Efficient Indoor Depth Completion Network Using Mask-adaptive Gated Convolution

Tingxuan Huang[1], **Jiacheng Miao**[1], **Shizhuo Deng**[1, 2], **Tong Jia**[1, 2], **Dongyue Chen**[1, 2, 3, *]

[1]College of Information Science and Engineering, Northeastern University, China
[2]Foshan Graduate School of Innovation, Northeastern University, China
[3]National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, China
{cangshuhuang, mjc24500817}@gmail.com, dengshizhuo@mail.neu.edu.cn, {jiatong, chendongyue}@ise.neu.edu.cn

Abstract

Most indoor depth completion tasks rely on convolutional auto-encoders to reconstruct depth images, especially in areas with significant missing values. While traditional convolution treats valid and missing pixels equally, Partial Convolution (PConv) has mitigated this limitation. However, PConv fails to distinguish the varying degree of invalidity across different missing areas, which highlights the need for a more refined strategy. To solve this problem, we propose a novel system for indoor depth completion tasks that leverages Mask-adaptive Gated Convolution (MagaConv). MagaConv utilizes gated signals to selectively apply convolution kernels based on the characteristics of missing depth data. These gating signals are generated using shared convolution kernels that jointly process depth features and corresponding masks, ensuring coherent weight optimization. Additionally, the mask undergoes iterative updates according to predefined rules. To improve the fusion of depth and color information, we introduce a Bi-directional Aligning Projection (Bid-AP) module, which utilizes a bi-directional projection scheme with global spatial-channel attention mechanisms to filter out depth-irrelevant features from other modalities. Extensive experiments on popular benchmarks, including NYU-Depth V2, DIML, and SUN RGB-D, demonstrate that our model outperforms state-of-the-art methods in both accuracy and efficiency.

Code — https://github.com/htx0601/MagaConv.

Introduction

Depth completion, or depth inpainting, is vital for filling missing pixels in depth images, crucial for applications such as 3D reconstruction (Bascle and Deriche 1993), virtual reality (Newcombe et al. 2011), and autonomous vehicles (Liu et al. 2019). It aims to efficiently replace missing pixels in raw depth maps acquired from sensors like Time-of-flight, structured light, Lidar, and binocular vision. In indoor environments, inherent limitations of these sensors, such as sensor noise, reflections, absorption, or sharp boundaries often result in incomplete data. Overcoming these challenges and developing robust depth completion algorithms is essential for obtaining accurate depth maps.

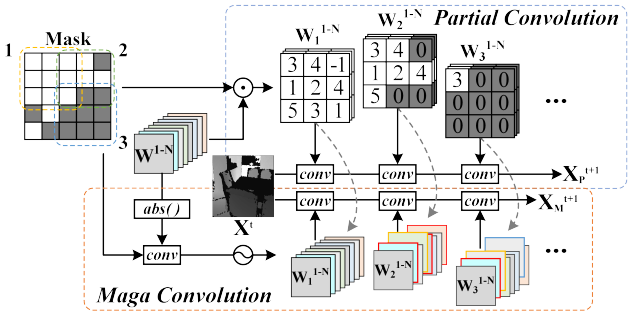

Figure 1: The comparison between Partial Convolution and Maga Convolution, designed to encode incomplete depth images using associated masks. Here, X^t is the input/output feature at encoding step t. W_i is the specific convolution kernels applied at position i. While PConv ensures output from valid pixels, it overlooks the challenge of using the same kernels for various invalidity levels, as it may mask out crucial parameters in W. MagaConv addresses this by selecting kernels tailored to specific invalid patterns.

Recent methods use encoder-decoder architectures like U-Net and its variants (Ronneberger, Fischer, and Brox 2015) to predict depth. However, the vanilla convolution, which treats all pixels equally, including missing ones, can lead to inaccuracies and error propagation in neighboring regions. Approaches like dilated convolutions (Yu and Koltun 2015), partial convolutions (Liu et al. 2018), gated convolutions (Yu et al. 2019), and attention-guided gated convolutions (Chen et al. 2023) aim to improve accuracy by handling missing data by adjusting kernel positions or suppressing invalid features related to missing pixels. However, they have not fully exploited the potential impact of the invalid pixel in extracting depth features.

Take the Partial Convolution (PConv) as an example, as shown in Fig. 1. It employs a binary mask to distinguish between valid and invalid data during convolution. Throughout each convolutional operation, this mask interacts with input features across all layers, ensuring that the resulting outputs only derive from valid pixels, thus guaranteeing reliability. However, this reliability assumption has two key limitations. Firstly, the convolutional receptive field contains

*Corresponding Author

varying numbers of invalid pixels with diverse distributions, and simply discarding these pixels overlooks crucial details. Secondly, employing identical convolution kernels across different invalid contexts lacks adaptability. The convolution kernel's parameters are intended to capture crucial features and patterns. Partially masking key parameters disrupts the learned features, preventing the full utilization of learned information and leading to unreliability.

We introduce Mask-adaptive Gated Convolution (MagaConv), a novel convolutional operation modulated by iteratively updated masks to solve these challenges. It enhances depth feature extraction by selecting convolution kernels based on the specific characteristics of incomplete depth data. It is achieved by dynamically generating gating signals to evaluate each convolutional operation. By employing shared convolution kernels that process both depth features and corresponding masks, MagaConv can determine the degree of invalidity at each position within every channel. This information is then converted into a gating signal, through a unique activation function, to selectively choose kernels in a manner that prevents disruption of their essential parameters. Additionally, MagaConv iteratively updates masks to gradually complete depth features, effectively filling large holes and enabling precise extraction.

After depth coarsely completion and encoding, the next step involves integrating them with color information and decoding. Researchers have explored various RGB-guided approaches (Cheng, Wang, and Yang 2019; Cheng et al. 2020; Zhang et al. 2023) that typically fuse features by concatenating them and applying standard convolutions. However, these approaches face limitations. Firstly, they neglect the differences between color and depth modalities: depth captures geometry, while RGB depicts appearance and texture (Chen et al. 2023). Simply concatenating risks introduces depth-irrelevant features and misses complementary information. Secondly, localized convolution operations fail to capture the global context, crucial for understanding spatial relationships between distant objects.

To tackle these issues, we have considered using transformers for cross-attention mechanisms as a potential remedy (Vaswani et al. 2017). However, due to the limited availability of labeled indoor RGB-D pairs in most public datasets, transformers may struggle to learn the complex relationship between the two modalities (Liu et al. 2021b). Therefore, inspired by spatial-adaptive normalization (Park et al. 2019), we introduce a novel module named Bi-directional Aligning Projection (Bid-AP), facilitating a comprehensive alignment of these modalities from a global perspective.

In general, our main contribution can be summarized as follows:

- We develop an efficient convolutional encoder-decoder network that utilizes our newly proposed MagaConv and Bid-AP to generate high-quality completion of the indoor depth image.
- A Mask-adaptive Gated Convolution (MagaConv) is proposed to extract reliable depth features while considering the degree of invalidity in missing regions. MagaConv utilizes a shared convolution operation and iteratively updated masks to modulate the encoding process.
- A Bi-directional Aligning Projection module (Bid-AP) is proposed, leveraging MLP-based spatial-adaptive normalization to align with color data while filtering out depth-irrelevant features.
- Experimental results demonstrate that our model outperforms the state-of-the-art on three popular benchmarks, including NYU-Depth V2, DIML, and SUN RGB-D datasets.

Related Works

Depth Completion

The task of depth completion aims to generate dense depth maps from incomplete depth images. (Ma and Karaman 2018; Qu, Nguyen, and Taylor 2020; Wang et al. 2023; Yan et al. 2023b,a; Wang et al. 2024) employed encoder-decoder networks to obtain dense depth maps. S2DNet and Deepdnet (Hambarde and Murala 2020; Hegde et al. 2021) proposed a two-stage network, focusing on acquiring approximate depth images and enhancing the primary results.(Gu et al. 2021; Liu et al. 2021a; Zhu et al. 2022) introduced residual depth map completion networks, which utilize residual maps to enhance the initial completion image, resulting in sharper edges. SPN, CSPN, CSPN++, NLSPN, GraphSPN, DySPN (Liu et al. 2017; Cheng, Wang, and Yang 2018; Cheng et al. 2020; Park et al. 2020; Liu et al. 2022; Lin et al. 2022) optimized the SPN algorithms to enhance the prediction of unfamiliar depth values by effectively incorporating known depth information. However, the predicted depth maps still exhibit blurriness attributed to the limitations of vanilla convolution operations in encoding depth features.

Feature Extraction

The presence of missing or unreliable depth pixels in the raw depth map poses a challenge when using VConv to extract features. To overcome the artifacts, researchers have proposed a series of convolutions (Liu et al. 2018; Yu et al. 2019; Chi, Jiang, and Mu 2020; Xie et al. 2019) to avoid the impact of missing value. Besides PConv, (Yu et al. 2019) introduced a Gated Convolution (GConv) that generalizes partial convolution by providing a learnable dynamic feature selection mechanism. (Chi, Jiang, and Mu 2020) proposed Fast Fourier Convolution, which has a larger receptive field and a cross-scale fusion within the convolution. (Xie et al. 2019) introduced a Learnable Bidirectional Attention Maps (LBAM) module that learns feature re-normalization and mask-updating in an end-to-end manner. Nevertheless, these methods did not adequately exploit the use of masks that mark that marks the invalid pixels. To tackle this issue, we propose the MagaConv, a new convolution operation that is modulated by iteratively updated masks.

Multi-modal Data Fusion

Multi-modal feature fusion is another essential aspect compared to depth feature extraction. (Xu et al. 2019; Imran et al. 2019) used direct channel-wise concatenation to fuse features. (Zhong et al. 2019; Li et al. 2020; Zhou and Dong

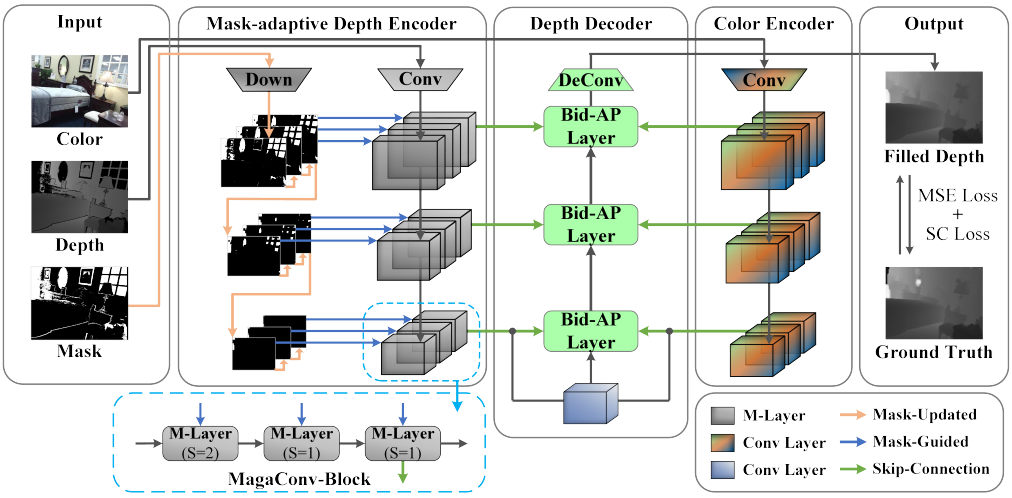

Figure 2: Pipeline of our depth completion model, including the MagaConv architecture and the Bid-AP module we proposed.

2022; Yan et al. 2022a) designed adaptive modules to realize global feature fusion throughout the encoding and decoding procedures.(Chen et al. 2023) proposed Attention Guided Gated-Convolution (AG-GConv) to fuse depth and color features at different scales, effectively reducing the negative impacts of invalid depth data on the reconstruction. Additionally, Rignet, GuideNet, and Ssgp (Yan et al. 2022b; Tang et al. 2020; Schuster et al. 2021; Tang et al. 2024) adopt dual-modal encoder-decoder networks, enabling a multi-level fusion within the network architecture. In this paper, We propose the Bi-directional Aligning Projection (Bid-AP) module, which aims to comprehensively align depth-relevant cues from the two modalities, and fuse the features in a global perspective.

Methods

In this section, we present our overall depth completion network architecture and its two key components: Mask-adaptive Gated Convolution (MagaConv) and Bi-directional Aligning Projection (Bid-AP). Additionally, we introduce the overall loss function, including an MSE loss and a Structure-Consistency loss.

Overall Network Architecture

The pipeline of our model is shown in Fig. 2, it aims to fill all the missing depth pixels in raw depth images with the guidance of color images. The network consists of three components: (i) Mask-adaptive Depth Encoder, (ii) Color Encoder, and (iii) Depth Decoder with Bid-AP decoding layer.

(i) The Mask-adaptive Depth Encoder is designed to extract reliable depth features while addressing missing data issues. The encoding procedure operates in three levels: MagaConv-Blocks (M-Blocks), MagaConv-Layers (M-Layers), and MagaConv. The encoder consists of three M-Blocks, each of which downsamples the depth feature by half. Within each block, input data undergoes three sequential M-Layers, each steered by distinct masks that are refreshed per block and layer. The initial layer strides by 2, while the subsequent layers use a stride of 1. Only the output features from the final layer of each block are then forwarded to the decoder via skip connections. Nevertheless, each M-Layer consists of three MagaConv heads to facilitate feature extraction using diverse kernel sizes.

(ii) The Color Encoder takes RGB images as inputs to extract depth-relevant features. Its architecture mirrors that of the Depth Encoder, with three convolutional layers in each block. However, it utilizes standard convolutional layers instead of MagaConv, integrates residual connections, and does not rely on pre-trained parameters.

(iii) The Depth Decoder, enhanced with the Bid-AP module, leverages multi-scale, skip-connected pathways to reconstruct a complete depth image. The Bid-AP module integrates features from both encoders, ensuring thorough alignment and capturing complementary details. By applying the Bid-AP module across different scales, the Depth Decoder is capable of generating high-quality depth completion results.

Mask-adaptive Gated Convolution

In tackling the challenges posed by invalid pixels during convolution operations, the introduction of PConv partially mitigates their negative effects. To address these issues more effectively, we introduce the MagaConv operation. It employs a convolution kernel selection mechanism to handle invalid patterns without compromising the essential parameters, achieved through the utilization of masks to regulate the convolution process.

MagaConv Operation. Considering a raw depth map $X^t \in \Re^{h\times w}$ and a vanilla convolution kernel W with the size of $k \times k$ that processes a group of pixels. The output O^{Conv} at the position (i, j) can be defined as follows:

$$O^{Conv}_{(i,j)} = \sum_{m=-k}^{k} \sum_{n=-k}^{k} W_{(i+m,j+n)} * X_{(i+m,j+n)}. \quad (1)$$

Then, denote $M \in \Re^{h\times w}$ as the corresponding mask of

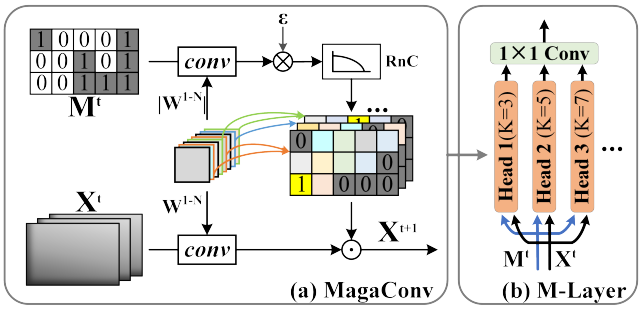

Figure 3: Details of the MagaConv and M-Layer.

the X, in which 1 of the mask map represents a missing depth pixel. It can be defined as follows:

$$M_{i,j} = \begin{cases} 1 & \text{if} \quad X_{i,j} \leq 0 \\ 0 & \text{if} \quad X_{i,j} > 0 \end{cases} . \tag{2}$$

This mask can be used to mark invalid pixels and then measure the suitability of each pixel within the receptive fields of W. Specifically, we suppose that the larger absolute parameter in W is likely to be an important reference. If the missing value is related to that parameter, the output of the convolutional kernel becomes less reliable at that position. Based on the observation, we adopt the same convolution kernel with absolute parameters $|W|$ to conduct convolution operation with the mask M, quantitatively measuring the suitability in a specific position. This operation is defined as follows:

$$O^{Mask}_{i,j} = \epsilon * \sum_{m=-k}^{k} \sum_{n=-k}^{k} |W_{(i+m,j+n)}| * M_{(i+m,j+n)}, \tag{3}$$

where $O^{Mask}_{i,j}$ indicates the convolution kernel is unsuitable at the position (i, j) when it encounters invalid areas. $\epsilon \in (0, 1)$ is a learnable parameter used to normalize $O^{mask}_{i,j}$, enhancing the training robustness. Notably, the mask M is replicated along the channel axis before the convolution operation, ensuring that its shape remains the same with the depth features. The value $O^{Mask}_{i,j}$ can be transformed into a penalty term, effectively punishing convolutional kernels that heavily rely on invalid areas. The penalty term is defined as follows:

$$X^{t+1}_{i,j} = RnC(O^{Mask}_{i,j}) \otimes O^{Conv}_{i,j}, \tag{4}$$

where $\otimes$ denotes element-wise multiplication, and the RnC (Reverse-and-Cut) is an activation function we proposed, which is defined as follows:

$$RnC(x) = \left[ReLU(e^{-x} - 0.5)\right] \times 2, x > 0. \tag{5}$$

The RnC function acts as a control mechanism, modulating the convolution process based on the suitability of the key parameters at each position. It is evident that $RnC(x) \in [0, 1]$. Higher values indicate MagaConv behaves more like vanilla convolution, while lower values lead it to act more like Partial Convolution. Additionally, similar to other activation functions, RnC also incorporates a non-linear and threshold-based activation during guidance, enabling convolution kernels to learn complex relationships.

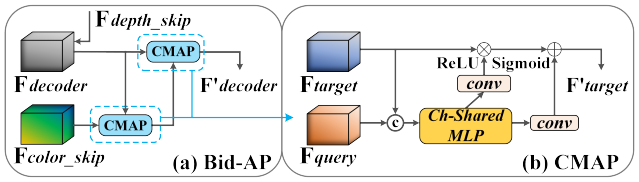

Figure 4: Details of the Bid-AP and CMAP.

Mask Update. To provide particular instructions for each MagaConv-Layer at different scales, masks are updated in every block and layer according to distinct rules. In layer l of block b, the mask updating rule is defined as follows:

$$\begin{aligned} (1 - M(b, l+1)) &= MP((1 - M(b, l)), s = 1) \\ (1 - M(b+1, l)) &= MP((1 - M(b, l+2)), s = 2) \end{aligned}, \tag{6}$$

where MP denotes the MaxPooling2D operation. It is to provide accurate localization of the boundary regions and prevent the convolution process from rapidly filling large holes. s denotes the stride parameter, which is aligned with the stride of the M-Layer. In this manner, a specific mask pixel is updated if surrounding areas contain at least one valid mark, and gradually the mask will become fully valid.

Bi-directional Aligning Projection (Bid-AP)

As previously mentioned, aligning cross-modalities is crucial for effectively integrating color and depth features while filtering out irrelevant aspects like surface appearance or textures. Therefore, we introduce Bid-AP, which aims to identify color features relevant to depth, thereby making precise adjustments to the encoding feature and enriching the overall depth representation.

The architecture of the Bid-AP module is illustrated in Fig. 4 (a). This module aligns the encoder features of the two modalities, $F_{decoder}$&F_{depth_skip} and F_{color_skip}, from the depth (D) and color (C) respectively. It consists of two parallel streams ($D \rightarrow C$ and $C \rightarrow D$) to perform a bi-directional information exchange through the Cross-modality Aligning Projection (CMAP). Initially, after combining $F_{decoder}$&F_{depth_skip} by VConv, $D \rightarrow C$ filters out depth-irrelevant features from the color information to emphasize vital aspects like geometric properties. Subsequently, $C \rightarrow D$ refines and enriches the representation of depth features. Finally, the decoding features undergo upsampling via de-convolution operations.

The CMAP is depicted in Fig. 4 (b). It takes the target feature $F_{target} \in \Re^{h \times w \times c}$ and the query feature $F_{query} \in \Re^{h \times w \times c}$ as inputs, with the target and query representing depth and color interchangeably during the bidirectional fusion. In each step, the CMAP projects F_{target} to align the feature into an enriched pattern. This projecting signal is derived from the concatenated feature $F_c = (F_{target}, F_{query})$,

generated through an MLP-based spatial-adaptive normalization. Specifically, firstly, F_c is fused through a 1×1 convolution across the channels, resulting in $F_c' \in \Re^{h \times w \times c}$. Secondly, it enters a channel-shared MLP layer with two hidden layers to attain a unified representation from a global perspective. The adoption of a channel-shared pattern aims to limit the number of learnable parameters and mitigate the risk of over-fitting. Thirdly, the output embeddings pass through two distinct convolution kernels that produce two modulated signals: $\gamma \in \Re^{h \times w \times c}$ and $\beta \in \Re^{h \times w \times c}$. The overall projection process can be defined as follows:

$$F'_{target} = ReLU(\gamma) \otimes F_{target} + Sigmoid(\beta), \quad (7)$$

where $\otimes$ denotes element-wise multiplication.

In general, the Bid-AP module achieves thorough fusion as Fig. 5 through three key advantages. **(i) Adaptive Feature selection:** The Bid-AP is to align the features from two modalities, including coarsely complete depth and color features. This alignment surpasses direct feature fusion, representing a learnable selection process that emphasizes the most informative elements from each modality. By avoiding the negative impact caused by directly concatenating both features or introducing depth-irrelevant features, Bid-AP realizes a learnable selection process that emphasizes the most informative elements from each modality. **(ii) Bi-directional Aligning:** The $D \rightarrow C$ acts as a filtering mechanism, converting color skip features into essential features like outlines and semantics. Conversely, $C \rightarrow D$ enriches the target features with necessary attributes without overwhelming them with irrelevant contexts from other modalities. **(iii) Global Perspective with Limited Resources** The CMAP module within the Bid-AP facilitates a global perspective on feature alignment while being resource-efficient. The spatial-channel attention mechanism, implemented via the channel-shared MLP, adapts to crucial contexts across every position, allowing for effective interaction through customized normalization parameters. Compared to other fusion mechanisms like cross-attention, our model achieves efficient alignment with less reliance on training data.

Loss

It is worth noting that depth maps often contain crucial boundary information that may not be effectively captured by the Mean Squared Error (MSE). Therefore, we employed Structure-Consistence (SC) loss function to address this limitation. The Structure-Consistence loss function can be formulated as follows:

$$\mathcal{L}_{sc} = \frac{1}{N} \sum_{i=1}^{N} \left| \nabla D_{pred}^{(i)} - \nabla D_{gt}^{(i)} \right|_2^2. \quad (8)$$

$\mathcal{L}_{sc}$ represents the structure-consistence loss, N is the number of samples in the training process, $Dpred^{(i)}$ is the predicted depth map for the i-th sample, $D_{gt}^{(i)}$ is the corresponding ground truth depth map, ∇ denotes the Laplacian operator for extracting edge information, and $|\cdot|_2^2$ denotes the squared Euclidean norm.

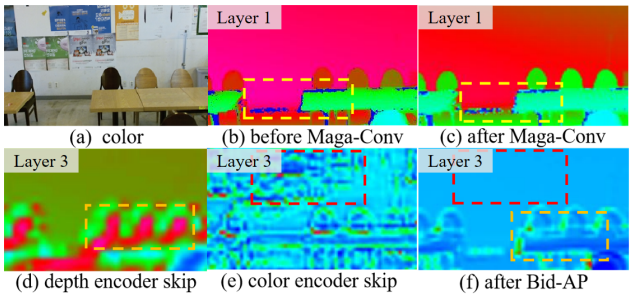

Figure 5: Visualizations of typical features to demonstrate effectiveness on DIML dataset. The missing area in (b) shrinks to (c) after a MagaConv. The green "coarsely complete depth" with unclear boundaries in (d) and the red"depth-irrelevant" features in (e) disappeared after being combined by a Bid-AP (f).

The overall loss function is given by:

$$\mathcal{L}_{all} = \mathcal{L}_{mse} + \mathcal{L}_{sc}. \quad (9)$$

By incorporating the SC loss with MSE loss, the model is encouraged to minimize not only the pixel-wise depth errors but also to preserve the structural integrity and edge information. This leads to more visually accurate and detailed depth completion results.

Experiments

Experimental Setup

We conducted comprehensive experiments on three popular benchmark datasets: NYU-Depth V2, DIML, and SUN RGB-D to validate the performance of the model.

NYU-Depth V2 (Silberman et al. 2012) is the most authoritative and widely used benchmark dataset for depth image completion, which contains 408,473 images collected in 464 different indoor scenes, and 1449 officially labeled images for evaluation.

DIML (Cho et al. 2019) This dataset includes images with typical edge shadows and irregular holes, providing a robust evaluation benchmark for assessing the adaptability of our model to various invalid patterns. We utilize 2000 pairs of labeled samples from the indoor part of the datasets according to the official split.

SUN RGB-D (Song, Lichtenberg, and Xiao 2015) is an extensive dataset comprising 10,335 densely captured RGB-D images obtained from four different sensors. The dataset covers 19 primary scene categories, providing a diverse range of scenes for evaluation. Following the default protocol, we partitioned the datasets into 4,845 images for training and 4,659 ones for testing.

Metrics. The evaluation of indoor depth completion results is based on three criteria: Root Mean Squared Error (RMSE), Relative Error (Rel), and Threshold Accuracy (δ_t) with thresholds $t = 1.10, 1.25, 1.25^2, 1.25^3$.

Implementation Details. Our model was implemented using the PyTorch framework and trained on NVIDIA GTX 2080ti GPU for a total of 100 epochs. We adopted the SGD

Scheme	MagaConv	M-Layer	PConv	GConv	Bi-direction	CMAP	MLP-based	Concat	$\mathcal{L}_{mse}$	$\mathcal{L}_{sc}$	RMSE↓	Rel↓	$\delta_{1.10}$ ↑
A (baseline)	-	-	-	-	-	-	-	-	✓	✓	0.188	0.028	95.6
B (w/ MagaConv)	✓	✓	-	-	-	-	-	-	✓	✓	**0.109**	**0.015**	**97.3**
C	✓	-	-	-	-	-	-	-	✓	✓	0.114	0.016	97.0
D	-	✓	✓	-	-	-	-	-	✓	✓	0.134	0.018	96.4
E	-	✓	-	✓	-	-	-	-	✓	✓	0.127	0.017	96.7
F (w/ Bid-AP)	-	-	-	-	✓	✓	✓	-	✓	✓	**0.113**	**0.016**	**97.1**
G	-	-	-	-	-	✓	✓	-	✓	✓	0.139	0.018	96.1
H	-	-	-	-	✓	✓	-	-	✓	✓	0.125	0.017	96.8
I	-	-	-	-	✓	-	-	✓	✓	✓	0.135	0.018	96.4
J	✓	✓	-	-	✓	✓	✓	-	✓	-	0.087	0.012	98.2
K (complete)	✓	✓	-	-	✓	✓	✓	-	✓	✓	**0.083**	**0.011**	**98.7**

Table 1: Ablation study results for different schemes of the pipeline on the NYU-Depth V2 datasets. RMSE is the main metric.

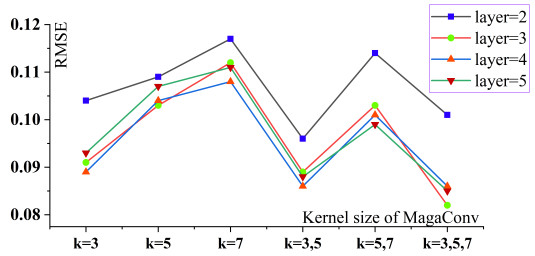

Figure 6: Analyzing the performance across various configurations involving different numbers of down-sampling layers and sets of kernel sizes for MagaConv. The best performance is observed with kernel sizes of 3, 5, and 7 alongside 3 down-sampling layers.

optimizer for training, with a momentum term of 0.95 and a weight decay term of 10^{-4}. The initial learning rate was set to 1×10^{-3} and was halved during the plateau period. The model was trained using end-to-end training methodology.

Ablation Studies

To optimize the proposed framework and evaluate its performance, ablation experiments were conducted on the NYU-Depth V2 datasets. At first, a baseline model (Scheme A) was constructed to resemble the proposed framework, retaining the encoder-decoder architecture but using vanilla convolution in place of the MagaConv operation, and employing direct concatenation of depth and color features at the bottleneck instead of the Bid-AP module. Based on this baseline, three categories with nine protocols (Schemes B to K) were designed by combining different configurations for each module. Schemes B–E evaluated the effectiveness of MagaConv, Schemes F–I examined the impact of the Bid-AP module, and Schemes J and K explored the influence of different loss functions. The details of these protocols are summarized in Tab. 1.

(i) On MagaConv. In the first group, we investigated the impact of different depth encoding methods while maintaining the remaining settings identical to the baseline. Scheme B incorporated the complete MagaConv module, which demonstrated improved performance compared to the baseline (0.109 v.s. 0.188 RMSE). Scheme C involved a modification where the three parallel MagaConv in each M-layer were replaced with a single MagaConv with a 5×5 kernel. This results in a minor performance decrease across all metrics (0.114 v.s. 0.109 RMSE), and indicates that the M-Layer is essential in capturing multi-scale features. In Scheme D and E, substituting the MagaConv with Partial convolution and Gated convolution led to a performance decline across all metrics (0.134 and 0.127 v.s. 0.109 RMSE), indicating that MagaConv indeed enhances the depth features' reliability. Furthermore, the parameters' ablation experiments are depicted in Fig. 6, and the visualization before and after MagaConv is provided in Fig. 5 (b) and (c). This also suggests that the MagaConv module effectively filters out invalid features, resulting in a more reliable feature representation.

(ii) On Bid-AP. In the second group, various fusion schemes were integrated into the baseline. Scheme F, featuring our novel Bid-AP module, displayed a significant performance boost compared to Scheme A (0.113 v.s. 0.188 RMSE), showcasing that the alignment of depth and color features by Bid-AP enhances depth map reconstruction accuracy. In Scheme G, replacing the bi-directional module with a unidirectional approach ($C \rightarrow D$) resulted in a performance decline (0.139 v.s. 0.113 RMSE), implying that Bid-Aligning aids in the filtering and fusion process. In scheme H, the channel-shared MLP within the CMAP was removed to facilitate a localized fusion process. This resulted in a noticeable decrease across all metrics (0.125 v.s. 0.113 RMSE), indicating that the global perspective plays an important role in the comprehensive alignment features. In scheme I, conventional concatenation and convolution structures replaced CMAP. The findings suggest that the fusion strategy centered around normalization, a core concept employed by CMAP, proves to be more effective (0.135 v.s. 0.113 RMSE). Additionally, the visualization of features related to Bid-AP is presented in Fig. 5 (d-f), demonstrating that depth-irrelevant features are effectively filtered out in the color encoder and seamlessly fused with the extracted depth features, further validating its effectiveness.

(iii) On loss function. In the last group, two different settings of loss functions are evaluated. Scheme K, which integrated both the losses significantly outperforms scheme J which only employed the MSE loss (0.083 v.s. 0.087 RMSE). It demonstrates the effectiveness of integrating the structure-consistency loss into our approach, leading to enhanced performance in depth map completion.

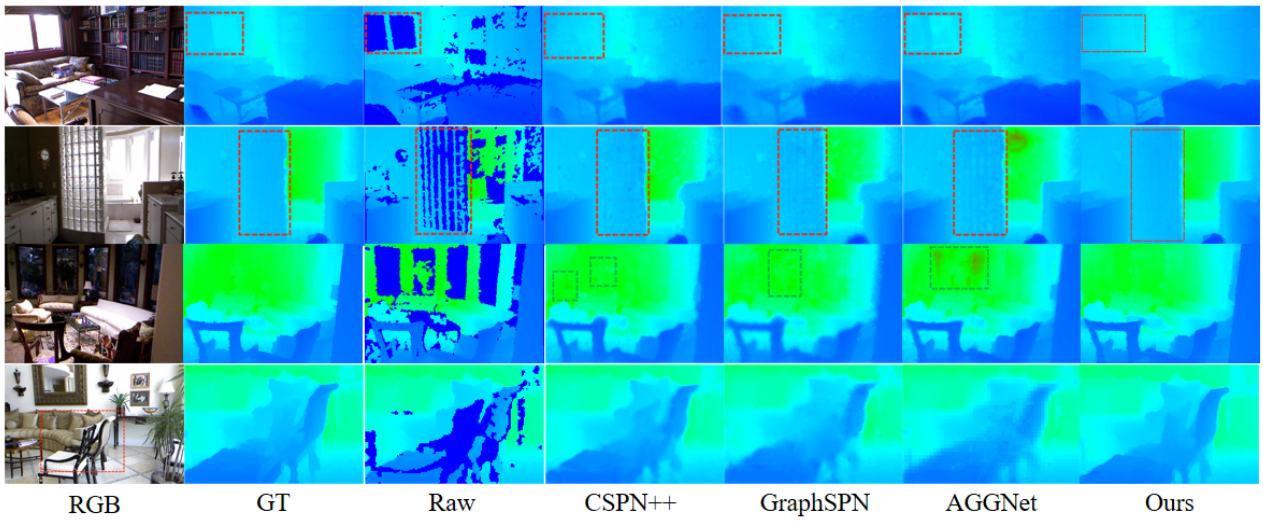

Figure 7: Depth completion comparison results with different methods on NYU-Depth V2.

Method	Params	RMSE↓	Rel↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
CSPN++	17.4 M	0.173	0.02	96.3	98.6	99.5
NLSPN	25.8 M	0.153	0.015	98.6	99.6	99.9
RDF-GAN	- M	0.139	0.013	98.7	99.6	99.9
GraphCSPN	- M	0.133	0.012	98.8	99.7	99.9
AGG-Net	129.1 M	0.092	0.014	99.4	99.9	100.0
CFormer	146.7 M	0.091	0.012	99.6	99.9	100.0
TPVD	31.2 M	0.086	0.010	99.7	99.9	100.0
Ours	30.1 M	**0.083**	0.011	**99.7**	**99.9**	**100.0**

Table 2: Quantitative evaluation on NYU-Depth V2 dataset.

Benchmark	Method	RMSE↓	Rel↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
DIML	CSPN++	0.162	0.033	96.1	98.7	99.6
	DfuseNet	0.143	0.023	98.4	99.4	99.9
	DM-LRN	0.149	0.015	99.0	99.6	99.9
	NLSPN	0.114	0.013	99.2	99.7	99.9
	AGG-Net	0.086	0.011	99.6	99.9	100.0
	Ours	**0.060**	**0.010**	**99.8**	**99.9**	**100.0**
SUN RGBD	CSPN++	0.295	0.137	95.6	97.5	98.4
	NLSPN	0.267	0.063	97.3	98.1	98.5
	RDF-GAN	0.255	0.059	96.9	98.4	99.0
	AGG-Net	0.202	0.038	98.5	99.0	99.4
	Ours	**0.197**	0.039	**98.5**	**99.2**	**99.6**

Table 3: Quantitative comparison results with competing methods on DIML and SUN RGB-D datasets.

Comparison to State-of-the-art

To evaluate the performance of our proposed model, we conducted comparative experiments against state-of-the-art depth completion methods.

On NYU-Depth V2. The quantitative comparison results with other state-of-the-art methods (Cheng et al. 2020; Park et al. 2020; Wang et al. 2022; Liu et al. 2022; Chen et al. 2023; Zhang et al. 2023; Yan et al. 2024) on NYU-Depth V2 datasets are shown in Tab. 3. Our model performs well across all metrics while maintaining relatively low parameter counts. Visual results in Fig. 7 further emphasize its effectiveness, demonstrating clearer details in challenging scenarios. For instance, our method recovers missing window regions with greater clarity in the first two rows than competitors. In the bottom row, it captures finer details, such as sharper chair edges, and avoids the unrealistic artifacts observed in other methods. Additionally, efficiency tests on a single RTX 3090 GPU at 192×320 resolution show our model achieves 101.7 GFlops, 41ms runtime, and 24.4 FPS. These results demonstrate its suitability for real-time applications with reduced computational demand and improved efficiency.

On DIML and SUN RGB-D. Regarding the dataset DIML, our model is also compared to state-of-the-art methods (Cheng et al. 2020; Shivakumar et al. 2019; Senushkin et al. 2021; Park et al. 2020; Chen et al. 2023). Our model outperforms these competing methods in all three metrics, with a remarkable 20% improvement in RMSE. For the datasets SUN RGB-D, our model is evaluated against comparative methods including (Cheng et al. 2020; Park et al. 2020; Wang et al. 2022; Chen et al. 2023), and also achieved competing performance.

Conclusion

In our research, we introduced a novel method for indoor depth completion by integrating the MagaConv and Bid-AP modules to improve accuracy and reliability. The MagaConv architecture strategically selects convolution kernels based on updated masks, facilitating precise depth feature extraction. Bid-AP aligns features from two modalities using a global bi-directional projection approach. Our model outperformed current state-of-the-art methods on datasets with relatively low parameter counts. Looking ahead, while our focus lies on indoor depth completion with TOF cameras rather than sparse depth completion in the future, we aim to extend this innovative technique to diverse applications, to better contribute to further tasks.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (62202087, 62173083, U22A2063), Guangdong Basic and Applied Basic Research Foundation 2024A1515010244, Fundamental Research Funds for the Central Universities (N2404008, N2404011), and the 111 Project B16009.

References

Bascle, B.; and Deriche, R. 1993. Stereo matching, reconstruction and refinement of 3D curves using deformable contours. In *ICCV*, 421–430.

Chen, D.; Huang, T.; Song, Z.; Deng, S.; and Jia, T. 2023. AGG-Net: Attention Guided Gated-convolutional Network for Depth Image Completion. 8853–8862.

Cheng, X.; Wang, P.; Guan, C.; and Yang, R. 2020. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10615–10622.

Cheng, X.; Wang, P.; and Yang, R. 2018. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European conference on computer vision (ECCV)*, 103–119.

Cheng, X.; Wang, P.; and Yang, R. 2019. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2361–2379.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33: 4479–4488.

Cho, J.; Min, D.; Kim, Y.; and Sohn, K. 2019. A Large RGB-D Dataset for Semi-supervised Monocular Depth Estimation. *CoRR*.

Gu, J.; Xiang, Z.; Ye, Y.; and Wang, L. 2021. DenseLiDAR: A real-time pseudo dense depth guided depth completion network. *IEEE Robotics and Automation Letters*, 6(2): 1808–1815.

Hambarde, P.; and Murala, S. 2020. S2DNet: Depth estimation from single image and sparse samples. *IEEE Transactions on Computational Imaging*, 6: 806–817.

Hegde, G.; Pharale, T.; Jahagirdar, S.; Nargund, V.; Tabib, R. A.; Mudenagudi, U.; Vandrotti, B.; and Dhiman, A. 2021. Deepdnet: Deep dense network for depth completion task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2190–2199.

Imran, S.; Long, Y.; Liu, X.; and Morris, D. 2019. Depth coefficients for depth completion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12438–12447. IEEE.

Li, A.; Yuan, Z.; Ling, Y.; Chi, W.; Zhang, C.; et al. 2020. A multi-scale guided cascade hourglass network for depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 32–40.

Lin, Y.; Cheng, T.; Zhong, Q.; Zhou, W.; and Yang, H. 2022. Dynamic spatial propagation network for depth completion. In *Proceedings of the aaai conference on artificial intelligence*, volume 36, 1638–1646.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; and Catanzaro, B. 2018. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 85–100.

Liu, L.; Song, X.; Lyu, X.; Diao, J.; Wang, M.; Liu, Y.; and Zhang, L. 2021a. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *AAAI*, volume 35, 2136–2144.

Liu, S.; De Mello, S.; Gu, J.; Zhong, G.; Yang, M.-H.; and Kautz, J. 2017. Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems*, 30.

Liu, S.; Liu, L.; Tang, J.; Yu, B.; Wang, Y.; and Shi, W. 2019. Edge computing for autonomous driving: Opportunities and challenges. *Proceedings of the IEEE*, 107(8): 1697–1716.

Liu, X.; Shao, X.; Wang, B.; Li, Y.; and Wang, S. 2022. Graphcspn: Geometry-aware depth completion via dynamic gcns. In *European Conference on Computer Vision*, 90–107. Springer.

Liu, Y.; Sangineto, E.; Bi, W.; Sebe, N.; Lepri, B.; and Nadai, M. 2021b. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34: 23818–23830.

Ma, F.; and Karaman, S. 2018. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE international conference on robotics and automation (ICRA)*, 4796–4803.

Newcombe, R. A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A. J.; Kohi, P.; Shotton, J.; Hodges, S.; and Fitzgibbon, A. 2011. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE international symposium on mixed and augmented reality*, 127–136.

Park, J.; Joo, K.; Hu, Z.; Liu, C.-K.; and So Kweon, I. 2020. Non-local spatial propagation network for depth completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 120–136. Springer.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

Qu, C.; Nguyen, T.; and Taylor, C. 2020. Depth completion via deep basis fitting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 71–80.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Schuster, R.; Wasenmuller, O.; Unger, C.; and Stricker, D. 2021. Ssgp: Sparse spatial guided propagation for robust and generic interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 197–206.

Senushkin, D.; Romanov, M.; Belikov, I.; Patakin, N.; and Konushin, A. 2021. Decoder modulation for indoor depth completion. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2181–2188. IEEE.

Shivakumar, S. S.; Nguyen, T.; Miller, I. D.; Chen, S. W.; Kumar, V.; and Taylor, C. J. 2019. Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 13–20.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *Computer Vision – ECCV 2012*, 746–760.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 567–576.

Tang, J.; Tian, F.-P.; An, B.; Li, J.; and Tan, P. 2024. Bilateral Propagation Network for Depth Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9763–9772.

Tang, J.; Tian, F.-P.; Feng, W.; Li, J.; and Tan, P. 2020. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30: 1116–1129.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NIPS*, 30.

Wang, H.; Wang, M.; Che, Z.; Xu, Z.; Qiao, X.; Qi, M.; Feng, F.; and Tang, J. 2022. Rgb-depth fusion gan for indoor depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6209–6218.

Wang, Y.; Li, B.; Zhang, G.; Liu, Q.; Gao, T.; and Dai, Y. 2023. Lrru: Long-short range recurrent updating networks for depth completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9422–9432.

Wang, Y.; Zhang, G.; Wang, S.; Li, B.; Liu, Q.; Hui, L.; and Dai, Y. 2024. Improving Depth Completion via Depth Feature Upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21104–21113.

Xie, C.; Liu, S.; Li, C.; Cheng, M.-M.; Zuo, W.; Liu, X.; Wen, S.; and Ding, E. 2019. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8858–8867.

Xu, Y.; Zhu, X.; Shi, J.; Zhang, G.; Bao, H.; and Li, H. 2019. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2811–2820.

Yan, Z.; Li, X.; Wang, K.; Chen, S.; Li, J.; and Yang, J. 2023a. Distortion and Uncertainty Aware Loss for Panoramic Depth Completion. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 39099–39109. PMLR.

Yan, Z.; Li, X.; Wang, K.; Zhang, Z.; Li, J.; and Yang, J. 2022a. Multi-modal Masked Pre-training for Monocular Panoramic Depth Completion. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 378–395. Cham: Springer Nature Switzerland. ISBN 978-3-031-19769-7.

Yan, Z.; Lin, Y.; Wang, K.; Zheng, Y.; Wang, Y.; Zhang, Z.; Li, J.; and Yang, J. 2024. Tri-Perspective View Decomposition for Geometry-Aware Depth Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4874–4884.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, J.; and Yang, J. 2022b. RigNet: Repetitive image guided network for depth completion. In *European Conference on Computer Vision*, 214–230. Springer.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, J.; and Yang, J. 2023b. DesNet: Decomposed Scale-Consistent Network for Unsupervised Depth Completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3): 3109–3117.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *CVPR*, 4471–4480.

Zhang, Y.; Guo, X.; Poggi, M.; Zhu, Z.; Huang, G.; and Mattoccia, S. 2023. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.

Zhong, Y.; Wu, C.-Y.; You, S.; and Neumann, U. 2019. Deep RGB-D canonical correlation analysis for sparse depth completion. *Advances in Neural Information Processing Systems*, 32.

Zhou, Z.; and Dong, Q. 2022. Self-distilled feature aggregation for self-supervised monocular depth estimation. In *ECCV*, 709–726.

Zhu, Y.; Dong, W.; Li, L.; Wu, J.; Li, X.; and Shi, G. 2022. Robust depth completion with uncertainty-driven loss functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3626–3634.