Contrastive Learning for Depth Prediction

Rizhao Fan* Matteo Poggi Stefano Mattoccia
Department of Computer Science and Engineering
University of Bologna, Italy
*rizhao.fan@unibo.it

Abstract

Depth prediction is at the core of several computer vision applications, such as autonomous driving and robotics. It is often formulated as a regression task in which depth values are estimated through network layers. Unfortunately, the distribution of values on depth maps is seldom explored. Therefore, this paper proposes a novel framework combining contrastive learning and depth prediction, allowing us to pay more attention to depth distribution and consequently enabling improvements to the overall estimation process. Purposely, we propose a window-based contrastive learning module, which partitions the feature maps into non-overlapping windows and constructs contrastive loss within each one. Forming and sorting positive and negative pairs, then enlarging the gap between the two in the representation space, constraints depth distribution to fit the feature of the depth map. Experiments on KITTI and NYU datasets demonstrate the effectiveness of our framework.

1. Introduction

Depth prediction aims at recovering the 3D structure of a scene and is a crucial task at the core of various computer vision applications, such as robot/vehicle navigation and augmented reality, to name a few. Methods for dense depth prediction can be divided into two main classes: active and passive techniques. Representatives of the former are LiDAR, structured light, and time-of-flight sensors. They perturb the sensed environment with a signal, according to different technologies, and measure its behavior or appearance to infer the depth of the sensed scene. For instance, LiDAR and Time-of-flight measure the traveling time a signal needs from the emitter to the receiver, both located in the sensing device. However, due to their intrinsic limitations, they generally generate sparse/low-resolution measurements. Thus, various approaches to densify such data have been proposed, typically in setups coupling such sensors with high-resolution RGB images [11, 18, 27, 29, 38, 47, 77, 81]. On the other hand, deep learning enabled pure passive sensing techniques to rely on single images [17, 21, 36] without any auxiliary aid to achieve the outlined goal despite being very challenging due to its ill-posed nature [52].

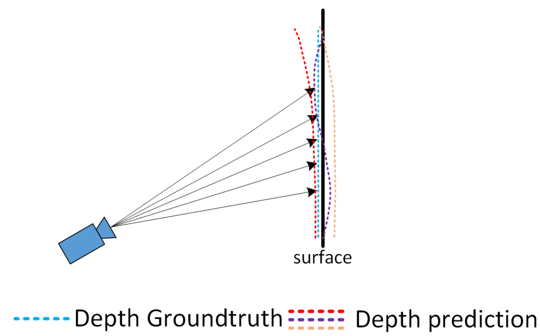
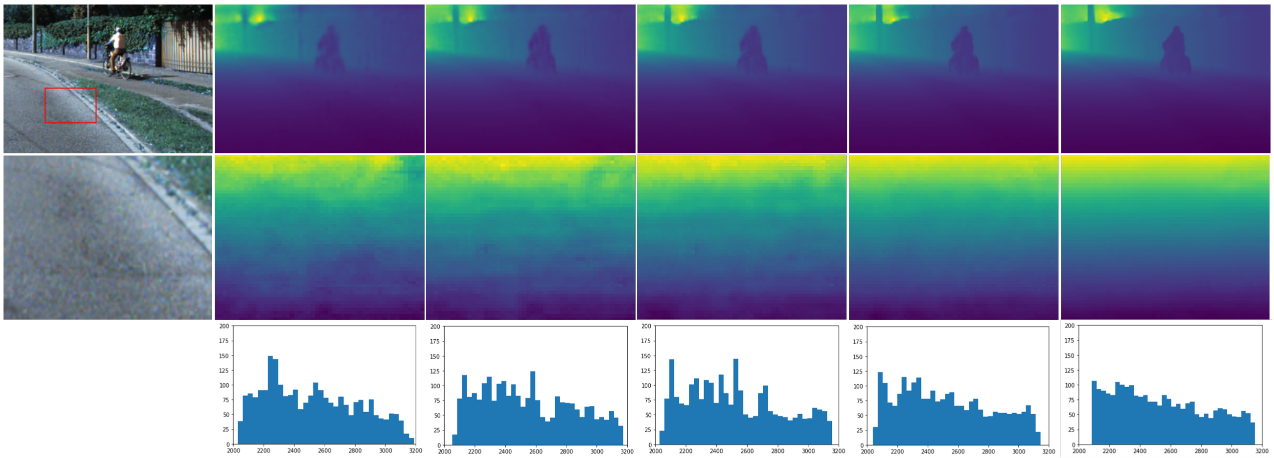


Figure 1. **Imaging system and depth prediction process.** The depth changes smoothly within adjacent pixels belonging to the same portion of the object, while this is not always the case for depth predicted by neural networks. Best viewed in color.

Regardless of the adopted setup, little effort in the literature focuses on analyzing the distribution of depth data in a statistical manner, as we propose in this paper. In the real world, the depth changes smoothly within adjacent pixels belonging to the same object's surface. However, when depth prediction models infer the depth, they may output different results, as shown in Fig. 1. These models formulate depth prediction as a regression task, and the output value within an extremely small region changes slowly. When a weak model predicts the depth, it is common to find that the result is represented by peaks in a narrow range of depth values, which can be observed from depth predictions and depth histograms in Fig. 2.

A depth histogram is a graphical representation of the value distribution in a depth map. By focusing on the depth distribution analysis in Fig. 2, we can notice how histograms change considerably: indeed, in poor-quality depth maps, depth histogram is usually concentrated in limited intensity values, whereas high-quality depth maps have a more regularized distribution spread into a broader interval.

To further confirm this observation, we synthesize three



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 2. **Illustration of different quality depth maps.** From left to right, rows 1 and 2: (a) the color image (first row) and one region in the next two rows corresponding to the red box in the original image. Columns (b, c, d, e, f) show depth maps of different quality, from worse to better. Row 3 reports depth histograms for the depth maps corresponding to the area within the red box. We took (b-e) from four training stages of CDCNet [18] on the KITTI depth completion task [20].

ideal depth maps for two smooth surfaces and one discontinuity – i.e. common structures in the real world – and construct their depth histogram, as shown in Fig. 3. We can observe that the distributions of depth values in the histograms are much more regular. Therefore, we argue that regularizing such distributions can yield higher-quality depth maps. Accordingly, by focusing on the structure of objects from a microscopic perspective, any object can be seen as the composition of several small, smooth surfaces interleaved by depth discontinuities, whose depth distributions can be regularized.

Thus, we inquire whether a learning method can regularize the distribution of the depth predictions by a deep network. The recent contrastive learning approaches [4, 9, 24] proposed to contrast positive pairs – i.e., elements sharing the same label – against negative pairs – elements with different labels – in the representations space, which looks suitable for our purpose. Therefore, we introduce a contrastive learning framework tailored for depth prediction. Specifically, we propose a Window-based Contrastive Learning module (WCL) which segments the depth feature maps into non-overlapping windows and computes a contrastive loss only within each one, similarly to how recent Vision Transformers [44] compute self-attention on local windows. It allows us, in a more tractable manner rather than acting on the whole depth features, to contrast the depth values in small regions and expand their distribution locally by constructing positive and negative pairs in the depth representation and enlarging the gap between them. To the best of our knowledge, we are the first to apply contrastive learning for depth prediction, focusing on expanding depth distribution.

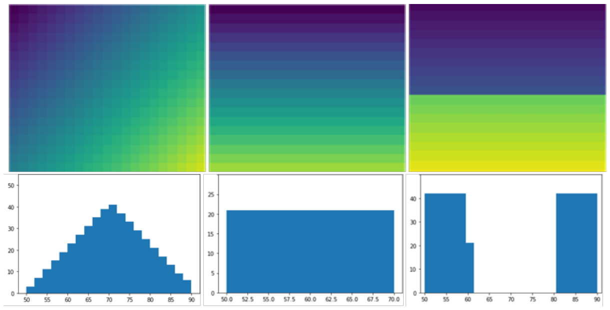
Our main contributions can be summarized as follows:

- We propose a novel method combining contrastive learning with depth prediction, contrasting depth distribution to improve accuracy.
- We propose a Window-based Contrastive Learning (WCL) module for depth prediction by learning similarity, forming positive and negative pairs of features, and computing contrastive loss within a window.
- Experimental results and a detailed ablation study with different depth prediction models demonstrate the effectiveness of our proposal.

2. Related Work

2.1. Depth prediction

Depth prediction is a challenging computer vision task, faced according to different methodologies and setups such as depth completion, depth estimation, and self-supervised depth estimation with stereo pairs and videos. Depth completion recovers dense depth maps from sparse measurements and a high-resolution image. Uhrig *et al.* [62] propose a sparsity-invariant convolution layer to consider the location of missing data while addressing data sparsity within deep networks. Ma *et al.* [45, 46] utilize early fusion to combine sparse depth with a color image and feed them into an encoder-decoder CNN, which boosts the performance of depth completion. Multi-branch network architecture is an effective approach to fuse multi-modal data as reported in [18, 32, 38, 55, 63]. Spatial propagation networks (SPN) [43] is another popular depth-refinement approach [10–12, 42, 48]. DeepLiDAR [55] introduces pixel-



(a) (b) (c)

Figure 3. **Three synthetic depth maps and corresponding histograms.** (a), (b) are two smooth surfaces and (c) is a depth discontinuity.

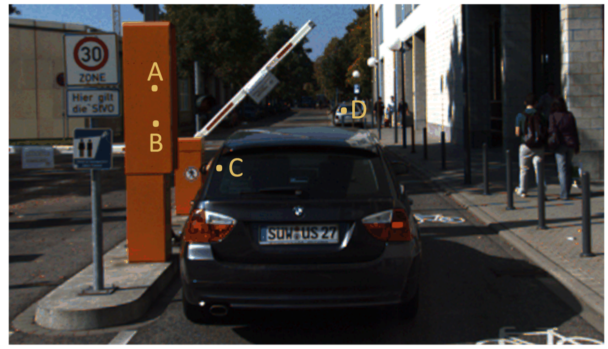


Figure 4. **Example of different points in the scene.** A and B are two points on the same, front-parallel surface, while C and D are points on two distinct cars. Although both A-B and C-D pairs of points belong to similar objects, A-B expose stronger similarity for the depth prediction task – i.e., they are very close, while C-D points are at very different distances.

wise surface normals as geometric constraints and proposes multiple branches to generate dense depth maps jointly. GuideFormer [56] and CompletionFormer [78] introduce transformers in this task as well. In [7, 73, 81], graph representation is utilized to model more helpful information from sparse and irregular point clouds.

Regarding depth estimation, supervised monocular approaches gained much interest in the literature in recent years. Eigen *et al.* [16, 17] proposed a multi-stage, coarse-to-fine network to estimate depth from a single image, and Laina *et al.* [35] a fully convolutional architecture for depth estimation. Some works introduce attention mechanisms to achieve significant performance improvements [1, 2, 37, 39]. Other works estimate depth by predicting probability distributions and discrete bins [5, 40, 59]. A high-order geometric constraint is also employed to reconstruct depth prediction in [36, 74]. Some multi-task works predict depth maps by jointly learning with other vision tasks [41, 54, 79]. DANet [59] proposes to utilize depth distribution as supervision for the prediction.

Self-supervised monocular depth estimation consists of two principal training methodologies using stereo images or monocular videos [52]. Garg *et al.* [19] propose the first self-supervised depth estimation, using an image reconstruction loss in stereo images to train a monocular depth model. Zhou *et al.* [82] employ a depth and a pose network to employ a photometric loss in monocular videos. Many following researchers follow these paths improving self-supervised depth estimation [21,22,30,34,49–51,53,60,80]. ViP-DeepLab [54] proposes a multi-task network for jointly learning self-supervised depth estimation and panoptic segmentation from videos, while PackNet [22] leverages 3D convolutions to learn geometric representations.

2.2. Contrastive learning

Contrastive learning has achieved remarkable progress employing discriminative learning by contrasting positive pairs against negative pairs in representations space [23] and some works target visual representation learning [9, 15, 24, 33, 69]. SimCLR [9] implements contrastive learning in a simple network framework, where positive pairs are from data augmentation of the same image, while negative ones are from different images. MoCo [24] maintains a queue of negative samples and turns one branch of a Siamese network into a momentum encoder to improve the queue consistency. Some recent works [6,68,72] have introduced contrastive learning for dense prediction. ReSim [71] learns regional representations from a pair of views originating by sliding windows from the same image. DenseCL [68] optimizes a pairwise contrastive loss at the pixel level between two different image views. Ke *et al.* [31] propose a weakly supervised segmentation method that utilizes contrastive relationships between pixels and segments in the feature space. Alonso *et al.* [3] utilize a memory bank to contrast the features from labeled and unlabeled data employing end-to-end training. Some works [70, 76] use a contrastive loss to generate high-frequency details for image super-resolution tasks. Shen *et al.* [58] propose contrastive differential learning in image translation and use it for depth-to-depth synthesis.

2.3. Window-based Approaches

Long-range dependency is a notable cue and Transformers [64], Markov Random Fields (MRFs) [57] and Conditional Random Fields (CRFs) [8, 67] use it to boost their learning ability. However, these methods have a severe drawback in the computational complexity, increasing quadratically with image size. Purposely, some works aim at addressing this issue. Dosovitskiy *et al.* [14] apply a transformer on sequences of image patches for image classification tasks. Pyramid ViT [65] uses a progressive shrinking pyramid and spatial-reduction attention to reduce computations of the transformer on large feature maps. Swin Transformer [44] proposes a novel architecture that computes attention within a patch-based window and uses a shifting window approach to capture attention in non-overlapping regions. Yuan *et al.* [75] employ a window-based CRFs approach for monocular depth estimation. CSWin [13] proposes a self-attention within a cross-shaped window in different directions, which yields strong representation learning with limited computation cost.

3. Contrastive learning for Depth Prediction

In this section, we first present the motivation for our work and how WCL can improve depth prediction.

3.1. Motivation

In the imaging system, the image is a perspective projection of a 3D scene, and the corresponding depth map reflects the distance between each point in the scene and the viewpoint. Almost all surfaces of observed objects have discrete depth values and adjacent pixels, even within small regions, have similar but different depth values. As shown in Fig. 2, the histograms of different depth images with different accuracy change dramatically, and high-quality depth maps have a more regular depth distribution than low-quality ones. Hence, an intuitive idea to model this assumption consists in regularizing such distribution. For this purpose, applying contrastive learning by clustering the different pixels in the representation space seems a promising strategy for possibly regularizing the depth distribution. Moreover, it has been recently applied in an unsupervised manner [6,9,24,68,72], thus not making use of labels. Some recent works [6, 26, 31, 66] focus on dense predictions, such as semantic segmentation, and use similarity or affinity between pixels, images, or features to construct contrastive loss for the pixels with the same label that share similar low-level features (color, texture) or high-level representation. Specifically, [26, 66] propose to use the semantic similarities among labeled pixels to contrast representation. Ke *et al.* [31] explore different types of contrastive relationships, such as low-level image similarity and feature affinity in weakly supervised segmentation. Chaitanya *et al.* [6] use contrastive learning at the level of local and global features. All these methods assume that pixels belonging to the same object share the same representation and label since the target task is semantic segmentation. However, this assumption does not hold when facing depth prediction tasks, since even pixels from the same object category may have different representations and depth labels. Indeed, distinguishing the positive and negative pairs in the depth map is challenging, and some of the main issues about deploying a contrastive loss in depth prediction tasks are:

- Discriminating between positive and negative examples for all pixels is challenging. For example, in Fig. 4, we can easily define the set (A, B) as a positive pair since the two are close in distance and representation and the set (A, D) as a negative pair being far in distance and representation. However, for the set (B, C), it is hard to define whether they are positive or negative pairs since they are close in distance but belong to different objects.
- Constructing and computing a global relationship graph for all pixels is highly resource-consuming. For instance, if we consider an H $\times$ W image, its relationship map results in size HW $\times$ HW, thus forming positive and negative pairs and contrasting them yields massive memory footprints, computation, and time.
- It is hard to contrast further the long-range sets, such as set (A, D), even in low-quality maps since there is already a significant distance between the two points.

For the reasons outlined, employing contrastive learning on the whole image is challenging. Purposely, we introduce the concept of local similarity in our method. As already pointed out, the output depth of a deep network changes smoothly within a small region. Therefore, a pixel has more similarities with its surrounding pixels. The closer the two pixels are, the stronger the similarity. For instance, The similarity between A and B is stronger than the similarity between A and others in Fig. 4. To deal with these issues, inspired by some works [13, 44, 75], we propose for depth prediction a cost-effective window-based approach for contrastive learning. We limit the similarity calculation and construct the positive and negative pairs within small regions rather than on the entire feature map, which is a more reasonable and resource-saving strategy. Thus, we set pairs with strong similarity as positive pairs and ones with weak similarity as negative pairs. Enlarging the gap between them also makes depth distribution better regularized and yields more accurate results.

3.2. Window-based Contrastive Learning (WCL)

Fig. 5 depicts the architecture of the proposed WCL module. It segments one H$\times$W$\times$C feature map into multiple tiles of the same size, each containing N $\times$ N elements. In our approach, these windows are the domain where contrastive learning comes into play.

Given a generic features map $F \in \mathbb{R}^{H \times W \times C}$, for instance, the output of a convolutional layer, we partition it into windows $X \in \mathbb{R}^{N \times N \times C}$. For each, we extract query $Q \in \mathbb{R}^{N \times N \times \frac{C}{2}}$ and key $K \in \mathbb{R}^{N \times N \times \frac{C}{2}}$ features by linear projections of the input X as

$$\begin{aligned} Q &= XW_Q \\ K &= XW_K \end{aligned} \tag{1}$$

The similarity map $T \in \mathbb{R}^{N^2 \times N^2}$ between each element

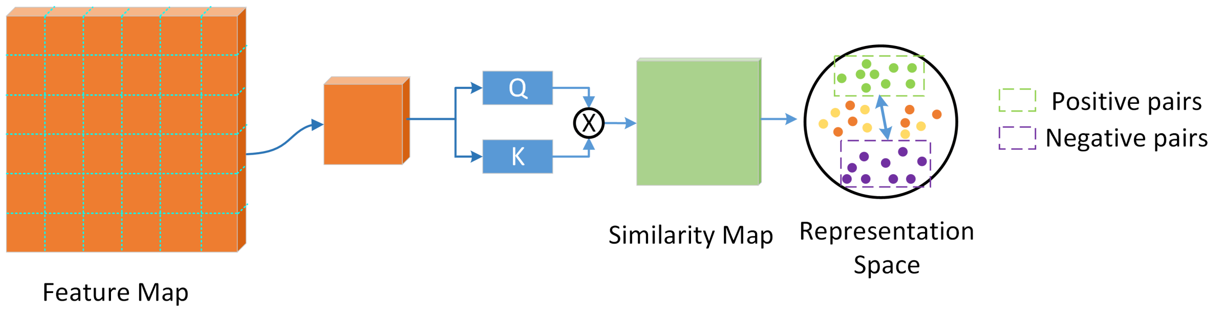


Figure 5. **Illustration of WCL module.** Our module segments one $H \times W \times C$ feature map into multiple tiles of the same size, each containing $N \times N$ elements. Then, it computes a similarity map and constructs positive and negative pairs within the window. By enlarging the gap between positive and negative pairs, the features representation becomes more meaningful of the depth distribution in the scene.

in the window is computed by means of dot product and normalization as

$$T = \frac{Q^T}{\|Q\|} \frac{K}{\|K\|} \tag{2}$$

We employ the exponential function $exp()$ to make the similarity map non-negative. Then, we sort T in descending order:

$$T^{'} = sort(T) \tag{3}$$

and use contrastive learning to enlarge the gap between positive and negative pairs in the representation space. We sample the first N_1 of $T^{'}$ as positive pairs P_T, and the last N_1 negative pairs as N_T to compute the contrastive loss L_{closs} as

$$L_{closs} = \frac{1}{N_w} \sum_{i=1}^{N_w} - \log \frac{\sum_{j=1}^{N_1} P_T/N_1}{\sum_{j=1}^{N_1} P_N/N_1} + a \tag{4}$$

with a being a constant set to 1; N_1 being set to the 20% of the total N^2; N_w being the number of windows. Accordingly, the total loss function of a depth prediction network employing window-based contrastive learning is defined as:

$$L_{total} = L_{depth} + w * L_{closs} \tag{5}$$

with L_{depth} the original depth loss from the depth prediction network, and w a weighting term for the contrastive loss.

Our WCL module is specifically designed for depth prediction tasks and can be seamlessly integrated into many networks. As shown in Fig. 6, the WCL block works on feature maps between two layers. This way, the module enables the contrast between the positive and negative pairs in the representation space.

4. Experiments

In this section, we evaluate our method on three main depth prediction tasks, i.e., depth completion, monocular

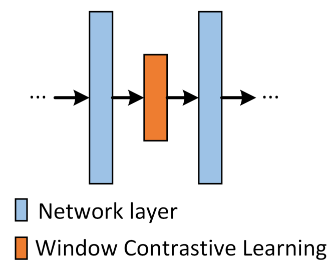


Figure 6. **WCL positioning.** Illustration of a network with a WCL module embedded in between two conventional layers.

depth estimation, and self-supervised monocular depth estimation. We start from existing networks, assumed as baselines over which we want to improve. We retrain them both with and without our module, allowing for a fair comparison under the same experimental setting (i.e., data, hardware support). This comes with little effort since our WCL module is a plug-and-play component easily embeddable in any depth prediction architecture. All the experiments are conducted using the PyTorch framework, on a single NVIDIA RTX 3090 GPU.

4.1. Dataset

KITTI dataset. KITTI [20, 62] is a popular outdoor dataset providing sparse depth maps captured by Velodyne LiDAR HDL-64e, color images and corresponding semi-dense ground truth. The sparse depth maps provide 5.9% valid depth values on all pixels, while the ground truth maps contain 16% valid depth values over the whole image. The dataset contains 85 895 training frames, with 1000 more selected validation frames, as well as 1000 and 500 test samples for which ground truth is withheld, respectively, for depth completion and depth prediction tasks.

The Eigen split [17] is the standard portion of KITTI used for evaluating self-supervised monocular depth prediction, after being filtered [21, 82] of static frames, which are unsuited for training from videos [82]. In this split, 39 910 images and 4 424 images are used for training and valida-

Method	RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)
MSG-CHN	820.145	223.987	**2.425**	0.979
MSG-CHN+WCL	**810.273**	**222.719**	2.428	**0.973**

Table 1. **Quantitative results on KITTI [62] – Depth Completion .** Comparison between MSG-CHN and its WCL counterpart.

Method	RMSE (mm)	REL	$\delta_{1.25^1}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Sparse-to-Dense [45]	0.1097	0.0185	99.39	99.91	99.98
Sparse-to-Dense [45]+WCL	**0.1038**	**0.0149**	**99.41**	**99.92**	**99.99**

Table 2. **Quantitative results on NYUv2 [61] – Depth Completion .** Comparison between Sparse-to-Dense and its WCL counterpart.

tion, respectively, with 697 further images used for testing.

NYUv2 dataset. The NYU Depth v2 dataset [61] (NYUv2) consists of 120K RGB images and depth maps at 480×640 resolution from 464 indoor scenes captured by a Microsoft Kinect sensor. For training, we use a subset of 50K images from the official training split.

4.2. Evaluation Metrics

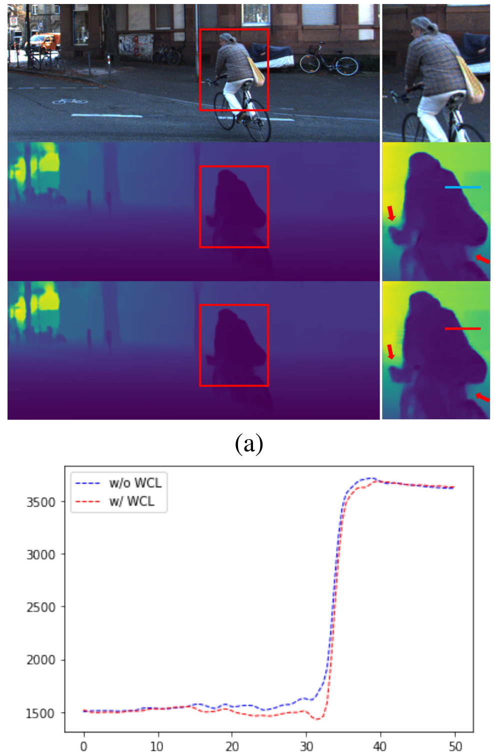
Following [21,36,38,48], we use common metrics in the field: Mean absolute error (MAE, mm), Root Mean Squared Error (RMSE, mm), Mean Absolute Error of the inverse depth (iMAE, 1/km), Root Mean Squared Error of the inverse depth (iRMSE, 1/km), Mean Absolute Relative Error (REL), Absolute relative difference (Abs Rel), Square relative difference (Sq Rel) and percentage of predicted pixels where the relative error is within a threshold (δ_i).

4.3. Experimental Results

We select a set of models representative of the three specific tasks, to which we apply our method to improve their accuracy consistently. Any model and its WCL variant are trained using the standard hyper-parameters reported in the original papers.

4.3.1 Depth completion

We consider two depth completion methods [38, 45], respectively MSG-CHN and Sparse-to-Dense. MSG-CHN [38] is a multi-branch guided cascade hourglass network for depth completion. Sparse-to-Dense [45] is built with an early-fusion network for multi-modal data. We insert our module between the last two layers of each branch in MSG-CHN and train it on the KITTI dataset [62]. In Sparse-to-Dense, we use ResNet18 [25] as the backbone to extract features, we insert our module between the last two layers of the decoder and train it on NYU Depth v2 [61]. We use the same training parameters, except for the batch size, for



(a)

(b)

Figure 7. **Qualitative comparison on the KITTI depth completion dataset** [20]. In (a), from top to bottom: RGB image, results of MSG-CHN [38], and MSG-CHN [38]+WCL, respectively. We zoom within the red line regions at the right; WCL achieves more precise object boundaries where the red arrow points. In (b), We show a depth slice from the results in (a). Dashed lines in the plot refer to blue and red lines in crops from (a). With WCL, the prediction results in more reasonable boundaries.

both networks. The batch size is set to 8 and 12, respectively. For the two networks, we set $w = 0.2$, $N = 7$ and $w = 0.1$, $N = 7$ respectively. Tab. 1 and Tab. 2 show that both MSG-CHN and Sparse-to-Dense can benefit from our WCL module. Fig. 7 shows a qualitative comparison between MSG-CHN and its counterpart using WCL on a sample from the KITTI dataset. We can notice how our module allows for more precise boundaries at depth discontinuities. Fig. 8 instead compares Sparse-to-Dense models on the NYUv2 dataset, highlighting the same behavior observed for MSG-CHN.

4.3.2 Monocular Depth estimation

For this task, we select BTS [36] as a baseline. It is a state-of-the-art method using local planar guidance layers as geometric constraints to guide the features to depth upsampling in the decoding phase. We use BTS variants with different backbones, i.e., ResNet50 [25] and DenseNet-121 [28], on NYUv2 for the monocular depth estimation task. We use

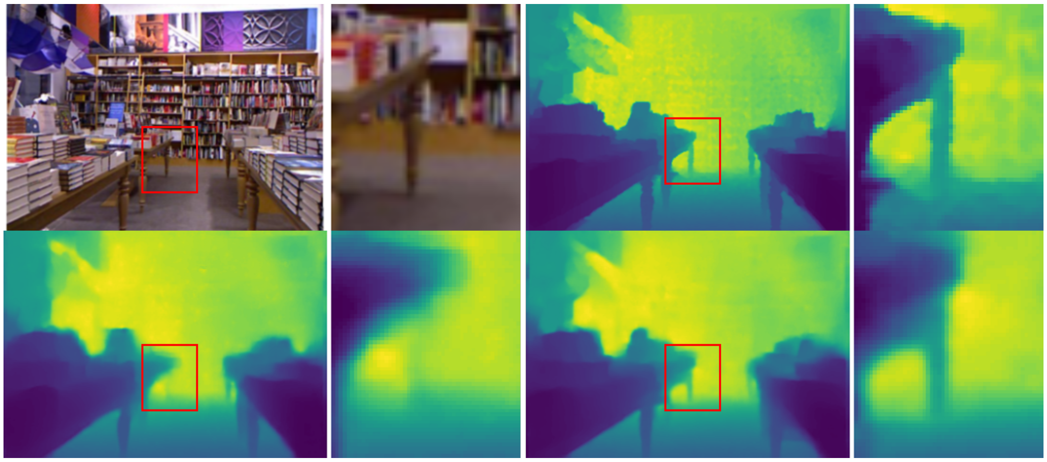


Figure 8. **Qualitative comparison on the NYUv2 depth completion dataset** [61]. From left to right, top to bottom, RGB image, ground truth, results of Sparse-to-Dense [45] and Sparse-to-Dense [45]+WCL. On the right of each image, we zoom into the red rectangle. With WCL, predicted depth maps expose more precise structures and boundaries.

Methods	*higher is better*			*lower is better*				
	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$	AbsRel	Sq Rel	RMSE	RMSE *log*	log10
BTS-ResNet50	0.865	0.975	0.993	0.119	0.075	0.419	0.152	0.051
BTS-ResNet50+WCL	**0.871**	**0.977**	**0.994**	**0.117**	**0.072**	**0.409**	**0.149**	**0.050**
BTS-DenseNet-121	0.865	0.976	**0.995**	0.120	0.075	0.421	0.152	0.051
BTS-DenseNet-121+WCL	**0.869**	**0.977**	0.994	**0.117**	**0.072**	**0.413**	**0.149**	**0.050**

Table 3. **Quantitative results on NYUv2** [61] – **Monocular Depth estimation.** Comparison between BTS variants and their WCL counterparts.

the same training parameters suggested by the authors [36], except for the batch size that is set to 10. For this task, we set $w = 0.1, N = 7$ for both the backbones. Tab. 3 shows that our WCL module allows consistent improvements over both BTS variants.

4.3.3 Self-Supervised Monocular Depth Estimation

Self-supervised monocular depth estimation eliminates the need for ground truth depth labels, which are usually hard to source. Supervision can be obtained in the form of monocular videos or stereo images. We select MonoDepth2 [21] as a state-of-the-art baseline for this task, simultaneously learning for depth and relative poses between consecutive frames in a video to implement the aforementioned self-supervised training scheme. We test it on the KITTI dataset [20] using the Eigen split [17]. Its training is realized by minimizing the photometric re-projection errors, either between temporally adjacent frames or stereo images. We use ResNet18 [25] as the backbone, process images resized to 192×640, and keep the same training parameters detailed by the authors [21]. In both cases, we evaluate our module setting $w = 0.01, N = 7$. Tab. 4 confirms that our method can also boost the accuracy of self-supervised depth estimation frameworks.

4.4. Ablation Study

In this section, we conduct an ablation study to verify the impact of the different hyper-parameters of our WCL module. For the experiments in this section, we focus on the depth completion task; we use a subset of 10000 samples from the KITTI depth completion dataset for training and evaluate the performance on the validation split. Images are center cropped to 1216×256, to focus on regions with available LiDAR points. We use Sparse-to-Dense [45] as the baseline network and adopt ResNet-18 as the backbone. All the parameters are optimized using Adam (β_1 = 0.9, β_2 = 0.99) and the weight decay factor is set to 0.0001. The learning rate is initialized to 0.001, decayed by {0.5, 0.2, 0.1, 0.01} at epoch {10, 15, 20, 25}. The network is trained for 30 epochs using a batch size of 10 samples. RMSE, MAE, and iMAE are used as the evaluation metrics.

Window size. We measure how the window size over which the contrastive loss is applied impacts the final results. Tab. 5 collects the outcome of this experiment. We evaluate our module with window sizes, *3, 5, 7, 9, 11, 13, 15*. w is fixed instead, to *0.1*. We can observe that window with $N = 7$ outperforms others according to the main evaluation metric, RMSE. In contrast, bigger window sizes cannot improve further while increasing the computational requirements. A small window size ($N = 3$) yields negli-

Methods	Train	*higher is better*			*lower is better*			
		$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$	AbsRel	Sq Rel	RMSE	RMSE *log*
Monodepth2	M	0.871	0.957	0.980	0.118	0.912	4.911	0.196
Monodepth2 + WCL	M	**0.873**	**0.959**	**0.981**	**0.116**	**0.852**	**4.837**	**0.194**
Monodepth2	S	0.865	0.949	**0.975**	**0.109**	0.909	5.015	0.208
Monodepth2 + WCL	S	**0.867**	**0.951**	**0.975**	**0.109**	**0.892**	**4.961**	**0.207**

Table 4. **Quantitative results on KITTI Eigen split [17] – Self-Supervised Monocular depth estimation.** All methods are trained and tested with 192×640 images. The best results in each category are in **bold**; M: Monocular supervision; S: Stereo supervision.

Window size (N)	RMSE (M)	MAE (mm)	iMAE (1/km)
baseline	930.326	269.866	2.575
3	927.818	266.248	2.536
5	925.169	267.290	2.813
7	**922.512**	264.044	2.030
9	925.358	267.614	2.045
11	926.511	**263.053**	2.639
13	925.072	265.628	2.429
15	925.074	264.810	**2.022**

Table 5. **Ablation results on the different window sizes on KITTI depth completion validation set.**

Shift number (N)	RMSE (M)	MAE (mm)	iMAE (1/km)
baseline	930.326	269.866	2.575
0	**922.512**	**264.044**	**2.030**
1	923.461	266.447	2.692
2	926.473	265.348	2.228
3	929.037	266.340	2.083
4	927.404	264.493	2.324

Table 6. **Ablation results on the shift number on KITTI depth completion validation set.**

gible improvements, probably because most 3×3 regions are not significant enough for applying contrastive learning effectively.

Shifted windows When using WCL, all windows are non-overlapped. Thus, distribution optimization occurs locally. Previous works exploiting windows processing as well [44,75] use effective shifted window partitioning to introduce connections between neighboring non-overlapping windows. We ran an ablation study about whether shifted window partitioning can bring improvement in our method or not. Following [44,75], we shift the windows by $(\frac{N}{2}, \frac{N}{2})$ pixels in the feature map and calculate the contrastive loss after computing the loss of the previous windows. We set $w = 0.1$ and $N = 7$. We shift the windows *1, 2, 3, 4* times. From the results in Tab. 6, we can conclude that shifting the windows does not bring improvement while increasing computational cost.

Embedded module Location Our WCL module can be easily embedded in between network layers, allowing to

Location	RMSE (M)	MAE (mm)	iMAE (1/km)
baseline	930.326	269.866	2.575
-1	922.512	264.044	2.030
-2	928.279	267.821	2.150
-3	929.924	264.688	2.594
-4	932.127	270.720	**1.992**
-5	940.022	270.118	2.120
-1 & -2	**912.286**	**263.886**	6.604

Table 7. **Ablation results on the the different locations on KITTI depth completion validation set.** -1 denotes between *layer1* and *layer0*; -2 denotes between *layer2* and *layer1*; -3 denotes between *layer3* and *layer2*; -4 denotes between *layer4* and *layer3*; -5 denotes between *layer5* and *layer4*.

contrast pixels in the representation space. The baseline network has six layers in the decoder stage. We define them as *layer5, layer4, layer3, layer2, layer1,* and *layer0* from bottleneck to final layer. We performed an ablation study to determine how much improvement our module can bring when placed at different locations within the network. We set $w = 0.1$ and $N = 7$ in all the experiments except the last one. In the last one, we set $w_1 = 0.1$ for the contrastive loss between *layer1* and *layer0* and $w_2 = 0.0001$ between *layer2* and *layer1*. From the results in Tab. 7, we can find that the closer to the final layer, the better the results. The outputs from the layers near the final layer present features strongly related to final depth maps, while the outputs near the bottleneck show higher-level, semantical information. Partitioning and contrasting the outputs near the bottleneck break semantic information and cause degradation.

5. Conclusion

We presented a Window-based Contrastive Learning (WCL) module for depth prediction. Our approach partitions the image into windows, and the contrastive loss is implemented within each. Accordingly, it constructs and sorts positive and negative pairs, then enlarges the gap between the two in feature space, which makes depth distribution more meaningful of the real depth in the scene. We evaluate our method on multiple depth prediction tasks, such as depth completion, depth estimation, and self-supervised depth estimation, reporting consistent improvements.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. *arXiv preprint arXiv:2210.09071*, 2022. 3

[2] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11746–11752. IEEE, 2021. 3

[3] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021. 3

[4] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 2

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 3

[6] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020. 3, 4

[7] Hu Chen, Hongyu Yang, and Yi Zhang. Depth completion using geometry-aware embedding. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8680–8686. IEEE, 2022. 3

[8] 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. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 4

[10] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10615–10622, 2020. 2

[11] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018. 1, 2

[12] Andrea Conti, Matteo Poggi, and Stefano Mattoccia. Sparsity agnostic depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5871–5880, January 2023. 2

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 4

[14] 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. 3

[15] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 3

[16] 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

[17] 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. 1, 3, 5, 7, 8

[18] Rizhao Fan, Zhigen Li, Matteo Poggi, and Stefano Mattoccia. A cascade dense connection fusion network for depth completion. In *The 33rd British Machine Vision Conference*, 2022. 1, 2

[19] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016. 3

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 5, 6, 7

[21] 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, 3, 5, 6, 7

[22] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 3

[23] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3, 4

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[26] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16291–16301, 2021. 4

[27] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13656–13662. IEEE, 2021. 1

[28] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[29] Xiaowen Jiang, Valerio Cambareri, Gianluca Agresti, Cynthia Ifeyinwa Ugwu, Adriano Simonetto, Fabien Cardinaux, and Pietro Zanuttigh. A low memory footprint quantized neural network for depth completion of very sparse time-of-flight depth maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2687–2696, 2022. 1

[30] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12642–12652, 2021. 3

[31] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. *arXiv preprint arXiv:2105.00957*, 2021. 3, 4

[32] Yanjie Ke, Kun Li, Wei Yang, Zhenbo Xu, Dayang Hao, Liusheng Huang, and Gang Wang. Mdanet: Multi-modal deep aggregation network for depth completion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4288–4294. IEEE, 2021. 2

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 3

[34] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020. 3

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

[36] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 1, 3, 6, 7

[37] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1873–1881, 2021. 3

[38] Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, Chong Zhang, et al. A multi-scale guided cascade hourglass network for depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–40, 2020. 1, 2, 6

[39] 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

[40] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 3

[41] Lukas Liebel and Marco Körner. Multidepth: Single-image depth estimation via multi-task regression and classification. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1440–1447. IEEE, 2019. 3

[42] Yuankai Lin, Tao Cheng, Qi Zhong, Wending Zhou, and Hua Yang. Dynamic spatial propagation network for depth completion. *arXiv preprint arXiv:2202.09769*, 2022. 2

[43] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[44] 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. 2, 3, 4, 8

[45] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3288–3295. IEEE, 2019. 2, 6, 7

[46] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4796–4803. IEEE, 2018. 2

[47] Danish Nazir, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Semattnet: Towards attention-based semantic aware guided depth completion. *arXiv preprint arXiv:2204.13635*, 2022. 1

[48] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision*, pages 120–136. Springer, 2020. 2, 6

[49] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5848–5854. IEEE, 2018. 3

[50] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[51] Matteo Poggi, Fabio Tosi, Filippo Aleotti, and Stefano Mattoccia. Real-time self-supervised monocular depth estimation without gpu. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12, 2022. 3

[52] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5314–5334, 2022. 1, 3

[53] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *6th International Conference on 3D Vision (3DV)*, 2018. 3

[54] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021. 3

[55] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019. 2

[56] Kyeongha Rho, Jinsung Ha, and Youngjung Kim. Guideformer: Transformers for image guided depth completion. In *CVPR*, pages 6250–6259, 2022. 3

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

[58] Yuefan Shen, Yanchao Yang, Youyi Zheng, C Karen Liu, and Leonidas J Guibas. Dcl: Differential contrastive learning for geometry-aware depth synthesis. *IEEE Robotics and Automation Letters*, 7(2):4845–4852, 2022. 3

[59] Fei Sheng, Feng Xue, Yicong Chang, Wenteng Liang, and Anlong Ming. Monocular depth distribution alignment with low computation. *arXiv preprint arXiv:2203.04538*, 2022. 3

[60] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 3

[61] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 6, 7

[62] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 2, 5, 6

[63] Wouter Van Gansbeke, Davy Neven, Bert De Brabandere, and Luc Van Gool. Sparse and noisy lidar completion with rgb guidance and uncertainty. In *2019 16th international conference on machine vision applications (MVA)*, pages 1–6. IEEE, 2019. 2

[64] 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. 3

[65] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 3

[66] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 4

[67] Xiaoyan Wang, Chunping Hou, Liangzhou Pu, and Yonghong Hou. A depth estimating method from a single image using foe crf. *Multimedia Tools and Applications*, 74(21):9491–9506, 2015. 3

[68] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3, 4

[69] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[70] Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. Efficient non-local contrastive attention for image super-resolution. *arXiv preprint arXiv:2201.03794*, 2022. 3

[71] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10539–10548, 2021. 3

[72] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 3, 4

[73] Xin Xiong, Haipeng Xiong, Ke Xian, Chen Zhao, Zhiguo Cao, and Xin Li. Sparse-to-dense depth completion revisited: Sampling strategy and graph construction. In *European Conference on Computer Vision*, pages 682–699. Springer, 2020. 3

[74] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019. 3

[75] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022. 4, 8

[76] Jiahui Zhang, Shijian Lu, Fangneng Zhan, and Yingchen Yu. Blind image super-resolution via contrastive representation learning. *arXiv preprint arXiv:2107.00708*, 2021. 3

[77] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018. 1

[78] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. CVPR. 3

[79] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4106–4115, 2019. 3

[80] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *International Conference on 3D Vision*, 2022. 3

[81] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing*, 30:5264–5276, 2021. 1, 3

[82] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 3, 5