. For CubeCNN, we pre-train it with 3D detection supervision. For all zero-shot experiments, the methods are first trained on Omni3D-ARKitscenes for 3D detection and then directly tested on Omni3D-SUNRGBD dataset. "2D GT" means we use ground-truth 2D bounding box to crop ROI image features. The results are reported for overlapped 14 classes between Omni3D-SUNRGBD and Omni3D-ARKiSscenes dataset.

To further demonstrate the transferrability of 3DiffTection, we train the models for 3D detection on Omni3D-ARkitscenes and directly test it on Omni3D-SUNRGBD datset without any further tuning. The results are shown in Column 3 and column 4 of Tab. 2. We observe that if we have ground truth 2D bounding boxes, 3DiffTection with semantic-ControlNet can even achieve the best performance. Without ground truth 2D bounding boxes, 3DiffTection is also able to outperform DIFT-SD and CubeRCNN by 5.90% and 5.83%, respectively. These results demonstrate the notable transferrability of our 3DiffTection.

### 4.3. Label Efficiency

We hypothesize that our usage of semantic ControlNet for tuning 3DiffTection towards a target dataset should maintain high label efficiency. We test this by using 50% and 10% labels from the Omni3D-ARKitscene datasets. The results are shown in Tab. **??** of supplementary materials. In low-data regime (for both 50% and 10% label setting), 3DiffTection demonstrates significantly better performance, and more modest degradation than baselines. Notably, even with 50% of the labels, our proposed 3DiffTection achieves 2.28 AP3D-N improvement over previous methods trained on 100% label. Additionally, when tuning only the 3D head 3DiffTection performs better than CubeRCNN and DreamTeacher with tuning all parameters.

### 4.4. Analysis and Ablation

**Feature correspondence fidelity (Fig. 2).** As described in 3.1, we conducted a feature correspondence experiment. We hypothesize that if our model is 3D aware, it should be find 3D correspondences. As can be seen, our method yields a more accurate point-matching result, primarily because our geometric ControlNet is trained to infer 3D correspondences through our Epipolar warp operator to successfully generate novel views. To provide further insights, we visualize a heatmap demonstrating the similarity of target image features to the reference key points. Notably, our 3DiffTection features exhibit better concentration around the target point. Furthermore, we quantitatively evaluate the correspondence performance on ScanNet dataset, which is never accessed by both our 3DiffTection and DIFT for fair comparison. The experiment results are shown in supplementary material. The results also demonstrate our hypothesis.
**Novel-view synthesis visualization (Fig. 5).** To further validate our geometric ControlNet ability to maintain geo-

| Backbone | NVS Train Views | Geo-Ctr | Sem-Ctr | NV-Ensemble | AP2D | AP3D↑ | AP3D@15↑ | AP3D@25↑ | AP3D@50↑ |
|---|---|---|---|---|---|---|---|---|---|
| VIT-B (MAE) | - | - | - | - | 26.14 | 25.23 | 36.04 | 28.64 | 8.11 |
| Res50 (DreamTchr) | - | - | - | - | 25.27 | 24.36 | 34.16 | 25.97 | 7.93 |
| StableDiff. (DIFT) | - | - | - | - | 29.35 | 28.86 | 40.18 | 32.07 | 8.86 |
| StableDiff. (Ours) | 1 | ✓ | - | - | 29.51 | 26.05 | 35.81 | 29.86 | 6.95 |
| StableDiff. (Ours) | 2 | ✓ | - | - | 30.16 | 31.20 | 41.87 | 33.53 | 10.14 |
| StableDiff. (Ours) | 2 | ✓ | ✓ | - | 37.12 | 38.72 | 50.38 | 42.88 | 16.18 |
| StableDiff. (Ours) | 2 | ✓ | ✓ | ✓ | 37.19 | 39.22 | 50.58 | 43.18 | 16.40 |

Table 3. **Analysis of 3DiffTection Modules on Omni3D-ARKitScenes testing set**. We first compare different backbones by freezing the backbone and only training the 3D detection head. Then, we perform ablative studies on each module of our architecture systematically. Starting with the baseline vanilla stable diffusion model, we incrementally incorporate improvements: Geometry-ControlNet (**Geo-Ctr**), the number of novel view synthesis training views (**NVS Train Views**), Semantic-ControlNet (**Sem-Ctr**), and the novel view synthesis ensemble (**NV-Ensemble**).

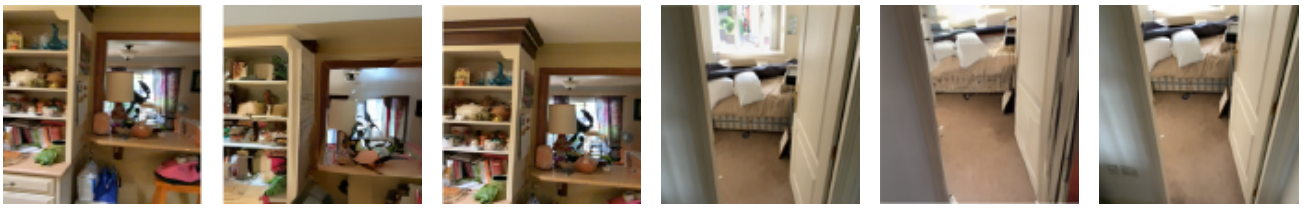


Figure 5. **Novel-view synthesis visualization on Omni3D-ARKitScenes testing set.** Our model with Geometry-ControlNet synthesizes realistic novel views from a single input image.

metric consistency of the source view content, we visualize novel-view synthesis results. The results demonstrate that our proposed epipolar warp operator is effective in synthesizing the scene with accurate geometry and layout compared to the ground truth images. We note that scene-level NVS from a single image is a challenging task, and we observe that our model may introduce artifacts. While enhancing performance is an interesting future work, here we utilize NVS as an auxiliary task which is demonstrated to effectively enhance our model's 3D awareness.

**3DiffTection modules**. We analyze the unique modules and design choices in 3DiffTection: the Stable Diffusion backbone, geometric and semantic ControlNets targeting NVS and detection, and the multi-view prediction ensemble. All results are reported using the Omni3D-ARKitscenes in Tab. 3. We first validate our choice of using a Stable Diffusion backbone. While diffusion features excel in 2D segmentation tasks [24, 56], they have not been tested in 3D detection. We analyze this choice independently from the other improvements by keeping the backbone frozen and only training the 3D detection head. The vanilla Stable Diffusion features achieve a 28.86% AP3D, exceeding CubeRCNN-VIT-B (MAE pretrained) by 3.63% and ResNet-50 DreamTeacher by 4.5% in AP30. This performance is mirrored in AP2D results, affirming Stable Diffusion's suitability for perception tasks. Our geometric ControlNet, is aimed at instilling 3D awareness via NVS training. A performance boost of 2.34% on AP3D and 0.81% on AP2D indicates that the geometric ControlNet imparts 3D awareness knowledge while preserving its 2D knowledge. To ensure our improvement is attributed to our view synthesis training, we limited the geometric ControlNet to single-view data by setting the source and target views to be identical (denoted by '1' in the NVS train view column of Tab. 3), which reduces the training to be *denoising training* [6]. This indicate a 2.81% decrease in AP3D compared to the standard Stable Diffusion, affirming our hypothesis. Further, the semantic ControlNet, co-trained with the 3D detection head enhances both AP2D and AP3D by around 7% confirming its efficacy in adapting the feature for optimal use by the detection head. Lastly, using NVS-ensemble results in additional 0.5% increase in AP3D.

## 5. Conclusion and Limitations

3DiffTection, utilizing a 3D-aware diffusion model, enables efficient 3D detection from single images, overcoming large-scale data annotation challenges. With its geometric and semantic tuning strategies, it surpasses previous benchmarks, showing high label efficiency and cross-domain adaptability. 3DiffTection has limitations, including the need for image pairs with accurate camera poses and challenges in handling dynamic objects from in-the-wild videos. Additionally, its use of the Stable Diffusion architecture demands substantial memory and runtime, achieving about 7.5 fps on a 3090Ti GPU. Suitable for offline tasks, it requires further optimization for online detection.

**Acknowledgements.** Or Litany is a Taub fellow and is supported by the Azrieli Foundation Early Career Faculty Fellowship. We thank Qianqian Wang, David Acuna, and Jonah Philion for the insightful discussion.

## References


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

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 2

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 5

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[5] 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, 2023. 1, 2, 5, 6

[6] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4175–4186, 2022. 2, 8

[7] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models, 2023. 1

[8] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 2

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

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[11] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*, 2022. 2

[12] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 2

[13] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs, 2023. 4

[14] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 2

[15] Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, 2023. 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 2

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *arXiv preprint arXiv:2106.15282*, 2021. 2

[20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[21] 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. In *Advances in Neural Information Processing Systems*, pages 206–217, 2018. 2

[22] Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. In *The Eleventh International Conference on Learning Representations*, 2023. 2

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

[24] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models, 2023. 1, 2, 6, 8

[25] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2

[26] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 2

[27] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 3

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and

Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021. 2

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[31] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[32] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris M. Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[33] 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

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 6

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 5

[37] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022. 2, 6

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[40] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J. Fleet. Monocular depth estimation using diffusion models, 2023. 2

[41] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[42] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792*, 2022. 3

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015. 2

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[45] Weimin Tan, Siyuan Chen, and Bo Yan. Diffss: Diffusion model for few-shot semantic segmentation. *arXiv preprint arXiv:2307.00773*, 2023. 2

[46] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion, 2023. 1, 2, 3, 6

[47] Shubham Tulsiani, Saurabh Gupta, David Fouhey, Alexei A. Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 2

[48] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 5

[49] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[50] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2

[51] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[52] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[53] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022. 2

[54] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection, 2023. 1, 2, 6

[55] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360° views. 2022. 3

[56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 1, 2, 3, 8

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

[58] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 1, 3, 4