Robust Category-Level 3D Pose Estimation from Diffusion-Enhanced Synthetic Data

Jiahao Yang[1] Wufei Ma[2] Angtian Wang[2] Xiaoding Yuan[2]
Alan Yuille[2] Adam Kortylewski[3,4]
[1] Peking University [2] Johns Hopkins University
[3] University of Freiburg [4] Max Planck Institute for Informatics

Abstract

*Obtaining accurate 3D object poses is vital for numerous computer vision applications, such as 3D reconstruction and scene understanding. However, annotating real-world objects is time-consuming and challenging. While synthetically generated training data is a viable alternative, the domain shift between real and synthetic data is a significant challenge. In this work, we aim to narrow the performance gap between models trained on synthetic data and fully supervised models trained on a large amount of real data. We achieve this by approaching the problem from two perspectives: 1) We introduce P3D-Diffusion, a new synthetic dataset with accurate 3D annotations generated with a graphics-guided diffusion model. 2) We propose **C**ross-domain **3D C**onsistency, CC3D, for unsupervised domain adaptation of neural mesh models. In particular, we exploit the spatial relationships between features on the mesh surface and a contrastive learning scheme to guide the domain adaptation process. Combined, these two approaches enable our models to perform competitively with state-of-the-art models using only 10% of the respective real training images, while outperforming the SOTA model by a wide margin using only 50% of the real training data. By encouraging the diversity of synthetic data and generating the images with an OOD-aware manner, our model further demonstrates robust generalization to out-of-distribution scenarios despite being trained with minimal real data. The code is available at* https://github.com/YangYY06/synthetic_3d.

1. Introduction

Object pose estimation is a fundamentally important task in computer vision with a multitude of real-world applications, e.g., in autonomous driving, 3D reconstruction, or in virtual and augmented reality applications. Pose estimation has been studied in depth on the instance level [14, 17, 19, 25, 38], and on the category-level for very specific object classes like cars [11] and faces [26]. However, it remains unclear how to learn category-level 3D pose estimation for general object categories. The main reason is that current models require large-scale annotated data, but annotating data with 3D poses is prohibitively expensive.

We aim to approach this problem by developing models that learn from limited manual annotation and large-scale synthetic data with automated annotations. In particular, we build on recent results that develop a render-and-compare approach to category-level pose estimation [17,34] and demonstrated more efficient learning from few examples [35] compared to standard deep neural network-based methods, due to their inherent 3D-aware network architecture. However, these methods still suffer from lower pose prediction accuracy when learned from few examples, compared to models learned from large-scale annotated data.

In this work, we aim to close the performance gap between models trained on a limited number of annotated real images and fully supervised models. To achieve this, we first introduce diffusion-enhanced synthetic data with realistic images coupled with accurate 3D annotations, and second, we develop an unsupervised domain adaptation method that achieves strong few-shot performance on both in-distribution and out-of-distribution data.

The major obstacle that prevents the community from using generated data rendered using computer graphics is that most current object pose estimation approaches [21, 30, 32, 41] are sensitive to domain shift. This means that their performance degrades significantly when trained on synthetic images and then evaluated on real-world images. To address this issue, we create and develop P3D-Diffusion, a diffusion-enhanced synthetic dataset with high-quality realistic images and accurate 3D annotations with no manual efforts. As outlined in Figure 1, the dataset generation begins with the rendering the CAD models with a graphics-based renderer. To narrow the gap between synthetic images and natural images, we propose a graphics-guided style transfer module that utilizes a pre-trained diffusion model

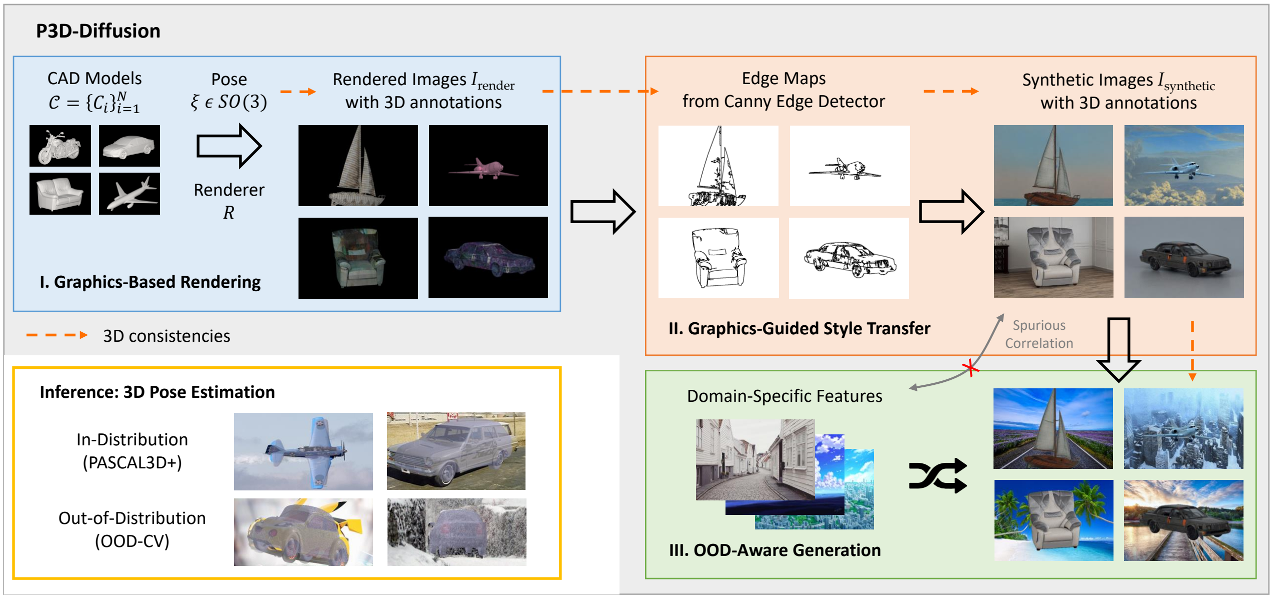

Figure 1. Our approach learns 3D pose estimation from P3D-Diffusion where CAD models are rendered under randomly sampled viewpoints. Additonally, we apply diffusion models to transfer the style of synthetic images while maintaining 3D consistencies. To address out-of-distribution (OOD) challenges, our OOD-aware generation method breaks the spurious correlations between task-related semantic information and domain-specific background features. Using P3D-Diffusion, we propose CC3D that allows for accurate 3D pose estimation on real data, even in challenging domains considered to be out-of-distribution for standard benchmarks.

to produce high-quality images while maintaining 3D consistency. We also introduce an out-of-distribution (OOD)-aware generation design that can effectively break the spurious correlations between task-related semantic information and domain-specific features. P3D-Diffusion can improve model's robustness in OOD scenes with only a negligible degradation in in-distribution benchmark performance.

As a second contribution, we develop a domain robust object pose estimation approach based on prior work on neural mesh models [34] that use inverse rendering on discriminative neural features for pose estimation. In particular, our approach represents an object category as a cuboid mesh and learns a generative model of neural feature activations at each mesh vertex for pose estimation via differentiable rendering. The feature representations at each vertex are trained to be invariant to instance-specific details and changes in 3D pose using contrastive learning. We extend the model to achieve better domain generalization by enhancing the consistency among vertex features across domains, and reweighting predictions to depend more on reliable features. To better adapt our model to real-world image domains, we fine-tune it on unlabeled real-world images using pseudo-labels from unlabeled data.

We summarize the contributions of our paper as follows:

- We create the diffusion-enhanced P3D-Diffusion dataset by rendering CAD models in various poses and lighting conditions, and then feeding the rendered images into a graphics-guided diffusion model to produce high-quality realistic images with 3D annotations. As a result, models trained on P3D-Diffusion dataset achieve better accuracy on real-world images and generalize well to out-of-distribution scenarios.
- We introduce a novel training and inference process for neural mesh models that enables them to perform unsupervised domain adaptation via feature consistency.
- Our results show that our proposed model combined with our synthetic data generalizes almost as well as fully supervised models, when using only 50 training samples per class. Using 10% of the annotated data it can even outperform fully supervised models. Moreover, our model generalizes more robustly in realistic out-of-distribution scenarios.

2. Related Works

Category-level 3D pose estimation. Category-level 3D pose estimation estimates the 3D orientations of objects in a certain category. A classical approach was to formulate pose estimation as a classification problem [21, 32]. Subsequent works can be categorized into keypoint-based methods and render-and-compare methods [5, 36]. Keypoint-based methods [24, 41] first detect semantic keypoints and then predict the optimal 3D pose by solving a Perspective-n-Point problem. Render-and-compare methods [5, 36] predict the 3D pose by fitting a 3D rigid transformation to minimize a reconstruction loss. Recently, NVSM [35] proposed a semi-supervised approach and investigated pose estimation in few-shot settings. Annotations of 3D poses are hard to obtain, and most previous works are largely limited by the number and quality of 3D annotations on real images. In this work, we propose to incorporate synthetic images generated from CAD models to address this challenge.

Unsupervised domain adaptation. Unsupervised domain adaptation (UDA) leverages both labeled source domain data and unlabeled target domain data to learn a model that works well in the target domain. One approach is to learn domain-invariant feature representations by minimizing domain divergence in a latent feature space [20, 28, 31]. Another line of work adopts adversarial loss [15, 33] to extract domain invariant features, where a domain classifier is trained to distinguish the source and target distributions. Recent works have also investigated UDA in downstream tasks, such as human pose estimation [4] and parsing deformable animals [22]. However, previous works often limited their scope to improving pose estimation or segmentation performance on i.i.d. data by involving synthetic images during training. In this work, we demonstrate that our proposed approach can both effectively improve benchmark performance on i.i.d. data, as well as enhancing model robustness in o.o.d. scenarios.

Self-training. Self-training has been found effective in self-supervised settings where we utilize unlabeled target domain data to achieve domain adaptation. Since generated pseudo-labels are noisy, several methods [10,18,42,43] were proposed to address this problem. [42, 43] formulated self-training as a general EM algorithm and proposed a confidence regularized framework. [18] proposed a self-ensembling framework to bootstrap models using unlabeled data. Moreover, [10] extended the previous work to unsupervised domain adaptation and investigated self-ensembling in closing domain gaps. In this work, we introduce an approach that leverages 3D cross-domain consistency in a contrastive learning framework.

3. P3D-Diffusion Dataset

We generate realistic-looking synthetic images with 3D annotations for training to reduce the domain generalization gap. Given CAD models $\mathcal{C} = \{C_i\}_{i=1}^N$ and 2D background images $\mathbf{B} = \{B_j\}_{j=1}^K$, our synthetic image generation can be formulated as

$$I_{\text{render}} = R(C_i, \xi), \quad I_{\text{synthetic}} = I_{\text{render}} \oplus B_j \tag{1}$$

where $\xi \in SO(3)$ represents a randomized object pose, R is an off-the-shelf renderer, and $\oplus$ overlays the rendered object image onto the background image B_j.

Although the image generation pipeline we employ yields image samples with 3D annotations at no additional cost, there is a significant domain gap between synthetic and real images. This gap presents great challenges for deep learning models to apply knowledge learned from synthetic data to natural images. Moreover, the generation of synthetic data is often biased towards the domain style of the testing benchmark, leading to models trained on the abundant synthetic data overfitting on domain-specific features. The overfitting can result in a drop in performance when evaluated on out-of-distribution (OOD) datasets.

To address these issues, we propose two novel designs for our P3D-Diffusion dataset that improve both the in-distribution and out-of-distribution performance. In Section 3.1 we demonstrate how we utilize a graphics-guided diffusion model to produce realistic synthetic data with accurate 3D annotations. Then we show how we can improve the out-of-distribution robustness of models trained on synthetic data in Section 3.2.

3.1. Diffusion-Enhanced Synthetic Data

Rendering photo-realistic images from CAD models is a challenging task, despite the plentiful CAD models available online and the technological advancements in modern renderers. Achieving high levels of realism requires detailed object materials and textures, which are not available in most CAD models publicly available [3, 37]. Moreover, simulating authentic lighting conditions demands professional expertise to set up various types of lights and largely increases the rendering time of synthetic images. In fact, modern generative models [29,39] are capable of generating high-resolution, detailed images with realistic textures and specular reflections. We propose to utilize such models to generate high-quality training images with 3D annotations.

Therefore, we design a graphics-guided style transfer module that can rapidly generate photo-realistic images without relying on high-quality CAD models. As demonstrated in Figure 1, we start by rendering the image $I_{\text{render}} = R(C_i, \xi)$ with the graphics-based renderer. Then we use a Canny edge detector [2] to produce a 2D edge map E encoding the semantic and structural information of the object C in the 2D lattice. The edge map is used to guide the generation of a high-quality image $I'_{\text{synthetic}}$ using a pretrained style transfer generative model Ψ [39]. The generative model Ψ takes an edge map as input and generates a high-quality realistic image consistent with the semantics provided in the edge map. By leveraging the edge map input, our approach effectively retains the semantic and structural information of C, enabling us to obtain 3D annotations for high-quality image $I'_{\text{synthetic}}$ directly from the rendering parameters. Formally this module is given by

$$\begin{aligned} I_{\text{render}} &= R(C_i, \xi) \\ E &= \text{CannyEdge}(I_{\text{render}}) \\ I'_{\text{synthetic}} &= \Psi(E) \oplus B_j \end{aligned} \tag{2}$$

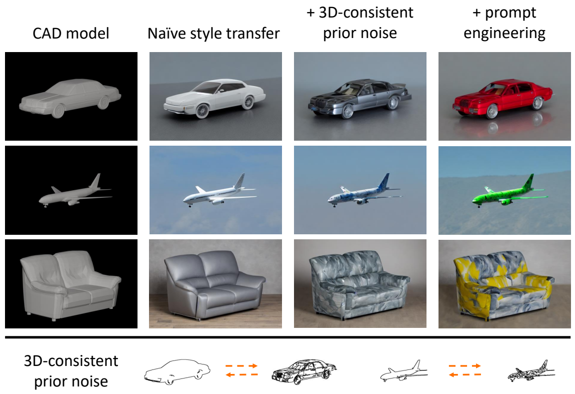

Figure 2. **Top:** Visualizations of the P3D-Diffusion. The naïve approach yields textureless objects with similar colors. We promote diverse textures and colors with 3D-consistent prior noise and simple prompt engineering. **Bottom:** Visualizations of 3D consistent prior noise for diverse texture generation.

Note that the style transfer generative model can be trained with abundant 2D images from the Internet. The high-quality synthetic training data with 3D annotations come at no extra cost with the help of our graphics-guided module.

Encouraging diverse outputs. Early experiments revealed that the style transfer network exhibits mode collapse, resulting in textureless objects with similar colors (see Figure 2). We propose two approaches that address this issue. First, to promote varied textures from the style transfer generative model, we render the CAD models with textures from the Describable Texture Dataset (DTD) [7]. This strategy introduces 3D-consistent prior noise into the edge maps, which compels the model to generate a variety of textures and colors. One the other hand, we adopt prompt engineering when applying the style transfer network and add random colors in the prompts in the form of "[color] [category]", e.g., "red car" and "green aeroplane". This approach allows us to produce a wide range of colors while maintaining 3D consistencies.

3.2. OOD-Aware Generation

From a causal perspective, the OOD robustness problem can be attributed to the spurious correlation between task-related semantic features, such as object parts and their locations, and domain-specific features, such as backgrounds [16]. Models trained on real images would inevitably learn from such spurious correlation, resulting in a high in-distribution benchmark performance (largely due to overfitting) and poor OOD robustness. Previous methods struggled to break the spurious correlation in real images [12, 16], which involves complex data augmentations or swapping features as a regularization.

The fully controllable generation of our synthetic dataset allows us to disentangle task-related semantics of foreground objects, including CAD models and poses, from domain-specific features such as background images. To this end, we collected 100 images from the Internet, and during our synthetic data generation process, we fully randomized the selection of B_j, independent of the foreground object category. In Section 5.6, we demonstrate that our OOD-aware design significantly enhances our model's OOD robustness while only marginally degrading in-distribution performance.

4. Domain Consistent 3D Pose Estimation via Render-and-Compare

Our work builds on and significantly extends neural mesh models (NMMs) [34] that learn generative models for feature activations and solve object 3D poses with analysis-by-synthesis. We propose a novel unsupervised domain adaptation approach, Cross-domain 3D Consistency (CC3D), that effectively adapt models trained on synthetic images to real data. In Section 4.1, we provide a review of neural mesh models, introducing the background and mathematical notations. Then we present our unsupervised domain adaptation methods in Section 4.2, where we propose a cross-domain feature consistency loss.

4.1. Background: Neural Mesh Models

Neural Mesh Models (NMMs) represent objects as a neural mesh $\mathfrak{N} = \{\mathcal{V}, \mathcal{C}\}$ with a set of vertices that represent a cuboid mesh $\mathcal{V} = \{V_r \in \mathbb{R}^3\}_{r=1}^R$ and learnable features for each vertex $\mathcal{C} = \{C_r \in \mathbb{R}^c\}_{r=1}^R$, where c is the number of channels and R is the number of vertices, and $\mathcal{C}$ is learned with a running average of neural features collected from training images. During training, we first extract feature map $F = \Phi_W(I)$, where Φ_W is the feature extractor with weights W and I is the RGB image. The feature extractor is trained with the contrastive loss that increases features' spatial distinguishability from each other [1]:

$$\mathcal{L}_{\text{con}}(F) = -\sum_{i\in\mathcal{FG}}\Big(\sum_{j\in\mathcal{FG}\setminus\{i\}}\|f_i - f_j\|^2 + \sum_{j\in\mathcal{BG}}\|f_i - f_j\|^2\Big),$$

where $\mathcal{FG}$ and $\mathcal{BG}$ indicate pixels in the foreground or background and i, j traverse all 2D locations on the feature map.

At test time, we can infer the object pose m by minimizing the feature reconstruction error w.r.t. the pose m with gradient descent

$$\begin{aligned}\mathcal{L}_{\text{rec}}(F, \mathfrak{N}, m, b) &= -\ln p(F \mid \mathfrak{N}, m, b)\\ = -\sum_{i\in\mathcal{FG}}&\left(\ln\left(\frac{1}{\sigma_r\sqrt{2\pi}}\right) - \frac{1}{2\sigma_r^2}\|f_i - C_r\|^2\right)\\ -\sum_{i'\in\mathcal{BG}}&\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\|f_{i'} - b\|^2\right). \qquad (3)\end{aligned}$$

where $\mathcal{FG}$ and $\mathcal{BG}$ indicates pixels assigned as foreground or background respectively, b is learnt features that represent backgrounds, and σ is the variance. Note that the correspondence between the feature map vector f_i and the mesh vertex feature C_r is given by the 2D projection of neural mesh $\mathfrak{N}$ with camera parameters m. $\mathcal{L}_{\text{rec}}$ is also used in training to train the neural features on the mesh.

4.2. Domain Consistency 3D Pose Estimation via Render-and-Compare

Domain-consistency loss. Although the analysis-by-synthesis design of NMMs leads to a superior performance on in-distribution data, directly applying the models to a new domain presents great challenges. New domain-specific features yield a domain shift on the feature representation $F = \Phi_W(I)$, yet the model synthesize source-domain features from unchanged vertex features $\mathcal{C} = \{C_r \in \mathbb{R}^c\}_{r=1}^R$. Such shifts in feature space makes the optimized 3D pose biased and inaccurate. Therefore, to improve the domain generalization ability of the NMMs, we require the features $\mathcal{C}$ to be invariant to the variations between synthetic and real images.

To achieve this, we introduce a domain-consistency loss that encourages features in real and synthetic data to become similar to each other:

$$\mathcal{L}_{\text{domain}}(C, \{\tilde{F}\}) = \sum_{r=1}^{R} \|\tilde{f}_r - C_r\|^2, \tag{4}$$

where $\tilde{f}_r$ are *corresponding* features for the vertex r on the neural mesh in $\tilde{F}$. $\tilde{F}$ is the feature map of the real image. The correspondence between the neural mesh $\mathfrak{N}$ and the real data is obtained with pseudo labels introduced below. Finally, our full model is trained by optimizing the joint loss:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{domain}}, \tag{5}$$

with α being a weight parameter that ensures that both losses are approximately on the same scale.

Unsupervised domain adaptation with pseudo labels. The core challenge of our approach lies in finding the corresponding features for every vertex on the neural mesh in the real data without access to any pose annotations. To resolve this problem, we first train a neural mesh from synthetic data where we have ground-truth annotations. We train the parameters of the neural texture $\mathcal{C}$ through maximum likelihood estimation (MLE) by minimizing the negative log-likelihood of the feature representations over the whole training set. The correspondence between the feature vectors f_i and vertices r in the synthetic data is computed using the annotated 3D pose. To reduce the computational cost of optimizing Equation 3, we follow [1] and update $\mathcal{C}$ in a moving average manner.

Given a synthetically trained neural mesh, we start by estimating the 3D poses $\{m_{\text{est}}\}$ of the unlabeled real data by optimizing the pose parameters m to maximize the likelihood $p(F \mid \mathfrak{N}, m, b)$. Since predicted pose labels are noisy, we proposed a criteria which leverages the advantage of render-and-compare method to do self-consistency examination for the quality of the predicted labels. Specifically, with unlabeled real image I, extracted feature map F and predicted label m_{est}, we project the neural mesh $\mathfrak{N}$ to the 2D lattice with m_{est} to estimate the $\mathcal{FG}$ and to find the corresponding C_r for each f_i. Then we compute the self-consistency confidence score $S_{\text{confidence}}$

$$S_{\text{confidence}} = \frac{1}{\gamma} \sum_{i \in \mathcal{FG}} f_i \cdot C_r \tag{6}$$

where γ is the sum of $\mathcal{FG}$ pixels. Note that all f_i and C_r are normalized. For each category, we choose 100 real images with the highest $S_{\text{confidence}}$ and their predicted poses as pseudo-labels. Finally, by utilizing these pseudo labels, we fine-tune our model through optimizing Equation 5.

In the following section, we demonstrate that our proposed unsupervised domain adaptation approach is highly efficient at bridging the domain gap between real and synthetic data, giving accurate predictions on real data without using any real annotations, and outperforming state-of-the-art models when fine-tuned with few annotated real data.

5. Experiments

In this section, we present our main experimental results. We start by describing the experimental setup in Section 5.1. Then we study the performance of approach on 3D pose estimation under unsupervised and semi-supervised settings in Section 5.2 and 5.3. We also report experimental results on out-of-distribution data in Section 5.4 to demonstrate the generalization ability of our model.

5.1. Experimental Setup

Datasets. We first evaluate 3D pose estimation by our model and baseline models on PASCAL3D+ dataset [37]. The PASCAL3D+ dataset contains 11045 training images and 10812 validation images of 12 man-made object categories with category and object pose annotations. We evaluate 3D pose estimation under 5 different settings – unsupervised, semi-supervised with 7, 20, and 50 images [35], as well as the fully-supervised setting. To investigate model robustness in out-of-distribution scenarios, we evaluate our method on the OOD-CV dataset [40]. The OOD-CV dataset includes out-of-distribution examples from 10 categories of PASCAL3D+ and is a benchmark to evaluate out-of-distribution robustness to individual nuisance factors including pose, shape, texture, context and weather.

Metric	$ACC_{\frac{\pi}{6}}\uparrow$				$ACC_{\frac{\pi}{18}}\uparrow$				$MedErr\downarrow$			
Num Annotations	7	20	50	Mean	7	20	50	Mean	7	20	50	Mean
Res50-General	36.1	45.2	54.6	45.3	14.7	25.5	34.2	24.8	39.1	26.3	20.2	28.5
StarMap [41]	30.7	35.6	53.8	40.0	4.3	7.2	19.0	10.1	49.6	46.4	27.9	41.3
NeMo [34]	38.4	51.7	69.3	53.1	17.8	31.9	45.7	31.8	60.0	33.3	22.1	38.5
NVSM [35]	53.8	61.7	65.6	60.4	27.0	34.0	39.8	33.6	37.5	28.7	24.2	30.1
P3D-Diffusion + NeMo	78.2	79.3	82.6	80.0	55.3	55.9	60.2	57.1	**15.8**	15.4	10.6	13.9
P3D-Diffusion + CC3D	**79.1**	**80.6**	**83.5**	**81.1**	**56.6**	**57.2**	**61.9**	**58.6**	16.2	**15.0**	**9.7**	**13.6**
NeMo Full Sup. [34]	—	—	—	89.3	—	—	—	66.7	—	—	—	7.7

Table 1. Few-shot pose estimation results on 6 vehicle classes of PASCAL3D+ following the evaluation protocol in [35]. We indicate the number of annotations during training for each category and evaluate all approaches using Accuracy (in percent, higher better) and Median Error (in degree, lower better). We also include the fully supervised baseline [34] (Full Sup.) which is trained from the full dataset (hundreds of images per category).

Evaluation. 3D pose estimation aims to recover the 3D rotation parameterized by azimuth, elevation, and in-plane rotation of the viewing camera. Following previous works [34, 41], we evaluate the error between the predicted rotation matrix and the ground truth rotation matrix: $\Delta\left(R_{pred}, R_{gt}\right) = \frac{\left\|\log \mathrm{m}\left(R_{pred}^{T} R_{gt}\right)\right\|_F}{\sqrt{2}}$. We report pose estimation accuracies under common thresholds, $\frac{\pi}{6}$ and $\frac{\pi}{18}$.

Training Setup. We use an ImageNet [9] pre-trained ResNet50 [13] as feature extractor. The dimensions of the cuboid mesh $\mathfrak{N}$ are defined such that for each category most of the object area is covered. Which takes around 1 hour per category on a machine with 2 RTX Titan Xp GPUs. We implement our approach in PyTorch [23] and apply the rasterisation implemented in PyTorch3D [27].

P3D-Diffusion. We sample the synthetic training data using the CAD models provided in the PASCAL3D+ and OOD-CV datasets. We use Blender [8] as our renderer to generate the synthetic images. We sample 7000 images per class and randomize the texture of the CAD model by sampling textures from the describable texture database [6]. The background images are sampled from a collection of 100 images that we collected from the internet by searching for the keywords “wallpaper”+[“street, jungle, market, beach”]. We provide detailed statistics and examples of our synthetic dataset in the supplementary materials.

Baselines. We compare our model to fully supervised methods for category-level 3D pose estimation, including StarMap [41] and NeMo [34] using their official implementation and training setup. Following common practice, we also evaluate a popular baseline that formulates pose estimation as a classification problem. We follow the implementation in [41] and trained a ResNet50 classifier, which shares the same backbone as NeMo.

Few-shot Learning. We further compare our approach at a recently proposed semi-supervised few-shot learning setting [35]. In this setting, 7, 20, and 50 annotated images from Pascal3D+ are used for training. We follow [35] and evaluate 6 vehicle categories (aeroplane, bicycle, boat, bus, car, motorbike) with a relatively even distribution of the azimuth angle. In order to utilize the unlabelled images, a common pseudo-labelling strategy is used for all baselines. Specifically, we first train a model on the annotated images, and use the trained model to predict a pseudo-label for all unlabelled images in the training set. We keep those pseudo-labels with a confidence threshold $\tau = 0.9$, and we utilize the pseudo-labeled data as well as the annotated data to train the final model. The previous state-of-the-art model in this few-shot setting is NVSM [35].

5.2. Few-Shot 3D Pose Estimation

Table 1 shows the performance of our approach and all baselines at semi-supervised few-shot 3D pose estimation on 6 vehicle classes of the PASCAL3D+ dataset. All models are evaluated using 7, 20, and 50 (per class) training images with annotated 3D pose and a collection of unlabelled training data (as described in Section 5.1). Among the models trained without our P3D-Diffusion dataset, the ResNet50 classification baseline and NeMo achieve a comparable performance using few annotated images. Notably, NVSM is by far the best performing baseline when using only 7 or 20 annotated images per object class. However, when using 50 annotated images, the NeMo baseline outperforms NVSM by a margin of 3.7%.

Using our P3D-Diffusion dataset, our proposed **CC3D outperforms all baselines across all few-shot data regimes**. Remarkably, our model constantly outperforms the prior arts by a margin of $> 20\%$ in both $\frac{\pi}{6}$ and $\frac{\pi}{18}$ accu-

Evaluation Metric	$ACC_{\frac{\pi}{6}} \uparrow$	$ACC_{\frac{\pi}{18}} \uparrow$	$MedErr \downarrow$
Res50-General	88.1	44.6	11.7
StarMap [41]	89.4	59.5	9.0
NeMo [34]	86.1	61.0	8.8
P3D-Diffusion + Res50	53.5	13.2	26.7
P3D-Diffusion + NeMo	71.8	39.5	17.6
P3D-Diffusion + CC3D	76.3	41.4	15.5
P3D-Diffusion + CC3D + 10%	86.7	62.4	8.4
P3D-Diffusion + CC3D + 50%	**90.7**	**71.4**	**6.9**

Table 2. Pose estimation results on PASCAL3D+. We evaluate all models using pose accuracy and median pose error. We compare the state-of-the-art fully supervised baselines (StarMap, NeMo, Res50) to models learned on synthetic data and transferred to real (P3D-Diffusion + Res50, P3D-Diffusion + NeMo, P3D-Diffusion + CC3D) and P3D-Diffusion + CC3D + fine-tuned with 10% and 50% of annotated data. CC3D outperforms other approaches when trained without real annotations, and even outperforms the SOTA methods with only 10% of the annotated data.

racy. We further observe that the NeMo model trained using our P3D-Diffusion dataset (P3D-Diffusion +NeMo) and domain adapted as described in Section 5.1 also significantly outperforms the original NeMo baseline, demonstrating the effectiveness of our synthetic data. Nevertheless, it does not match the performance of our proposed 3D-aware contrastive consistency approach. Finally, with only 50 annotated images our CC3D model even performs competitively to the fully supervised trailing it by only by 8.2%@$\frac{\pi}{6}$ and 8.1%@$\frac{\pi}{18}$, hence significantly closing the gap between fully supervised models and models trained on synthetic data.

5.3. Comparison to Supervised Approaches

Table 2 summarizes our results when comparing to fully supervised models trained on the full annotated dataset. In the experiment P3D-Diffusion +CC3D, we first pre-train with synthetic data and then use $\mathcal{L}_{\text{domain}}$ for fine-tuning with unlabeled real images. In experiments named "P3D-Diffusion +CC3D+X%", we additionally use labelled data for a final fine-tuning, where X% denotes the number of available real image labels. These real image samples were randomly selected after shuffling the dataset.

When annotations of real images are not available, our proposed CC3D outperforms the NeMo and ResNet50 baselines that use same training data (P3D-Diffusion + Res50, P3D-Diffusion + NeMo) by a significant margin. Notably, P3D-Diffusion + NeMo can bridge the synthetic-to-real domain gap much better compared to P3D-Diffusion + Res50, outperforming it by $> 10\%$ at $\frac{\pi}{6}$ and $\frac{\pi}{18}$. Our CC3D further outperforms P3D-Diffusion + NeMo by 4.5% and 1.9% at $\frac{\pi}{6}$ and $\frac{\pi}{18}$ respectively, while also reducing the median prediction error by 2.1%. All these results are achieved without access to any real image annotation, which demonstrates the effectiveness of our proposed approach.

When annotations of real images are available, our proposed **CC3D widely outperforms the fully supervised state-of-the art using only** 50% **of the annotated data** that is available to the fully supervised methods by 1.3%@$\frac{\pi}{6}$ and 10.4%@$\frac{\pi}{18}$. This shows that our method can effectively leverage the accurate annotations in synthetic data to learn a representation that benefits real data well. Remarkably, even when using only 10% of the data that is available to the SOTA supervised methods, our approach can match their performance and even outperform them in terms of the finer $\frac{\pi}{18}$ accuracy by a fair margin. This demonstrates the enhanced efficiency that our proposed CC3D approach enables for 3D pose estimation.

5.4. Robust 3D Pose Estimation

In Table 3 we illustrate the performance of our CC3D approach and several baselines at 3D pose estimation on the OOD-CV dataset to investigate their their robustness under domain shifts to shape, pose, texture, context, and weather. We observe that the fully supervised ResNet50 baseline has on average a similar performance under OOD shifts as the NeMo model. We note that the NeMo model achieves a higher accuracy on the Pascal3D+ data (Table 2) and hence indicating less robustness compared to the ResNet50.

All models trained without real annotations achieve a lower performance compared to the fully supervised baselines. However, the performance gap between the fully supervised and unsupervised baselines is lower compared to the PASCAL3D+ dataset. This can be attributed to the much larger variability in the synthetic data regarding texture, context, pose and background. Notably, there only remains a large performance gap in terms of OOD robustness to texture and weather shifts in the data between supervised and unsupervised models, indicating that the variability in the texture of the synthetic data is not sufficiently realistic. We also note that our **CC3D achieves the highest OOD robustness among the unsupervised models**.

When fine-tuned with 10% of real data the performance of the unsupervised models is enhanced significantly. Notably, our **CC3D is able to close the gap to fully supervised models in terms of OOD robustness** due to the large variability in the synthetic data and its ability to transfer this knowledge to real images.

We provide some qualitative results in the supplementary materials to visualize our model's predictions on PASCAL3D+ and OOD-CV.

5.5. Ablation Study

As shown in Table 4, we evaluate the contribution of each proposed component. Specifically, we evaluate various settings on five categories (aeroplane, boat, car, mo-

Evaluation Metric	$ACC_{\frac{\pi}{6}} \uparrow$						$ACC_{\frac{\pi}{18}} \uparrow$					
Nuisance	Context	Pose	Shape	Texture	Weather	Mean	Context	Pose	Shape	Texture	Weather	Mean
Res50-General	50.6	20.8	48.5	**64.6**	57.5	46.9	12.3	0.2	12.1	23.3	23.2	14.3
NeMo [34]	53.3	37.5	51.1	62.1	**57.8**	51.6	23.7	7.1	21.6	**39.1**	**33.8**	24.5
P3D-Diffusion + Res50	37.2	32.8	33.5	32.5	37.3	33.5	6.8	5.3	6.3	6.7	9.9	6.3
P3D-Diffusion + NeMo	48.1	43.7	46.0	39.9	38.8	42.6	12.8	10.9	14.7	11.5	13.3	12.7
P3D-Diffusion + CC3D	54.6	**45.8**	52.3	51.0	44.5	48.2	12.1	12.3	16.1	16.6	16.3	14.8
P3D-Diffusion + Res50 + 10%	37.0	4.8	34.8	47.0	40.3	34.8	9.5	0.0	7.4	13.9	12.9	7.4
P3D-Diffusion + NeMo + 10%	62.2	42.5	**62.0**	60.4	56.5	**55.2**	25.3	**13.0**	29.5	35.2	31.6	26.3
P3D-Diffusion + CC3D + 10%	**62.5**	42.8	61.8	60.7	55.2	54.9	**26.5**	12.9	**30.5**	35.4	33.5	**27.3**

Table 3. Robustness of pose estimation methods on the OOD-CV dataset. We report the performance on OOD shifts in the object shape, 3D pose, texture, context and weather. We compare fully supervised baselines (NeMo, Res50) to models learned on synthetic data and transferred to real (P3D-Diffusion + Res50, P3D-Diffusion + NeMo, P3D-Diffusion + CC3D) and when fine-tuning these models with 10% real annotated data. Note how our CC3D model achieves higher robustness compared to other models trained without real annotation. When fine-tuned on 10% of the training data in OOD-CV (+10%) it performs on par at $\frac{\pi}{6}$ and outperforms all baselines at $\frac{\pi}{18}$.

PASCAL3D+	$ACC_{\frac{\pi}{6}} \uparrow$	$ACC_{\frac{\pi}{18}} \uparrow$	$MedErr \downarrow$
full model	79.2	52.0	14.1
- style transfer	75.9 (-3.3)	47.8 (-4.2)	17.1 (+3.0)
- unsup adaptation	76.5 (-2.7)	49.0 (-3.0)	16.0 (+1.9)
- style transfer - unsup adaptation	70.6 (-8.6)	46.5 (-5.5)	23.6 (+9.5)

Table 4. Ablation study on the unsupervised domain adaptation and graphics-guided style transfer module on the PASCAL3D+ dataset (aeroplane, boat, car, motorbike, and train).

torbike, and train) of the PASCAL3D+ dataset. We use "P3D-Diffusion + CC3D" as the full model. The graphics-guided style transfer, denoted "style transfer", produced high-quality synthetic data with diverse textures and colors using a style transfer network. The unsupervised domain adaptation, denoted "unsup adaptation", adapts the synthetically trained model to real data with a domain-consistency loss on pseudo-labels (Eq 4).

5.6. Breaking Spurious Correlation with Domain-Nonspecific Synthetic Data

From the causal perspective, the OOD robustness problem is mainly due to the spurious correlation between domain-specific features and task-related semantic features [16]. Our proposed OOD-aware generation can effectively break such spurious correlation by generating synthetic data with domain-nonspecific backgrounds and demonstrate large improvements on OOD-CV dataset. As an ablation study, we re-generate the synthetic dataset with domain-specific backgrounds (e.g., cars have backgrounds on roads), denoted P3D-Diffusion-Spurious. As shown in Table 5, models trained on P3D-Diffusion achieves much better OOD robustness with a negligible degradation on the in-distribution benchmark (i.e., PASCAL3D+).

PASCAL3D+	$ACC_{\frac{\pi}{6}} \uparrow$	$ACC_{\frac{\pi}{18}} \uparrow$	$MedErr \downarrow$
P3D-Diffusion +CC3D	76.3	41.4	15.5
P3D-Diffusion-Spurious+CC3D	77.0 (+0.7)	42.8 (+1.4)	14.8 (-0.7)
OOD-CV	$ACC_{\frac{\pi}{6}} \uparrow$	$ACC_{\frac{\pi}{18}} \uparrow$	$MedErr \downarrow$
P3D-Diffusion +CC3D	48.2	14.8	37.0
P3D-Diffusion-Spurious+CC3D	42.7 (-5.5)	16.4 (+1.6)	45.7 (+8.7)

Table 5. Ablation study on the OOD-aware generation with which we can effectively break spurious correlation between domain-specific features and task-related semantic features. Our method with P3D-Diffusion demonstrate much better OOD robustness at the cost of a small degradation on in-distribution dataset.

6. Conclusion

In this work, we narrowed the performance gap between category-level 3D pose estimation models trained on synthetic and real data. To bridge the domain gap, we proposed diffusion-enhanced synthetic data with realistic synthetic images and accurate 3D annotations. Moreover, we developed a new domain adaptation algorithm, CC3D, that leverages the 3D mesh geometry to obtain consistent pseudo-correspondences between synthetic and real images. Experimental results demonstrate that CC3D can largely reduce the domain gap to fully-supervised models when trained without any real annotations, and performs competitively with previous SOTA models when fine-tuned with very few annotated real data. Lastly, we show that with the help of our diffusion-enhanced synthetic data and our CC3D, we can effectively improve model robustness in very challenging out-of-distribution scenarios.

Acknowledgements. AK acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under Grant No. 468670075. AL acknowledges support via Office of Naval Research with grants N00014-23-1-2641 and N00014-21-1-2812.

References

[1] Yutong Bai, Angtian Wang, Adam Kortylewski, and Alan Yuille. Coke: Contrastive learning for robust keypoint detection. *WACV*, 2023. 4, 5

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 3

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[4] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016. 3

[5] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. *arXiv preprint arXiv:2008.08145*, 2020. 2

[6] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 4

[8] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017. 3

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[12] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5078–5088, 2022. 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3

[16] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*, pages 4555–4562. PMLR, 2021. 4, 8

[17] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M. Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3303–3312, October 2021. 1

[18] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[19] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[20] Xiaofeng Liu, Yuzhuo Han, Song Bai, Yi Ge, Tianxing Wang, Xu Han, Site Li, Jane You, and Jun Lu. Importance-aware semantic segmentation in self-driving with discrete wasserstein training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11629–11636, 2020. 3

[21] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 1, 2

[22] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12395, 2020. 3

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6

[24] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2011–2018. IEEE, 2017. 2

[25] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[26] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017. 1

[27] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 6

[28] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018. 3

[29] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[30] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2686–2694, 2015. 1

[31] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 3

[32] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 1, 2

[33] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 3

[34] Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In *International Conference on Learning Representations*, 2021. 1, 2, 4, 6, 7, 8

[35] Angtian Wang, Shenxiao Mei, Alan L Yuille, and Adam Kortylewski. Neural view synthesis and matching for semi-supervised few-shot learning of 3d pose. *Advances in Neural Information Processing Systems*, 34:7207–7219, 2021. 1, 2, 5, 6

[36] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2

[37] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014. 3, 5

[38] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1

[39] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3

[40] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 5

[41] Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. Starmap for category-agnostic keypoint and viewpoint estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 1, 2, 6, 7

[42] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 3

[43] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 3