Egocentric Scene Interpretation

Cristhian Forigua
Universidad de los Andes

Daniel Chacón
Universidad de los Andes

Gabriel Pérez
Universidad de los Andes

Jordi Pont-Tuset
Google Research

Kevis-Kokitsi Maninis
Google Research

Pablo Arbeláez
Universidad de los Andes

Abstract

In this paper, we present Egocentric Scene Interpretation, an integrated framework that combines depth and semantic maps for 3D scene geometry, semantics, and appearance reconstruction. Furthermore, we introduce a new metric, called the Egocentric Scene Interpretation (ESI) Score, to evaluate the alignment between geometry and semantics for 3D scene understanding. In addition to the experimental framework, we propose EgoScene, a diffusion-based approach that leverages temporal consistency in feature extraction for egocentric downstream tasks. We compare EgoScene against transformer-based methods. Our method outperforms transformer-based methods by 26% in the ESI score, establishing the state-of-the-art for Egocentric Scene Interpretation.

1. Introduction

Understanding the intrinsic properties of 3D scenes from monocular images is a foundational challenge of computer vision. The pioneering formulation by Barrow and Tenenbaum [1] aimed at extracting scene characteristics such as depth, surface orientation, reflectance, and incident illumination from intensity images. Hoiem *et al.* [18] proposed to decompose 3D scenes into oriented surfaces, occlusion boundaries, viewpoints, and objects. More recently, Malik *et al.* [32] defined visual understanding as the joint solution to the problems of recognition, 3D reconstruction, and perceptual re-organization. Today, the advent of augmented reality (AR), powered by the development of groundbreaking spatial computing capture devices, provides an unprecedented opportunity to revisit this classical computer vision problem from an egocentric perspective. For instance, Liu *et al.* [26] explored egocentric activity recognition and localization within a 3D map, while Mai *et al.* [31] studied 3D object localization in egocentric videos using visual queries. This growing body of work highlights a renewed interest and significant advancements in understanding and interpreting scenes from an egocentric viewpoint.

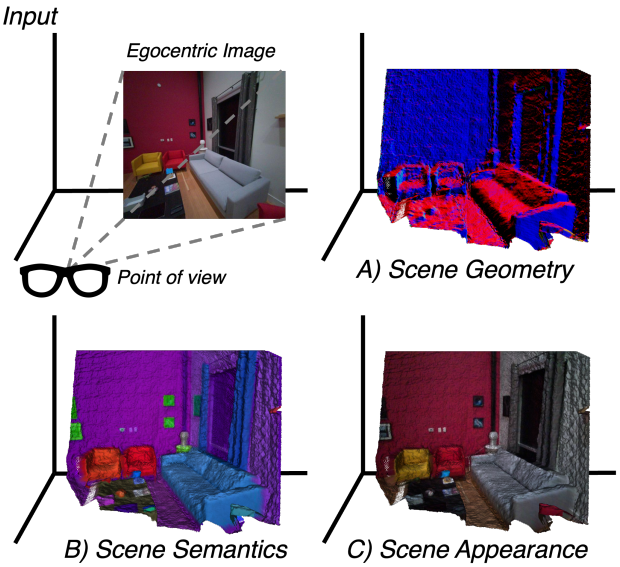

Figure 1. **Egocentric Scene Interpretation.** Starting with an egocentric image and the camera pose (Input), our framework aims to infer the 3D scene's geometry *(A)*, semantic composition *(B)*, and appearance *(C)*, thereby providing a comprehensive interpretation of the surrounding space. We show 3D reconstructions obtained by EgoScene.

This paper presents an integrated framework for egocentric scene interpretation, combining metric depth estimation and semantic segmentation to form a nuanced understanding of 3D spaces from a first-person perspective. Our goal is not only to predict the geometry of the scene, but also to understand the semantics and appearance of all visible entities, translating RGB egocentric images into a detailed and actionable 3D interpretation of the surrounding environment, as shown in Figure 1. We build on recent advances

on spatial computing devices, such as Meta's Aria glasses [42], that enrich an egocentric video feed with precise synchronization of a user's 3D position and orientation. We leverage the annotations of the Aria Digital Twin (ADT) dataset [36], which possesses the unique characteristic of pairing real first-person footage with per-frame semantic, depth, localization, and orientation data. Thus, we create more than 690.000 egocentric ground-truth reconstructions that are densely labeled in space and time into 118 semantic categories.

We introduce the Egocentric Scene Interpretation (ESI) score as a novel metric for assessing the quality of 3D scene comprehension from a first-person viewpoint. ESI extends the principle of the Jaccard Index, traditionally applied to 2D segmentation, into 3D spatial understanding, focusing on the alignment between geometry reconstruction and semantic accuracy.

With this metric in mind, we benchmark leading transformer-based models against our newly developed diffusion-based approach, which leverages language pretraining and temporal consistency across video frames. The results highlight our performance, especially in semantic segmentation, where it consistently surpasses the results of transformer-based frameworks. This result indicates the significant role that the fusion of visual and linguistic data plays in interpreting complex scenes. Enhanced by this integration, our method holds the potential to transform augmented reality experiences, providing users with a deeper, more nuanced interaction with their environment.

Overall, our contributions are as follows:

- We introduce Egocentric Scene Interpretation, a new formulation that integrates depth estimation and semantic segmentation for 3D scene understanding from an egocentric perspective.
- We establish an experimental framework to study this problem. We introduce the Egocentric Scene Interpretation (ESI) Score metric that measures the alignment between 3D surfaces and semantic segments.
- Along with transformer-based baselines, we introduce a diffusion-based model that incorporates temporal consistency and language-vision alignment for feature extraction for downstream egocentric visual tasks.

We will publicly release our source code, dataset, and pretrained models to boost research in egocentric scene interpretation further.

2. Related Work

3D egocentric scene understanding benchmarks and methods Augmented Reality's (AR) ascent has amplified the demand for refined 3D scene interpretation from an egocentric perspective, prompting the development of targeted datasets. EpicKitchens [8], Ego4D [16], and HOI4D [27] have broadened the range of daily-life scenarios for analysis, while EgoPAT3D [25] and the Aria Digital Twin (ADT) dataset [36] provide dense annotations tailored for egocentric interpretation. Innovations in 3D comprehension have been propelled by methods from Do *et al.* [9], Nagarajan *et al.* [34], Liu *et al.* [26], and others, extending from spatial rectification to probabilistic mapping of actions in 3D. However, the fusion of language and vision for in-depth 3D egocentric scene interpretation is a frontier our research seeks to explore. We fill this gap by integrating diffusion models with linguistic data, offering a comprehensive approach to egocentric scene analysis that could reshape interaction in AR environments.

Diffusion models and visual perception tasks Denoising diffusion probabilistic models have revolutionized image synthesis with their ability to generate high-quality images from a reverse noising process [17]. Recent endeavors, such as VPD [50] and TADP [22], have adapted these models for visual tasks by encoding images into latent spaces and using cross-attention to integrate text [39]. Nonetheless, these models traditionally lack a mechanism for incorporating temporal information and have yet to be specifically designed or tested for egocentric data interpretation. Our approach diverges by harnessing the temporal continuity in video sequences, utilizing frame-to-frame information to enhance the fidelity of egocentric scene interpretation.

Monocular depth estimation Monocular depth estimation is an ill-posed problem because a single 2D image does not provide sufficient information to determine objects' depth uniquely. The inherent difficulty arises from depth-scale ambiguity, which remains unsolvable with just a single input image. Supervised methods [10, 13, 20, 23, 28, 38] assume a perfect matching between visual inputs and depth ground truth maps. However, collecting diverse real-world datasets for depth estimation is challenging. To overcome this limitation, some works address depth estimation in a self-supervised manner [14, 15, 48, 52]. Nevertheless, all these methods are designed for either autonomous driving [5, 33] or third-person perspective [35] pipelines, which left unexplored the potential of egocentric depth estimation and its potential for 3D scene interpretation and human-object interactions. Also, in contrast, monocular depth estimation is an intermediate task of our complete framework.

Semantic segmentation Semantic segmentation helps to recognize each semantic category in an image spatially and densely, which is crucial for scene interpretation. The first deep learning methods were CNN-based architectures that estimate semantic segmentation maps through convolution operations [29, 40]. More recently, transformer-based

methods, such as Segformer [45] and DPT [38], leverage the self-attention mechanism for semantic segmentation [3, 4, 24, 49]. Cheng *et al.* proposed a masked-attention mechanism to transformers [4] to improve MaskFormer [3] performance. In addition, Li *et al.* proposed Mask DINO [24], a unified object detection and segmentation framework that extends from DINO [47] by adding a mask prediction branch. Similarly to depth estimation, these methods ignore the egocentric perspective and its potential for human-centric scene interpretation. Similarly, akin to monocular depth estimation, semantic segmentation is an intermediate task within our comprehensive framework.

Static 3D reconstruction benchmarks Advances in 3D object reconstruction have traditionally been driven by comprehensive datasets, such as ShapeNet [2] and ModelNet [44], which provide abundant 3D models for object reconstruction. In the vicinity of these datasets, the DTU Robot Image dataset [19] offers real-world high-resolution scans for evaluating multi-view stereo algorithms. Meanwhile, the scene analogous task has benchmarks such as ScanNet [7], SUN RGB-D [43], and NYUv2 [35], enriching the field with detailed annotations of indoor scenes. However, our research deviates from these static benchmarks by focusing on dynamic video scenes where depth maps and semantic segmentation can extrapolate comprehensive scene understanding. Instead of competing with static reconstruction tasks that utilize 3D meshes or point clouds as ground truth, our approach, tailored to the complexities of egocentric video data, prioritizes the accurate prediction of dense depth maps and semantic labels per frame. We can indirectly infer 3D scene structure, appearance, and semantics through this. Thereby, the concept of 3D reconstruction is adapted by sampling a single point in time of a fluid 4D space.

3. Egocentric Scene Interpretation Benchmark

This section formally introduces Egocentric Scene Interpretation and its experimental framework, our novel framework for 4D egocentric scene understanding.

3.1. Task Definition

Given an input RGB egocentric image I and the user's point of view v, the goal is to perform a partial 3D reconstruction of a scene with geometric, semantic, and appearance awareness. Thus, the objective is to develop a function F such that,

$$F\left(I, v\right) \rightarrow \left(G, S, A\right) \tag{1}$$

where G, S, A are the reconstructed 3D meshes of the same visible scene with geometrical, semantic, and appearance information, respectively.

We approach this task using an intermediate function H_1 that acts on the image I to obtain two corresponding intermediate outputs: a depth metric map D, encoding the distance of each pixel from the camera, and a semantic segmentation mask Y, where each pixel value corresponds to its semantic label. Hence,

$$H_1(I) \rightarrow (D, Y) \tag{2}$$

with $I \in \mathbb{R}^{H\times W\times 3}$, $D \in \mathbb{R}^{H\times W}$ and $Y \in \mathbb{R}^{H\times W}$. By integrating D and Y from the user's point of view v, we can ultimately reconstruct a 3D partial representation from an egocentric perspective via a second function H_2:

$$H_2(H_1(I), v) \rightarrow (G, S, A) \tag{3}$$

In order to un-project the metric depth maps in the 3D space and approximate the G, S, A meshes, reconstruction algorithms such as [6, 11, 51] use the following projection given an RGB-D image, camera pose, and intrinsics:

$$\mathbf{p}_{3D} = \mathbf{K}^{-1} \cdot \mathbf{D}(u, v) \cdot \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \tag{4}$$

where $\mathbf{p}_{3D}$ is the point in 3D world coordinates, $\mathbf{K}$ is the intrinsic camera matrix, $\mathbf{D}(u, v)$ is the depth value at the pixel (u, v). The whole expression is finally multiplied by the pose of the camera to obtain the point in 3D space.

3.2. Egocentric Scene Interpretation (ESI) Score

We propose a new metric called Egocentric Scene Interpretation (ESI) Score to measure the alignment between 3D surfaces and semantic segments for Egocentric Scene Interpretation. ESI aims to measure how accurate is the mapping of the function H_1 defined in Eq. 2 in a joint manner for both outputs.

ESI integrates segmentation quality with 3D geometry reconstruction as follows: given a predicted depth value $\hat{d}_{i,j}$ and its corresponding ground-truth $d_{i,j}$ for a pixel $p_{i,j}$, we compute the absolute difference as:

$$\delta_{i,j} = |\hat{d}_{i,j} - d_{i,j}| \tag{5}$$

Then, we compute performance for tolerances up to one meter, $\epsilon \in [0, 1]$. To determine if the pixel $p_{i,j}$ can be considered a true positive. Thus, we define for a given class y and pixel $p_{i,j}$,

$$p_{i,j} \in TP^* \leftrightarrow (\delta_{i,j} \leq \epsilon \wedge y_{i,j} = \hat{y}_{i,j} = y) \tag{6}$$

where $y_{i,j}$ and $\hat{y}_{i,j}$ are the ground-truth and predicted semantic class of the pixel (i, j), respectively. We also define the false positives for a given class y as follows,

$$\begin{aligned} p_{i,j} \in FP^* \leftrightarrow &(\delta_{i,j} > \epsilon \wedge y_{i,j} = \hat{y}_{i,j} = y) \\ &\vee (y_{i,j} \neq y \wedge \hat{y}_{i,j} = y) \end{aligned} \tag{7}$$

Moreover, following semantic segmentation standards, we compute the Intersection over Union (IoU) to measure the segmentation quality as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives pixels, respectively.

Therefore, the $mIoU^*$ is:

$$mIoU^* = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k^*}{TP_k^* + FP_k^* + FN_k} \tag{9}$$

with N being the number of semantic classes. Finally, following this formulation, the ESI is defined as:

$$ESI = \int_0^1 mIoU^* \, d\epsilon \tag{10}$$

This formulation allows us to evaluate the quality of each semantic class for different threshold values, which provides a complete understanding of the impact of the 3D geometry reconstruction on the scene semantics.

3.3. Dataset

We build on the Aria Digital Twin (ADT) Dataset [36], a real-world egocentric dataset released by Meta Reality Labs using Aria glasses [42]. ADT aims to supply a benchmark dataset that facilitates the assessment and progression of egocentric research. ADT includes wide-ranging ground truth information for 2D depth maps and 2D instance segmentations. We obtained 118 semantic categories from these segmentations in more than 20.2 million masks.

ADT contains 200 real-world video sequences recorded by Aria wearers at 30Hz in 2 indoor scenes, an apartment and an office, featuring 398 object instances of stationary and dynamic elements. In these sequences, the Aria wearers can perform different activities such as room decoration, meal preparation, work, object examination, room cleaning, partying, and dining table cleaning. Some of these activities can include two people in the same scene simultaneously, which causes the number of video captures to be greater than the number of sequences. Considering these human interactions, ADT provides ground truth depth maps and semantic segmentations with and without human skeleton occlusions. Since we are mainly interested in the scene interpretation from an egocentric perspective, we use use the depth and semantic ground truth data that disregard the human occlusions.

In total, we curated 242 captures from ADT. We transformed each frame from a fish-eye camera model to a pinhole model. We partitioned the dataset into training and validation sets, maintaining consistent distributions across scenarios (apartment and office) and activities. The training set comprises 120 captures, while the validation set consists of 122 captures. This way, we ensure that consecutive frames from the same capture are distinctly assigned to the training or validation set.

4. Methods

In this section, we introduce our diffusion-based approach: EgoScene. We also present the task-specific baselines for depth estimation and semantic segmentation, which we compare against.

4.1. EgoScene

Inspired by the capacity to capture language and visual relations of diffusion models [39, 50], we use a pretrained text-to-image diffusion model ϵ_θ as backbone for egocentric downstream tasks. As shown in Figure 2, given a frame x_t and its preceding frame x_{t-1}, we extract their latent representations $z_{0,t}$ and $z_{0,t-1}$ using a visual encoder $\mathcal{E}$ such that,

$$\mathcal{E}(x_t) = z_{0,t} \ , \ \mathcal{E}(x_{t-1}) = z_{0,t-1} \tag{11}$$

Then, we make a forward pass through the last step of the denoising autoencoder, such that we can obtain a set of multi-scale feature maps F and cross-attention maps A from the UNet:

$$\begin{aligned} \epsilon_\theta(z_{0,t}, 0, C_t) &\rightarrow F_t, A_t \\ \epsilon_\theta(z_{0,t-1}, 0, C_t) &\rightarrow F_{t-1}, A_{t-1} \end{aligned} \tag{12}$$

where C_t works as the text conditioner at timestamp t. We use the same text conditioning C for both frames.

Specifically, we extract 4 feature maps from F,

$$F_i \in \mathbb{R}^{H_i, W_i}; H_i = W_i = 2^{i+2}, i = 1, 2, 3, 4 \tag{13}$$

Then, we calculate V by concatenating the feature and attention maps as follows:

$$V = [(F_t \oplus A_t) \oplus (F_{t-1} \oplus A_{t-1})] \tag{14}$$

V is then passed to the task specific head to compute the final prediction $\mathcal{H}(V) = \hat{p}$.

Furthermore, to generate text conditioning C, we use the average text representation of 80 templates by using CLIP text encoder [50] for each class of interest. For depth estimation, we pass conditioning $C_d \in \mathbb{R}^{1\times768}$ representing the scene type, where 768 is the text embedding dimension of CLIP [37]. For semantic segmentation, we pass $C_s \in \mathbb{R}^{N\times768}$ where N is the number of classes in ADT. Since we get an embedding for each semantic class, we can extract cross-attention maps for each. During training, ϵ_θ and $\mathcal{H}$ are optimized.

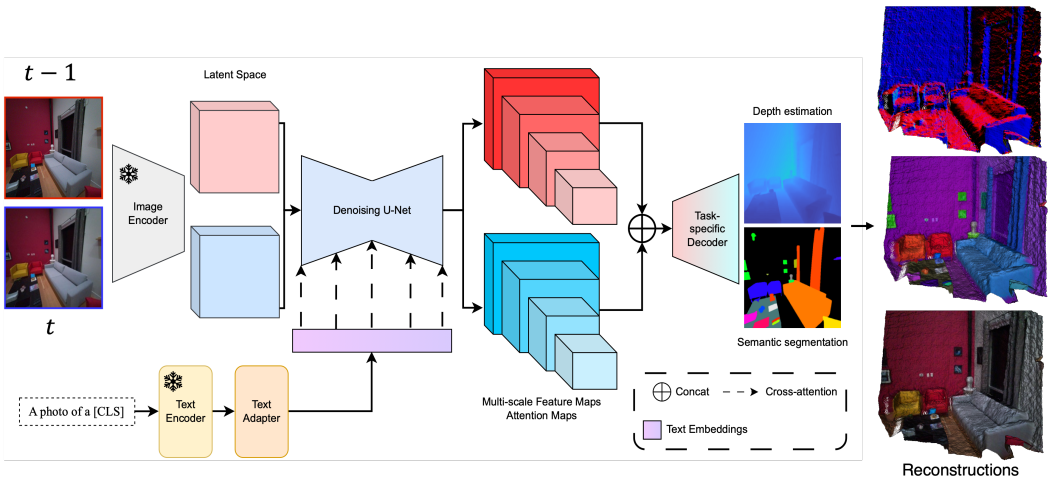

Figure 2. **Overview figure of Egoscene** Our method leverages language-vision alignment of diffusion models to tackle depth estimation and semantic segmentation for 4D Egocentric Scene Interpretation. First, we use a visual encoder to extract latent representations for the current and previous frames t and $t-1$. We extract text embeddings through CLIP text encoder [37] to guide the feature extraction. Then, we pass both multi-scale features as input to the task-specific head. Ultimately, we extract 3D geometry, semantics, and appearance reconstructions from the intermediate depth and semantic maps.

4.2. Task-specific baselines

In our comparative analysis, we combine independent state-of-the-art transformer-based models as baselines for depth estimation and semantic segmentation tasks. For depth estimation, we utilize Swin Transformer V2 [28] and Global-Local Path Networks (GLP) [20]. The Swin Transformer V2, developed by Liu *et al.*, represents an advancement in transformer architectures, scaling up capacity and resolution to effectively handle complex visual data. The GLP model, proposed by Kim *et al.*, integrates global and local contextual information, offering an innovative approach to monocular depth estimation with its vertical cut-depth feature.

For semantic segmentation, we employ SegFormer [45] and DPT (Dense Prediction Transformer) [38]. SegFormer, introduced by Xie *et al.*, emphasizes a simple and efficient design optimized for semantic segmentation tasks. DPT, by Ranftl *et al.*, harnesses the power of vision transformers for dense prediction, demonstrating effectiveness in semantic segmentation tasks.

4.3. Implementation details

Our method, for both depth estimation and semantic segmentation, maintains the VQGAN [12] encoder and the CLIP [37] encoder constant during training. With these two architectures we obtain an embedded representation of the RGB images and the captions, respectively. We first train the task-specific decoder head and the diffusion UNet for depth estimation, followed by a separate independent training phase for semantic segmentation. We also set the learning rate of the backbone to be $\frac{1}{10}$ of the task-specific head learning rate to leverage pre-training. For the text adapter, we use $\gamma = 1e-4$.

4.3.1 Depth Estimation

We use the pretrained weights in NYUv2 [35] dataset provided by [50] as starting point. During training, we employ random cropping of images to a size of 480×480. We fine-tune the model for 25 epochs with a batch size of 7, utilizing the AdamW optimizer [30] with a learning rate set at $5e-4$ and a weight decay of $5e-2$. The learning rate follows a polynomial strategy, starting at $5e-4$ and decreasing with a factor of 0.9 until it reaches a minimum value of $3e-5$. The layer decay is set to $0.9/0.85$, and the DropPath is set at $0.3/0.5$. The depth estimation head upsamples the last features of the backbone using deconvolutions as described in [46]. For training, we use the Scale-Invariant loss [10] to optimize the network's parameters.

As metrics, we report the absolute relative error (REL), root mean squared error (RMSE), and the average $\log 10$ error between the predicted depth and the actual ground truth depth. We also report the threshold accuracy δ_n for $n = 1, 2, 3$ defined as:

$$\delta_n = \frac{\text{\# of pixels where } \max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) < 1.25^n}{\text{Total \# of pixels}} \times 100\%$$

where d_i represents the ground truth depth and $\hat{d}_i$ is the predicted depth.

4.3.2 Semantic Segmentation

We use a semantic FPN [21] head with cross-entropy loss. We train our model for 80000 iterations with a batch size of

	RMSE (cm) ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	REL ↓	log_{10} ↓	# params
SwinV2-L [28]	**11.43**	0.9874	<u>0.9956</u>	**0.9980**	0.0307	0.0123	195.7 M
GLP [20]	11.88	0.9868	0.9953	0.9978	**0.0275**	**0.0109**	61.2 M
EgoScene (Ours)	12.65	**0.9883**	**0.9956**	0.9978	0.0284	0.0110	928 M

Table 1. **Depth estimation on ADT.** Main results for depth estimation on ADT dataset. We compare our method against a transformer-based method.

	mIoU ↑	mAcc ↑	aAcc ↑	# params
Segformer [45]	35.61	48.48	94.36	24.7 M
DPT [38]	43.71	56.8	94.64	109.7 M
EgoScene (Ours)	**73.62**	**83.31**	**97.28**	902 M

Table 2. **Semantic segmentation on ADT.** Main results for semantic segmentation on ADT. We compare our method against a transformer-based method.

7, utilizing a learning rate of $8e-5$. The training process employs the AdamW optimizer [30] with a weight decay of $1e-3$ and incorporates a warm-up phase of 1500 iterations. We implement a polynomial learning rate scheduler with a power of 1.0, ensuring a minimum learning rate of 0.

To evaluate semantic segmentation performance, we use the mIoU, mAcc, representing the average pixel-wise accuracy over all classes, and aAcc, which provides insights into the accuracy across all pixels.

4.3.3 3D Reconstruction from Egocentric Data

To visualize the 3D reconstructions from ADT, we integrate RGB images, trajectory data, depth metric maps, and semantic segmentation masks from the evaluated models into Open3D's RGBD volume integration pipeline. We retrieve the odometry from the Aria glasses, giving us the necessary translation and rotation data in 4x4 homogeneous matrices and ensuring the alignment of our data streams. These streams are then coalesced into a unified 3D structure using Open3D's ScalableTSDFVolume method. This method implements the Dense Scene Reconstruction algorithm [51], chosen for its scalability and robustness, which is ideal for the vast ADT data. To finalize the reconstruction, we perform detailed post-processing, including point gluing, quadric decimation, and removal of small isolated components, duplicate faces, and vertices to ensure that the 3D models accurately and efficiently represent the intricacies of the egocentric scenes.

5. Experiments

This section showcases our framework's effectiveness in egocentric scene interpretation, comparing our diffusion-based methods with transformer models using ADT and assessing performance through the newly proposed Egocentric Scene Interpretation Score alongside detailed 3D visualizations. Also, we show ablations on temporal consistency and vision-language integration in our approach. We subsample the video captures every 20 frames for metric computations during evaluation.

5.1. Depth and Segmentation Results

In Table 1 , we present the results of depth estimation tasks comparing our method, EgoScene, with SwinV2-L [28] and GLP [20]. The results indicate a competitive landscape: SwinV2-L leads with the lowest RMSE, demonstrating minimal deviation from the ground truth depth on average. GLP shows its strength in minimizing the absolute relative error, indicating a reliable depth estimation over the dataset. Our method, EgoScene, shines in threshold accuracy, achieving the highest scores for δ_1 and δ_2, suggesting it consistently estimates depth within tighter bounds for most pixels. This balanced performance across different metrics underscores the capability of each method to handle depth estimation effectively, with each excelling in different aspects of the task.

Table 2 shows the semantic segmentation results of Egoscene, Segformer [45] and DPT [38] on the ADT. Our Egoscene model outperforms competing transformer-based methods. It achieves a remarkable mIoU of 73.62, significantly higher than DPT's 43.71 and Segformer's 35.61. Moreover, our model leads with a mAcc of 83.31, compared to the runner-up DPT's 56.8, and excels in aAcc at 97.28, over DPT's 94.64. All this demonstrates Egoscene's capability to understand and classify complex scene semantics. We also present the number of parameters for each.

This superiority in segmentation, paired with competitive performance in depth estimation, can be further observed in the qualitative results. As depicted in Figure 3, EgoScene's proficiency in depth estimation and semantic segmentation can be observed. Particularly notable is its precision in handling objects at close range. Additionally, the model's ability to reconstruct occluded surfaces is evident, especially in scenarios like the obscured chair base in the fifth frame. This capability, likely honed through

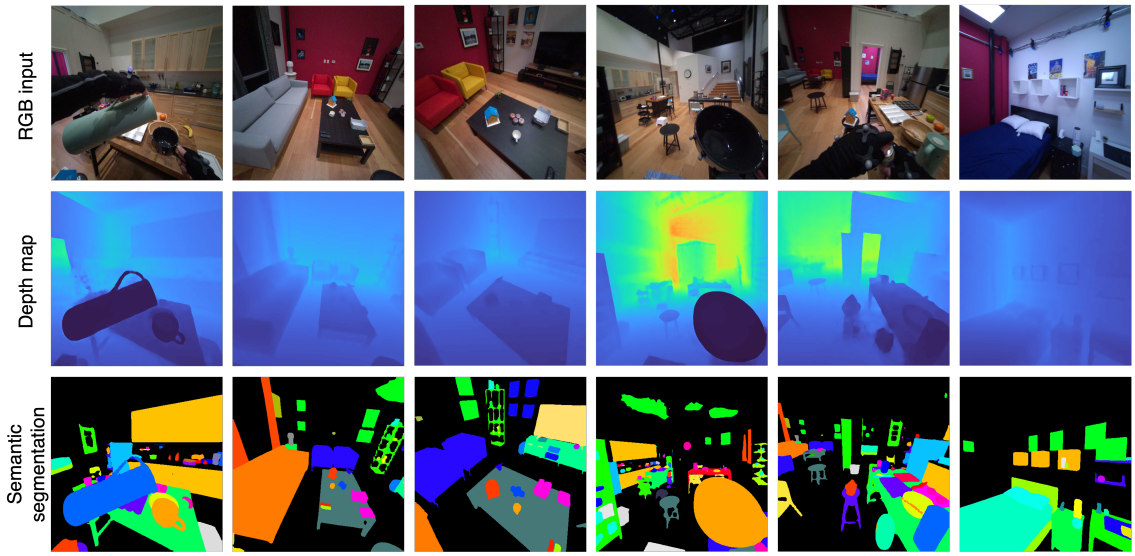

Figure 3. **Qualitative results of EgoScene**. We show the results of depth and semantic segmentation across six different frames with our method. We observe the amodal completion in frames 1 and 5 for semantic segmentation. Our method can correctly hallucinate semantic objects occluded in the egocentric image.

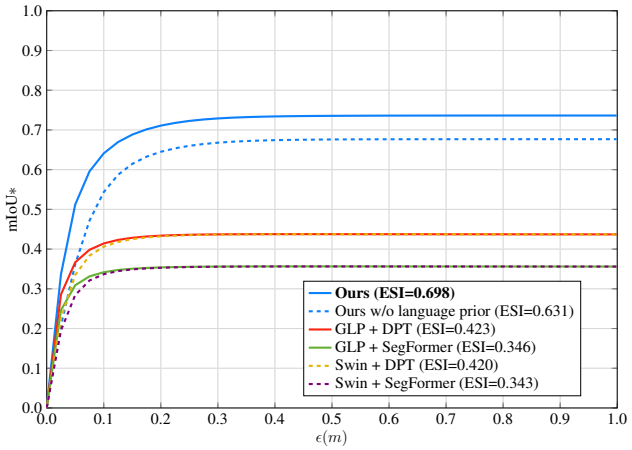

Figure 4. **Main Results.** Comparison of mIoU∗ scores across different methods for egocentric scene interpretation. EgoScene with an ESI score of 0.698 significantly outperforms the transformer-based architectures by 26%.

training on repetitive object sequences, is enhanced by the model's design that integrates temporal information from successive frames, thereby ensuring a coherent scene structure even in the presence of occlusions.

These qualities of our model not only highlight its adeptness in interpreting 3D scenes but also its potential to significantly improve the user experience in AR environments. Its ability to perform amodal completions and maintain depth and semantic accuracy underlines the model's sophistication and utility in dynamic, real-world interactions, promising a more immersive and intuitive interface for AR applications.

5.2. ESI Results

The initial observation from Figure 4 confirms the expected behavior where all methods begin with an ESI near zero, as an ϵ of zero allows only for perfectly predicted depth pixels to be true positives, a stringent and telling sanity check. As ϵ increases, each method's curve trends upwards, reflecting the diminishing strictness of depth precision; this transition highlights the trade-off between depth accuracy and semantic correctness. Seeing the different rates at which this increase occurs, we can extrapolate how better depth prediction capabilities translate to quicker attainment of corresponding peak performance in ESI — a measure of saturation speed. Analogously, each curve's ultimate ESI plateau corresponds with its semantic segmentation performance (IoU), reflecting its max saturation potential.

More notably, Figure 4 also illustrates that our method outperforms even the best alternative approach that employs GLP [20] for depth prediction and DPT [38] for semantic segmentation. Our model achieves a higher ESI of 0.6988, which indicates superior overall performance in the joint task of scene interpretation. This larger ESI reflects a more consistent and accurate alignment between the predicted 3D surfaces and their corresponding semantic classes across various thresholds.

Figure 5 allows us to evaluate our approach against joint transformer methods like GLP [20] and DPT[38] in a 3D reconstructed setting. As is to be expected given the ESI scores, we can contrast in the semantic segmentation how, in the leftmost part, denoted by a pink ring, objects such as the vase or kettle are entirely disregarded by DPT. Also, observing the rightmost side of the figure, denoted by a black

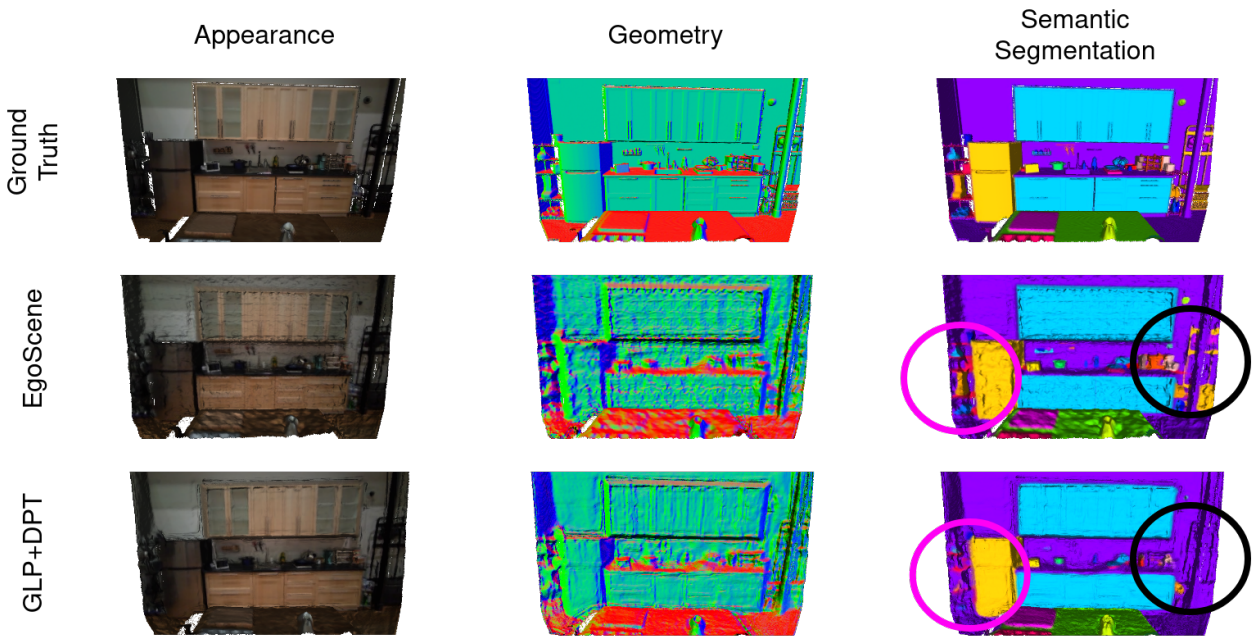

Figure 5. **Qualitative results for 3D scene reconstructions** We compare a scene's appearance, geometry and semantic segmentation. We show the results of EgoScene and GLP + DPT.

	RMSE (cm) ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	REL ↓	log_{10} ↓
w/o temporal	12.92	0.9878	0.9955	0.9977	0.0321	0.0129
w temporal	**12.65**	**0.9883**	**0.9956**	**0.9978**	**0.0284**	**0.011**

Table 3. **Depth estimation with temporal consistency.** Adding temporal consistency to our method leads to better performance across all metrics. The temporal consistency of the previous frame improves by $\sim 2\%$ in the RMSE metric.

ring, fine-grained kitchen utensils are segmented wrong, most of them being assigned the "background" semantic class. Following their corresponding ESI scores, the geometry reconstruction of GLP+DPT appears smoother and with more consistent normals, especially on flat surfaces like the refrigerator and the kitchen table.

Transformer-based methods show promise in depth prediction but fall short in semantic segmentation, significantly affecting their overall ESI score and highlighting the need for a method that equally addresses depth accuracy and semantic clarity.

5.3. Ablation Experiments

First, we study the effect of adding temporal consistency to our method via diffusion-based feature extraction into the task-specific head into the depth estimation task. Moreover, we explore the impact of using the pretrained weights of Stable-Diffusion [39] trained on LAION-5B [41] for the semantic segmentation task.

Temporal consistency. Unlike previous approaches [22, 50], our method exploits temporal consistency for enhanced performance. As detailed in Table 3, incorporating temporal consistency into our model leverages previous and current frame features, yielding significant improvements across all metrics. Specifically, the RMSE saw a reduction of 2%, the Relative Error (REL) decreased by 11.5%, and the logarithmic error (log_{10}) was reduced by 14.7%. These results corroborate that temporal redundancy benefits the model's performance, allowing the task-specific decoder to discern and utilize the temporal relationships within the features set more effectively.

Language Stable Diffusion Prior. We explore the impact of pretrained Stable-Diffusion model initialization on semantic segmentation using the Denoising U-Net. Initializing with Stable-Diffusion weights significantly boosts mIoU to 73.62, surpassing the 67.67 achieved with random weights. Surprisingly, random weights did not lead to a much lower mIoU. Notably, this ablation study focuses on initialization effects and the language prior learned by Stable-Diffusion, maintaining language consistency by using captions of the images in both cases.

6. Conclusions

We present Egocentric Scene Interpretation, a robust framework that integrates depth estimation and semantic segmentation for reconstructing a 3D scene's geometry, semantics, and appearance. We introduce the ESI score as a novel

paradigm to measure the alignment between 3D geometry and semantics for scene understanding. We also propose EgoScene, a diffusion-based model that leverages temporal consistency for depth estimation and semantic segmentation. Our method outperforms previous transformer-based architectures by 26% in the ESI Score. Our work paves the way for innovative applications that require a deep understanding of complex, real-world 3D environments.

References

[1] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2 (3-26):2, 1978. 1

[2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[6] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 3

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[9] Tien Do, Khiem Vuong, and Hyun Soo Park. Egocentric scene understanding via multimodal spatial rectifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2832–2841, 2022. 2

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2, 5

[11] Felix Endres, Jürgen Hess, Nikolas Engelhard, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the rgb-d slam system. In *2012 IEEE international conference on robotics and automation*, pages 1691–1696. IEEE, 2012. 3

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 5

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 2

[14] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016. 2

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[18] Derek Hoiem, Alexei A Efros, and Martial Hebert. Closing the loop in scene interpretation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1

[19] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 3

[20] Doyeon Kim, Woonghyun Ka, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 2, 5, 6, 7

[21] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 5

[22] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimarães, and Pietro Perona. Text-image alignment for diffusion-based perception. *arXiv preprint arXiv:2310.00031*, 2023. 2, 8

[23] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2

[24] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 3

[25] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20971–20980. IEEE, 2022. 2

[26] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In *European Conference on Computer Vision*, pages 621–638. Springer, 2022. 1, 2

[27] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 2

[28] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 2, 5, 6

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[30] I Loshchilov and F Hutter. " decoupled weight decay regularization", 7th international conference on learning representations, iclr. *New Orleans, LA, USA, May*, (6-9):2019, 2019. 5, 6

[31] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 45–57, 2023. 1

[32] Jitendra Malik, Pablo Arbeláez, Joao Carreira, Katerina Fragkiadaki, Ross Girshick, Georgia Gkioxari, Saurabh Gupta, Bharath Hariharan, Abhishek Kar, and Shubham Tulsiani. The three r's of computer vision: Recognition, reconstruction and reorganization. *Pattern Recognition Letters*, 72:4–14, 2016. 1

[33] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[34] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egocentric scene context for human-centric environment understanding from video. *arXiv preprint arXiv:2207.11365*, 2022. 2

[35] Pushmeet Kohli Nathan Silberman Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 3, 5

[36] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 2, 4

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5

[38] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2, 3, 5, 6, 7

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 8

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[41] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 8

[42] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2, 4

[43] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 3

[44] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 3

[45] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3, 5, 6

[46] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked im-

age modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14475–14485, 2023. 5

[47] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[48] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18537–18546, 2023. 2

[49] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 3

[50] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5729–5739, 2023. 2, 4, 5, 8

[51] Qian-Yi Zhou and Vladlen Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics (ToG)*, 32(4):1–8, 2013. 3, 6

[52] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2