Joint Depth Prediction and Semantic Segmentation with Multi-View SAM

Mykhailo Shvets
University of North Carolina
Chapel Hill
mshvets@cs.unc.edu

Dongxu Zhao
University of North Carolina
Chapel Hill
dongxuz1@cs.unc.edu

Marc Niethammer
University of North Carolina
Chapel Hill
mn@cs.unc.edu

Roni Sengupta
University of North Carolina
Chapel Hill
ronisen@cs.unc.edu

Alexander C. Berg
University of California
Irvine
bergac@uci.edu

Abstract

Multi-task approaches to joint depth and segmentation prediction are well-studied for monocular images. Yet, predictions from a single-view are inherently limited, while multiple views are available in many robotics applications. On the other end of the spectrum, video-based and full 3D methods require numerous frames to perform reconstruction and segmentation. With this work we propose a Multi-View Stereo (MVS) technique for depth prediction that benefits from rich semantic features of the Segment Anything Model (SAM). This enhanced depth prediction, in turn, serves as a prompt to our Transformer-based semantic segmentation decoder. We report the mutual benefit that both tasks enjoy in our quantitative and qualitative studies on the ScanNet dataset. Our approach consistently outperforms single-task MVS and segmentation models, along with multi-task monocular methods.

1. Introduction

Depth prediction and semantic segmentation are core tasks for visual understanding in robotic perception. The ability to recognize *what* the objects are and *where* they are in the scene, with respect to the robot's viewpoint, plays a key role in enabling effective navigation and interaction in complex environments.

The tremendous success of deep learning methods has significantly improved the performance of both tasks in recent years. Moreover, depth prediction and semantic segmentation have been shown to carry mutually beneficial information: depth complements the visual RGB information with geometric cues to help segmentation [16,20,27,44,57], and conversely, segmentation can help guide depth prediction [7, 18, 29, 60]. To take it one step further, rather than

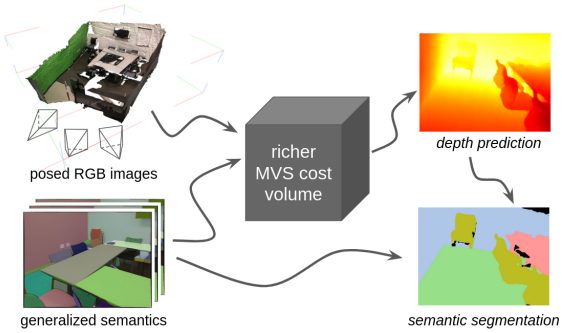

Figure 1. Generalized semantic features help to build a richer cost volume for MVS. In turn, the depth predicted from the cost volume serves as a rich prompt for semantic decoding.

treating tasks in isolation or solving them sequentially, a range of multi-task approaches were proposed to jointly solve depth estimation and semantic segmentation, especially in the monocular image domain [34, 37, 46, 49]. The discovery of this complementarity between the tasks improves the generalization capabilities of both [45]. In the single image domain, it is natural, from an architectural design perspective, to combine these two since both can be addressed using a 2D convolutional encoder-decoder architecture. Nevertheless, a significant amount of geometric and relational context is often missing from the monocular estimation algorithms since it is challenging to estimate geometry and semantics purely from a single view.

On the other end of the spectrum, full 3D recognition [24] and reconstruction [36,43] models showed impressive performance on a range of tasks. These approaches typically operate on video sequences [13] or complete 3D scans [9]. Such data can be expensive to acquire and/or process. Moreover, in robotics applications, scenarios arise that necessitate achieving visual understanding from just a few frames without a holistic scene analysis. This is partic-

ularly prominent in dynamic scenes that are impossible to measure completely at a single point in time.

Multi-View Stereo (MVS) approaches have the potential to bridge the gap between weaker monocular methods suffering from the absence of explicit geometry notions, and full video-based or 3D-based solutions. Indeed, the goal of MVS models is to successfully predict depth from just a few input views by leveraging principles and concepts from traditional camera geometry [23]. Modern MVS models [10, 17, 55] extract deep 2D geometry-aware features for cost volume construction. Unfortunately, while these MVS models do perform complex intra-frame reasoning, their feature extraction process is relatively simplistic. Due to this, the model predictions can suffer and may be highly error-prone, most notably in textureless regions. This challenge can only be mitigated to a limited extent with standard multi-scale and planar prior methods [51, 52].

We propose the use of semantic cues to help recover better from those errors. There are several challenges to overcome to achieve joint MVS depth prediction and segmentation in indoor settings. Our first challenge is that 2D features in the MVS pipeline are not naturally good at extracting broad semantics. A vast amount of models trained end-to-end for 2D semantic segmentation exist [2,4–6,39,44,59] that could be used to augment MVS 2D features with semantic cues. However, in our studies we show that their intermediate features might still lack the generalizability required to help MVS. One of the main reasons is that the sample diversity in existing indoor datasets [8, 42] is fairly limited—even with millions of images, those datasets typically only contain a limited number of scenes. Second, contrary to the single-task domain, performing joint prediction of MVS depth and semantics in a multi-task way is less straightforward due to the divergence of their models' architectural design: while semantics still follow the convolutional or attention-based design, MVS depth is naturally inferred from correlation or variance-based cost volumes [55]. This is the gap we plan on bridging in this work.

Additionally, MVS methods are not yet well explored for indoor environments and robotics scenarios. Popular datasets that are used to benchmark MVS methods [26, 30] often feature inward-facing views with a single object of interest captured from many perspectives, while in the indoor scenario, it is often the case that cameras are facing outwards and thus view baseline sampling options may be limited, as shown in Figure 2.

Our key contribution is to introduce a joint depth and segmentation prediction network using the multi-view stereo framework where we can reason about the 3D scene using a few images (3-5). The previous joint depth and segmentation approaches are either monocular [34, 37, 46, 49], or require a full dense scan of the environment [9, 19]. Our pipeline's concept is illustrated in Figure 1. We aim to exploit the advances in large pre-trained foundation models for image segmentation, The Segment Anything Model (SAM) [28], which offers strong generalizable semantic features. In this work, we use the SAM feature encoder to guide the construction of better cost volumes for MVS. We also propose a SAM-style decoder to extract semantic maps and use the depth predicted by the MVS branch as a dense decoder prompt. Our experimental results show improved performance from the use of these depth cues.

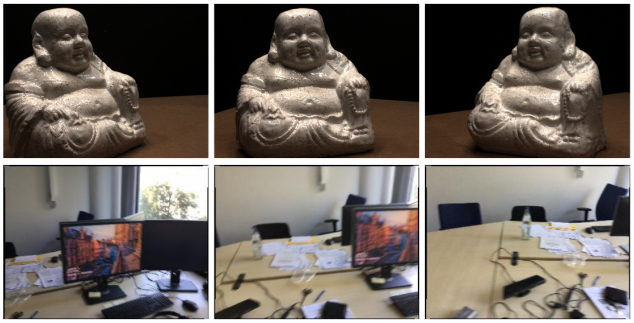

Figure 2. Challenges of view sampling in real-world scenarios. Top: DTU [26] dataset, commonly used for benchmarking MVS methods, has *inward* facing cameras: the views have wide baseline and the object of interest is featured from multiple angles. Bottom: ScanNet [8] indoor recognition dataset often features *outward* facing cameras with camera motion close to pure rotation.

We perform an extensive quantitative and qualitative evaluation on the large ScanNet [8] dataset of indoor scenes. We demonstrate considerable improvement over monocular multi-task methods (11% relative improvement in the semantic mIoU metric, and more notably, 40% improvement in the absolute depth error). Our method also performs better than the competing single-task MVS methods (with a 10% relative improvement in depth prediction over CasMVSNet), and better than the competing single-task segmentation approaches (17% relative improvement over AdapNet++ [44]), matching the performance of RGB-D algorithms, while requiring only RGB inputs.

Summarizing our contributions, we present:

1) A unified architecture that simultaneously solves the Multi-View Stereo and the Semantic Segmentation problems, filling the gap between monocular and full 3D multi-task methods, a first to our knowledge;
2) A method to enrich MVS-based cost volumes for depth prediction using rich SAM segmentation features;
3) A semantic decoder that takes advantage of both the rich SAM features, and the predicted depth;
4) Quantitative and qualitative studies of the proposed approach on the ScanNetv2 dataset show consistent improvement over both multi-view depth prediction and segmentation models and over algorithms that jointly predict depth and segmentation from a single image.

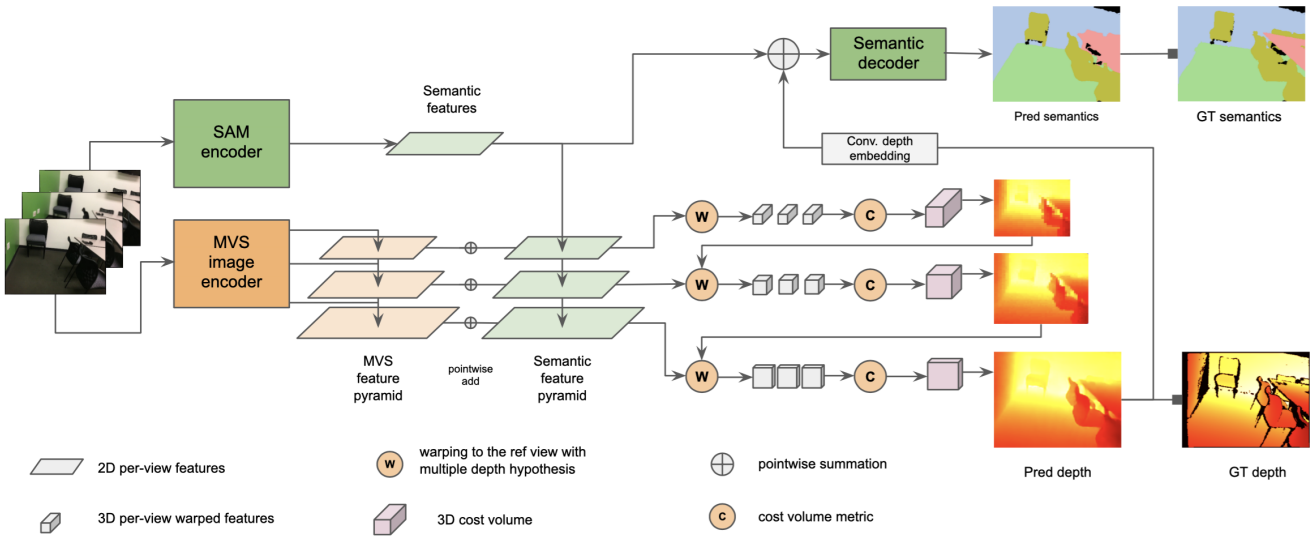

Figure 3. Architecture overview. Given a set of input RGB views (one reference and several other source images), we produce a depth map, and a semantic segmentation map for the reference view. The model first extracts per-view 2D features fusing the outputs of the convolutional MVS encoder and a transformer-based semantic encoder. The 2D features are warped onto the reference view with an array of depth hypotheses, and a 3D cost volume is constructed at a range of scales, where the initial depth prediction from a coarser scale allows to increase the resolution of the hypothesis step at a finer scale, focusing on refinement around the established coarse value. A 3D convolutional network decodes the cost volume into a depth prediction at each scale. The final depth prediction is used as a dense prompt for the transformer-based semantic decoder. The model is supervised with the ground truth depth and semantic labels during training.

2. Background and Related work

Monocular depth & segmentation Monocular semantic segmentation and depth estimation both perform dense predictions from a single image. Semantic segmentation requires a class label for each pixel [2, 4–6, 39, 44, 59], while depth estimation predicts depth values that lift the pixels to 3D points [1, 31, 33, 38, 40]. There have been works exploring sequential (one task guides the other) and joint learning of both tasks (multi-task). Casser *et al*. [3] use precomputed instance segmentation masks to handle objects in highly dynamic scenes. And a range of works exists that simultaneously learn to predict segmentation and depth [14, 22, 47, 48]. We compare with four monocular methods: MT RefineNet [37], PAD Net [49], MTI Net [46] and MTAN [34]. PAD Net leverages a set of intermediate tasks to guide the final segmentation and depth predictions. MTAN proposes task-specific attention modules. However, monocular approaches are naturally limited by the information encoded in a single image and struggle to recover complete geometry, so in this work we focus on reconstruction from multi-view images.

Full 3D segmentation and reconstruction To utilize information from multiple images, methods based on Truncated Signed Distance Functions (TSDF) have been proposed which can perform real-time processing and fusion of a sequence of images or a video. Atlas [36] and NeuralRecon [43] use back-projection to lift 2D features into 3D space and regress a TSDF in a voxel grid. Atlas [36] performs a running average aggregation and produces a scene-level voxel grid, while NeuralRecon [43] predicts fragments of the scene and fuses them together through an RNN.

Segmentation and reconstruction can benefit from each other by taking additional depth information into account or by being guided by semantic information. Hane *et al*. [19] simultaneously solve the tasks of dense 3D scene reconstruction from multiple images and image semantic segmentation by utilizing class-specific smoothness assumptions in place of standard smoothness assumptions. 3DMV [9] utilizes the reconstruction of a RGB-D scan to predict a 3D semantic segmentation.

While these methods achieve high quality reconstruction and segmentation, they require many frames to work on and sometimes other inputs such as an initial mesh reconstruction. Our work aims at jointly learning depth and segmentation prediction from just a few views.

Multi-View Stereo (MVS) The goal of Multi-View Stereo (MVS) is to reconstruct 3D scenes from multiple posed RGB images, usually by predicting depth maps and then fusing them together. One widely used approach builds plane-sweep volumes [10, 17, 25, 55, 58] on top of depth hypotheses. Many SOTA methods are built upon MVS-Net [55], which consists of extracting image features, warping features using homography to construct 3D cost volumes and applying a 3D CNN to estimate depth. CasMVS-Net [17] builds the cost volumes in a cascaded way to im-

prove time and memory efficiency. TransMVSNet [10] further proposes to improve the feature extraction and matching task by leveraging a transformer to aggregate long-range context information within and across images.

Semantic segmentation has been deployed to improve MVS quality in a range of works. Semantic features can be used as a consistency constraint that favors the reconstructions yielding a consistent segmentation across binocular [53] or multi-view images [50]. They also help with 3D plane fitting in textureless regions where photometric feature matching often fails [41, 56].

However, to the best of our knowledge, no existing method jointly learns MVS and segmentation in a multi-task manner. Our method simultaneously estimates depth and semantic maps. We show that the two tasks benefit from each other. Furthermore, only a few MVS methods [54] report their performance on indoor datasets, while others report their performance on table-top datasets (see Figure 2). We benchmark our method on ScanNet [8].

Foundation models for segmentation The Transformer model was initially proposed for natural language processing (NLP) [1] introducing the attention mechanism to infer dependencies between language tokens. Transformers were then adopted for vision tasks achieving great success especially with the availability of large datasets [11, 35]. Vision Transformer (ViT) [12] is now a widely used model, which splits an image into patches and applies a Transformer on the encoding of a sequence of image patches. Although ViT was first introduced for image classification, it can serve as a rich feature extraction backbone on a range of vision tasks, similar to its convolutional counterparts like ResNet [21].

Producing truly universal features for segmentation was not straightforward. In this work, we leverage the Segment Anything Model (SAM) [28] that was recently proposed as a foundation model for open-set segmentation. We use SAM to extract semantic features to construct cost volumes that are more efficient for depth prediction. SAM has ViT as its image encoder and shows good zero-shot performance on several tasks such as edge detection and instance segmentation inside a given box. SAM's mask decoder operates on the image embeddings and accepts prompts of two types: sparse (e.g. foreground and background points of the object of interest), and dense (e.g. coarse masks). We follow this style of the decoder and adapt it to our needs, so that MVS depth predictions serve as dense prompts and learned class embeddings as sparse prompts.

3. Method

We consider the problem of joint depth prediction and segmentation of a scene from a set of M RGB images. Given input images $I_i, \forall i = 1, ..., M$ of spatial size $W \times H$ and corresponding camera poses $P_i = \{K_i, (R_i, t_i)\}$ (intrinsic parameters K_i, rotation matrix R_i and translation vector t_i), we aim to predict a pixel-wise depth map D_i and a segmentation map S_i. We follow the Multi-View Stereo (MVS) problem set-up, where we consider one of the M images as a reference image and use the remaining images to predict the depth map and segmentation map for the reference image. We repeat this M times by choosing each one of the M images as the reference image. Note that this is different from monocular depth map and segmentation estimation techniques where each image is used in isolation for prediction.

We take inspiration from the state-of-the-art MVS architecture for depth prediction, CasMVSNet [17], and for segmentation, Segment Anything (SAM) [28]. We propose a novel approach, as shown in Figure 3, that unifies these techniques to improve both depth and segmentation prediction. The model extracts 2D features from individual views through the geometry-aware convolutional MVS encoder, as well as the semantic-aware transformer-based SAM encoder, and fuses these features. Section 3.1 describes the 2D feature pipeline in detail. A multiscale cost volume is constructed from these per-view features and decoded into a depth probability volume, as discussed in Section 3.2. The depth prediction of the MVS branch is used as a dense prompt for the transformer-based semantic decoder (Section 3.3). The model is then trained end-to-end for joint depth prediction and semantic segmentation.

3.1. 2D encoders and feature fusion for MVS

The first step of an MVS approach extracts deep geometry-aware features that are good for matching across M input views. We design this MVS-specific 2D encoder, as shown in Figure 3, following CasMVSNet [17]. This encoder follows the design in [32] producing a feature pyramid with increasing spatial resolutions at 3 scales.

We augment the geometry-aware features of the MVS encoder with semantics-aware features. The Segment Anything Model (SAM) [28] was very recently proposed as a foundation model for open set segmentation. It uses the Vision Transformer (ViT) [11] as a feature encoder. Given a high resolution (normally, with a long side of 1024 pixels) RGB image at the input, the ViT encoder breaks it into tokens that have a spatial size of 16×16 pixels, producing a grid of tokens. The transformer is applying multiple layers of non-local attention, keeping the spatial resolution fixed throughout all layers, producing a single rich segmentation-aware feature tensor. We note that the spatial resolution of these features roughly corresponds to the coarsest resolution of the MVS pyramid, and we resize the tensor with bilinear interpolation to that resolution. We additionally create a lightweight convolutional pyramid encoder for the SAM features to match the spatial and feature dimensions at all three scales of the MVS pyramid.

The geometry-aware MVS features and the SAM seman-

tic features are fused together with a simple pointwise summation at all scales of the pyramid, same as it is done in other works (e.g. in FuseNet [20] for geometry-aware depth features and the visual RGB features). In our preliminary experiments other approaches, such as stacking didn't show any performance improvement.

3.2. Cost volumes and depth prediction

Once the features are extracted from individual views, a range of depth hypotheses are created for each pixel of the reference view $i = 0$. Features from source view i are warped with each depth hypothesis d using a homogeneous coordinate mapping:

$$\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} \sim K_i R_i^T \left(R_0 K_0^{-1} d \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} + t_0 - t_i \right) \quad (1)$$

This mapping creates per-view 3D feature volumes V_i in the reference coordinate view. All V_i are aggregated into a cost volume C using the variance-based cost metric:

$$C = \frac{1}{M} \sum_{i=1}^{M} \left(V_i - \overline{V} \right)^2 , \quad (2)$$

where $\overline{V}$ is the average feature volume. Such a 3D cost volume is created at a range of scales, as proposed in CasMVSNet [17], and decoded into the single-channel 3D depth probability volume P with 3D convolutions. When the probability volume is normalized, it defines a distribution over the range of depth hypothesis, and the depth prediction for the reference frame at pixel (i, j) is inferred as:

$$D_0^{i,j} = \sum_t P^{i,j,t} d_t \, . \quad (3)$$

At the coarsest scale, the entire depth range is covered for some fixed $[d_{min}, d_{max}]$ with equally spaced hypotheses d_t. Once the coarser depth is predicted at one scale, it is interpolated to a finer scale to serve as a mean of the new (narrower) depth hypothesis range, allowing for higher resolution of the hypothesis interval. The final depth prediction D_0 is performed at the finest scale that has the same spatial $W \times H$ resolution as the input image.

3.3. Semantic segmentation

We use the SAM-pretrained encoder in Section 3.1 and modify the mask decoder since our semantic segmentation task is different from the open-set instance segmentation in SAM. The mask decoder transforms the input image embedding and a set of prompt embeddings into a set of output masks. As a prediction is required for each of the K semantic classes, we introduce K learned embeddings to serve as prompts, analogous to SAM's *sparse prompts*.

Additionally, the depth maps are available from the MVS prediction branch, which we use similarly to SAM's *dense prompts*. A dense embedding is created with a shallow convolutional network that aligns with the image feature representation and is summed point-wise with the semantic encoder's features as shown in Figure 3. We note that other learned dense prompts could be used as well in addition to our depth maps. For example, one may think of using the SAM's zero-shot edge maps, however such an approach would be very computationally expensive (SAM needs to decode 768 masks and run NMS and Sobel filter post-processing to produce the edge map). We limit this study to only our predicted depth prompt.

The decoder is represented by a lightweight two-way transformer that communicates information between the grid of image tokens and a set of queries, followed by a dynamic mask prediction block. It is depicted in Figure 4. Each of the two transformer blocks starts with token self-attention, followed by token-to-image cross-attention (tokens serving as attention queries), a point-wise MLP that updates the tokens, and finally image-to-token cross-attention (with image embeddings serving as attention queries). Thus, each of the transformer blocks updates both the tokens (entangling tokens via self-attention and aggregating image features through token-to-image attention), and the image features (infusing the tokens information through image-to-token attention). Standard residual connections and layer normalizations are employed after each of the four steps in each block that are omitted from Figure 4 to avoid clutter. Image features are also supplied with positional encodings that are added pointwise at the input of each attention layer.

The resulting image features are upscaled with a shallow 2-layer deconvolutional network to match the $W \times H$ resolution of the system's input image, and tokens are additionally embedded with a token-to-image attention block and an MLP to produce the semantic class queries. Semantic masks are decoded through a cross-product between the image embeddings and the class queries.

Loss function We minimize the joint loss function:

$$L = L_{seg} + \alpha L_{MVS}, \quad (4)$$

where L_{seg} is a cross-entropy semantic segmentation loss, and L_{MVS} is a smooth L1 depth loss.

4. Experiments

4.1. Training settings

Models are trained with the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $weight\_decay = 10^{-2}$), a learning rate of 10^{-3}, and a total batch size of 8 distributed over 4 GPUs, for 16 epochs. With most of the SAM encoder parameters frozen, we optionally tune the last $0 - 3$ blocks of the SAM

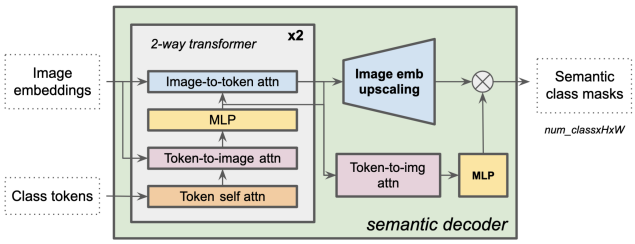

Figure 4. Semantic decoder. Two-way attention blocks are used to simultaneously update the class tokens and the image embeddings. Image embeddings are then upscaled, while the final MLP transforms the class tokens into queries that decode the semantic segmentation masks. Decoding is done by correlating the queries with the final image embeddings at each spatial location.

Method	#views	Abs (cm) ↓	Rel (%) ↓
MVS2D [54]	3	10.8	5.9
TransMVSNet [10]	3	13.2	8.3
CasMVSNet* [17]	3	10.0	6.0
Ours, MVS + UNet feat	3	9.5	5.8
Ours, MVS + SAM feat	3	9.3	5.5
Ours, full	3	9.0	5.6
Ours, full	5	8.3	5.2

Table 1. Multi-view depth estimation. We follow the same cost volume-based cascaded architecture as CasMVSNet. "MVS + UNet features" and "MVS + SAM features" are single-task depth prediction models that augment the 2D MVS geometric features with semantic features (see Sec 3.1) and do not have a semantic decoder. Our full model jointly predicts depth and semantic segmentation and fine-tunes the last layers of the SAM encoder.

encoder with a $10\times$ smaller learning rate (i.e. 10^{-4}). The input image size is set to 320×256, and both depth and segmentation maps are predicted at this resolution. We use a smooth L1 loss between the ground truth and the predicted depth with $\beta = 0.02$ (i.e. the threshold is set at the error value of 2 cm), ignoring values outside of the depth range. Semantic and depth loss contribute equally ($\alpha = 1$ in Eq. 4).

4.2. Architecture settings

ViT-base is used as a semantic feature extractor for SAM.

A cascade of cost volumes is constructed to predict depth at 3 scales. The depth is predicted across the range of $[0.1m, 5.0m]$, assuming 192 virtual depth bins. This corresponds to the base interval of roughly $2.5cm$ between the depth hypotheses. The network spaces the intervals with a ratio of $[4\times, 1\times, 0.5\times]$ of the base interval at its three scales, evaluating $[48, 32, 8]$ depth hypotheses at each pixel in the corresponding scales. Unless explicitly specified, our networks use $M = 3$ input RGB views.

Method	mIoU ↑	Abs ↓	Rel ↓	RMSE ↓
MT RefineNet [37]	42.0	32.4	21.7	44.3
PAD Net [49]	42.2	26.9	17.4	37.8
MTI Net [46]	53.7	22.4	14.0	32.2
MTAN [34]	56.1	22.7	14.6	32.6
Ours, 3 views	62.5	9.0	5.6	14.9

Table 2. Comparisons to monocular multitask methods for joint depth prediction and semantic segmentation.

4.3. Datasets

We evaluate our results on the ScanNet [8] dataset. ScanNet is a large-scale established indoor recognition benchmark. It contains 1513 scans across 707 locations. Semantic segmentation tasks includes 20 categories. For most of our studies (e.g. Tables 1, 2, 4) we use the train/validation split provided by MVS2D [54] to benchmark the results of the methods on the same sets of views. For other experiments, the official split of the ScanNet v2 scenes into training and validation sets is used.

4.4. Metrics

We use the standard metrics for evaluating the predicted depth maps [15]: mean absolute depth error (Abs err), mean relative depth error (Rel err), root mean square error (RMSE). Absolute depth error and RMSE can are in centimeters, relative depth error is measured with respect to the ground truth. Lower is better for all depth evaluation metrics. Segmentation outputs are evaluated by the mean intersection-over-union (mIoU). Higher is better.

4.5. Depth prediction

CasMVSNet [17] and TransMVSNet [10] serve as our MVS baseline models for depth prediction. We use the train/validation split released in MVS2D [54] that includes both the scene partitions, and the view sampling, in order to perform a fair comparison of the M-view MVS methods against each other. MVS2D is trained using the parameters reported in [54]. The paper's codebase is publicly available and results are reported on ScanNet. CasMVSNet and TranMVSNet are trained using our custom settings described above. In Table 1 we analyze the benefits of using semantic features to augment the geometric MVS features during cost volume construction. Our baseline CasMVSNet has an absolute depth error of 10.0, and MVS2D shows as 10.8. TransMVS, although reported as a superior model on the standard MVS datasets [26, 30], shows a worse generalization to ScanNet with an absolute depth error of 13.2.

We first demonstrate the benefit of using standard encoder-decoder semantic features. A U-Net [39] with a ResNet [21] backbone is pre-trained on the same ScanNet

Method	Modality	#views	mIoU ↑
SegNet [2]	RGB	1	27.5
DeepLab v2 [5]	RGB	1	43.9
DeepLab v3 [6]	RGB	1	50.1
AdapNet++ [44]	RGB	1	53.0
FuseNet [20]	RGB-D	1	63.8
EMSAFormer [16]	RGB-D	1	63.8
SSMA [44]	RGB-D	1	67.4
Ours	RGB	5	62.1

Table 3. 2D semantic segmentation on ScanNet validation set.

train split, and the three scales of the decoder pyramid that correspond to the MVS decoder spatial scales are used instead of our SAM encoder in Figure 3. The semantic prediction heads are discarded in this experiment, and all the remaining U-Net weights are tuned during the training, improving the depth prediction performance to 9.5.

Augmenting 2D features with the SAM encoder output brings the error down to 9.3cm even without any encoder tuning, while training our full model with the segmentation branch and tuning of the last 3 SAM encoder layers demonstrates an additional improvement with an error of 9.0. Note that although the segmentation decoder outputs are not influencing the MVS branch during inference, during training the gradients still propagate back to the SAM encoder features, serving as an additional regularizer. Note that the the evaluation is performed against the ground truth depth maps that are also noisy and imprecise, and thus the lower bound for a method's performance is greater than zero, and unknown, as discussed below in Section 4.6. Interpreting the errors' absolute values is therefore not straightforward without looking at some qualitative results.

In Table 2 we compare with monocular multitask methods for joint depth prediction and semantic segmentation with open-source implementation. We re-train the methods on our train/val split of ScanNet to report the results. We notice that the results are similar to those reported in the papers for NYUD2, since the datasets have similar data distributions. For example, Multi-Task RefineNet [37] reports 42.02 mIoU and 56.5 RMSE on NYUD2, while our performance on ScanNet shows 42.0 mIoU and 44.3 RMSE.

4.6. Qualitative studies

Table 6 shows the depth estimation results for three methods: multitask monocular MTAN [34], multi-view CasMVSNet [8], and our model. Both CasMVSNet and our model are trained and tested using only $M = 3$ views. While from Table 2 the improvement of our method is obvious compared to monocular methods, there is still a question of how much we really improve over the baseline MVS in terms of depth prediction (i.e. is 9.0 Ours vs 10.0 CasMVSNet in Table 1 a considerable difference or not). As we already stated above, due to the fact that the ground truth is imprecise, the lower bound for such metrics as the absolute depth error are nonzero and unknown. Comparing our method to the baseline MVS visually we notice that it performs consistently better across the dataset, with a considerable difference in some cases. In the first column the effect on flat matte structures, such as walls, is pronounced, where semantic cues may help to recover the known MVS failures. In the last column, an unconventional view and the glossy floor surface make it hard to infer the depth from geometric patterns. Semantic meaning identifies the objects as table and chair legs, making it easier to determine boundaries.

Method	Modality	mIoU ↑
UNet (ResNet-34)	RGB	50.9
UNet (ResNet-50)	RGB	34.5
SAM (frozen) + conv decoder	RGB	27.2
SAM (frozen) + our decoder	RGB	47.2
SAM (tuned) + our decoder	RGB	59.0
Our full model	RGB	62.5
SAM (frozen) + our dec. (GT depth)	RGB-D	52.4
SAM (tuned) + our dec. (GT depth)	RGB-D	63.4

Table 4. 2D semantic segmentation ablations. Conventional methods such as UNet overfit easily when only using ImageNet-pretrained backbones. SAM pretrained features are much more robust, but produce poor performance when decoded using convolutional layers. Using a transformer-based decoder improves the quality considerably, especially with encoder feature tuning. Our full model uses the depth predicted from the MVS branch and comes close to matching the performance of the RGBD version that uses ground truth depth maps as a decoder prompt.

Method	Abs (cm) ↓	Rel (%) ↓	mIoU ↑
CasMVSNet	10.0	6.0	NA
MVS + SAM features	9.3	5.5	NA
Semantic SAM (frozen)	NA	NA	47.2
Semantic SAM (tuned)	NA	NA	59.0
Ours, no SAM→MVS	10.0	6.1	60.0
Ours, full	9.0	5.6	62.5

Table 5. Model ablations. CasMVSNet is the baseline depth prediction approach. MVS + SAM features is our model without the segmentation head (i.e. *semantic features help depth prediction*). Semantic SAM is our semantic decoder on top of the SAM encoder with no depth prediction and no depth prompting. "Ours, no SAM→MVS" cuts the SAM to MVS feature connection, i.e. depth prediction is performed without SAM features, while still running segmentation with the depth prompt.

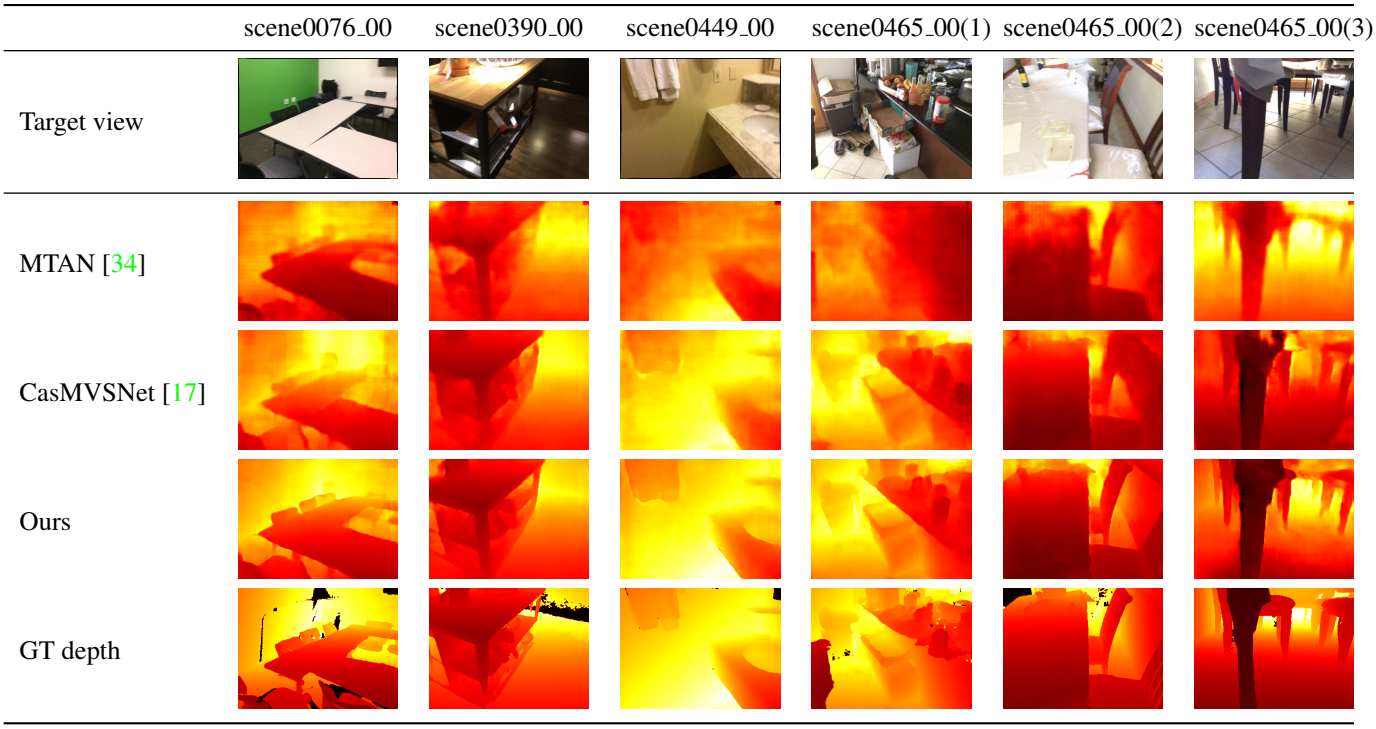

Table 6. Depth estimation results on ScanNet. Existing MVS techniques, MTAN and CasMVSNet, perform poorly on textureless flat surfaces like walls (scene0076_00, scene0449_00) and tables (scene0390_00,scene0465_00(2)) and on thin structures with strong occlusions like legs of tables (scene0390_00,cene0465_00(2)) and clutter on kitchen counter top (scene0465_00(1)). In contrast, our proposed technique produces significantly better performance on these regions by leveraging semantic understanding of the scene. Both CasMVSNet and Ours are trained using $M = 3$ views on the MVS2D [54] split.

4.7. Segmentation

Table 4 offers a study of our SAM-based semantic segmentation. We first establish a benchmark with UNet-type encoder-decoder architectures using ResNet as a backbone. We notice that deeper models overfit on ScanNet since the data variety is limited. In particular, the ResNet-34 based method outperforms the deeper ResNet-50 version. The SAM-pretrained transformer proves to be much more robust. However, a simple convolutional decoder delivers poor performance (27.2 mIoU). The transformer-based decoder described in Section 3.3 significantly improves the mIoU to 47.2 even with the frozen encoder, while additional feature tuning reaches the quality of 59.0 mIoU. In our full model that jointly predicts depth and segmentation maps, the predicted depth is used as a dense prompt to the decoder, for which the performance of 62.5 mIoU is reported in the table. The methods marked as RGB-D in Table 4 use the ground truth depth as the input. Our multi-view RGB model gets close to matching the corresponding RGB-D setting.

4.8. Ablation studies

Table 5 summarizes the effects of individual components of our framework. CasMVSNet sets the absolute depth error baseline of 10.0 without the use of any semantic information; our Semantic SAM is trained in isolation (with no MVS branch and no depth prompt) setting the segmentation baseline at 47.2 mIoU and 59.0 with encoder tuning. MVS + SAM features represent our model without the semantic head, improving the depth prediction. Our full model shows an improvement in both depth prediction (9.0 absolute error) and semantic segmentation (62.5 mIoU).

5. Conclusion

In this work, we introduced a novel approach to joint depth and segmentation estimation in the multi-view setting, the first to our knowledge. We show that jointly estimating depth and segmentation is better than separately predicting them in multiview settings in Table 1 and 3 respectively. This discovery in itself is not new and has been previously explored for multi-task learning in single-view monocular depth and segmentation estimation. However, we show that multi-task learning in multiview is significantly better than single-view approaches in Table 2, highlighting the importance of this work to extend multi-task learning from single-view to multi-view. Multi-view-based 3D scene understanding approaches that require only 3-5 images are significantly more practical in many robotics applications compared to approaches that require a full-dense 3D scan of the room.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5861–5870, 2023. 3, 4

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2, 3, 7

[3] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2, 3

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2, 3, 7

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 3, 7

[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. 1

[8] 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. 2, 4, 6, 7

[9] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 1, 2, 3

[10] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 2, 3, 4, 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[13] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15324–15333, 2021. 1

[14] 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. 3

[15] 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. 6

[16] Söhnke Benedikt Fischedick, Daniel Seichter, Robin Schmidt, Leonard Rabes, and Horst-Michael Gross. Efficient multi-task scene analysis with rgb-d transformers. *arXiv preprint arXiv:2306.05242*, 2023. 1, 7

[17] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2, 3, 4, 5, 6, 8

[18] 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. 1

[19] Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3d scene reconstruction and class segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 97–104, 2013. 2, 3

[20] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, pages 213–228. Springer, 2017. 1, 5, 7

[21] 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. 4, 6

[22] Lei He, Jiwen Lu, Guanghui Wang, Shiyu Song, and Jie Zhou. Sosd-net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing*, 440:251–263, 2021. 3

[23] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 2

[24] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14373–14382, 2021. 1

[25] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 3

[26] 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. 2, 6

[27] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 1

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4

[29] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 582–600. Springer, 2020. 1

[30] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2, 6

[31] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 3

[32] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[33] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *arXiv preprint arXiv:2302.06556*, 2023. 3

[34] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 1, 2, 3, 6, 7, 8

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[36] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020. 1, 3

[37] Vladimir Nekrasov, Thanuja Dharmasiri, Andrew Spek, Tom Drummond, Chunhua Shen, and Ian Reid. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7101–7107. IEEE, 2019. 1, 2, 3, 6, 7

[38] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21477–21487, 2023. 3

[39] 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, 3, 6

[40] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3

[41] Elisavet Konstantina Stathopoulou, Roberto Battisti, Dan Cernea, Fabio Remondino, and Andreas Georgopoulos. Semantically derived geometric constraints for mvs reconstruction of textureless areas. *Remote Sensing*, 13(6):1053, 2021. 4

[42] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[43] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1, 3

[44] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020. 1, 2, 3, 7

[45] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021. 1

[46] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 527–543. Springer, 2020. 1, 2, 3, 6

[47] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 541–550, 2020. 3

[48] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2809, 2015. 3

[49] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 1, 2, 3, 6

[50] Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3030–3038, 2021. 4

[51] Qingshan Xu, Weihang Kong, Wenbing Tao, and Marc Pollefeys. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4945–4963, 2022. 2

[52] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019. 2

[53] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 636–651, 2018. 4

[54] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8574–8584, 2022. 4, 6, 8

[55] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2, 3

[56] Zhenlong Yuan, Jiakai Cao, Hao Jiang, Zhaoqi Wang, and Zhaoxin Li. Tsar-mvs: Textureless-aware segmentation and correlative refinement guided multi-view stereo. *arXiv preprint arXiv:2308.09990*, 2023. 4

[57] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 1

[58] Dongxu Zhao, Daniel Lichy, Pierre-Nicolas Perrin, Jan-Michael Frahm, and Soumyadip Sengupta. Mvpsnet: Fast generalizable multi-view photometric stereo. *arXiv preprint arXiv:2305.11167*, 2023. 3

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 3

[60] 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. 1