SHED Light on Segmentation for Depth Estimation

Anonymous ICCV submission

Paper ID *****

Abstract

Monocular depth estimation is a dense prediction task that infers per-pixel depth from a single image, fundamental to 3D perception and robotics. Although real-world scenes exhibit strong structure, most methods treat it as an independent pixel-wise regression problem, often resulting in structural inconsistencies in depth maps, such as ambiguous object shapes. We propose SHED, a novel encoder-decoder architecture that incorporates segmentation into dense prediction. Inspired by the bidirectional hierarchical reasoning in human perception, SHED improves upon DPT by replacing fixed patch tokens with segment tokens, which are hierarchically pooled in the encoder and unpooled in the decoder to reverse the hierarchy. The model is supervised only at the final output, and the intermediate segment hierarchy emerges naturally without explicit supervision. SHED offers three key advantages over DPT. First, it improves depth boundaries and segment coherence while reducing computational cost. Second, it enables features and segments to better capture global scene layout. Third, it enhances 3D reconstruction and reveals part structures that conventional pixel-wise methods fail to capture.

1. Introduction

Images are 2D projections of the 3D world, where surfaces, regions, and boundaries form a coherent structure. Many vision tasks aim to recover this structure by predicting semantic or geometric values at each pixel, a process known as dense prediction [20]. Among them, monocular depth estimation is one of the most studied, inferring depth from a single RGB image [70]. Despite the inherent structure of real-world scenes, most models, including the Dense Prediction Transformer (DPT) [60], treat the task as independent pixel-wise regression. Although their outputs may appear plausible, they often lack structural consistency, resulting in ambiguous object shapes (Fig. 1, row 1).

This limitation stems from a disconnect between depth estimation and scene organization. Depth encodes geometric structure, while segmentation captures semantically coherent regions. Though serving different purposes, the two are closely related: segment boundaries align with depth discontinuities, and depth gradients with semantic boundaries. This relationship has long been recognized in classical vision literature [48], yet recent models such as Depth Anything [77] and Segment Anything [61] treat them as independent tasks, largely overlooking their connection.

In contrast, the human visual system integrates depth and segmentation through a bidirectional hierarchical process [28], where part-whole segmentation informs depth estimation, and depth in turn guides segmentation. It first infers a global layout by grouping segments from fine to coarse, then refines depth from coarse to fine, adding detail within smaller regions while preserving the overall structure. This organization supports part-whole reasoning and yields depth maps with sharp boundaries and smooth intra-object variations (Fig. 1, row 2).

To realize this idea, we propose a novel architecture called SHED, which performs dense prediction using a bidirectional segment hierarchy. SHED follows the design of DPT [60], an encoder-decoder framework built on the Vision Transformer (ViT) [13], but replaces fixed-size patch tokens with hierarchical segment tokens. These tokens are organized from fine to coarse and learned in an *unsupervised* manner, guided solely by dense prediction objectives.

Our encoder builds on CAST [37], a ViT-based model for hierarchical segmentation in recognition tasks. It replaces patch tokens with superpixel tokens and merges them iteratively based on feature similarity to construct a hierarchy of segment tokens. The decoder inverts this hierarchy to produce dense predictions, leveraging both the segment maps and their features. It unpools finer segments from coarser ones using soft assignments computed in the encoder, and concatenates them with tokens from the corresponding encoder layer. Each segment token is projected into a spatial map by distributing its features over the associated region, producing sharp boundaries and smooth transitions. The resulting features from multiple segment levels are fused with pixel-level features from a convolutional encoder to produce outputs that preserve global layout while capturing fine detail.

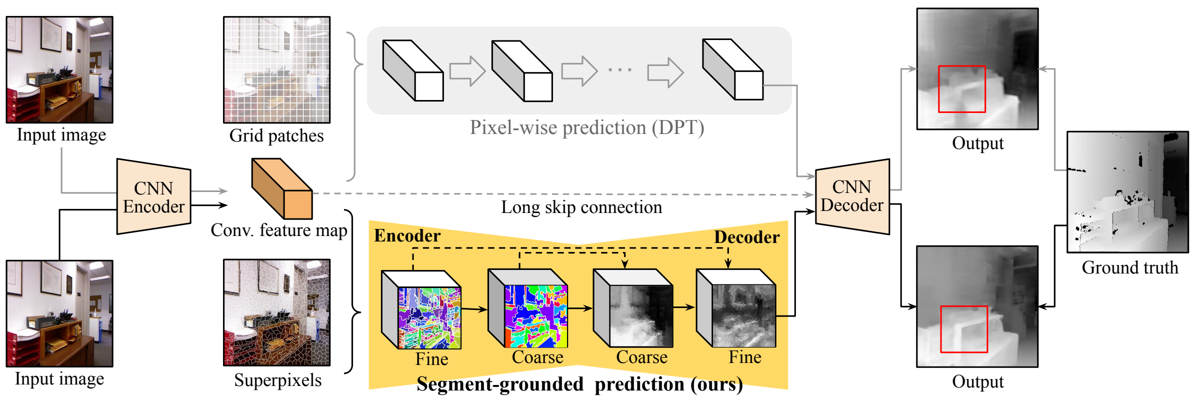


Figure 1. **Segment hierarchy for estimating depth (SHED).** Conventional methods such as DPT [60] perform pixel-wise prediction without considering structure, often resulting in blurry object shapes. SHED addresses this by leveraging a hierarchy of segment tokens to guide prediction. Unlike DPT, which uses fixed grid tokens across all layers, we adapt its ViT [13] blocks into two stages: the encoder pools superpixel tokens into coarser segment tokens, and the decoder progressively refines predictions from coarse to fine segments, producing depth maps with structural coherence.

We highlight the main differences between SHED and CAST [37]. First, while CAST is encoder-only, SHED extends it to an encoder-decoder for dense prediction. Second, CAST treats segmentations solely as outputs, whereas SHED also uses segment-associated features as decoder inputs to produce dense representations. Third, CAST relies on image-level supervision and produces segmentations guided by visual cues, while SHED is trained with dense supervision (e.g., depth), resulting in segmentations guided by geometric cues. Finally, CAST links reorganization to recognition in the "3Rs" [48], whereas SHED links reorganization to reconstruction.

By looping hierarchical segmentation into dense prediction, SHED offers three key advantages over DPT. **1)** Segmentation enhances depth estimation by enforcing object-level structure, yielding sharper boundaries and coherence within segments. SHED reduces the boundary error by 54% (1.64→0.76). It also improves efficiency through coarse-to-fine decoding, lowering GFLOPs by 26%. **2)** Depth supervision leads to structured representations that better capture scene layout. As a result, SHED retrieves layout-similar images more accurately, increasing top-1 recall by 34% (45.2→60.5). **3)** Accurate depth maps from SHED improve 3D reconstruction, producing smooth surfaces aligned with the ground truth. Its hierarchy also enables unsupervised 3D part discovery, which DPT cannot achieve as it predicts depth holistically without structural understanding.

2. Related Work

Dense prediction is a core problem in computer vision, aiming to assign pixel-level outputs across an image [20]. It includes tasks such as semantic segmentation [47], depth estimation [16], optical flow [69], and image editing [32]. Modern approaches typically adopt encoder-decoder architectures, such as U-Net [62] and DPT [60], trained using task-specific supervision. These models perform well on benchmarks focused on per-pixel accuracy, as demonstrated by large-scale systems like Segment Anything [61] and Depth Anything [77]. However, recent studies show that even top-performing models often lack structural consistency [17, 49]. We argue that dense prediction should move beyond local estimation toward structured reasoning guided by region-level abstraction.

Monocular depth estimation is a representative dense prediction task, that infers per-pixel depth from a single image. It is widely used in 3D reconstruction [67], autonomous driving [22], and robotic perception [68]. Early approaches relied on hand-engineered features [64, 70], while deep learning methods later became dominant [16, 23, 24, 30, 42, 43, 59, 83]. Recent ViT [13]-based models such as DPT [60] have shown strong performance, leveraging foundation models pretrained on diverse data [4, 35, 76]. However, these models still struggle with structural consistency in complex scenes.

Structural cues in depth estimation have been extensively explored to enhance geometric coherence. Existing approaches can be broadly categorized into four types: **1)** Representation approaches modify how depth is encoded, such as by discretizing depth values [3, 21, 44] or modeling spatial dependencies [10, 45, 81]. **2)** Regularization imposes geometric constraints through loss functions that promote smooth surfaces [5, 23, 82], consistent normals [78], or planar regions [72, 79]. **3)** Multi-task learning jointly estimates depth with auxiliary signals, such as scene geometry [15, 80] or semantics [7, 25, 38, 54, 84]. **4)** Post-processing refines predictions using off-the-shelf techniques [8, 41].

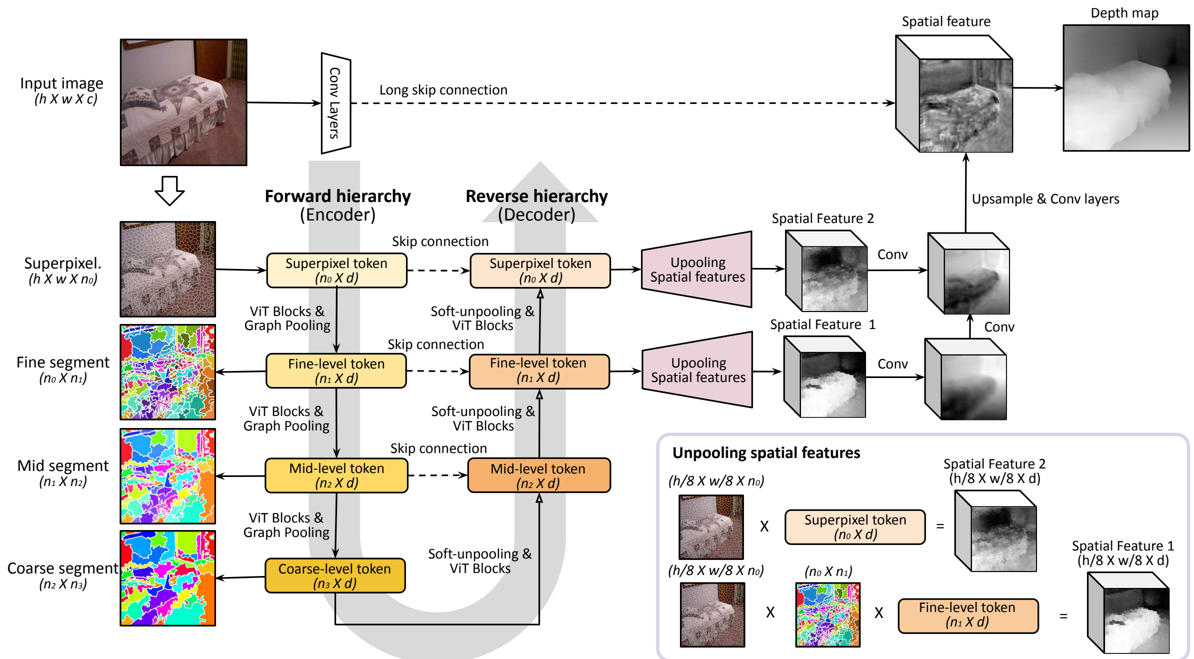


Figure 2. **SHED integrates a forward and reverse segment hierarchy into the ViT blocks of DPT.** While following the overall architecture of DPT, including convolutional layers, we adapt the ViT into two stages. **1)** The encoder converts the input image into superpixel tokens and applies graph pooling to form coarser segments, following the hierarchical clustering strategy of CAST [37]. **2)** The decoder reverses this hierarchy by unpooling segment tokens from coarse to fine and fusing them with encoder features at corresponding levels via skip connections. The tokens are projected into 2D maps according to their regions. These multi-level maps are fused with pixel-level features from early convolutional layers to recover fine details and produce the final depth map.

Several multi-task approaches have explored segmentation as an auxiliary signal to improve depth estimation. Early works used segmentation as an additional supervision signal [38, 54], while more recent ones leveraged segment regions or boundaries to guide depth discontinuities [7, 25, 84]. SHED follows this principle but integrates segmentation and depth estimation into a unified process, enabling them to benefit from each other. Moreover, it discovers hierarchical segmentation in an unsupervised manner, eliminating the need for costly human annotations.

Although structural cues offer clear benefits, most existing methods do not scale well to modern architectures. Representation-based approaches often require architectural changes that are incompatible with transformers, while regularization and multi-task methods rely on additional annotations, limiting scalability. In contrast, SHED integrates seamlessly into ViT-based models such as DPT and learns structural segmentation solely from depth supervision. By design, it inherently produces sharp, segment-aligned boundaries, reducing the need for post-processing.

Perceptual grouping is a key mechanism in human vision that organizes low-level elements into coherent global structures [50, 73]. This principle has inspired a broad range of computer vision research, including perception [12, 33, 46, 53, 58], segmentation [2, 31, 36, 75], and generation [27, 29, 52]. In particular, CAST [37] recently applied it to ViTs for concurrent segmentation and recognition. While most of these methods, including CAST, consider only a *forward hierarchy*, constructing representations and segmentations in a bottom-up manner, we adopt the complementary concept of a *reverse hierarchy* [28], where global structures guide and refine local parts through top-down feedback. We leverage this principle to design an encoder-decoder that accounts for both hierarchies.

While some prior works [1, 14, 66] have explored reverse hierarchies for recognition, they do not address dense prediction. Other studies [18, 63, 65] apply similar ideas to encoder-decoder architectures, but focus on object-centric representations, lacking the ability to model segment hierarchies and often producing blurry outputs. To the best of our knowledge, this is the first work to leverage bidirectional segment hierarchies to enhance dense prediction within a modern ViT framework.

3. Method

We propose SHED, which integrates a bidirectional segment hierarchy into the ViT blocks of DPT [60]. While retaining DPT's overall architecture, we modify only the

ViT blocks between the convolutional encoder and decoder. Unlike DPT, which uses fixed-size patch tokens across all layers, our model constructs a hierarchy of segment tokens: the encoder builds a forward hierarchy by grouping features from fine to coarse, while the decoder applies a reverse hierarchy to refine predictions from coarse to fine, guided by the learned segment tokens. This design, illustrated in Fig. 2, enables the model to progressively reorganize and reconstruct structured scene information.

3.1. Grouping segments via forward hierarchy

Our encoder builds on CAST [37], which **1)** replaces square patch tokens with superpixel tokens, and **2)** progressively clusters them into coarser segment tokens by token similarity. This process produces a fine-to-coarse hierarchy of segment tokens. CAST was originally developed as an encoder-only model for image-level recognition. We extend it into an encoder-decoder, where the segment hierarchy not only guides dense prediction but is also refined through dense supervision.

Tokenization. Given an input image $X \in \mathbb{R}^{h\times w\times c}$, the encoder produces hierarchical segmentations $S_0, S_1, \dots$ and corresponding embeddings $Z_0, Z_1, \dots$, ordered from fine to coarse. This process begins by dividing the image into n_0 superpixels (e.g., using the SEEDS algorithm [71]), which yields a one-hot assignment matrix $S_0 \in \mathbb{R}^{(h\times w)\times n_0}$ that maps each pixel to a superpixel.

We extract a convolutional feature map $F_{\text{conv}} \in \mathbb{R}^{(h_0\cdot w_0)\times d}$ with spatial stride 8 ($h_0 = h/8$, $w_0 = w/8$), add fixed sinusoidal positional embeddings, and average-pool features within each superpixel to obtain initial embeddings $Z_0 \in \mathbb{R}^{n_0\times d}$. To enable global context modeling, we append a class token to form $\bar{Z}_0 \in \mathbb{R}^{(n_0+1)\times d}$, which is passed to the first ViT block.

Hierarchical clustering. We construct coarser segment tokens by alternating ViT blocks with graph pooling [37]. At each level l, given Z_{l-1} and S_{l-1} from the previous layer, we append a class token to form $\bar{Z}_{l-1}$, apply ViT blocks, and obtain updated features, excluding the class token.

To form coarser tokens $Z_l \in \mathbb{R}^{n_l\times d}$, we compute a soft assignment matrix $P_l \in \mathbb{R}^{n_{l-1}\times n_l}$ based on cosine similarity between fine- and coarse-level tokens:

$$P_l(i \to j) \propto \text{sim}(Z_{l-1}[i], Z_l[j]), \quad \text{for } i \in [n_{l-1}],\ j \in [n_l],$$

where $[n] := \{0, \dots, n-1\}$. The coarse tokens Z_l are initialized via farthest point sampling [56] from Z_{l-1}, and refined by aggregating fine-level features weighted by P_l, followed by an MLP and a residual connection:

$$Z_l \leftarrow Z_l + \text{MLP}(P_l^\top Z_{l-1} \oslash P_l^\top \mathbf{1}),$$

where $\oslash$ denotes element-wise division for normalization.

To propagate segmentation labels through the hierarchy, we compute coarser segmentations by composing the assignment matrices:

$$S_l = S_{l-1}\,\bar{P}_l, \quad l = 1, 2, \dots, l_{\max},$$

where $\bar{P}_l$ is a hard assignment matrix obtained by taking the argmax over each row of P_l.

3.2. Predicting outputs via reverse hierarchy

The decoder reconstructs spatial feature maps by reversing the encoder's segment hierarchy, progressively unpooling segment tokens $Z_{l_{\max}}, \dots, Z_0$. This involves two steps: **1)** computing decoder features Z'_l by unpooling from Z'_{l+1} and fusing them with encoder features Z_l via skip connections; and **2)** projecting Z'_l to the image space to obtain a spatial feature map F_l of size (h_l, w_l).

Unpooling segment tokens. We reverse the encoder's clustering in a coarse-to-fine manner. At each level $l = l_{\max} - 1, \dots, 0$, we compute

$$Z'_l \leftarrow P_{l+1}^\top\, Z'_{l+1},$$

which distributes coarse features to finer segments. We then concatenate the unpooled features with the corresponding encoder output:

$$Z'_l \leftarrow \text{MLP}(\text{Concat}(Z'_l, Z_l)),$$

followed by ViT blocks with class tokens.

Unpooling spatial features. We convert the segment tokens Z'_l into spatial feature maps by composing the soft assignment matrices:

$$P_{0\to l} = P_1 \cdots P_l \quad \in \mathbb{R}^{n_0\times n_l},$$

and applying them to the initial superpixel-to-pixel map S_0 to obtain soft segmentations $S_{0\to l} = S_0 P_{0\to l}$. The spatial feature map is then reconstructed as

$$F_l = S_{0\to l}\, Z'_l, \quad F_l \in \mathbb{R}^{(h_l\cdot w_l)\times d}.$$

The set of spatial maps $\{F_l\}_{l=1}^{l_{\max}}$ is fused using convolutional layers, combined with F_{conv}, and further refined through final convolution and upsampling to produce the final dense prediction. DPT reduces the spatial resolution of feature maps F_l at each level by a factor of 2^l, with $h_l = h_0/2^l$, $w_l = w_0/2^l$, producing coarse maps in early ViT layers that are progressively refined. This forms a *spatial hierarchy* similar to U-Net [62], improving global coherence and reducing computation. However, it relies on local aggregation, which lacks fine-grained structure, and reduces computation only in the final decoder, not in the ViT blocks where most of the cost arises.

In contrast, our *segment hierarchy* groups segment regions, providing a stronger inductive bias that promotes

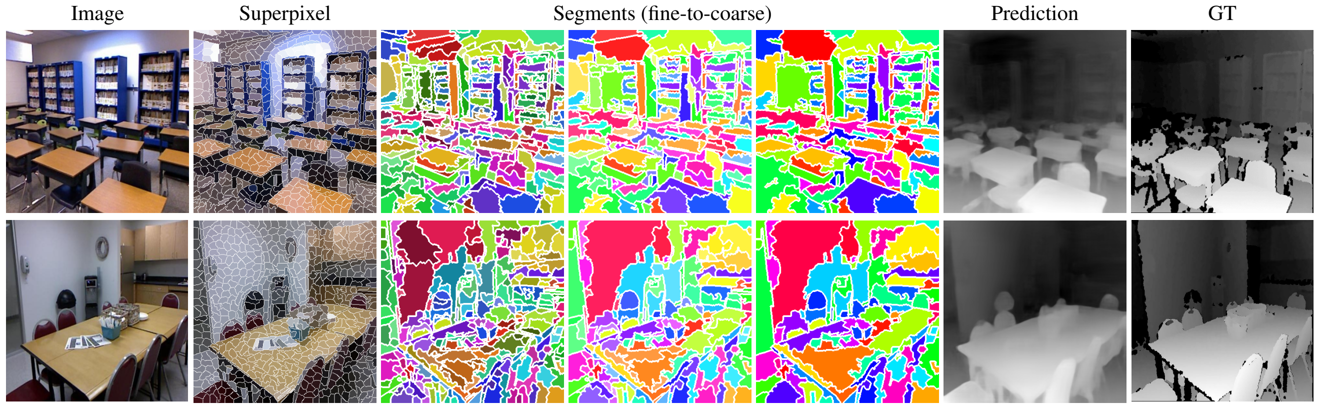


Figure 3. **SHED produces consistent and continuous structures across segmentation and depth.** We visualize the fine-to-coarse segments and corresponding depth maps from SHED, along with ground truth (GT) depth for comparison. Examples are from the NYUv2 [55] test set. SHED captures fine structures through its segments, such as desks in a classroom, which allow the depth map to clearly separate them from the background (row 1). It also decomposes large objects, such as a table, into multiple parts, leading to smooth depth variations toward the back (row 2).

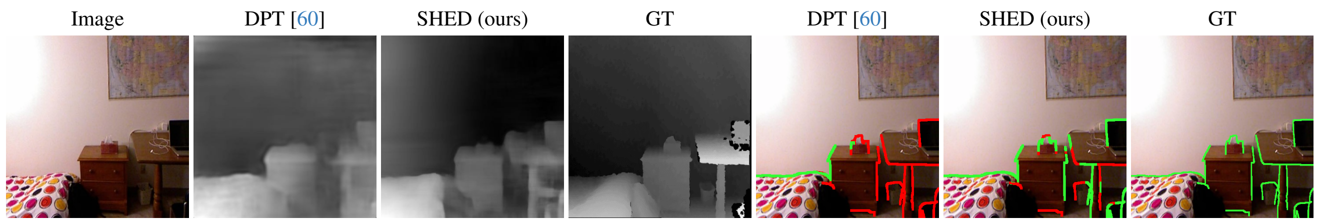


Figure 4. **SHED generates sharper object contours, clearer occlusion boundaries, and more coherent values within segments.** We compare depth maps (cols 2-4) and occlusion boundaries (cols 5-7) from DPT, SHED, and the ground truth (GT) on the NYUv2-OC++ dataset [57]. Boundaries are extracted using a Canny edge detector [6] and evaluated against GT, with correct edges shown in green and errors in red. SHED more accurately captures object edges and produces smoother depth within segments. Its predicted boundaries also align more closely with the ground truth.

structural consistency and reducing computation in the ViT blocks. As a result, applying spatial downsampling in SHED was not beneficial: it yielded minimal efficiency gains in the decoder while degrading boundary quality by projecting coarse segments onto low-resolution maps. Therefore, we omit spatial reduction in SHED and simply set $h_l = h_0, w_l = w_0$.

4. Experiments

We demonstrate the benefits of SHED by integrating segmentation into the loop for dense prediction: **1)** Segment-consistent depth estimation that preserves occlusion boundaries and intra-segment coherence, leading to improved accuracy and efficiency; **2)** Structure-aware representation learning through dense supervision, enabling layout-aware features and segmentations; **3)** 3D scene reconstruction from predicted depth maps, yielding globally coherent and part-aware structures.

Table 1. **SHED improves boundary accuracy and intra-segment coherence.** We evaluate the structural quality of SHED depth maps using two metrics: **1)** Occlusion boundary accuracy [40], evaluated on the NYUv2-OC++ dataset [57]. Occlusion boundaries are extracted using a Canny edge detector, and the average Chamfer distance is computed in both directions: from prediction to ground truth and vice versa. SHED significantly improves over DPT, reducing recall error by 54% (1.64→0.73). **2)** Intra-segment coherence, computed on the NYUv2 dataset [55]. This metric measures how well the predicted depth values within each object segment align with the ground-truth distribution. We compute the average Wasserstein-1 distance using object-level annotations, evaluating both precision and recall. SHED outperforms DPT on both metrics.

Method	OBs (ϵ_a / ϵ_c) ↓	Intra-segment coherence ↓
DPT [60]	6.429 / 1.637	0.631 / 0.664
SHED (ours)	**5.875 / 0.732**	**0.614 / 0.614**

Table 2. **SHED improves computational efficiency while matching per-pixel metrics.** We evaluate standard depth accuracy and error metrics, along with GFLOPs, on the NYUv2 [55] test set. SHED reduces GFLOPS by 26% (302→223) by predicting depth at the segment level, avoiding redundant processing of similar pixels and reducing token counts in internal layers. Its per-pixel metrics are similar to DPT, which is not surprising since these metrics reflect only local plausibility and ignore structural quality. Both models are trained with the same supervision, resulting in similar losses despite clear differences in visual fidelity, especially in fine details.

	Efficiency	Depth Accuracy			Depth Error		
Method	GFLOPs ↓	$\delta > 1.25$ ↑	$\delta > 1.25^2$ ↑	$\delta > 1.25^3$ ↑	AbsRel ↓	RMSE ↓	log10 ↓
DPT [60]	302.0	0.664	0.886	0.954	1.011	0.731	0.100
SHED (ours)	**223.1**	**0.674**	**0.890**	**0.955**	**0.991**	**0.720**	**0.097**

4.1. Setup

We implement SHED on top of DPT [60], adopting its overall training setup. Specifically, we use the DPT-Hybrid variant, which combines ResNet-50 [26] and ViT-Base [13], and refer to it simply as DPT throughout the paper. We primarily train and evaluate on NYUv2 [55], a standard benchmark for indoor depth estimation.

Tokenization. Input images of size 640×480 are randomly cropped to 416×416 during preprocessing. We generate 676 superpixels using the SEEDS algorithm [71], matching the 26×26 token grid of DPT, which corresponds to 16×16 patches. Features are extracted from intermediate ResNet-50 blocks at 1/4 and 1/8 of the input resolution; the latter initializes segment token embeddings, while both are passed to the final decoder via skip connections.

Architecture. We modify the ViT encoder-decoder in DPT by inserting graph pooling and unpooling layers. The encoder consists of three stages, each with two ViT blocks followed by graph pooling, progressively reducing the number of segment tokens to 256, 128, and 64. The decoder mirrors this structure with graph unpooling and receives skip connections from the corresponding encoder stages.

Training. We train both SHED and DPT from scratch on NYUv2 for a fair comparison, using a batch size of 16 for 50 epochs with the Adam optimizer [39] and a learning rate of 5e-5. Both models could be further improved by using pretrained ResNet and ViT backbones, as done in the original DPT. We follow DPT's default training recipe, including the scale-invariant logarithmic loss computed against ground-truth depth. At inference time, predicted depth maps at 416×416 resolution are bilinearly upsampled to 640×480 to match the ground-truth size.

4.2. Segment-consistent depth estimation

SHED generates structured depth maps by leveraging a learned segment hierarchy. We begin by visualizing the hierarchy and predicted depth to illustrate their structural alignment. Next, we evaluate quality in terms of boundary accuracy and intra-segment coherence. Finally, we show that hierarchical decoding improves efficiency without compromising pixel-wise accuracy.

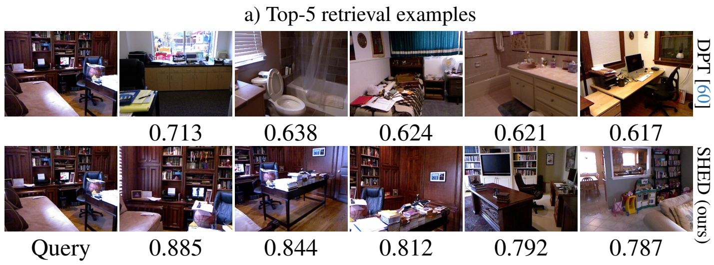
Fig. 3 shows that the segment hierarchy in SHED yields depth maps with coherent object geometry. The learned segments capture contours of objects, such as desks in a classroom, allowing the depth to clearly separate them from the floor. They also decompose larger structures, like tables, into parts, enabling smooth depth transitions from front to back. This suggests that structure guides depth prediction toward more accurate and interpretable results.

Boundary accuracy. We assess the structural quality of SHED comparing its boundary predictions to those of DPT. Fig. 4 shows predicted depth maps and their occlusion boundaries, extracted using a Canny edge detector [6], on samples from the NYUv2-OC++ dataset [57]. For quantitative evaluation, we follow the standard protocol [40] and compute the average Chamfer distance [19] in two directions: from prediction to ground truth, and vice versa. SHED produces sharper contours and outperforms DPT on both metrics, with particularly large gains in recall, likely due to its fine-grained segmentation. However, oversegmentation may introduce spurious edges that reduce precision, highlighting the importance of accurate segmentation.

Intra-segment coherence. Beyond boundary, we evaluate how coherently depth values vary within each segment. We use a metric called intra-segment coherence, which measures the similarity between the predicted and ground-truth depth distributions within each segment, treating the latter as structural references. It is computed as the average Wasserstein-1 distance [34], using ground-truth segmentations from the NYUv2 dataset [55]. As shown in Fig. 4, SHED produces smoother depth variations within segments. This is reflected quantitatively in Tab. 1, where it outperforms DPT in both precision and recall.

Per-pixel metrics and efficiency. Beyond structure-aware metrics, we compare SHED and DPT using standard per-pixel depth metrics and computational cost, as shown in Tab. 2. SHED achieves comparable accuracy with significantly lower cost by predicting at the segment level, which reduces token count and avoids redundant computation. Although both models perform similarly on these metrics, this is expected, as they capture only local plausibility. This highlights the need for structure-aware evaluation to assess

ICCV
#*****



b) Top-K retrieval accuracy (%)

Method	Top-1	Top-3	Top-5
Scene retrieval			
DPT [60]	45.2	69.7	77.2
SHED (ours)	**60.5**	**78.7**	**87.0**
Frame retrieval (k=5)			
DPT [60]	18.5	31.0	38.3
SHED (ours)	**30.5**	**42.3**	**48.3**

Figure 5. **SHED learns layout-aware representations through depth supervision.** We evaluate image retrieval on NYUv2 [55] based on cosine similarity between class tokens from the final ViT block. **a)** Top-5 results (ranked left to right), with similarity scores shown below. SHED retrieves images with similar layouts, such as a central desk and a rear bookshelf, while DPT retrieves unrelated scenes. **b)** Top-K accuracy at the scene and frame level ($k = 5$), where the targets are different views from the same scene or nearby frames. SHED significantly outperforms DPT in all settings, indicating that our depth-guided segmentation effectively encodes spatial layout.

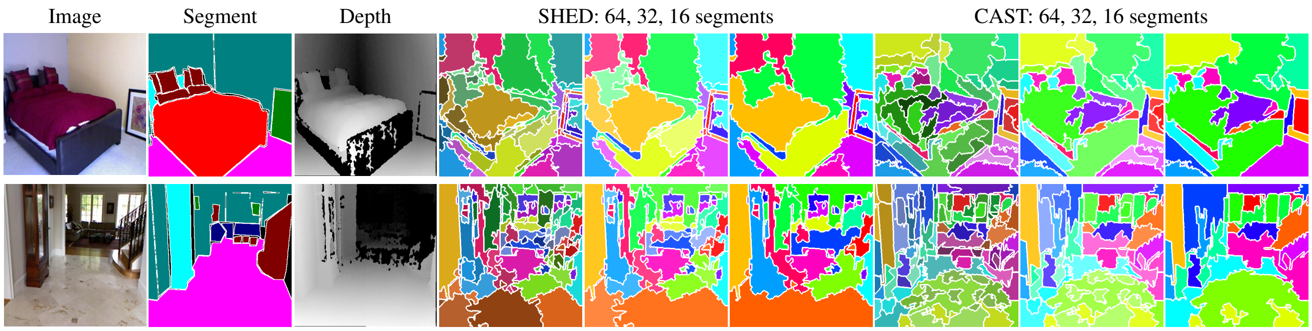


Figure 6. **SHED learns depth-aware segment hierarchies, while CAST relies on visual cues.** We compare segmentations from SHED and CAST [37] at the same hierarchy levels: 64, 32, and 16 segments. SHED captures meaningful part structures, such as separating the blanket and pillow from the bed (row 1). It also decomposes large structures like the floor based on depth, grouping nearby regions into a single large segment while dividing distant areas into smaller ones (row 2). In contrast, CAST relies on appearance cues and fails to capture geometric structure. For instance, it groups white floor regions by color but divides them arbitrarily, ignoring depth. These results highlight the value of depth supervision in learning 3D-aware segmentations.

depth more comprehensively.

4.3. Structure-aware representation learning

Our structured architecture not only improves depth prediction but also facilitates structure-aware representation learning. First, SHED learns features that reflect scene layout, enabling more accurate layout-aware image retrieval than DPT [60]. Second, its segment hierarchy captures geometric cues informed by depth supervision, whereas CAST [37] relies on visual cues.

Layout-aware image retrieval. We assess the structural understanding of learned representations by performing layout-aware image retrieval on the NYUv2 dataset [55], using 120K video frames collected from 206 scenes. These frames serve as queries, and we define two retrieval settings. In scene retrieval, all frames from the same sequence are valid targets. For finer-grained evaluation, we also consider frame-k retrieval, where only frames within k time steps of the query are included. Given a query image, we rank other images by the cosine similarity of their class tokens from the final ViT decoder block. Fig. 5 presents both qualitative and quantitative results. The left side shows that SHED retrieves images with similar spatial layouts, such as a central desk and a rear bookshelf, while DPT returns unrelated scenes. The right side shows that SHED significantly outperforms DPT in both scene- and frame-level metrics, improving Top-1 recall in scene retrieval from 45.2 to 60.5. Additional analysis, including trends across varying values of k, is provided in supplementary.

Depth-aware image segmentation. We analyze the segment hierarchy learned by SHED by comparing it to CAST, an encoder trained for image recognition using segment-based representations. We use CAST-B, trained on ImageNet [11] with the MoCo-v3 objective [9], a self-supervised learning method based on instance discrimination [74] that clusters visually similar images. Following CAST's setup, we use 224×224 images and extract 196 superpixels, clustered into 64, 32, and 16 segments. For comparison, we adapt the graph pooling layers of our pretrained SHED to produce the same number of segments, while keeping the original input resolution and superpixels. Despite the shift in token configurations, SHED produces

ICCV
#*****

Figure 7. **SHED produces more accurate and structured 3D reconstructions.** We visualize 3D point clouds reconstructed from single-view depth maps, following the semantic scene completion protocol [67], using predictions from DPT, SHED, and the ground truth on NYUv2 [55] examples. Frontal views (cols 2-4) show that DPT fails to preserve planar structures, producing curved wall boundaries, whereas SHED more accurately recovers straight lines. This difference is even more apparent in the bird's-eye views (cols 5-7): DPT yields warped surfaces, while SHED produces flatter layouts that better match the ground truth.

consistent and meaningful segmentations.

Fig. 6 shows qualitative results. SHED learns hierarchical structures that align with scene geometry: it separates objects like blankets and decomposes large structures such as floors into segments that reflect their spatial extent. In contrast, CAST groups regions based on appearance. For example, it clusters white floor areas by color but fails to account for geometric cues. We attribute this difference to the training objective: CAST learns segments through image-level recognition, while SHED is guided by dense prediction. Although our focus here is depth, the ability to learn segment hierarchies grounded in 3D structure opens possibilities for other dense prediction tasks as well.

4.4. 3D scene reconstruction with part structures

We conclude by demonstrating SHED's capability for 3D scene understanding. While plausible pixel values may suffice for 2D depth estimation, accurate and structured depth is particularly critical when projected into 3D space. Accordingly, SHED enables high-quality 3D reconstruction and supports unsupervised 3D part discovery through concurrent segmentation.

3D scene reconstruction. To evaluate the structural quality of predicted depth maps, we project them into 3D point clouds on the NYUv2 dataset [55], following the semantic scene completion protocol [67] and using NYUv2 camera intrinsics. For interpretability, all depth values are scaled by 1/1000. Fig. 7 shows that SHED produces cleaner reconstructions with sharper boundaries and flatter surfaces that better align with ground truth geometry, whereas DPT yields curvier, less faithful shapes. We quantify reconstruction accuracy using the average Chamfer distance [19] in both directions. Tab. 3 shows that SHED consistently achieves lower distances than DPT, confirming its advantage in structured 3D prediction.

3D part discovery. By jointly predicting segmentation and depth, SHED lifts 2D parts into 3D space, enabling part-level decomposition of scenes. Tab. 4 shows an example from NYUv2 [55], where segments corresponding to objects like beds and carpets form coherent 3D structures in the point cloud. This demonstrates SHED's potential for unsupervised 3D part reasoning, a key capability for interactive and dynamic scene understanding [51].

Table 3. **SHED improves 3D alignment.** We compute the average Chamfer distance [19] between point clouds reconstructed from the predicted and ground-truth depths. SHED achieves lower errors than DPT.

Method	Precision / Recall ↓
DPT [60]	0.171 / 0.251
SHED (ours)	**0.158 / 0.244**

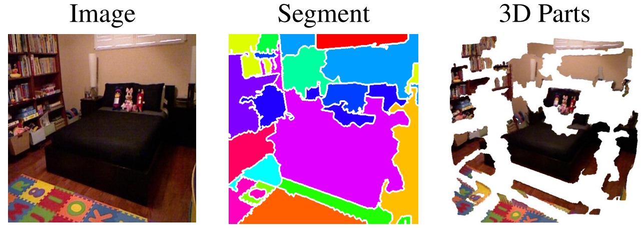


Table 4. **SHED discovers 3D part structures.** Concurrent segmentation and depth estimation enable part-level decomposition of the reconstructed 3D point clouds.

5. Conclusion

We shed light on the role of segmentation in depth estimation. SHED learns a segment hierarchy in the encoder and reverses it in the decoder to predict dense maps. This results in depth maps with segment-consistent structure, layout-aware representations, and coherent 3D scenes with interpretable parts. Our principle of unifying reconstruction and reorganization offers a new direction for 3D vision and robotics, particularly for tasks that require fine-grained interaction with physical components.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3

[2] Pablo Arbeláez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik. Semantic segmentation using regions and parts. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3378–3385. IEEE, 2012. 3

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021. 2

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[5] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019. 2

[6] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 5, 6

[7] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2624–2632, 2019. 2, 3

[8] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 2

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 7

[10] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 103–119, 2018. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[12] Zhiwei Deng, Ting Chen, and Yang Li. Perceptual group tokenizer: Building perception with iterative grouping. *arXiv preprint arXiv:2311.18296*, 2023. 3

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 6

[14] Ainaz Eftekhar, Kuo-Hao Zeng, Jiafei Duan, Ali Farhadi, Ani Kembhavi, and Ranjay Krishna. Selective visual representations improve convergence and generalization for embodied ai. *arXiv preprint arXiv:2311.04193*, 2023. 3

[15] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 2

[16] 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

[17] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2

[18] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in neural information processing systems*, 29, 2016. 3

[19] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 6, 8

[20] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002. 1, 2

[21] 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

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[23] 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

[24] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 2

[25] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. 2, 3

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

[27] Xingzhe He, Bastian Wandt, and Helge Rhodin. Ganseg: Learning to segment by unsupervised hierarchical image

generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1225–1235, 2022. 3

[28] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002. 1, 3

[29] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018. 3

[30] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019. 2

[31] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7334–7344, 2019. 3

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[33] Hyunwoo Kang, Sangwoo Mo, and Jinwoo Shin. Oamixer: Object-aware mixing layer for vision transformers. *arXiv preprint arXiv:2212.06595*, 2022. 3

[34] Leonid V Kantorovich. The mathematical method of production planning and organization. *Management Science*, 6 (4):363–422, 1939. 6

[35] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2

[36] Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2571–2581, 2022. 3

[37] Tsung-Wei Ke, Sangwoo Mo, and X Yu Stella. Learning hierarchical image segmentation for recognition and by recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 4, 7

[38] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 2, 3

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

[40] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 5, 6

[41] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 2

[42] 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

[43] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[44] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 2024. 2

[45] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015. 2

[46] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 3

[47] 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

[48] 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, 2

[49] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024. 2

[50] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 3

[51] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 8

[52] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018. 3

[53] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, 34:12251–12264, 2021. 3

[54] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 611–619. IEEE, 2016. 2, 3

[55] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5, 6, 7, 8

[56] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4

[57] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14657, 2020. 5, 6

[58] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5571–5584, 2023. 3

[59] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 2

[60] 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. 1, 2, 3, 5, 6, 7, 8

[61] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025. 1, 2

[62] 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, 4

[63] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd Van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in neural information processing systems*, 35:9512–9524, 2022. 3

[64] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 2

[65] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022. 3

[66] Baifeng Shi, Trevor Darrell, and Xin Wang. Top-down visual attention from analysis by synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2102–2112, 2023. 3

[67] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2, 8

[68] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6243–6252, 2017. 2

[69] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[70] Antonio Torralba and Aude Oliva. Depth estimation from image structure. *IEEE Transactions on pattern analysis and machine intelligence*, 24(9):1226–1238, 2002. 1, 2

[71] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin De Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, pages 13–26. Springer, 2012. 4, 6

[72] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2162–2171, 2019. 2

[73] Max Wertheimer. Laws of organization in perceptual forms. 1938. 3

[74] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 7

[75] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144, 2022. 3

[76] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[77] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1, 2

[78] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from

videos with edge-aware depth-normal consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[79] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5684–5693, 2019. 2

[80] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2

[81] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3916–3925, 2022. 2

[82] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 340–349, 2018. 2

[83] 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

[84] Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13116–13125, 2020. 2, 3