Dionysus: Recovering Scene Structures by Dividing into Semantic Pieces

Likang Wang[1], Lei Chen[1,2]
[1]Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
[2]Data Science and Analytics Thrust
The Hong Kong University of Science and Technology (Guangzhou)
lwangcg@connect.ust.hk, leichen@cse.ust.hk

Abstract

Most existing 3D reconstruction methods result in either detail loss or unsatisfying efficiency. However, effectiveness and efficiency are equally crucial in real-world applications, e.g., autonomous driving and augmented reality. We argue that this dilemma comes from wasted resources on valueless depth samples. This paper tackles the problem by proposing a novel learning-based 3D reconstruction framework named Dionysus. Our main contribution is to find out the most promising depth candidates from estimated semantic maps. This strategy simultaneously enables high effectiveness and efficiency by attending to the most reliable nominators. Specifically, we distinguish unreliable depth candidates by checking the cross-view semantic consistency and allow adaptive sampling by redistributing depth nominators among pixels. Experiments on the most popular datasets confirm our proposed framework's effectiveness.

1. Introduction

Recovering 3D structures from 2D images is one of the most fundamental computer vision tasks [7, 25, 26, 68] and has wide applications in various scenarios (e.g., autonomous driving, metaverse, and augmented reality). Thanks to the popularity of smartphones, high-quality RGB videos are always obtainable. Since each image describes only a tiny piece of the whole scenario [31, 33], reconstruction from multiple frames [44, 50] is more attractive than from a single image [13, 14]. Although model quality and real-time response are equally essential, the existing 3D reconstruction methods have difficulty performing well in both aspects. For example, multi-view stereo (MVS) models [63, 73] consume seconds on each image, while real-time approaches [9, 57, 64] lead to missing details or large-area mess.

The mainstream reconstruction methods [15, 63, 72] generally sample a set of candidate depths or voxels, then use neural networks (e.g., CNNs [21] and transformers [58]) to evaluate each candidate's reliability. The candidate number is generally small because of the evaluation networks' high computational demand. Consequently, the reconstruction quality becomes unsatisfying because the ground truth is less likely to be sampled.

Pursuing a higher candidate sampling density, CasMVSNet [15] shrinks the depth range in a coarse-to-fine fashion; IS-MVSNet [63] searches for new candidates according to the estimated error distribution; NeuralRecon [57] prunes the voxels thought to be unreliable in the previous predictions. All these methods select candidates according to the initial depth estimations. Notably, small objects and delicate structures are hard to recover in the beginning low-resolution phase because they require a high resolution to distinguish. As a result, the final reconstruction often suffers from missing objects and crude details because the actual depth can be outside the final search range when the initial estimation is unreliable. However, decent reconstruction quality on delicate objects is crucial to many real-world applications. Specifically, traffic accidents may occur if any pedestrian or warning post fails to recover; frequent collisions and even fire disasters might happen if cleaning robots cannot well reconstruct table legs and electric cables. Besides the defects in meticulous structures, coarse-to-fine models also have room to improve efficiency. As mentioned, the mainstream coarse-to-fine methods sample and then evaluate initial candidates multiple times to locate the most valuable candidate depths. Notably, examining the preliminary candidates may be more expensive (usually two times more [15, 78]) than assessing the final nominators.

In addition to the costly evaluation networks widely adopted in coarse-to-fine architectures, another natural so-

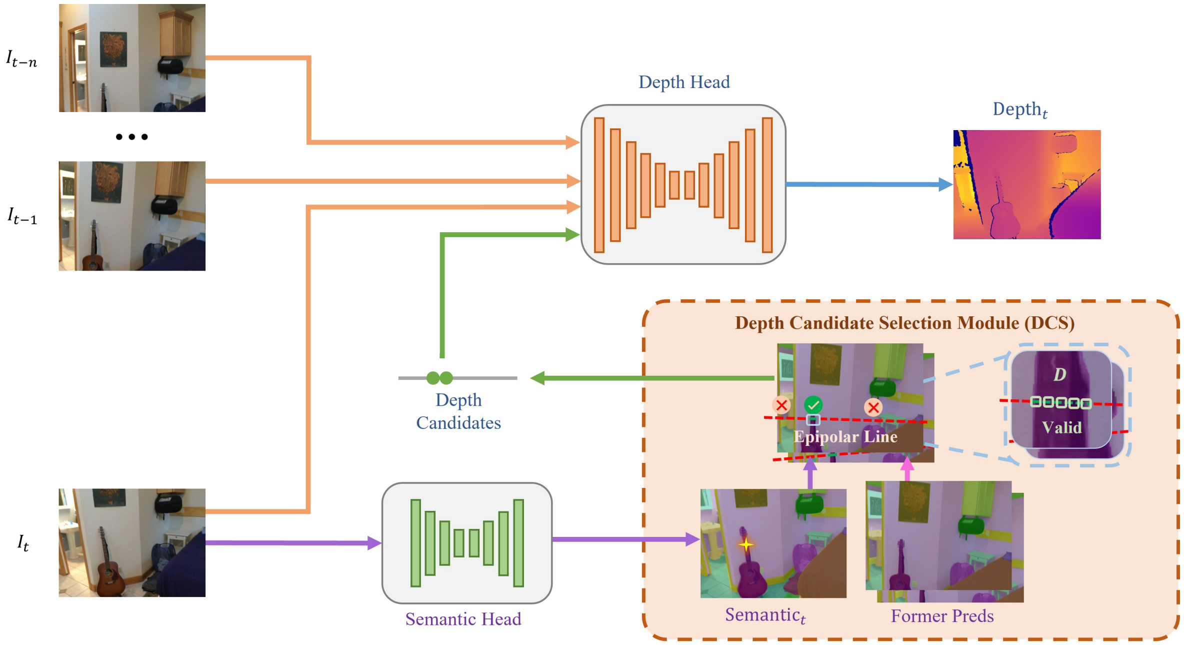

Figure 1. **The overall architecture of our framework.** We first estimate a semantic map for the current frame and then retrieve the former semantic estimations. After that, we locate the most valuable depth candidates through the semantic maps. Finally, we predict the depth map for the current frame by examining each depth candidate.

lution is to measure each depth candidate's reliability according to the photometric consistency among RGB images or feature maps from different views. The basic logic behind these methods is that pixels from distinct perspectives should own the same color if they are the correct projections from the same 3D point. However, a 3D point may look distinct in different views due to illumination, transparency, and reflection. Moreover, candidates close to but not precisely coinciding with the ground truth may have different colors for delicate objects, thus leading to low photometric consistency. Consequently, the found candidate may correspond to a pixel distant from the ground truth but of a similar color.

To get rid of detail defects and keep efficiency, it becomes crucial to accurately and efficiently find the most promising candidate depths. This paper proposes a novel high-quality real-time 3D reconstruction framework named Dionysus, which looks for depth hypotheses based on semantic estimations. Precisely, we distinguish each depth candidate's reliability according to its semantic consistency among warped views. Fig. 1 illustrates the overall architecture of our framework. Our intuition is that two pixels should share the same semantic label if they are projections of the same 3D point. We argue that selecting depth candidates according to the semantic tags results in various benefits:

1. **Consistent:** A 3D point may have distinct colors when observing from different perspectives, but its semantic label never changes.
2. **Efficient:** Semantic estimation requires a meager cost. A semantic segmentation model may spend only milliseconds on each frame [11,47] while evaluating each cost volume in coarse layers generally takes ten times more cost. In addition, most hierarchical methods shrink the depth range only by half in each stage, while our method may significantly reduce the depth range on delicate objects (e.g., desk legs).
3. **Dense:** Tiny objects always get missing in hierarchical frameworks because the initial samples are too sparse. However, the semantic maps are highly dense, thus retaining fine details.
4. **Adaptive:** Semantics indicate various properties of the pixel. For example, an electric cable may demand highly dense samples to recover, but a wall may not. In other words, we can assign each pixel the most suitable sampling density according to its semantic estimation.

Depth Reassignment: A bed pixel may have much more valid depth candidates than a pen pixel after the semantic-based pruning because there are likely many bed pixels in other views. Consequently, we cannot form a regular cost volume based on the pruned depth candidates because pixels may have different numbers of valid depth candidates. However, most GPUs are not designed for irregular inputs and are inefficient in such cases. Thus, we further propose reallocating depth samples from pixels with excess valid candidates to those in the opposite. In this way, all pixels finally have the same number of candidates and thus can be conveniently computed by GPUs. Moreover, we sample more densely for delicate structures, otherwise more sparsely, because tiny objects have a narrower depth range after semantic pruning, and all pixels have the same depth number.

To summarize, this paper has two significant contributions:

1. Instead of densely evaluating all depths or sparsely evaluating limited depths, we efficiently select the most promising depth candidates based on the cross-view semantic consistency.

2. We reallocate depth samples among pixels to make our model adaptive to objects and efficient on GPUs.

The mentioned contributions significantly benefit the effectiveness while retaining efficiency. Our extensive experiments on the most popular 3D reconstruction datasets [5] further verify our proposed method's validity.

2. Related Work

We first introduce the development of 3D reconstruction, then present the progress in semantic segmentation, and finally discuss the existing semantic-based 3D reconstruction networks.

2.1. 3D Reconstruction

Recovering 3D structures from images has been widely studied for many decades. Despite the great success of traditional methods [1,12,52,53], most recent solutions [9,15,73] are based on neural networks [21,35–38,62] because of the latter's popularity [27,32,71,75,76] in the last decades. Thanks to the popularity of smartphones, continuous video sequences can now be easily captured. Compared to independent images, video sequences [28, 80] contain rich correlations among frames, thus leading to more consistent reconstruction quality. In consequence, learning-based video reconstruction [44,56,69] is becoming a hot spot. However, due to their high computation demand, most video reconstruction methods are limited in offline applications.

Real-time video reconstruction is attracting wide attention from the academy and the industry, considering its vital position in autonomous driving and augmented reality. The mainstream real-time methods can be categorized into depth-based [9, 16, 40, 61] and volume-based [57]. Specifically, the former estimates depth maps and fuses them into 3D representations. In contrast, the latter directly regresses 3D representations from the RGB inputs. Depth-based methods generally lead to lower latency because they immediately process each new frame without waiting. Conversely, volumetric methods estimate 3D representations from a group of frames, thus cannot respond to new frames at once and are less suitable for online scenarios. The mainstream depth-based approaches usually uniformly sample depth candidates from an overly broad range [9] or gradually shrink the search range by repeated predictions [15,72]. Our main difference from the existing depth-based methods is that we utilize semantics to select depth candidates.

2.2. Semantic Segmentation

Semantic segmentation focuses on estimating each pixel's object category (e.g., whether it represents a dog). Most semantic segmentation networks [2,51,67,74,77] inherited a learning-based framework from FCN [39]. Real-time solutions generally put their most effort into designing a lightweight backbone considering its contribution to the whole model's complexity. The existing online methods generally own three backbone types [43]: classification-based [22,24,60], segmentation-based [23,34,41], and two-branch-based [48,74,79].

The first class encodes features using image classification networks and then decodes to output semantics. The second class designs segmentation-specific backbones considering the task difference between segmentation and classification. The third class processes macro and micro information separately to reduce the overall cost. State-of-the-art segmentation models can infer hundreds of frames per second and retain high quality simultaneously. In comparison, the fastest depth estimation networks are bounded to dozens of frames per second.

2.3. Semantic-based 3D Reconstruction

Quite a few papers focus on integrating semantics into depth estimation networks. Most related works belong to single-image depth estimation and stereo matching.

In single-image depth estimation, view correspondence does not exist; thus, semantic information becomes more vital. [66] simultaneously estimates depth and semantic maps by multi-task learning. [65] shares depth weights among pixels of the same semantic label. [10] jointly estimates depth, semantics, and surface normal. [20] introduces a synergy network to share features between the semantic and depth branches. [30] restricts depth predictions by semantic priors (e.g., the sky is distant, and the ground is horizontal). [59] infers invisible regions from semantics.

In stereo matching, two views with fixed relative poses are given, and the goal is to match pixels in the two views. These papers can be generally divided into two categories. The first class treats semantic maps as additional channels. [8] refines disparities by concatenating semantic and disparity features along the channel dimension. [70] constructs cost volumes for semantics and disparity and then

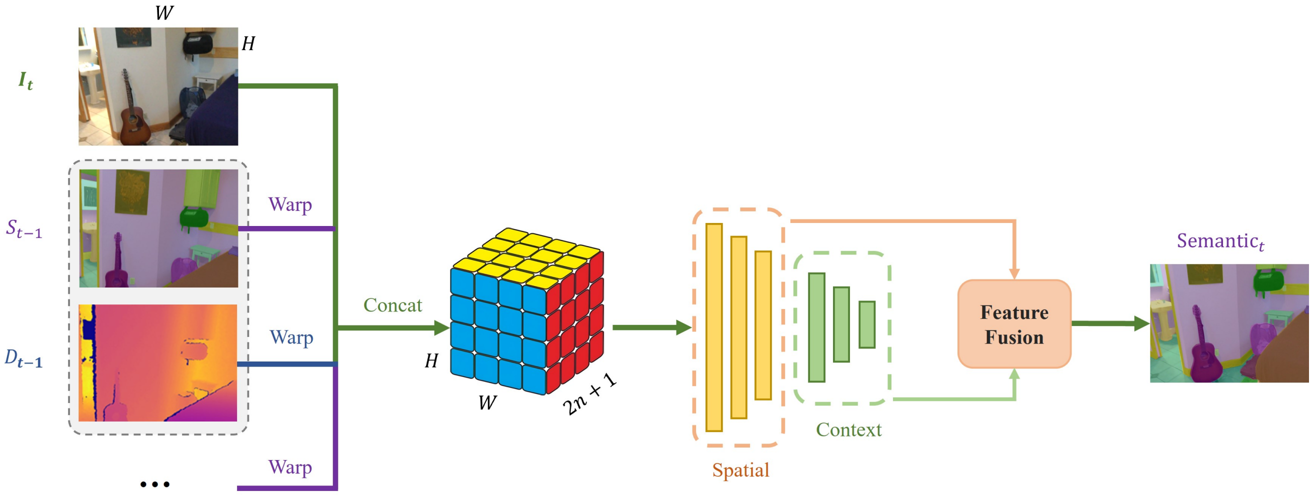

Figure 2. **Architecture of the semantic head.** Former semantic and depth estimations are warped and then integrated with the current frame. After that, we extract hierarchical information from the concatenated feature map. The semantic map for the current frame is inferred from the feature volume composed of spatial and context information.

fuses them by summing and global pooling. [3] constructs cost volume using semantic consistency between left and right images and estimates disparity residuals according to semantic category. The second class revises the disparity using priors in semantics. [49] filters depth maps based on semantic priors. For example, ground and water pixels with heights higher than 90% get pruned. [46] refines edge and texture-less regions using semantic cues. [4] estimates disparity confidence using semantic consistency and refines disparity in each semantic region.

The related works in multi-view stereo are limited. [18] recovers a segmented 3D representation from 2D images and semantics. [42] augments depth edges and guides cost aggregation using semantics.

Unlike the mentioned works, we select depth candidates by explicitly exploiting semantic consistency for multi-view inputs. We also balance depth candidates among pixels and adaptively attend to pixels in need.

3. Methodology

We first introduce the overall architecture of our model in Sec. 3.1, then elaborate on each sub-component in Sec. 3.2 and 3.3. Our model inputs an RGB video sequence divided into frames and the camera parameters of each frame, which are estimated by off-the-shelf tools (e.g., ARKit). The output is expected to be the depth map of each frame.

3.1. Overall Architecture

Fig. 1 exhibits the overall framework. We have a semantic head to predict a semantic map S_t for the current frame I_t. At the same time, we retrieve the semantic maps $\{S_i\}_{i=t-n}^{t-1}$ predicted in the previous timestamps from a semantic queue.

After that, we look for the most promising depth candidates for each pixel in I_t using our novel semantic-based depth candidate selection module, which is our main contribution. Finally, we examine each depth nominator using a depth head to output the final depth estimation. We sum losses from the depth and semantic heads as the overall loss.

3.2. Semantic-based Depth Candidate Selection

This section introduces our main contributions. Sec. 3.2.1 first shows how to prune unreliable depth candidates using the semantic information, and then Sec. 3.2.2 demonstrates how to concentrate on pixels in need and become efficient on GPUs.

3.2.1 Semantic-based Depth Pruning

To estimate the depth of a pixel in the current frame I_t, we sample a set of depth candidates and evaluate their reliability one by one. The candidate number must be high enough to hit the ground truth.

However, the existing learning-based methods can only examine a small number of samples because they use a giant neural network to assess each candidate. Consequently, small objects distant from the background and delicate structures get defective. The small object problem cannot be solved via coarse-to-fine architectures because significantly wrong predictions in the coarse stages cannot be recovered in finer stages.

Unlike the existing methods, we first predict the semantic labels for each pixel at a low cost; then, we use the se-

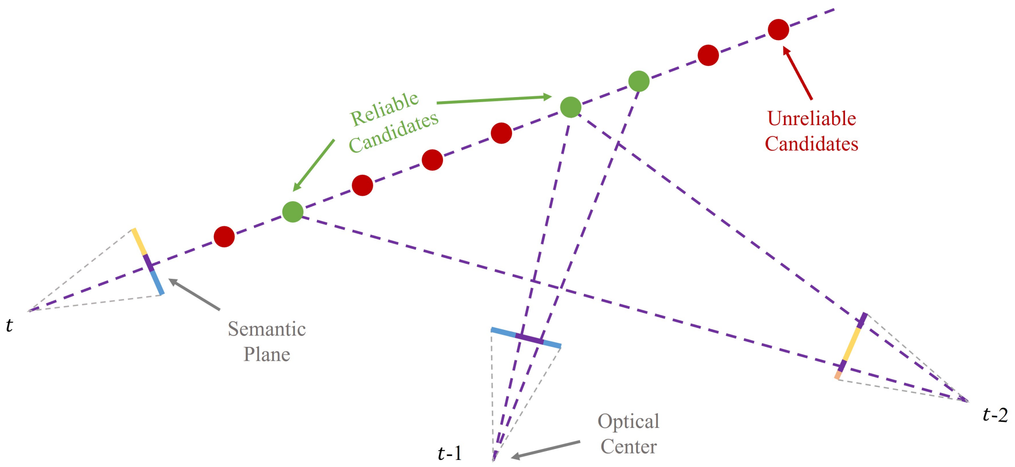

Figure 3. **Illustration of depth pruning.** We uniformly sample depth candidates for a given pixel (purple) in the current frame, then evaluate each candidate to determine whether it leads to a pixel of the same semantic label projected on the former frames.

mantic maps to restrict the depth candidate sampling only to the most reliable regions.

I/O: The inputs include the current frame I_t and retrieved previous semantic $\{S_i\}_{i=t-n}^{t-1}$ and depth $\{D_i\}_{i=t-n}^{t-1}$ estimations. Besides, similar to most existing methods [9, 57, 73], we assume the camera parameters from time $t-n$ to t are available. In fact, the camera parameters can be obtained from off-the-shelf tools (e.g., ARKit).

The outputs are the most promising depth candidates for each pixel in the current frame I_t.

Semantic Head: Instead of outputting the correct label, semantic consistency among frames is more crucial because we examine depth candidates by cross-view conformity. Thus, we need to consider previous semantic estimations when processing new frames.

We propose a highly efficient 2D convolutional network to predict the semantic labels for each pixel in I_t, as shown in Fig. 2. The inputs of the semantic prediction head include the current frame I_t, the semantic and depth maps predicted from timestamp $t-n$ to $t-1$, and the corresponding camera parameters.

The former depth and semantic predictions are warped to the current view according to the relative poses and then fused by concatenating along the channel dimension. Inspired by STDC-Seg [11] and BiSeNet [74], we adopt a two-pathway architecture to efficiently extract spatial and context information, which are then fused to generate the final semantic estimation.

Depth Pruning: After obtaining the semantic maps, we start to prune depth candidates. According to the camera parameters from time $t-n$ to t, we can draw a straight line (i.e., the epipolar line) on each previous frame $\in \{I_i\}_{i=t-n}^{t-1}$ for each pixel $p_t \in$ the current frame I_t and guarantee that the correct correspondences can only lie on the epipolar line.

Then, we sample N_s (a significant number) depth candidates for each pixel in I_t, as shown in Fig. 3. Each depth corresponds to one pixel on the epipolar line in each former frame. After that, we check whether the projected pixel's semantic label is exactly the same as p_t. Only the candidate depths leading to a positive answer get retained, otherwise pruned.

Consequently, each former frame provides a set of reliable depth candidates. These sets of candidate overlap yet may differ due to occlusion and field views. We use the union of all the sets as the aggregated depth candidate set because each pair of matched pixels may imply the existence of a 3D point.

Since we judge the reliability based on semantics, a question might be whether our framework relies heavily on each pixel's precise prediction. The answer is no because we only require semantics in different frames to be consistent instead of strictly being the ground truth. Moreover, the existing semantic segmentation architectures [11, 47] can provide impressive enough label estimations.

3.2.2 Semantic-based Depth Reallocation

After pruning unreliable candidates, pixels may own inconsistent numbers of depth nominators because different objects might occupy distinct amounts of space, as shown in the left of Fig. 4. Since the initial depth candidates are uniformly sampled, a pixel owning many reliable depth nominators likely belongs to a large object (e.g., walls and floors). These objects usually do not require dense samples

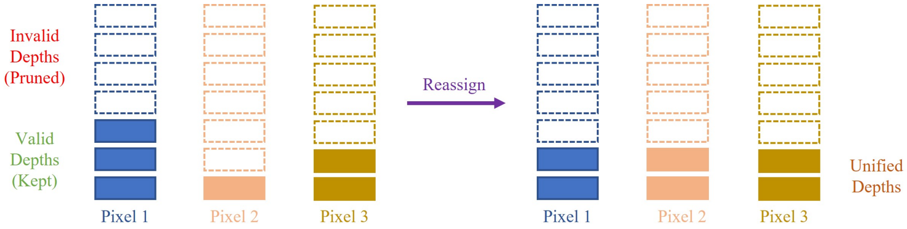

Figure 4. **Reassignment among pixels.** Pixels may have a different number of candidate depths after semantic pruning. However, the difference disappears after reallocation.

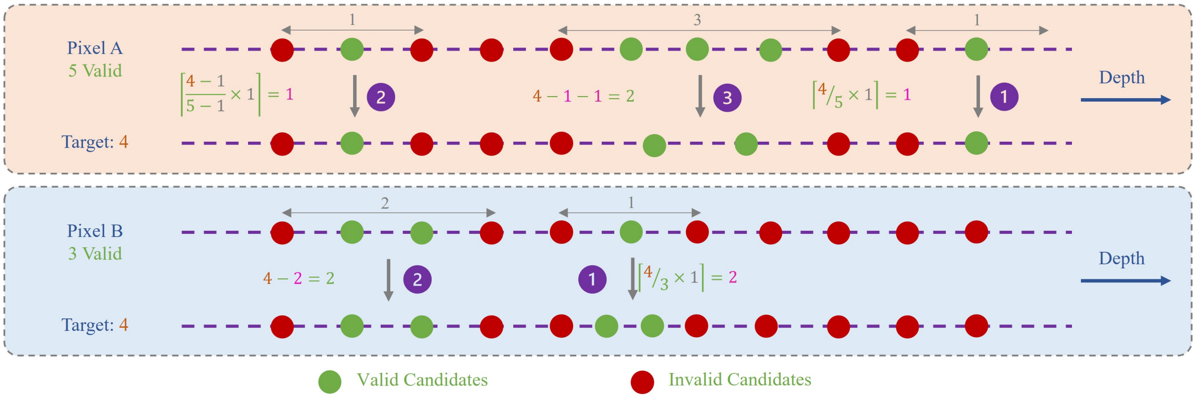
Pixel A
5 Valid
Target: 4
1
3
1
$\left\lceil \frac{4-1}{5-1} \times 1 \right\rceil = 1$
2
$4 - 1 - 1 = 2$
3
$\left\lceil {}^{4}/_{5} \times 1 \right\rceil = 1$
1
Depth
Pixel B
3 Valid
Target: 4
2
1
$4 - 2 = 2$
2
1
$\left\lceil {}^{4}/_{3} \times 1 \right\rceil = 2$
Depth
Valid Candidates
Invalid Candidates

Figure 5. **Illustration of candidate reallocation. Top:** pixel with too many candidates. We reduce the nominator number by downsampling. **Bottom:** pixel with too few candidates.

because they have smooth surfaces, and their depths can be easily inferred from neighbors. In contrast, coarser sampling may help the network overcome repetitive patterns and non-textured regions by collecting information from a larger field of view.

However, delicate objects (e.g., toothbrushes and nail clippers) eagerly demand higher depth resolution for detail recovery. Thus, we should put more effort into pixels holding fewer candidates.

Reallocate Candidates: We propose to set a unified depth number N_u for all pixels by transfer payment, as shown in Fig. 4. Specifically, if a pixel occupies too many candidates, we lower its sampling density; otherwise, we promote.

As shown in Fig. 5, we uniformly position N_u candidates within the depth range bounded by the N_v^i valid nominators for each pixel p_i. Segmented depth ranges may occur when multiple objects of the same semantic label appear in the same image, as shown in Fig. 3. In such cases, candidates in different fragments are reallocated separately, and the quota of new candidates in each fragment is proportional to the number of valid candidates inside the fragment.

To avoid fractional new candidate numbers, we prioritize the narrowest fragments and offer them slightly more candidates by the ceiling function. After that, we assign the remaining quotas for the rest fragments. This way, attention transfers from large surfaces to small objects because fragments containing fewer valid candidates get promoted.

A more significant unified depth number N_u offers higher effectiveness while lowering the model efficiency. In default, the value of N_u is set to ensure the whole model runs as fast as the state-of-the-art real-time methods.

Efficiency: Our reallocation strategy benefits effectiveness by attending more to delicate structures and contributes to efficiency.

We take Fig. 4 as an example. Although these three pixels have the same number of initial depth hypotheses, they get different valid candidate numbers after the semantic pruning. Notably, most GPUs are designed to process regular inputs and are inefficient on irregular data.

Thus, without reassignment, we have to pad pixels 2 and 3 with zeros to form a three (pixel number) by three (the max valid candidate number among pixels: $\max_i \{N_v^i\}$) matrix as the input of the depth head. In comparison, a three-by-two matrix is enough after reallocation. The reduction of input matrix size directly lowers the computation demand of the costly depth head.

Although sparsification may sound like a good solution,

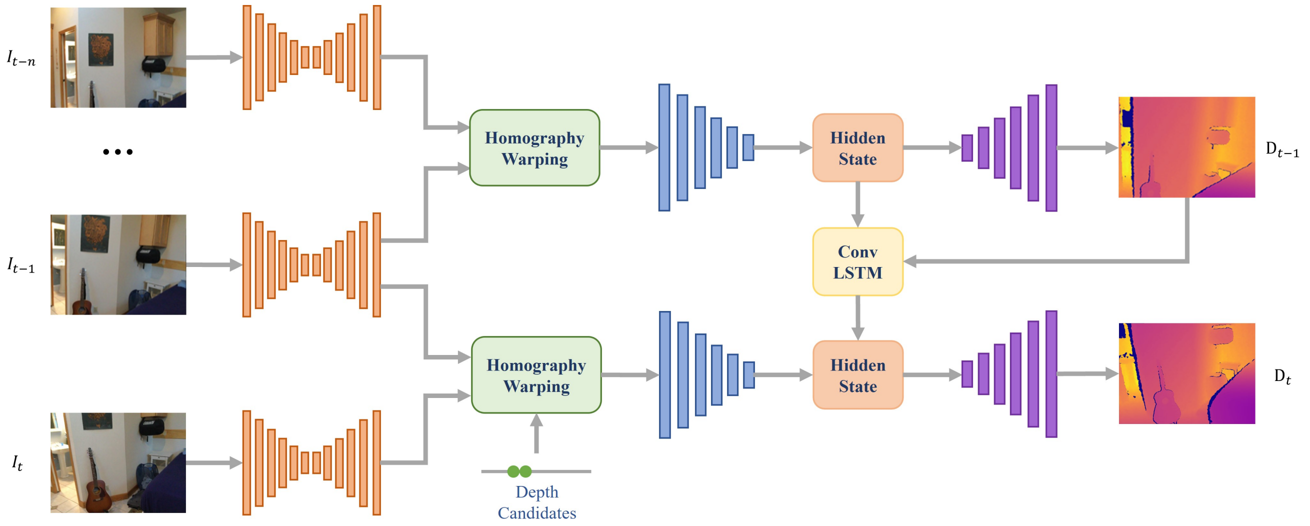

Figure 6. **Architecture of the depth head.** The outputs of the semantic-based depth selection module are used to warp the previous video frames to the current view for cost volume calculation. The depth predictions are regressed based on the current cost volume and former estimations.

it heavily relies on specialized hardware and is usually less efficient on inputs of irregular zero distributions. In addition, it leads to lower sampling density on small objects.

Spatial Correlation: Since our samples are pixel-wisely determined, a possible doubt might be whether the spatial correlation still holds. The answer is yes because neighboring pixels generally share the same semantic label. As a result, they should match adjacent pixels in other views, thus leading to similar depth candidates after the semantic pruning.

Moreover, our candidate reassignment module evenly distributes candidates in the valid depth range; thus, neighboring pixels should have similar final depth samples. The local spatial correlation still holds as illustrated in Fig. 7. In other words, we can exploit convolutions to evaluate the candidate depths.

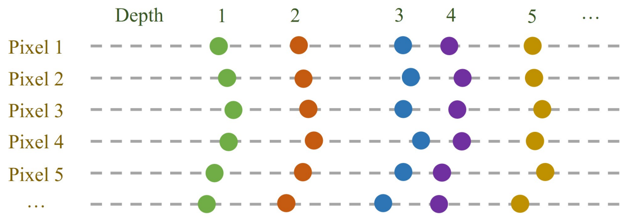

Figure 7. **Spatial correlation after depth selection.** Neighboring pixels generally have similar sets of depth candidates.

3.3. Depth Head

After determining the depth candidates for each pixel, we predict the depth map by assessing each depth candidate. The model starts by extracting image features from the current frame I_t using a feature pyramid network (FPN) [29]. After that, we warp the image features from previous time stamps to the current view and then construct a cost volume by measuring the pixel-wise correlation between the current and former features.

Finally, we send the raw cost volume into a U-shape refining network to regress the depth prediction at time t. Notably, we keep the temporal consistency by introducing a ConvLSTM [54] to update the extracted features and by warping the previous depth estimations to the current viewpoint.

4. Experiments

Our model is implemented using PyTorch [45]. We optimize the model using MADGRAD [6] on eight A100 GPUs. The input image is cropped and scaled to 256 × 256. The weights for testing are selected according to their losses on the validation set. The depth metrics are borrowed from DeepVideoMVS [9].

Keyframe Selection: Considering the high similarity between adjacent frames, estimating depth maps for every raw frame is inefficient. Thus, we select the most iconic frames (i.e., keyframes) from the input video sequences. When a new frame comes, we examine its pose distance to the latest keyframe and only adopt the new frames leading to pose

	ScanNet					
	Abs Diff↓	Abs Rel↓	Sq Rel↓	$\delta < 1.05$↑	$\delta < 1.25$↑	Time (ms) ↓
DPSNet [19]	0.1548	0.0789	0.0292	49.39	93.31	308 ms
MVDepthNet [61]	0.1643	0.0844	0.0341	46.73	92.80	46 ms
DELTAS [55]	0.1492	0.0783	0.0271	48.68	93.81	83 ms
GPMVS [16]	0.1492	0.0753	0.0290	51.09	93.98	49 ms
DeepVideoMVS (Fusion) [9]	0.1182	0.0580	0.0187	60.23	96.78	**30 ms**
Ours	**0.1069**	**0.0506**	**0.0165**	**61.34**	**97.13**	**30 ms**

Table 1. **Depth evaluation** results on the ScanNet dataset. The best results are marked in bold. Our model is leading in both effectiveness and efficiency.

distances above a threshold. We define the pose distance as Eq. (1) and set the threshold as 0.35, following [9, 17]. Notably, we do not estimate depth for the first frame in a video sequence because it has no predecessors to provide multi-view information.

$$Dis_p = \sqrt{|transition|^2 + \frac{2}{3}Trace(I - rotation)} \quad (1)$$

Datasets: Following the mainstream settings [9, 57], we train and evaluate our model on ScanNet [5], which is a large-scale real-world video dataset containing 1201 scenes for training, 321 for validation, and 100 for testing.

Comparison to existing SOTAs Tab. 1 demonstrates that our method outperforms various state-of-the-art depth estimation approaches in effectiveness and efficiency.

Compared to real-time SOTAs [9], our method shows significant advantages in all effectiveness metrics while retaining comparable time efficiency. Compared to semi-real-time algorithms [16, 55, 61], our method dominates inference time and offers significant competitiveness in reconstruction quality. In addition, our real-time framework even generates more accurate depth maps than some offline solutions [19] while spending one magnitude less cost.

Moreover, our method benefits relative error more than absolute error and favors L^1 difference more than L^2. These results indicate that delicate structures and small objects win more attention.

Impacts of the Semantic Head To assess the effectiveness of our proposed semantic-based depth selection strategy, we fix all other variables and only examine the semantic head's impacts. Tab. 2 confirms that our semantic-based depth selection module offers significant benefits.

5. Limitations

Like most dense reconstruction models, our pipeline requires the camera parameters to be known, which may not always be feasible. Besides, our current model may have difficulty with unseen object categories, although we only require adjacent frames to have the same class prediction (even if it is incorrect).

	Abs Diff↓	Abs Rel↓	Sq Rel↓
No Semantic Head	0.1225	0.0633	0.0192
+ Semantic Head	0.1069	0.0524	0.0161

Table 2. **Improvements from our semantic head.**

6. Conclusion

This paper proposes a novel 3D reconstruction network that is leading in effectiveness and efficiency. Our core contribution is to locate the most promising depth candidates using semantic maps. Specifically, we prune unreliable candidates according to the cross-view semantic consistency and then resample the candidates to look after pixels in need. Consequently, our method produces highly delicate reconstructions of meticulous structures and generates smooth surfaces for large planes with low overhead. Experimental results on the most popular datasets verify the proposed method's advantages in real scenarios.

Acknowledgments

This work is partially supported by the National Science Foundation of China (NSFC) under Grant No. U22B2060, the Hong Kong RGC GRF Project 16209519, CRF Project C6030-18G, C2004-21GF, AOE Project AoE/E-603/18, RIF Project R6020-19, Theme-based project TRS T41-603/20R, China NSFC No. 61729201, Guangdong Basic and Applied Basic Research Foundation 2019B151530001, Hong Kong ITC ITF grants MHX/078/21 and PRP/004/22FX, Microsoft Research Asia Collaborative Research Grant, HKUST-Webank joint research lab grant and HKUST Global Strategic Partnership Fund (2021 SJTU-HKUST). In addition, we gratefully acknowledge the generous support from MindSpore.

References

[1] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008. 3

[2] 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. 3

[3] Shuya Chen, Zhiyu Xiang, Chengyu Qiao, Yiman Chen, and Tingming Bai. Sgnet: Semantics guided deep stereo matching. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 4

[4] Shuya Chen, Zhiyu Xiang, Chengyu Qiao, Yiman Chen, and Tingming Bai. Pgnet: Panoptic parsing guided deep stereo matching. *Neurocomputing*, 463:609–622, 2021. 4

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3, 8

[6] Aaron Defazio and Samy Jelassi. Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *J Mach Learn Res*, 23:1–34, 2022. 7

[7] Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training. In *CVPR*, 2023. 1

[8] Pier Luigi Dovesi, Matteo Poggi, Lorenzo Andraghetti, Miquel Martí, Hedvig Kjellström, Alessandro Pieropan, and Stefano Mattoccia. Real-time semantic stereo matching. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10780–10787, 2020. 3

[9] 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, 3, 5, 7, 8

[10] 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 (ICCV)*, December 2015. 3

[11] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9725, 2021. 2, 5

[12] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 3

[13] 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. 1

[14] 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. 1

[15] 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. 1, 3

[16] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019. 3, 8

[17] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019. 8

[18] Christian Häne, Christopher Zach, Andrea Cohen, and Marc Pollefeys. Dense semantic 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1730–1743, 2017. 4

[19] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 8

[20] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1, 3

[22] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019. 3

[23] Gen Li and Joongkyu Kim. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 259. BMVA Press, 2019. 3

[24] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2019. 3

[25] Haoquan Li, Laoming Zhang, Daoan Zhang, Lang Fu, Peng Yang, and Jianguo Zhang. Transvlad: Focusing on locally aggregated descriptors for few-shot learning. In *Computer*

Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX, pages 524–540. Springer, 2022. 1

[26] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *ECCV*, 2022. 1

[27] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In *NeuIPS*, 2022. 3

[28] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. *arXiv preprint arXiv:2210.15929*, 2023. 3

[29] 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. 7

[30] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1253–1260, 2010. 3

[31] Chengliang Liu, Jie Wen, Xiaoling Luo, Chao Huang, Zhihao Wu, and Yong Xu. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[32] Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view- and category-aware transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 3

[33] Chengliang Liu, Zhihao Wu, Jie Wen, Yong Xu, and Chao Huang. Localized sparse incomplete multi-view clustering. *IEEE Transactions on Multimedia*, 2022. 1

[34] Mengyu Liu and Hujun Yin. Feature pyramid encoding network for real-time semantic segmentation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 260. BMVA Press, 2019. 3

[35] Yue Liu, Wenxuan Tu, Sihang Zhou, Xinwang Liu, Linxuan Song, Xihong Yang, and En Zhu. Deep graph clustering via dual correlation reduction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7603–7611, 2022. 3

[36] Yue Liu, Jun Xia, Sihang Zhou, Siwei Wang, Xifeng Guo, Xihong Yang, Ke Liang, Wenxuan Tu, Z. Stan Li, and Xinwang Liu. A survey of deep graph clustering: Taxonomy, challenge, and application. *arXiv preprint arXiv:2211.12875*, 2022. 3

[37] Yue Liu, Xihong Yang, Sihang Zhou, and Xinwang Liu. Simple contrastive graph clustering. *arXiv preprint arXiv:2205.07865*, 2022. 3

[38] Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *Proc. of AAAI*, 2023. 3

[39] 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. 3

[40] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. In *European Conference on Computer Vision*, pages 640–657. Springer, 2020. 3

[41] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018. 3

[42] Vlad-Cristian Miclea and Sergiu Nedevschi. Real-time semantic segmentation-based stereo reconstruction. *IEEE Transactions on Intelligent Transportation Systems*, 21(4):1514–1524, 2020. 4

[43] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022. 3

[44] 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 *European conference on computer vision*, pages 414–431. Springer, 2020. 1, 3

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[46] Fang Peng, Yu Tan, and Cheng Zhang. Exploiting semantic and boundary information for stereo matching. *Journal of Signal Processing Systems*, pages 1–13, 2021. 4

[47] Juncai Peng, Yi Liu, Shiyu Tang, Yuying Hao, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Zhiliang Yu, Yuning Du, et al. Pp-liteseg: A superior real-time semantic segmentation model. *arXiv preprint arXiv:2204.02681*, 2022. 2, 5

[48] Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 289. BMVA Press, 2019. 3

[49] Rongjun Qin, Xu Huang, Wei Liu, and Changlin Xiao. Pairwise stereo image disparity and semantics estimation with the combination of u-net and pyramid stereo matching network. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4971–4974, 2019. 4

[50] Alexander Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer. 3dvnet: Multi-view depth prediction and volumetric refinement. In *2021 International Conference on 3D Vision (3DV)*, pages 700–709. IEEE, 2021. 1

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th*

International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 3

[52] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

[53] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 3

[54] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 7

[55] Ayan Sinha, Zak Murez, James Bartolozzi, Vijay Badrinarayanan, and Andrew Rabinovich. Deltas: Depth estimation by learning triangulation and densification of sparse points. In *European Conference on Computer Vision*, pages 104–121. Springer, 2020. 8

[56] Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021. 3

[57] 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, 5, 8

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[59] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Nießner, Stuart Golodetz, Victor A. Prisacariu, Olaf Kähler, David W. Murray, Shahram Izadi, Patrick Pérez, and Philip H. S. Torr. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 75–82, 2015. 3

[60] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1296–1305, 2021. 3

[61] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)*, pages 248–257. IEEE, 2018. 3, 8

[62] Likang Wang and Lei Chen. Ftso: Effective nas via first topology second operator. *Preprints*, 2023. 3

[63] Likang Wang, Yue Gong, Xinjun Ma, Qirui Wang, Kaixuan Zhou, and Lei Chen. Is-mvsnet:importance sampling-based mvsnet. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 668–683, Cham, 2022. Springer Nature Switzerland. 1

[64] Likang Wang, Yue Gong, Qirui Wang, Kaixuan Zhou, and Lei Chen. Flora: dual-frequency loss-compensated real-time monocular 3d video reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[65] 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 (CVPR)*, June 2020. 3

[66] 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 (CVPR)*, June 2015. 3

[67] Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. Prototype knowledge distillation for medical segmentation with missing modality. *arXiv preprint arXiv:2303.09830*, 2023. 3

[68] Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. *arXiv preprint arXiv:2303.10902*, 2023. 1

[69] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 3

[70] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7483–7492, 2019. 3

[71] Liu Xiaolong, Li Lujun, Li Chao, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. In *ICLR*, 2023. 3

[72] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 1, 3

[73] 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. 1, 3, 5

[74] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 3, 5

[75] Dingyi Zeng, Wanlong Liu, Wenyu Chen, Li Zhou, Malu Zhang, and Hong Qu. Substructure aware graph neural networks. In *Proc. of AAAI*, 2023. 3

[76] Dingyi Zeng, Li Zhou, Wanlong Liu, Hong Qu, and Wenyu Chen. A simple graph neural network via layer sniffer. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5687–5691. IEEE, 2022. 3

[77] Daoan Zhang, Chenming Li, Haoquan Li, Wenjian Huang, Lingyun Huang, and Jianguo Zhang. Rethinking alignment

and uniformity in unsupervised image semantic segmentation. *arXiv preprint arXiv:2211.14513*, 2022. 3

[78] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. 1

[79] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018. 3

[80] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. *arXiv preprint arXiv:2303.10826*, 2023. 3