DatasetNeRF: Efficient 3D-aware Data Factory with Generative Radiance Fields

Yu Chi[1,2], Fangneng Zhan[2], Sibo Wu[1],
Christian Theobalt[2], and Adam Kortylewski[2,3]

[1] Technical University of Munich
[2] Max Planck Institute for Informatics
[3] University of Freiburg

Abstract. Progress in 3D computer vision tasks demands a huge amount of data, yet annotating multi-view images with 3D-consistent annotations, or point clouds with part segmentation is both time-consuming and challenging. This paper introduces DatasetNeRF, a novel approach capable of generating infinite, high-quality 3D-consistent 2D annotations alongside 3D point cloud segmentations, while utilizing minimal 2D human-labeled annotations. Specifically, we leverage the semantic prior within a 3D generative model to train a semantic decoder, requiring only a handful of fine-grained labeled samples. Once trained, the decoder generalizes across the latent space, enabling the generation of infinite data. The generated data is applicable across various computer vision tasks, including video segmentation and 3D point cloud segmentation in both synthetic and real-world scenarios. Our approach not only surpasses baseline models in segmentation quality, achieving superior 3D-Consistency and segmentation precision on individual images, but also demonstrates versatility by being applicable to both articulated and non-articulated generative models. Furthermore, we explore applications stemming from our approach, such as 3D-aware semantic editing and 3D inversion. Code can be found at ⭘/GenIntel/DatasetNeRF.

Keywords: Efficient Synthetic Dataset Generation · 3D-Aware Generative Model · Neural Rendering

1 Introduction

In recent years, research on Large-Scale Models or Foundation Models has become a prevailing trend. Training these kinds of models demands vast amounts of 2D or 3D labeled data, which entails significant human effort. Based on this limitation, a critical question emerges: How do we efficiently generate a substantial volume of high-quality data-annotation pairs while minimizing human labor? Our paper introduces a method to generate an unlimited supply of high-quality, 3D-aware data by utilizing only a limited set of human-provided 2D annotations.

To efficiently scale datasets, recent approaches [3,34,68,76] utilize rich semantic features from 2D generative models as image representations for downstream
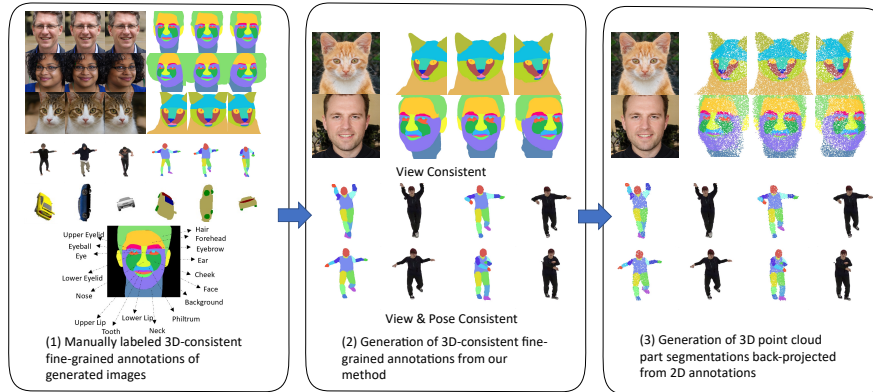
Fig. 1: DatasetNeRF Pipeline Overview: (1) The manual creation of a small set of multi-view consistent annotations, followed by the training of a semantic segmentation branch using a pretrained 3D GAN backbone. (2) Leveraging the latent space's generalizability to produce an infinite array of 3D-consistent, fine-grained annotations. (3) Employing a depth prior from the 3D GAN backbone to back-project 2D segmentations to 3D point cloud segmentations.

tasks, such as semantic segmentation. The remarkable representational capacity of generative models facilitates training segmentation models with only a minimal dataset. During inference, a randomly sampled latent code from the generator is capable of producing a corresponding high-quality annotation. This mechanism effectively transforms the generator into an inexhaustible source of data, enabling the creation of extensive datasets with significantly reduced labeling requirements. However, existing methods predominantly focus on 2D generation models, limiting their capability for 3D-aware tasks. Nevertheless, the emergence of geometry-aware 3D Generative Adversarial Networks (GANs) [8,9,51], which decouple latent code and camera pose, offers promising avenues.

In this paper, we introduce *DatasetNeRF*, an efficient 3D-aware data factory based on generative radiance fields. Our 3D-aware Data Factory is adept at creating extensive datasets, delivering high-quality, 3D-consistent, fine-grained semantic segmentation, and 3D point cloud part segmentation as shown in Figure Fig. 1. This is accomplished by training a semantic branch on a pre-trained 3D GAN, such as EG3D [8], leveraging the semantic features in the generator's backbone to enhance the feature tri-plane for semantic volumetric rendering. To improve the 3D consistency of our segmentations, we incorporate a density prior from the pre-trained EG3D model into the semantic volumetric rendering process. We further exploit the depth prior from the pre-trained model, efficiently back-projecting the semantic output to obtain 3D point cloud part segmentation. Our approach facilitates easy manipulation of viewpoints, allowing us to render semantically consistent masks across multiple views. By merging the back-projected point cloud part segmentations from different perspectives, we can achieve comprehensive point cloud part segmentation of the entire 3D rep-

resentation. Remarkably, our process for generating this vast array of 3D-aware data requires only a limited set of 2D data for training.

We evaluate our approach on the AFHQ-cat [13], FFHQ [27], AIST++ dataset [35] and Nersemble dataset [32]. We generate detailed annotations for these datasets, showing our method outperforms existing baselines by enhancing 3D consistency across video sequences and improving segmentation accuracy for single images. Additionally, we demonstrate that our method is also seamlessly compatible with articulated generative radiance fields [6] on AIST++ dataset.

In addition, we qualitatively demonstrate that models trained with our generated dataset can generalize well to real-world scans, such as those in the Nersemble dataset [32]. We also augment the point cloud semantic part segmentation benchmark dataset [72] using our method, with a specific focus on the ShapeNet-Car dataset [10]. Our work further analyzes potential applications like 3D-aware semantic editing and 3D inversion, demonstrating that the ability to generate infinite 3D-aware data from a limited number of 2D labeled annotations paves the way for numerous 2D and 3D downstream applications.

2   Related Work

2.1   Neural Representations and Rendering

In recent years, the emergent implicit neural representation offers efficient, memory-conscious, and continuous 3D-aware representations for objects [2, 21, 43, 53] and scenes [42, 44, 45, 61, 62, 64, 80] in arbitrary resolution. By combining implicit neural representation with volume render, NeRF [44] and its descendants [19, 24, 30, 36, 37, 39–41, 46, 63, 66, 70, 75] have yielded promising results for both 3D reconstruction and novel view synthesis applications. Along with image synthesis, the implicit representations are also used to predict semantic maps [33, 67, 77]. For example, Semantic-NeRF [77] augments the original NeRF by appending a segmentation renderer. NeSF [67] learns a semantic-feature grid for semantic maps generation. However, querying properties for each sampled point leads to a low training and inference speed. Considering the pros and cons of explicit representations and implicit representations, recent works [4, 5, 11, 14, 45, 73] propose hybrid representations to complement each other. In this work, we also use hybrid tri-plane representations for 3D modeling.

2.2   3D Generative Models

The Generative Adversarial Networks (GANs) [20] have demonstrated remarkable capabilities in generating photorealistic 2D images [17, 26, 28, 29, 74]. With this success, some works extended this setting to 3D domain. For instance, PrGANs [18] and VON [79] first learn to generate a voxelized 3D shape and then project it to 2D. BlockGAN [48] learns 3D features but separates one scene into different objects. However, these approaches encounter challenges in achieving photorealistic details due to the limited grid resolutions.

Recent works [8, 9, 14, 22, 49, 52, 59, 65] integrated neural implicit representation into GANs to enable 3D-aware image synthesis with multi-view consistency. Specifically, GRAF [59] combines NeRF for scene representation with an adversarial framework to enable the training on unposed images; pi-GAN [9] operates in a similar setting but makes some differences in network architecture and training strategy; EG3D [8] learns tri-plane hybrid 3D representation and interprets the aggregated features via volume rendering, ensuring expressive semantic feature extraction and high-resolution geometry-aware image synthesis. While the learned features in generative models are aggregated to generate 3D-aware images, there is still space to harness them for other proposes.

2.3   Synthetic Dataset Generation

Traditional dataset synthesis [16,54,57,58] relies on computer simulations for rendering images along with their corresponding labels, which can save annotation cost. However, models trained on such datasets often face challenges in generalizing to real-world datasets due to domain gaps. Unlike traditional methods of dataset synthesis, the use of generative models for dataset synthesis is favored due to their ability to produce a large number of high-quality and diverse images with similar distribution of natural data [71]. The family of generative models is extensive, with GANs [20], diffusion models [25], and NeRF [44] having achieved notable success in image synthesis. Specifically, many works leverage the rich semantic information learned by GANs to manipulate images [1, 38, 60]. Diffusion models benefit from a stationary training objective and demonstrate decent scalability, enabling the generation of high-quality images [15]. NeRF, as a recent and emerging generative model, has received widespread acclaim for maintaining multi-view consistency. Additionally, the capability of generative models to learn rich semantic information allows such methods to learn to generate new data and labels using only a few manually annotated images [34, 47, 68, 76]. For instance, DatasetGAN [76] leverages StyleGAN [28] as an image generator and synthesize accurate semantic labels with a few human labeled data. Nevertheless, these previous efforts focused on generating semantic maps for 2D datasets.

3   Method

We introduce *DatasetNeRF*, a framework designed to generate an extensive range of 3D-aware data. It efficiently produces fine-grained, multi-view consistent annotations and detailed 3D point cloud part segmentations from a limited collection of human-annotated 2D training images.

To address the challenge of generating a varied 3D-aware dataset, we employ a 3D GAN generator as the foundational architecture of our framework. We augment this 3D GAN with a semantic segmentation branch, enabling the production of precise annotations across diverse 3D viewpoints as well as detailed 3D point cloud part segmentations. Fig. 2 provides a comprehensive visualization of
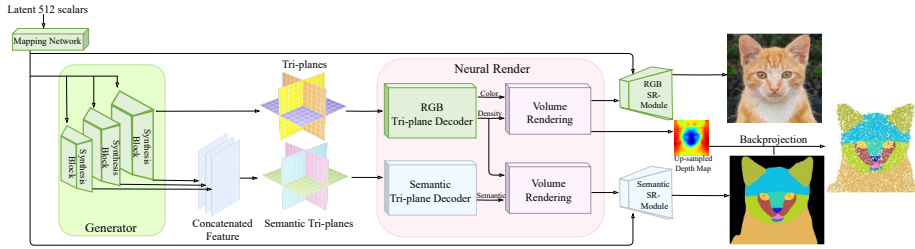
Fig. 2: Overall Architecture of DatasetNeRF. The DatasetNeRF architecture unifies a pretrained EG3D model with a semantic segmentation branch, comprising an enhanced semantic tri-plane, a semantic feature decoder, and a semantic super-resolution module. The semantic feature tri-plane is constructed by reshaping the concatenated outputs from all synthesis blocks of the EG3D generator. The semantic feature decoder interprets aggregated features from semantic tri-plane into a 32-channel semantic feature for every point. The semantic feature map is rendered by semantic volumetric rendering. We incorporate a density prior from the pretrained RGB decoder during the rendering process to enhance 3D consistency. The semantic super-resolution module upscales and refines the rendered semantic feature map into the final semantic output. The combination of the semantic mask output and the upsampled depth map from the pretrained EG3D model enables an efficient process for back-projecting the semantic mask, thereby facilitating the generation of point cloud part segmentation.

the entire model architecture. For a more in-depth understanding of the different components of our framework, we delineate the specific backbones used for various tasks in Sec. 3.1. Subsequently, in Sec. 3.2, we elaborate on the methodology employed to train the semantic segmentation branch. In Sec. 3.3, we provide a detailed presentation of both the generation process and the resulting 3D-aware data within the DatasetNeRF framework.

3.1 3D GAN Generator Backbone

We take EG3D [8] as our backbone model, which introduces a tri-plane architecture for efficient neural rendering at reduced resolutions. This tri-plane consists of reshaped feature representations derived from the output of the generator. To enhance the representational power of the triplane in our work, we take all feature maps $\{S^0, S^1, \ldots, S^k\}$ from the the output of consecutive synthesis block of the pretrained generator. These features are upsampled to match the highest output resolution, and subsequently concatenated into a feature tensor with dimensions $[N, N, C]$. Following this, we reshape the concatenated feature tensor into an augmented tri-plane format, similar to that of EG3D [8], to facilitate our semantic neural rendering pipeline. This tri-plane format, distinct from other work [8], represents a key innovation in our work. Its significance and impact are further validated through an ablation study detailed later in the text. Our enhanced tri-plane serves as the semantic feature volume for rendering
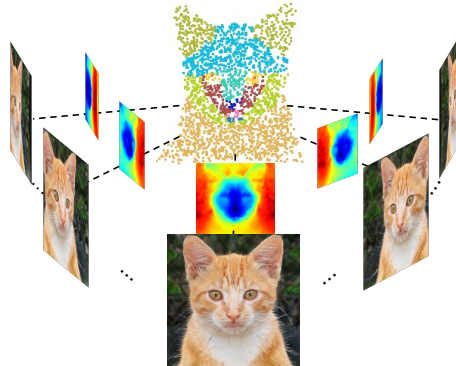
Fig. 3: The illustration of multi-view point cloud fusion.

semantically-rich features within our semantic segmentation branch, enabling accurate depiction of complex structures in images.

Notably, our methodology also exhibits compatibility with a range of 3D GAN architectures, whether articulated or inarticulated. This adaptability underscores the robustness of our approach in generating 3D-consistent segmentation. It facilitates not only multi-view consistency but also pose-consistency in segmentations, further proving the utility of our approach across diverse tasks.

3.2   Semantic Segmentation Branch Training

We query any 3D position x within our semantic tri-plane with enhanced format by projecting it onto each of the three feature planes, obtaining the respective feature vectors (F_{xy}, F_{xz}, F_{yz}) through bi-linear interpolation. These vectors are then aggregated via summation. This aggregated feature serves as the input to the subsequent semantic decoder, which outputs a 32-channel semantic feature. To harness the 3D consistency inherent in the pretrained EG3D model, we re-use the same density σ as the pretrained RGB decoder at the equivalent tri-plane point. For a majority of our experiments, the semantic feature map is rendered at a resolution of 128^2. Through semantic volumetric rendering, we derive a raw semantic map $\hat{I}_s \in \mathbf{R}^{128 \times 128 \times C}$ and a semantic feature map $\hat{I}_\phi \in \mathbf{R}^{128 \times 128 \times 32}$. Subsequently, a semantic super-resolution module U_s is utilized to refine the semantic map into a high-resolution segmentation $\hat{I}_s^+ \in \mathbf{R}^{512 \times 512 \times C}$:

$$\hat{I}_s^+ = U_s(\hat{I}_s, \hat{I}_\phi). \tag{1}$$

For a given ground-truth viewpoint P and corresponding latent code $\mathbf{z}$, we compare the ground-truth semantic mask I_s with our model's output semantic mask using cross-entropy loss, mathematically represented as:

$$\mathcal{L}_{\text{CE}}(I_s, \hat{I}_s^+) = -\sum_{c=1}^{C} I_{s,c} \log(\hat{I}_{s,c}^+), \tag{2}$$

where $I_{s,c}$ is the binary indicator of the ground-truth class label for class c and $\hat{I}_{s,c}^+$ is the predicted probability of class c for each pixel in the final output.

3.3   DatasetNeRF as 3D-Aware Data Factory

DatasetNeRF as Multi-view Consistent Segmentations Factory. Empowered by the geometric priors derived from 3D GAN, our DatasetNeRF naturally specializes in generating segmentations that maintain consistency across multiple viewpoints. Once trained, the model adeptly produces high-quality semantic segmentations from a randomly sampled latent code paired with any given pose. The generated multi-view consistent images are illustrated in Fig. 4. The easy generation of fine-grained, multi-view consistent annotations markedly diminishes the need for human effort.

DatasetNeRF as 3D Point Cloud Segmentation Factory. Initially, we render a depth map using the pretrained RGB branch. The depth maps are generated via volumetric ray marching. This method computes depth by aggregating weighted averages of individual depths along each ray. The depth map is upsampled to align with the dimensions of the semantic mask, allowing the semantic mask to be back-projected into 3D space. The final point cloud of the object is formed by merging back-projected semantic maps from various viewpoints, shown in Fig. 3. The efficient acquisition of fine-grained 3D point cloud segmentation significantly reduces the amount of human effort required. The visualization of point cloud part segmentation is illustrated in Fig. 1 (3).

Extension to Articulated Generative Radiance Field. We showcase how our method can also be applied to articulated generative radiance field. Instead of using EG3D, we adopt the generator of GNARF [5] as our backbone. GNARF [5] introduces an efficient neural representation for articulated objects, including bodies and heads, that combines the tri-plane representation with an explicit feature deformation guided by a template shape. The semantic branch is trained on top of the deformation-aware feature tri-plane. The training set contains 150 annotations from 30 different human samples and 60 different training poses. As shown in Fig. 4, the result can be well generalized on novel human poses.

4   A Small Dataset with Human Annotations

Our method necessitates a small dataset with annotation. To this end, we employ our backbone model to synthesize a small number of images, followed by a professional annotator for fine-grained annotation. Our fine-grained annotation protocol was applied to AFHQ-Cat [13], FFHQ [27], and AIST++ [35], with a simplified scheme utilized for ShapeNet-Car [10].

Annotation Details For the training set, we crafted 90 fine-grained annotations for each of the AFHQ-Cat [13] and FFHQ [27] datasets. These encompass 30 distinct subjects with three different views for each subject. The angular disposition for both training and testing spans from $-\frac{\pi}{6}$ to $\frac{\pi}{6}$ relative to the frontal view, holding all other degrees of freedom constant. Consequently, each subject is

depicted in a frontal stance, accompanied by both leftward and rightward poses. The AIST++ dataset [35] posed a greater challenge due to the diversity of human poses, prompting us to generate 150 annotations for 60 disparate poses across various human subjects. For the ShapeNet-Car dataset, our efforts yielded 90 annotations from 30 distinct samples, each from a unique viewpoint. The annotations for ShapeNet-Car, identifying parts like *hood, roof, wheels, other*, align with the standard labels used in the point cloud part segmentation benchmark [72]. The manually created dataset is visualized in Fig. 1(1).
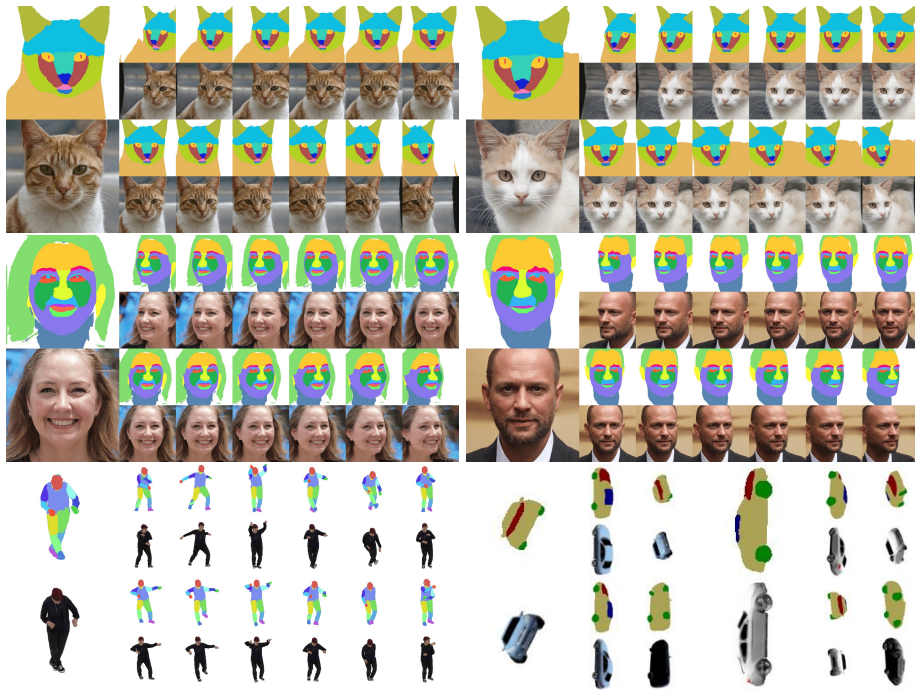

Fig. 4: Examples of synthesized image-annotation pairs from 3D-aware data factoty.

5   Experiments

We conduct extensive experiments with our approach. First, we assess the 2D part segmentation performance across two distinct object categories: cat and human faces. Furthermore, we demonstrate the efficacy of our method in generating 3D point cloud part segmentations for both cat faces and ShapeNet-Cars. Finally, we show a variety of 3D applications based on GAN inversion [78].

Table 1: Comparison of different approaches on AFHQ-Cat and FFHQ datasets. In the table, "Ind." refers to 90 unique individual image annotations representing arbitrary poses within the training range, and "Vid." denotes a video annotation comprising 61 frames.

Methods	AFHQ-Cat Ind.		AFHQ-Cat Vid.		FFHQ Ind.		FFHQ Vid.	
	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.
Transfer Learning	0.2995	0.6358	0.2605	0.5766	0.4083	0.7964	0.3895	0.8611
DatasetGAN	0.5381	0.8464	0.5971	0.8625	**0.6317**	0.8881	0.6390	0.9259
DatasetNeRF	**0.6057**	**0.8798**	**0.6756**	**0.9253**	0.6200	**0.8996**	**0.6561**	**0.9278**

5.1   2D Part Segmentation

2D Part Segmentation Network. Consistent with the approach used in DatasetGAN [76], we employ Deeplab-V3 [12] with ResNet101 [23] backbone as our 2D part segmentation network.

Experimental Setup. Our baselines include transfer learning and Dataset-GAN [76]. For the transfer learning approach, we fine-tune the last layer of a pre-trained (on ImageNet) network with our human-annotated data. For the DatasetGAN baseline, we retrain a DatasetGAN model with the annotated data to generate a dataset comprised of 10K image-annotation pairs, which serves as the training set for the Deeplab-V3 segmentation network. We adjust the size of the concatenated feature to match that of the DatasetNeRF's feature size, ensuring a fair comparison. For our DatasetNeRF, we generate 10K images with uniformly distributed angles which cover both frontal and horizontal views. The Deeplab-V3 segmentation network is then trained from scratch using this dataset. Each test dataset contains 90 unique individual image annotations for arbitrary poses within the training spectrum, as well as a sequential video annotation comprising 61 frames.

Quantitative Evaluation. Tab. 1 presents the segmentation results for the AFHQ-Cat and FFHQ datasets. DatasetNeRF outperforms the baseline models in terms of segmentation quality in the video test set on both AFHQ-Cat and FFHQ datasets. This enhancement underscores the informativeness of the data generated by our approach, which is attributed to the 3D consistency prior inherent from the pretrained generator backbone. Specifically, in the individual test set, DatasetNeRF achieves superior segmentation results for the AFHQ-Cat dataset and demonstrates comparable performance for the FFHQ dataset.

Qualitative Evaluation for 3D-Consistency. We show spatiotemporal line textures [7] of semantic masks from different poses in Fig. S10 in the supplementary material. The smoothness of these semantic textures, matching the corresponding RGB textures, demonstrates the 3D-consistency of our generated data. While modeling 3D-consistency in RGB space is challenging due to high-frequency details with 2D CNN upsampling module [69], semantic mask generation is comparably easier and performs well.

Table 2: Training results with different numbers of generated point cloud training samples on AFHQ-Cat dataset.

Experiments	Accuracy	mIoU
400 (Generated)	0.8788	0.6268
600 (Generated)	0.8809	0.6447
800 (Generated)	0.8828	0.6403
1100 (Generated)	**0.8950**	**0.6651**

Table 3: Evaluation of the generated point cloud on ShapeNet-Car dataset with Point-Net as the backbone model.

Experiments	Accuracy	mIoU
600 (ShapeNet)	0.8796	0.6773
600 (ShapeNet) + 725 (Generated)	**0.9073**	**0.7519**
1325 (ShapeNet)	0.9059	0.7412
1325 (ShapeNet) + 1325 (Generated)	**0.9104**	**0.7571**

5.2   3D Point Cloud Part Segmentation

We demonstrate the effectiveness of our generated point cloud part segmentation dataset by training PointNet [55] on the generated data. We assess the performance on AFHQ-Cat faces and Shape-Net Car based on mean Intersection-over-Union (mIoU) and accuracy metrics. We show that our approach not only enables the generation of high-quality new point cloud part segmentations dataset from self-annotated 2D images but also acts as a valuable augmentation to existing classical 3D point cloud part segmentation benchmark datasets.

AFHQ-Cat Face Point Cloud Segmentation. From 1200 generated samples, we create a fixed test set of 100 point clouds and train PointNet with varying numbers of training samples. Tab. 2 illustrates that increased training samples enhance model performance on the test set, which shows the effectiveness of our self-generated point cloud part segmentation dataset.

Augmentation of ShapeNet-Car. Our method's efficacy, detailed in Tab. 3, shows its potential as both a substitute and an augmentation for the original ShapeNet-Car dataset in different experimental setups. The original benchmark dataset consists of 1,825 ShapeNet-Car point cloud part segmentations. Out of these, 500 point clouds are designated as the test set, and the remaining 1,325 serve as the training set.

Real-world Human Face Point Cloud Segmentation. We further demonstrate qualitatively that models trained on point clouds generated by our method exhibit effective generalization capabilities when applied to real-world datasets. We trained a PointNet segmentation model on a dataset comprising 1,200 point clouds, synthesized by our proposed method. For evaluation purposes, we employed the Nersemble dataset [32], extracting point clouds from 2D multi-view

images as our test dataset. We show segmentation results for one sample in Fig. 5, showcasing the model's ability to achieve plausible segmentation on real-world human face, particularly in critical regions such as the neck, nose bridge, eyes, cheek, and forehead. Nonetheless, the model encounters challenges in correctly classifying certain areas, notably the ears, which are misidentified due to the model's limitations in distinguishing them from facial regions.
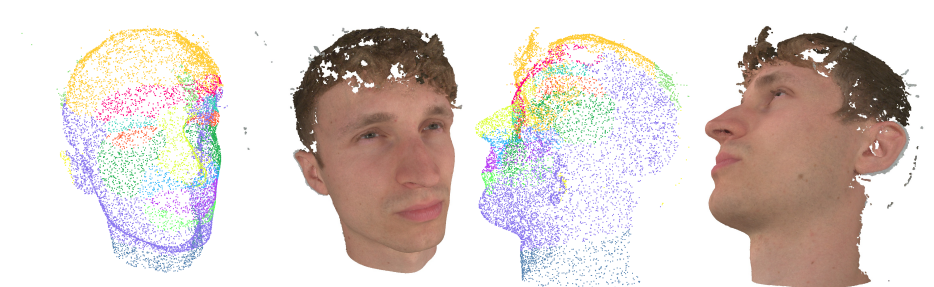

Fig. 5: Visualization of real-world human face point cloud segmentations.

Table 4: Ablation study comparing the impact of different settings on the dataset, focusing on mIoU and accuracy metrics.

Settings	mIoU	Accuracy
w/o Multiscale Feature	0.4014	0.7067
w/ Multiscale Feature	**0.4796**	**0.7884**
w/o Density Prior	0.4728	0.7813
w/ Density Prior	**0.4796**	**0.7884**
w/o Density Prior (Video)	0.6899	0.9188
w/ Density Prior (Video)	**0.6913**	**0.9268**

5.3   Ablation Study

In this section, we evaluate various aspects of our methodology. We ablate the experiments on AFHQ-Cat dataset. The testset is same as the testset used in Sec. 5.1. We employ GAN inversion [78] to initially optimize the latent code and pose of an input RGB test image, subsequently generating its corresponding semantic segmentation. We begin by examining the impact of the tri-plane architecture's size, as shown in Tab. 4. Enhancing the original tri-plane architecture from EG3D with multiscale features extracted from the generator's backbone

Table 5: Ablation study on the effect of training sample size on mIoU and accuracy metrics for individual images and video sequences.

Training Samples	AFHQ-Cat Individual		AFHQ-Cat Video	
	mIoU	Acc.	mIoU	Acc.
30 images	0.4394	0.7716	0.6138	0.8892
45 images	0.4588	0.7778	0.6752	0.9148
90 images	**0.4795**	**0.7884**	**0.6913**	**0.9268**

Fig. 6: 3D RGB Inversion. When presented with an arbitrarily posed input RGB image, our model concurrently optimizes the latent code **z** and pose code to develop a 3D representation. It effectively functions as a segmentation model, capable of rendering segmentations from various viewpoints for the given input image.

leads to a significant improvement in performance. Moreover, Tab. 4 shows incorporating a density prior from the pretrained RGB decoder into the semantic branch is also beneficial. Further, we investigate how the number of training samples affects performance in Tab. 5. Our findings suggest a moderate improvement when increasing the sample size from 30 to 90 images.

5.4 Applications

We explore a series of applications with our approach, including 3D inversion and 3D-aware editing.

3D RGB Inversion. DatasetNeRF functions effectively as a segmentation model. When given a arbitrary posed RGB image, GAN inversion techniques [78] are employed to jointly optimize the input latent code and pose parameters. The optimized latent code uncovers the underlying 3D structure, thereby allowing for precise rendering of semantic segmentation from multiple viewpoints. The optimization is supervised by MSE loss and Adam [31] optimizer is used. The inversion result is showed in Fig. 6.

3D Segmentation Inversion. Pix2pix3D [14] introduces a conditional GAN framework to infer a 3D representation from an input semantic mask. While effective, this approach requires extensive training annotations and significant computational time. DatasetNeRF offers an alternative for accomplishing the similar task. Utilizing an arbitrarily posed semantic mask, our model conducts GAN inversion through its semantic branch. In this process, we jointly optimize the input latent code **z** and the pose, employing cross-entropy loss and gradient

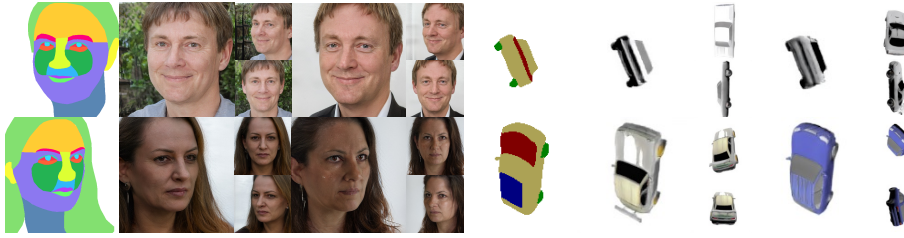**Fig. 7: 3D Segmentation Inversion.** Given an arbitrary posed input semantic mask, we jointly optimize the latent code **z** and pose code to construct a 3D representation. The inherent 2D-to-3D ambiguity in this process results in a significant diversity in the 3D reconstructions. This optimized representation allows rendering from various viewpoints.
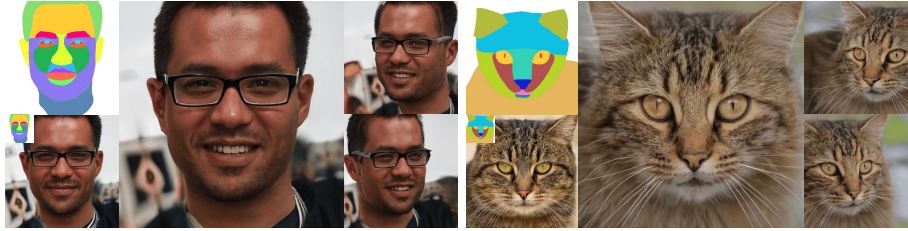

Fig. 8: Semantic Editing Results. Our 3D editing system enables users to modify input label maps and subsequently acquire the corresponding updated 3D representation. We can render the updated 3D representation from different views.

descent as our optimization strategies. The Adam optimizer [31] is employed in this process. The results of this process are illustrated in Fig. 7.

3D-aware Semantic Editing. Our 3D editing system enables users to modify input label maps and subsequently acquire the corresponding updated 3D representation. To accomplish this task, our system focuses on updating the semantic mask output to align with the edited mask while preserving the object's texture through GAN inversion. Initially, GAN inversion is employed to determine the initial latent code **z** from a given forward-oriented input image, which serves as the starting point for subsequent optimization, enhancing performance. Subsequently, this latent code is refined through GAN inversion to yield the optimized updated representation. We define the region of interest r as a binary mask which includes the union region of the label region before and after the edit. We define the loss function $\mathcal{L}(\mathbf{z}; r)$ to quantify the quality of an edit based on the latent code **z** and the region of interest r. It is given by:

$$\mathcal{L}(\mathbf{z}; r) = \lambda_1 \cdot \mathcal{L}_{\text{label}}(G_{\text{semantic}}(\mathbf{z}); M_{\text{edit}})$$
$$+ \lambda_2 \cdot \mathcal{L}_{\text{rgb}}(\overline{r} \odot G_{\text{rgb}}(\mathbf{z}); \overline{r} \odot I_{\text{rgb}})$$
$$+ \lambda_3 \cdot \mathcal{L}_{\text{vgg}}(\overline{r} \odot G_{\text{rgb}}(\mathbf{z}); \overline{r} \odot I_{\text{rgb}}),$$

where:

- $G_{\mathrm{semantic}}(\mathbf{z})$ is the rendered semantic mask from $\mathbf{z}$ with the semantic branch $G_{semantic}$.
- $G_{\mathrm{rgb}}(\mathbf{z})$ is the rendered RGB image with the RGB branch.
- M_{edit} is the edited semantic mask.
- $\mathcal{L}_{\mathrm{label}}$ is the cross-entropy loss for semantic consistency.
- $\bar{r}$ is the complement of the region r.
- $\odot$ is the element-wise product.
- I_{rgb} is the original RGB image.
- $\mathcal{L}_{\mathrm{rgb}}$ measures the RGB prediction's mean squared error.
- $\mathcal{L}_{\mathrm{vgg}}$ is the perceptual loss calculated using a VGG-based network.
- $\lambda_1, \lambda_2, \lambda_3$ balance the loss components.

When editing on the FFHQ dataset, an additional identity loss [56] is incorporated, which calculates the cosine similarity between the extracted features of both the input and edited faces. Fig. 8 shows the edited results.

6 Limitations and Future Work

One limitation of our approach is its reliance on the availability and suitability of pre-annotated data, which restricts its application to more general contexts, such as indoor scene segmentation. In addition, our current approach utilizes the generators of 3D GANs, such as EG3D [8] and GNARF [6], as the backbones of our model. While these 3D GANs adeptly handle single-category data distributions, we aim in our future work to expand this approach to diffusion models, taking advantage of their broader generative diversity. Moreover, although reshaping the concatenated feature into a tri-plane structure has been shown to significantly enhance segmentation quality, it presents challenges in terms of memory efficiency and computational demands. An intriguing avenue for future research lies in identifying the most representative semantic features within the generator backbone, thereby optimizing memory usage and reducing computational load.

7 Conclusions

We present an efficient and powerful approach to developing a 3D-aware data factory, requiring only a minimal set of human annotations for training. Once trained, the model is capable of generating multi-view consistent annotations and point cloud part segmentations from a 3D representation by sampling in the latent space. Our approach is versatile, compatible with both articulated and non-articulated generative radiance field models, making it applicable for a range of tasks such as consistent segmentation of human body poses. This method facilitates advanced tasks like 3D-aware semantic editing, and 3D inversions including segmentation and RGB inversions. The capability of our model to efficiently produce a wide range of 3D-aware data from a limited set of 2D labels is not only crucial for training data-intensive models but also opens up new possibilities in various 2D and 3D application domains.

Acknowledgements

References

1. Alaluf, Y., Tov, O., Mokady, R., Gal, R., Bermano, A.: Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In: Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition. pp. 18511–18521 (2022)
2. Atzmon, M., Lipman, Y.: Sal: Sign agnostic learning of shapes from raw data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2565–2574 (2020)
3. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models (2021)
4. Bautista, M.A., Guo, P., Abnar, S., Talbott, W., Toshev, A., Chen, Z., Dinh, L., Zhai, S., Goh, H., Ulbricht, D., et al.: Gaudi: A neural architect for immersive 3d scene generation. Advances in Neural Information Processing Systems **35**, 25102–25116 (2022)
5. Bergman, A., Kellnhofer, P., Yifan, W., Chan, E., Lindell, D., Wetzstein, G.: Generative neural articulated radiance fields. Advances in Neural Information Processing Systems **35**, 19900–19916 (2022)
6. Bergman, A.W., Kellnhofer, P., Yifan, W., Chan, E.R., Lindell, D.B., Wetzstein, G.: Generative neural articulated radiance fields (2023)
7. Bolles, R.C., Baker, H.H., Marimont, D.H.: Epipolar-plane image analysis: An approach to determining structure from motion. International Journal of Computer Vision **1**, 7–55 (1987)
8. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3d generative adversarial networks (2022)
9. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis (2021)
10. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository (2015)
11. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision. pp. 333–350. Springer (2022)
12. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs (2017)
13. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains (2020)
14. Deng, K., Yang, G., Ramanan, D., Zhu, J.Y.: 3d-aware conditional image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4434–4445 (2023)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

16. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
17. Fei, Z., Fan, M., Zhu, L., Huang, J., Wei, X., Wei, X.: Masked auto-encoders meet generative adversarial networks and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24449–24459 (2023)
18. Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: 2017 International Conference on 3D Vision (3DV). pp. 402–411. IEEE (2017)
19. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: High-fidelity neural rendering at 200fps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14346–14355 (2021)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
21. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. arXiv preprint arXiv:2002.10099 (2020)
22. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. arXiv preprint arXiv:2110.08985 (2021)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
24. Hedman, P., Srinivasan, P.P., Mildenhall, B., Barron, J.T., Debevec, P.: Baking neural radiance fields for real-time view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5875–5884 (2021)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
26. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021)
27. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)
28. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
29. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
30. Kellnhofer, P., Jebe, L.C., Jones, A., Spicer, R., Pulli, K., Wetzstein, G.: Neural lumigraph rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4287–4297 (2021)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
32. Kirschstein, T., Qian, S., Giebenhain, S., Walter, T., Nießner, M.: Nersemble: Multi-view radiance field reconstruction of human heads. ACM Transactions on Graphics **42**(4), 1–14 (Jul 2023). `https://doi.org/10.1145/3592455`, `http://dx.doi.org/10.1145/3592455`
33. Kohli, A.P.S., Sitzmann, V., Wetzstein, G.: Semantic implicit neural scene representations with semi-supervised training. In: 2020 International Conference on 3D Vision (3DV). pp. 423–433. IEEE (2020)
34. Li, D., Ling, H., Kim, S.W., Kreis, K., Barriuso, A., Fidler, S., Torralba, A.: Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations (2022)
35. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++ (2021)

36. Lin, K.E., Lin, Y.C., Lai, W.S., Lin, T.Y., Shih, Y.C., Ramamoorthi, R.: Vision transformer for nerf-based view synthesis from a single input image. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 806–815 (2023)

37. Lindell, D.B., Martel, J.N., Wetzstein, G.: Autoint: Automatic integration for fast neural volume rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14556–14565 (2021)

38. Ling, H., Kreis, K., Li, D., Kim, S.W., Torralba, A., Fidler, S.: Editgan: High-precision semantic image editing. Advances in Neural Information Processing Systems **34**, 16331–16345 (2021)

39. Liu, L., Gu, J., Zaw Lin, K., Chua, T.S., Theobalt, C.: Neural sparse voxel fields. Advances in Neural Information Processing Systems **33**, 15651–15663 (2020)

40. Liu, S., Zhang, Y., Peng, S., Shi, B., Pollefeys, M., Cui, Z.: Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2019–2028 (2020)

41. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblur-nerf: Neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12870 (2022)

42. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7210–7219 (2021)

43. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.: Implicit surface representations as layers in neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4743–4752 (2019)

44. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

45. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)

46. Neff, T., Stadlbauer, P., Parger, M., Kurz, A., Mueller, J.H., Chaitanya, C.R.A., Kaplanyan, A., Steinberger, M.: Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In: Computer Graphics Forum. vol. 40, pp. 45–59. Wiley Online Library (2021)

47. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation. arXiv preprint arXiv:2309.14303 (2023)

48. Nguyen-Phuoc, T.H., Richardt, C., Mai, L., Yang, Y., Mitra, N.: Blockgan: Learning 3d object-aware scene representations from unlabelled images. Advances in neural information processing systems **33**, 6767–6778 (2020)

49. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)

50. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024)

51. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation (2022)
52. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13503–13513 (2022)
53. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
54. Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A.: Virtualhome: Simulating household activities via programs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8494–8502 (2018)
55. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation (2017)
56. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation (2021)
57. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 102–118. Springer (2016)
58. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
59. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems **33**, 20154–20166 (2020)
60. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9243–9252 (2020)
61. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in neural information processing systems **33**, 7462–7473 (2020)
62. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019)
63. Srinivasan, P.P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., Barron, J.T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7495–7504 (2021)
64. Sucar, E., Liu, S., Ortiz, J., Davison, A.J.: imap: Implicit mapping and positioning in real-time. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6229–6238 (2021)
65. Sun, J., Wang, X., Shi, Y., Wang, L., Wang, J., Liu, Y.: Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis (2022)
66. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)

67. Vora, S., Radwan, N., Greff, K., Meyer, H., Genova, K., Sajjadi, M.S., Pot, E., Tagliasacchi, A., Duckworth, D.: Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. arXiv preprint arXiv:2111.13260 (2021)
68. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models (2023)
69. Xiang, J., Yang, J., Deng, Y., Tong, X.: Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds (2023)
70. Yang, H., Hong, L., Li, A., Hu, T., Li, Z., Lee, G.H., Wang, L.: Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16508–16517 (2023)
71. Yang, Z., Zhan, F., Liu, K., Xu, M., Lu, S.: Ai-generated images as data source: The dawn of synthetic era. arXiv preprint arXiv:2310.01830 (2023)
72. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3d shape collections. SIGGRAPH Asia (2016)
73. Zhan, F., Liu, L., Kortylewski, A., Theobalt, C.: General neural gauge fields. In: The Eleventh International Conference on Learning Representations (2023)
74. Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S., Liu, L., Kortylewski, A., Theobalt, C., Xing, E.: Multimodal image synthesis and editing: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
75. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)
76. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort (2021)
77. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15838–15847 (2021)
78. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold (2018)
79. Zhu, J.Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., Freeman, B.: Visual object networks: Image generation with disentangled 3d representations. Advances in neural information processing systems **31** (2018)
80. Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M.R., Pollefeys, M.: Nice-slam: Neural implicit scalable encoding for slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12786–12796 (2022)

Supplementary Material

8   Self-annotated Fine-grained Dataset

We show more in detail about the fine-grained annotation dataset of AFHQ-Cat [13] and AIST++ dataset [35] in Figure S9.
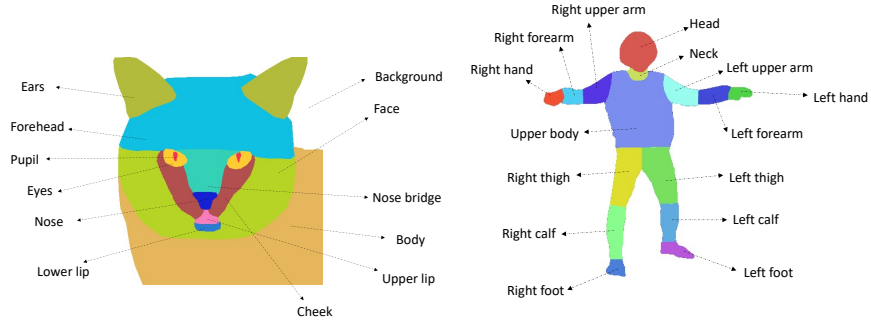

Fig. S9: Detailed label illustration for different datasets.

9   Analysis and Evaluation Metrics for 3D-Consistency

We show spatiotemporal line textures [7] of semantic masks from different poses in Fig. S10.

(a)   (b)

Fig. S10: Spatiotemporal line RGB and semantic textures. Zoom in to better visualize the green line segments in the RGB images.

10   Real-world Human Face Scan Segmentation Results

We present additional real-world human face scan segmentation results. Figure S11a displays a sample selected from the Nersemble dataset [32], along with its segmentation outcomes. Conversely, Figure S11b shows a sample obtained from a self-scanned human face using a smartphone. Both examples demonstrate segmentation results that are both decent and reasonable.

(a) Scan segmentation result for the sample from Nersemble dataset [32].

(b) Scan segmentation result for the self-created scan.

Fig. S11: Real-world human face scan segmentation results for samples from Nersemble dataset [32] and self-made scan.

11   ShapeNet-Car Point Cloud Segmentation Performance Saturation

ShapeNet-Car [10] Point Cloud Segmentation Performance Saturation: Table Tab. 3 in the main paper shows minimal performance gains due to a saturation in PointNet's capabilities, observed when training with approximately 1000 ShapeNet-Car samples (Fig. S12).

12   Segmentation with Foundation Features

We conduct the suggested experiment and use pretrained DINOv2 features [50] to train a segmentation model with our 90 annotated images. For testing, we feed the model frames from our test video sequences. As shown in Fig. S13, our method's segmentation (left) is less noisy and more accurate than the DINOv2-based segmentation (right).
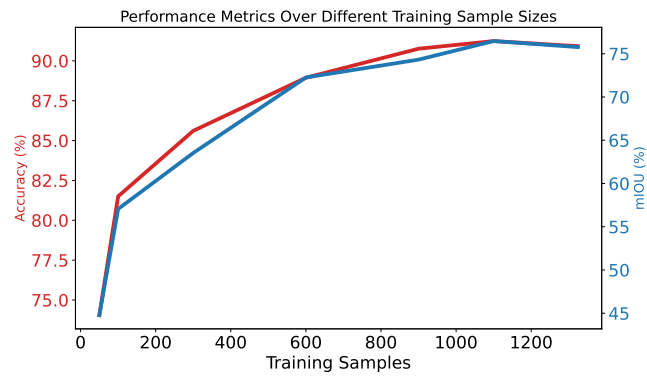
Fig. S12: Experiments showing PointNet segmentation performance saturation. The red curve represents accuracy and blue curve represents mIoU.

Fig. S13: The comparisons with our result (left) with DINOv2 based method (right).