Semi-Supervised Semantic Depth Estimation using Symbiotic Transformer and NearFarMix Augmentation

Md Awsafur Rahman, Shaikh Anowarul Fattah

Dept. of EEE, BUET, Bangladesh

{1706066,fattah}@eee.buet.ac.bd

Abstract

In computer vision, depth estimation is crucial for domains like robotics, autonomous vehicles, augmented reality, and virtual reality. Integrating semantics with depth enhances scene understanding through reciprocal information sharing. However, the scarcity of semantic information in datasets poses challenges. Existing convolutional approaches with limited local receptive fields hinder the full utilization of the symbiotic potential between depth and semantics. This paper introduces a dataset-invariant semi-supervised strategy to address the scarcity of semantic information. It proposes the Depth Semantics Symbiosis module, leveraging the Symbiotic Transformer for achieving comprehensive mutual awareness by information exchange within both local and global contexts. Additionally, a novel augmentation, NearFarMix is introduced to combat overfitting and compensate both depth-semantic tasks by strategically merging regions from two images, generating diverse and structurally consistent samples with enhanced control. Extensive experiments on NYU-Depth-V2 and KITTI datasets demonstrate the superiority of our proposed techniques in indoor and outdoor environments.

1. Introduction

The integration of depth information, crucial for 3D scene understanding, significantly contributes to applications such as bokeh, autonomous vehicles, robotics, and augmented-virtual reality [4, 8, 23, 28]. Combined with semantic information, it enables comprehensive scene interpretation and fosters a symbiotic relationship [6, 18, 21]. However, obtaining this data from sensors presents formidable challenges, including high costs, sparse data, and inflexibility. As a result, there is growing interest in multi-task systems for interpreting depth and semantics from a single image. Monocular depth estimation remains an ill-posed problem [13], diluting 3D perception in 2D images. The scarcity of semantic labels in datasets further

complicates multi-tasking. Thus, efficient solutions are a pivotal research domain.

Traditional methods for acquiring depth and semantic information have employed Self-Supervised Depth Estimation (SSDE) and Semi-supervised Semantic Segmentation (Semi-Seg) [2, 14, 18, 37, 40, 42] to reduce reliance on labeled data. However, SSDE requires supplementary data like stereo imagery and motion artifacts, limiting its applicability. In terms of performance, SSDE lags behind supervised learning alternatives [1, 3, 27, 29, 38, 39]. Paradoxically, state-of-the-art supervised approaches neglect the symbiotic relationship between depth and semantics. Current methodologies resort to CNN-based strategies [2, 14, 37, 40, 42], which suffer from limited global context awareness [24], leading to localized symbiosis and unresolved issues like blurred object boundaries and inadequate object-background contrast.

Augmentation has been employed to combat overfitting [25, 41]. However, applying it to depth estimation faces challenges, as conventional augmentations risk invalidating depth maps due to object scale and orientation alteration. Semantics-specific augmentations involving copy-pasting consider depth but compromise its integrity post-blending. Depth-specific augmentations aim to enhance depth but impair semantics due to discontinuity. In a depth-semantics multi-task system, a dual-favoring augmentation approach remains elusive.

This paper presents a novel approach to address the aforementioned issues, namely the absence of semantic labels, ineffective symbiosis, and lack of a dual-compensating depth-semantics augmentation. The key contributions of the proposed methodology can be distilled into the following points:

1. The introduced semi-supervised strategy not only tackles the prevalent issue of semantic label scarcity across datasets but also guarantees a dataset-agnostic architecture thus improving model's practicality.

2. The proposed Symbiotic Transformer effectively exploits the symbiotic relationship between depth and se-

mantics, facilitating information exchange across both local and global contexts, thereby fostering comprehensive mutual awareness.

3. A novel augmentation namely NearFarMix is proposed to address the critical issue of overfitting in depth estimation, while also improving semantic tasks. This technique tackles challenges common to both depth and semantics, including compromised integrity, limited diversity, and inconsistencies-discontinuities in semantics while promoting the generation of controlled samples.

2. Related Work

Supervised Depth Estimation: Eigen et al. [9] pioneered a coarse-to-fine approach, followed by Laina et al. [19] with a fully CNN-based residual network. Transformer-based supervised methods [1, 3, 7, 38, 39] have dominated due to their ability to capture long-range dependencies. iDisc [27] utilizes internal discretized representations for improved performance. Notably, state-of-the-art methods neglect the symbiosis between depth and semantics.

Semi-supervised Semantic Segmentation: Semi-supervised Learning [36] is employed in segmentation tasks. Approaches like AG-MMD, SIG, AffineMix, DepthMix [2, 14, 37, 42] use a pre-trained teacher model to obtain semantic pseudo labels. However, relying on convolutional teacher models limits the global semantic interpretation.

Depth-Semantics Symbiosis: Current methodologies [2, 14, 34, 42] explore depth-semantics symbiosis with self-supervised depth estimation and convolutional information sharing. SW-Map [40] utilizes dual encoder-decoder systems, SIG [42] incorporates decoded semantic features into depth features, AG-MMD [37] employs gated attention, and CCAM [2] uses SE-based channel-attention. Limited global context in spatial or channel dimensions constrains the potential of symbiosis.

Augmentation for Depth-Semantics: Semantics-centric augmentations like CutMix, ClassMix [25, 41] blend images, while DepthMix, AffineMix [2, 14] integrate depth information. However, integrity compromise and structural disparities limit their suitability. Depth-oriented augmentations like Cut-Depth, GLPDepth [15, 17] enhance depth at the expense of semantic performance, introducing discontinuity and inconsistency.

In essence, self-supervised depth estimation pales compared to transformer-based supervised counterparts. Semi-supervised semantic segmentation falls short, tethered by only locally-aware teacher model. Current depth-semantics symbiosis methods inadequately exploit the symbiotic potential due to a lack of global context due to their convolutional short-sighted nature. Finally, depth-semantics aug-

mentations often sacrifice depth or semantics, highlighting a struggle to uphold object integrity and structural coherence.

3. Methodology

The proposed research presents a unified framework for jointly modeling the depth map and generating a semantic mask of an RGB image. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the developed model $\mathcal{F}(\Theta)$ predicts probability maps: $\hat{p}_d \in [0, 1]^{H \times W \times 1}$ for maximum depth likelihood and $\hat{p}_s \in [0, 1]^{H \times W \times 150}$ for 150 distinct classes in the semantic mask. The depth map is obtained by, $\hat{D} = \hat{p}_d \odot max\_depth$. Finally, the semantic mask is obtained by, $\hat{S} = \text{Max\_Index}(\hat{p}_s, \text{axis} = 2)$.

3.1. Proposed Semi-supervised Semantic Depth Estimation

The proposed method (Figure 2) combines supervised and semi-supervised learning to generate depth maps and semantic masks simultaneously. The semi-supervised component employs OneFormer [16], a Transformer-based model pre-trained on ADE20K [43], as the teacher model. It produces a pseudo mask utilized by the student model to capture semantics, fostering symbiosis. Simultaneously, the student model learns depth information in a supervised manner using ground truth data. The architecture adapts to the number of classes, but the teacher model maintains a fixed number of classes across datasets, ensuring a dataset-invariant architecture for the student model. The semi-supervised nature enables its application to datasets with only depth information [11, 30, 33, 35], without requiring semantic annotations.

3.2. Architecture Overview

The proposed scheme employs a shared encoder-decoder structure enhanced by a Depth Semantics Symbiosis (DSS) module (Sec. 3.3) for concurrent depth-semantic map generation. The ImageNet-pretrained Max-ViT [32] encoder, integrated with Block and Grid attention mechanisms, captures spatial dynamics. Leveraging skip connections, the decoder refines multilevel encoder outputs ($\mathbf{E_i}$, $i \in [0, 4]$) prior to DSS application. Each decoder level involves upscaling the prior layer ($\mathbf{D_{i-1}}$), merging it with the corresponding encoder output ($\mathbf{E_i}$), and undergoing convolution, normalization, and GELU activation to yield $\mathbf{D_i}$, given by:

$$\mathbf{D'_{i-1}} = \text{Concat}(\mathbf{E_i}, \text{Upsample}(\mathbf{D_{i-1}}))$$
$$\mathbf{D_i} = \text{GELU}(\text{LayerNorm}(\text{Conv}_{3 \times 3}(\mathbf{D'_{i-1}}))) \quad (1)$$

3.3. Proposed Depth Semantics Symbiosis Module

The proposed Depth Semantics Symbiosis Module (DSS) leverages the Symbiotic Transformer (refer to Section 3.3.2) to exploit the symbiotic relationship between
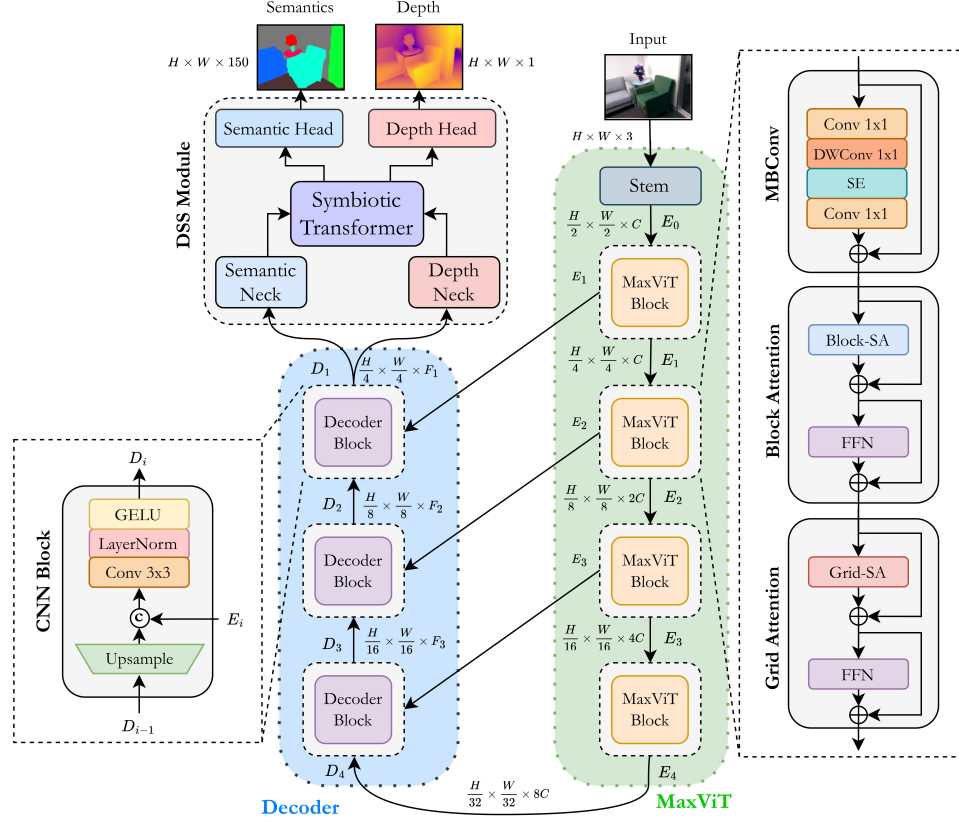
Figure 1. Proposed architecture overview. Input image is processed by Max-ViT encoder and CNN decoder. Then, resultant features undergo the Depth Semantics Symbiosis (DSS) module for joint Depth and Semantics generation via local-global information sharing.
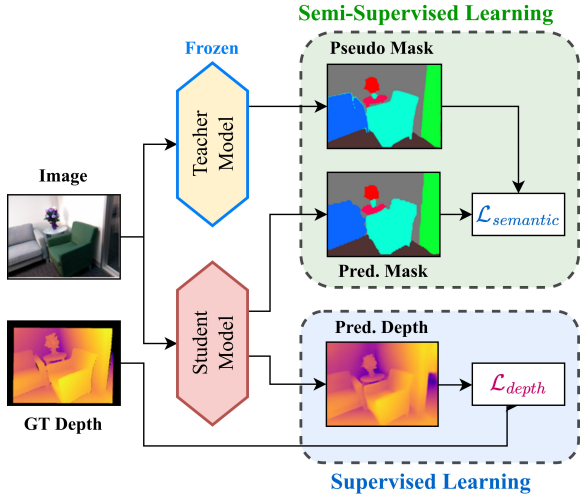

Figure 2. Proposed semi-supervised strategy. The Student model generates supervised Depth via ground-truth and semi-supervised Semantics via Teacher model.

depth and semantic features. In Figure 4, shared decoded features undergo individual processing through distinct Neck blocks, generating independent depth and se-

mantic features. These features are then fed into the Symbiotic Transformer for information exchange within local and global contexts, resulting in depth-aware-semantic and semantic-aware-depth features. Separate Head blocks refine these mutually-aware features, generating the depth map and semantic mask. The mathematical representation of the Depth Semantics Symbiosis Module (DSS) is encapsulated in the following equation:

$$\mathbf{F_s^{KV}} = \mathbf{F_s^{Q}} = \text{Semantic-Neck}(\mathbf{D_1})$$
$$\mathbf{F_d^{KV}} = \mathbf{F_d^{Q}} = \text{Depth-Neck}(\mathbf{D_1})$$
$$\mathbf{F_s} = \text{LG-CAT}(\mathbf{F_d^{Q}}, \mathbf{F_s^{KV}})$$
$$\mathbf{F_d} = \text{LG-CAT}(\mathbf{F_s^{Q}}, \mathbf{F_d^{KV}}) \qquad (2)$$
$$\mathbf{Semantics} = \text{Semantic-Head}(\mathbf{F_s})$$
$$\mathbf{Depth} = \text{Depth-Head}(\mathbf{F_d})$$

where $\mathbf{D_1}$ is the final decoder output. $\mathbf{F_s^{Q}}$ and $\mathbf{F_s^{KV}}$ are semantics' query and key-value features, respectively. $\mathbf{F_d^{Q}}$ and $\mathbf{F_d^{KV}}$ represent depth's query and key-value features. Enhanced features $\mathbf{F_s}$ and $\mathbf{F_d}$ capture depth-aware semantic and semantics-aware depth information, respectively.
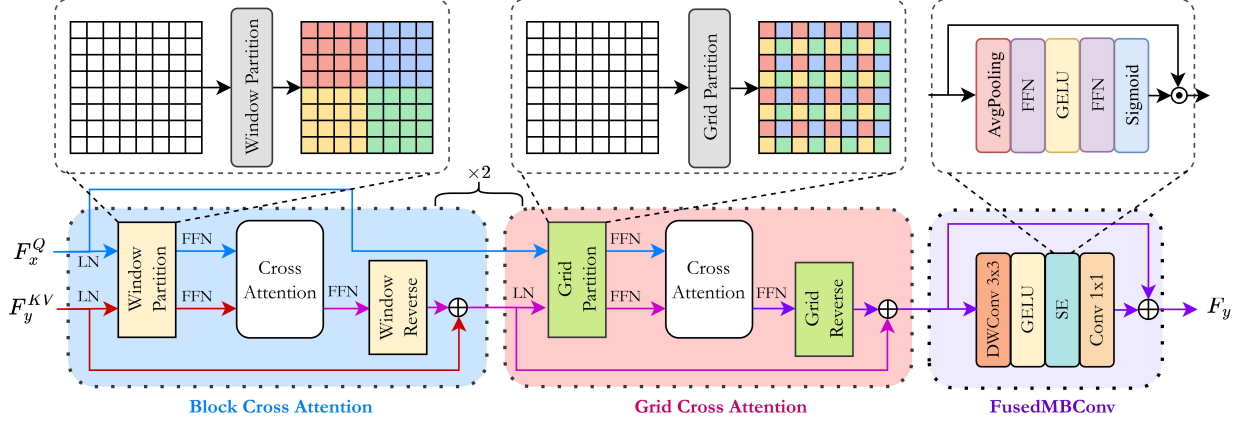
Figure 3. Proposed Local Global Cross Attention Transformer (LG-CAT). Input features undergo Block-Grid Cross Attentions for local-global spatial awareness and FusedMBConv block for channelwise awareness.
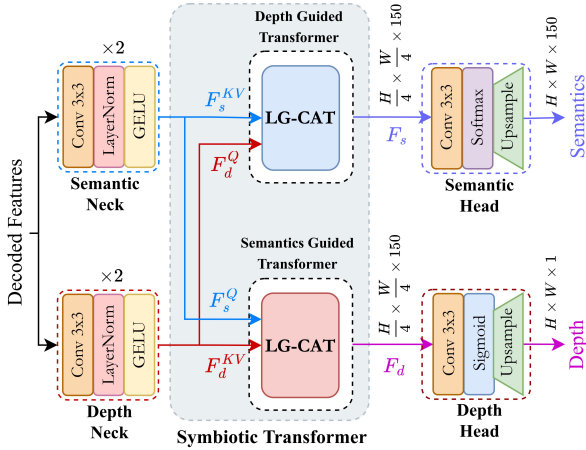

Figure 4. Proposed Depth Semantics Symbiosis (DSS) Module. The decoded features pass through Depth / Semantics Necks, Symbiotic Transformer for local-global sharing, producing Depth and Semantics with separate Heads.

3.3.1 Depth-Semantics Neck

The final decoded features (D_1) are fed into specialized Neck blocks for depth and semantics. These Neck blocks, with distinct weights, generate independent depth and semantic features. Each Neck block includes a $\mathrm{Conv}_{3\times3}$ layer, LayerNorm, and GELU. Mathematically:

$$\mathrm{Neck}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} = \mathrm{Conv}_{3\times3}(\mathbf{x}) \\ \mathbf{x} = \mathrm{LayerNorm}(\mathbf{x}) \\ \mathbf{x} = \mathrm{GELU}(\mathbf{x}) \end{bmatrix} \times 2 \quad (3)$$

3.3.2 Symbiotic Transformer

Figure 4 provides an overview of the proposed module that utilizes depth and semantic features as inputs. It consists of two specialized transformers: Semantics Guided Trans-

former (SGT) and Depth Guided Transformer (DGT), both instances of Local-Global Cross Attention Transformers (LG-CAT). SGT treats semantic features as query features ($\mathbf{F_s^Q}$) and depth features as key-value features ($\mathbf{F_d^{KV}}$), while DGT operates in the opposite manner. These features are transformed into Queries ($\mathbf{Q_s}$, $\mathbf{Q_d}$) and Key-Values ($\mathbf{K_s}$, $\mathbf{K_d}$, $\mathbf{V_s}$, $\mathbf{V_d}$) for cross-attention. The cross-attention operation can be expressed as:

$$\mathrm{CA}\left(\mathbf{Q_x}, \mathbf{K_y}, \mathbf{V_y}\right) = \mathcal{S}mx\left(\frac{\mathbf{Q_x}\mathbf{K_y}^T}{\sqrt{\mathbf{d}}} + \mathbf{B}\right)\mathbf{V_y} \quad (4)$$

Here, $\mathbf{Q_x}$ represents query features, $\mathbf{K_y}$ and $\mathbf{V_y}$ represent key and value features, $\mathcal{S}mx$ denotes softmax, and $\mathbf{d}$ is the query/key dimension. $\mathbf{B}$ represents relative positional bias, sampled similar to [12].

The LG-CAT block (Figure 3) comprises three modules: Block Cross Attention (BCA), Grid Cross Attention (GCA), and FusedMBConv (FMBC). BCA partitions query and key-value features into local dense windows, enabling local symbiotic enrichment. GCA partitions features into global sparse windows, facilitating global information sharing beyond limited receptive fields. FMBC enables inter-channel interactions while maintaining important properties.

The Symbiotic Transformer block can be mathematically expressed as:

$$\begin{bmatrix} {}_{\mathbf{i}}\mathbf{F_y^{KV}} = \mathrm{Block\text{-}Cross\text{-}Attention}(\mathbf{F_x^Q}, {}_{\mathbf{i-1}}\mathbf{F_y^{KV}}) \\ {}_{\mathbf{i+1}}\mathbf{F_y^{KV}} = \mathrm{Grid\text{-}Cross\text{-}Attention}(\mathbf{F_x^Q}, {}_{\mathbf{i}}\mathbf{F_y^{KV}}) \end{bmatrix}_{\times N_s} \quad (5)$$
$$\mathbf{F_y} = \mathrm{FusedMBConv}({}_{\mathbf{N+1}}\mathbf{F_y^{KV}})$$

Here, i denotes the iteration, ranging from 2 to N_s. $\mathbf{F_x^Q}$ represents query features of $\mathbf{x}$, $\mathbf{F_y^{KV}}$ denotes key-value features of $\mathbf{y}$, and $\mathbf{F_y}$ signifies the output features contextualized by $\mathbf{x}$. Specifically, for SGT, $\mathbf{x} = \mathbf{s}$ and $\mathbf{y} = \mathbf{d}$, while for DGT, $\mathbf{x} = \mathbf{d}$ and $\mathbf{y} = \mathbf{s}$.

3.3.3 Depth-Semantics Head

Enhanced depth and semantic features $(\mathbf{F_d}, \mathbf{F_s})$ from the DSS module pass through separate Depth and Semantics Head blocks. The Depth Head applies convolution ($\text{Conv}_{3\times3}$), sigmoid activation, and $4\times$ upsampling, yielding the final depth map ($(H, W, 1)$). The Semantics Head performs convolution, softmax activation, and $4\times$ upsampling, generating the semantic mask ($(H, W, 150)$). The Head blocks can be represented as:

$$
\text{Head}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} = \text{Conv}_{3\times3}(\mathbf{x}) \\ \mathbf{x} = \text{Activation}(\mathbf{x}) \\ \mathbf{x} = \text{Upsample}_{4\times4}(\mathbf{x}) \end{bmatrix} \times 1 \tag{6}
$$

3.4. Proposed NearFarMix Augmentation

The NearFarMix algorithm operates on input elements I_1 and I_2, each comprising an image, a semantic mask, and depth information: $I_i = \left[\text{Image}_i, \text{Mask}_i, \text{Depth}_i\right]$, where i denotes the input identifier. I_1 labels the $Near$ region below a depth threshold thr, while I_2 labels the pre-Far region exceeding the threshold. The algorithm combines the Near and pre-Far regions additively, eliminates the overlapping Far region, and fills the mutually exclusive region from I_2. Ultimately, the pre-Far, overlap, and exclusive regions contribute to only Far regions. The augmented output is represented as $I' = [\text{Image}', \text{Mask}', \text{Depth}']$. Mathematically, this can be expressed as:

$$
\begin{aligned}
M_1 &= (Depth_1 \leq thr); \quad M_2 = (Depth_2 > thr) \\
M_o &= -(M_1 \odot M_2); \quad M_e = (1 - M_1) \odot (1 - M_2) \\
I' &= (I_1 \odot M_1) + (I_2 \odot M_2) + (I_2 \odot M_o) + (I_2 \odot M_e)
\end{aligned} \tag{7}
$$

where $\odot$ denotes element-wise multiplication, M_1 and M_2 denote binary masks derived from the depth maps and the depth threshold thr. M_o represents the overlapped region, while M_e denotes the mutually exclusive region. Finally, the augmented output I' is obtained by summing the four masked components. More details in supplementary (suppl.) material.

In figures 8, and 7, augmented samples is presented generated by proposed NearFarMix. The resultant $Image'$, a synthesis of $Image_1$ and $Image_2$, exemplifies both structural and semantic consistency while upholding object integrity. Existing semantics-centric methods [2, 14] indiscriminately compare pixel depth values across two images, thereby risking object integrity due to their disregard for strategic blending and occlusion considerations. Conversely, depth-centric techniques [15, 17] often induce semantic discord during the image-depth fusion. In contrast, proposed NearFarMix augmentation strategically blends specific regions from two images, duly considering their depth values. This approach ensures both object integrity within the selected regions and overall semantic coherence. Governed by the depth threshold parameter, thr, Near-

FarMix presents flexible blending options and precise control over blend proportion, thereby elevating the diversity of the output.

3.5. Loss Function

The multi-task architecture proposed uses the Scale-Invariant (SI) loss [9] and Jaccard (IoU) loss [5] for depth estimation and semantic segmentation, respectively. Equation (8) represents the SI loss, with d_i as the ground truth depth, $\hat{d}_i$ as the estimated depth, T as the count of valid ground truth pixels, and $g_i = \log_e(\hat{d}_i) - \log_e(d_i)$. Parameters λ and α are 0.85 and 10.

$$
\mathcal{L}_{\text{depth}} = \alpha \sqrt{\frac{1}{T}\sum_i g_i^2 - \frac{\lambda}{T^2}\left(\sum_i g_i\right)^2} \tag{8}
$$

The IoU loss, used for semantic segmentation, is presented in Equation (9), with C representing classes (150 in total), N denoting pixel count, and m_{ij} and $\hat{m}_{ij}$ as the ground truth and predicted semantic labels for pixel j and class i.

$$
\mathcal{L}_{\text{semantic}} = \frac{1}{C}\sum_{i=1}^{C}\left(1 - \mathcal{L}_{iou}^i\right)
$$
$$
\mathcal{L}_{iou}^i = \frac{\sum_{j=1}^{N} m_{ij} \cdot \hat{m}_{ij}}{\sum_{j=1}^{N} m_{ij} + \sum_{j=1}^{N} \hat{m}_{ij} - \sum_{j=1}^{N} m_{ij} \cdot \hat{m}_{ij}} \tag{9}
$$

The overall loss, amalgamating depth and semantic losses, is as follows:

$$
\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{semantic}} \tag{10}
$$

4. Experiments

4.1. Datasets

The ADE20K [43] pre-training dataset contains 20K/2K (training/validation) images with 150 classes. The NYUv2 dataset [31], comprised of 120K RGB-depth pairs from 464 unique indoor scenes, provides a valuable asset for the study. The methodology benefits from a prescribed split of 50K training images and 654 test images, conforms to a depth limit of 10 meters, and is trained at a resolution of 480×640. In contrast, the KITTI dataset [11], based on Eigen et al. [9], features stereo images from 61 distinct outdoor scenes, outlines a 26K/697 training/testing split, and extends the depth threshold to 80 meters. Training for this methodology occurs at a resolution of 704×352.

4.2. Implementation Details

Evaluation Metrics: The proposed method evaluates depth estimation using metrics such as Average Relative error (Abs Rel), Root Mean Squared error (RMS), Average Logarithmic error ($\log_{10}$), Root Mean Squared Log error
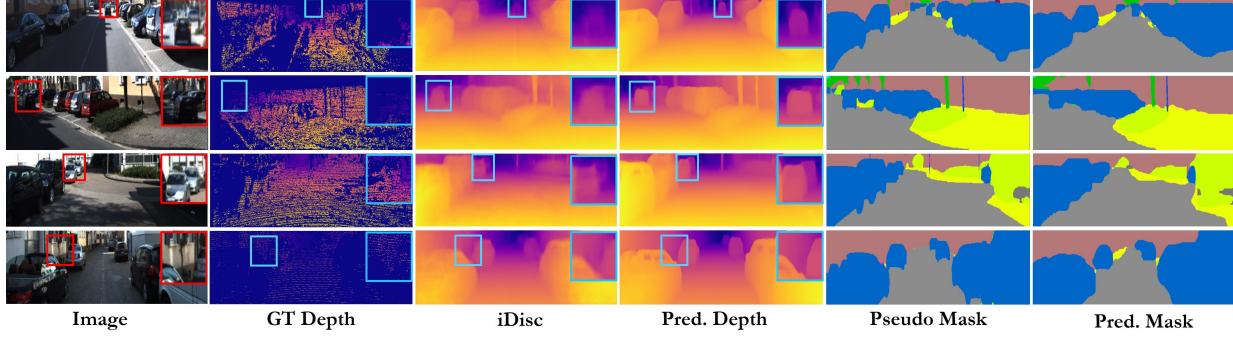
Figure 5. Qualitative results comparing proposed method to previous SOTA on KITTI dataset.
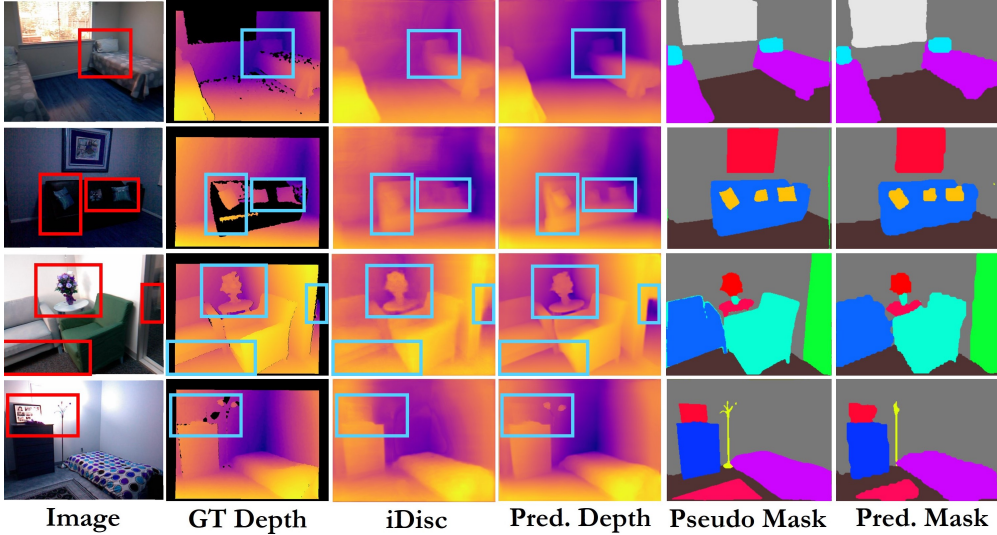

Figure 6. Qualitative results comparing proposed method to previous SOTA on NYUv2 dataset.

(RMS_{log}), and Threshold Accuracy (δ_i) [9]. For semantic segmentation, the mean Intersection over Union (mIoU) is employed [5], quantifying the overlap between predicted and ground truth semantic maps. A threshold of 0.5 is applied to obtain a binary representation before computing the mIoU score.

Network Architecture: The proposed method adopts an ImageNet pre-trained Max-ViT-Large [32] encoder with channels per level $C = 128$. Decoder output channels are $F_i = 2^{i+6}$, where $i \in [1, 3]$. The depth of the Symbiotic Transformer, regulated by N_s, signifies Grid-Block attention repetitions; $N_s = 2$ is found empirically. For Block-Grid attentions, $h = 4$ heads, and $w = 7$ window size is sued, with remaining parameters analogous to Max-ViT. More details in supplementary (suppl.) material.

Training: The encoder-decoder, excluding the Symbiotic Transformer module, is pre-trained on ADE20K [43] for 30 epochs to establish a favorable starting point and to prepare the encoder for feeding the decoder. The AdamW optimizer [22] is applied with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and

weight decay follows the scheduler. For scheduler, learning rate commenced at 4×10^{-6}, linearly increased to 3×10^{-5}, and then reduced with a Cosine function. The Symbiosis Training procedure is executed over 20 epochs using 8 NVIDIA V100 GPUs and a batch size of 4. Augmentations including random horizontal flipping, grayscale, and jitter are applied, complemented by the proposed NearFarMix. During NearFarMix, threshold, thr, is sampled from uniform distribution, $\mathcal{U}\left(max(d_{\min}, D_{\min}), min(d_{\max}, D_{\max})\right)$. Here, $d_{\min}, d_{\max}$ refer to the image depth and $D_{\min}, D_{\max}$ to dataset-specific depths (e.g., 20m and 60m for KITTI, and 1.5m and 6.5m for NYUv2).

4.3. Comparison with Existing Methods

4.3.1 Quantitative Comparison

Tables 1 and 2 quantitatively compare the proposed method against existing techniques on the KITTI outdoor and NYUv2 indoor datasets, respectively. The data substantiate the superior performance of proposed method, consistently

255

Table 1. Comparison on NYUv2 dataset with existing methods. The best result is indicated in **bold**, second best is <u>underlined</u>, and symbols ↑ or ↓ denote higher/lower values are preferable

Method	Abs Rel ↓	RMS ↓	$\log_{10}$ ↓	δ_1 ↑
Eigen et al. [9]	0.158	0.641	-	0.769
DORN [10]	0.115	0.509	0.051	0.828
BTS [20]	0.110	0.392	0.047	0.885
TransDepth [38]	0.106	0.365	0.045	0.900
DPT [29]	0.110	0.367	0.045	0.904
Adabins [3]	0.103	0.364	0.044	0.903
P3Depth [26]	0.104	0.356	0.043	0.898
NeWCRFs [39]	0.095	0.334	0.041	0.922
PixelFormer [1]	0.090	0.322	0.039	0.929
iDisc [27]	<u>0.086</u>	<u>0.313</u>	<u>0.037</u>	<u>0.940</u>
Proposed	**0.080**	**0.289**	**0.034**	**0.948**

Table 2. Comparison on KITTI dataset with existing methods. The best result is indicated in **bold**, second best is <u>underlined</u>, and symbols ↑ or ↓ denote higher/lower values are preferable

Method	Abs Rel ↓	RMS ↓	RMS_{log} ↓	δ_1 ↑
Eigen et al. [9]	0.203	6.307	0.270	0.702
DORN. [10]	0.072	2.727	0.120	0.932
BTS [20]	0.059	2.756	0.090	0.956
TransDepth [38]	0.064	2.755	0.098	0.956
Adabins [3]	0.058	2.360	0.088	0.964
DPT [29]	0.060	2.573	0.088	0.959
NeWCRFs [39]	0.052	2.129	0.079	0.974
PixelFormer [1]	0.051	2.081	0.077	0.976
iDisc	<u>0.050</u>	<u>2.067</u>	<u>0.077</u>	<u>0.977</u>
Proposed	**0.048**	**1.984**	**0.075**	**0.979**

outperforming competitors across all metrics due to its effective utilization of depth-semantics symbiosis. Notably, while many extant methods attempt to harness symbiotic potential through self-supervised depth estimation strategies, these generally underperform compared to supervised methods and are consequently excluded from the tables. Nevertheless, their symbiotic modules are evaluated against proposed method in the ablation study (Sec. 4.4.1).

4.3.2 Qualitative Comparison

Figures 5, and 6 visually contrasts the proposed method with existing techniques on the KITTI and NYUv2 datasets. The areas of exceptional performance by our method are spotlighted within red (image) and blue (depth) rectangles. This qualitative comparison is grounded on three criteria: 1) Edge sharpness, 2) Object-background contrast, and 3) Similarity between prediction and ground truth. As underscored by the marked rectangles, the proposed method evidently demonstrates superiority in these aspects.

Table 3. Ablation analysis of NearFarMix on KITTI and NYUv2 datasets. ↑ signifies improvement, ↓ degradation. Best outcomes are in **bold**; (*) denotes the proposed technique.

Dataset	Method	RMS ↓	Abs Rel ↓	δ_1 ↑	mIoU ↑
KITTI	Baseline	2.045	0.051	0.977	0.663
	CutMix [41]	2.441	0.059	0.962	0.671
	ClassMix [25]	2.302	0.056	0.967	0.693
	AffineMix [2]	2.102	0.053	0.969	0.711
	DepthMix [14]	2.104	0.054	0.968	0.708
	CutDepth [15]	2.068	0.050	0.973	0.640
	V-CutDepth [17]	2.065	0.050	0.975	0.645
	NearFarMix*	**1.984**	**0.048**	**0.979**	**0.731**
NYUv2	Baseline	0.296	0.082	0.945	0.568
	CutMix [41]	0.331	0.094	0.927	0.572
	ClassMix [25]	0.320	0.090	0.930	0.585
	AffineMix [2]	0.314	0.086	0.941	0.603
	DepthMix [14]	0.305	0.084	0.943	0.601
	CutDepth [15]	0.294	0.082	0.946	0.542
	V-CutDepth [17]	0.290	0.081	0.946	0.546
	NearFarMix*	**0.289**	**0.080**	**0.948**	**0.620**

Table 4. Ablation analysis of DSS on KITTI and NYUv2 datasets. ↑ signifies improvement, ↓ degradation. Best outcomes are in **bold**; (*) denotes the proposed technique

Dataset	Method	RMS ↓	Abs Rel ↓	δ_1 ↑	mIoU ↑
KITTI	Baseline	2.295	0.056	0.966	0.615
	SW-Map [40]	2.275	0.055	0.967	0.623
	SIG [42]	2.105	0.053	0.969	0.650
	AG-MMD [37]	2.096	0.052	0.973	0.677
	CCAM [2]	2.080	0.050	0.976	0.695
	DSS*	**1.984**	**0.048**	**0.979**	**0.731**
NYUv2	Baseline	0.323	0.089	0.934	0.519
	SW-Map [40]	0.314	0.087	0.933	0.522
	SIG [42]	0.309	0.086	0.937	0.555
	AG-MMD [37]	0.302	0.085	0.939	0.571
	CCAM [2]	0.300	0.083	0.943	0.589
	DSS*	**0.289**	**0.080**	**0.948**	**0.620**

4.4. Ablation Study

4.4.1 Depth Semantics Symbiosis Module

Table 4 compares the DSS approach and other symbiotic methodologies on KITTI and NYUv2 datasets using depth and semantic metrics, with NearFarMix augmentation included. The results underscore the need for symbiosis, as the baseline model underperforms. Approaches like SW-Map [40] and SIG [42] integrate semantics into depth at various scales, while others like AG-MMD [37] and CCAM [2] employ attention mechanisms yet fall short in global context consideration. In contrast, the DSS method, leveraging its transformer, excels in information exchange across contexts.
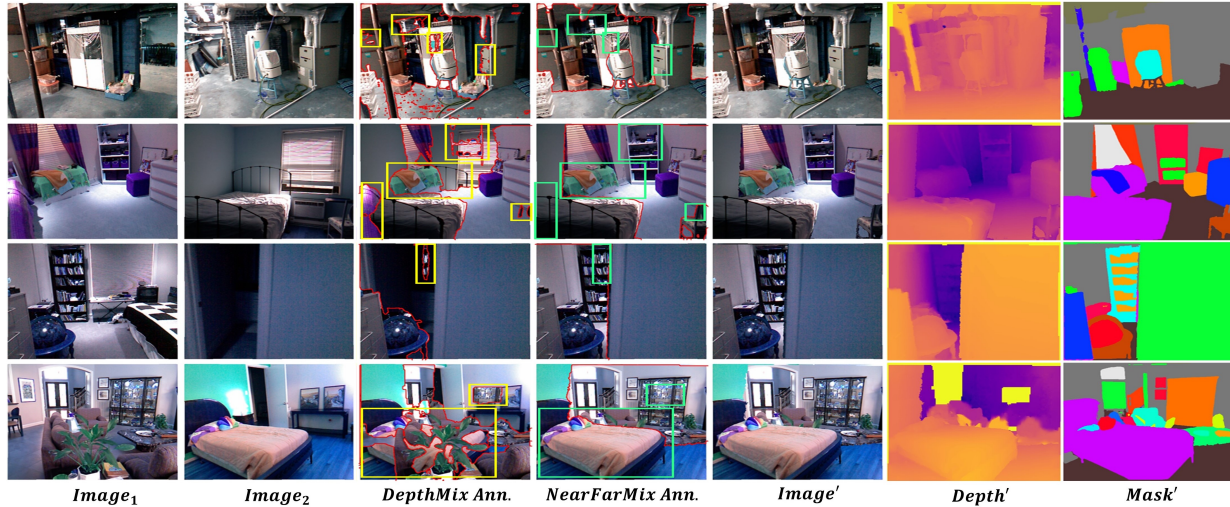
Figure 7. Visual contrasts between NearFarMix and DepthMix on NYUv2, with box areas showing of lost context or object integrity.
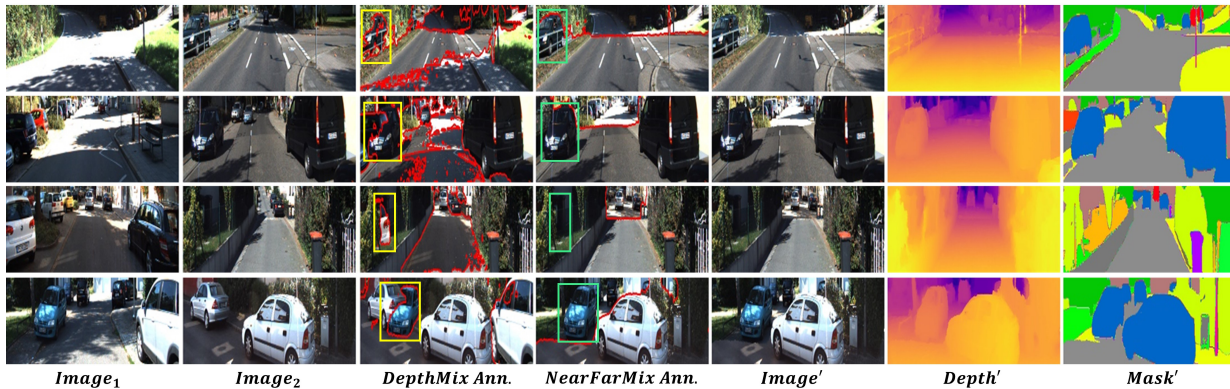

Figure 8. Visual contrasts between NearFarMix and DepthMix on KITTI, with box areas showing of lost context or object integrity.

4.4.2 NearFarMix Augmentation

Table 3 contrasts the proposed NearFarMix augmentation with prevailing methods, utilizing the DSS module as a standard. The 'baseline' denotes our method without NearFarMix. Existing augmentations, enhancing semantics or depth, compromise either object integrity (e.g., CutMix [41], ClassMix [25], AffineMix [2], DepthMix [14]) or create semantic discontinuity (CutDepth [15], V-CutDepth [17]). NearFarMix balances depth and semantics, preserves integrity, and controls sampling diversity. Visual evidence of NearFarMix's superiority over DepthMix [14] augmentation is provided in Fig. 8 and Fig. 7, where the rectangle regions emphasize the method's immunity to object integrity and context loss.

5. Conclusion

This paper brings forth an efficacious semi-supervised method designed to address the challenge of semantic information scarcity across datasets while establishing a dataset-agnostic, generic architecture. Furthermore, it introduces a novel transformer, adept at harnessing the symbiotic potential between depth and semantics, unlocking a richer, integrated scene understanding through comprehensive local-global information sharing. Also presented is a novel augmentation, NearFarMix, purpose-built to enhance both depth and semantics tasks, while upholding structural integrity and avoiding semantic inconsistency. Robust experimental validation underscores the superiority of our method over existing techniques, as clearly demonstrated by significant performance improvements on both NYUv2 and KITTI datasets.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proc. of the IEEE/CVF Winter Conf. on Appl. of Comput. Vision*, pages 5861–5870, 2023. 1, 2, 7

[2] Nitin Bansal, Pan Ji, Junsong Yuan, and Yi Xu. Semantics-depth-symbiosis: Deeply coupled semi-supervised learning of semantics and depth. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5828–5839, 2023. 1, 2, 5, 7, 8

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proc. of the IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 4009–4018, 2021. 1, 2, 7

[4] Benjamin Busam, Matthieu Hog, Steven McDonagh, and Gregory Slabaugh. Sterefo: Efficient image refocusing with stereo vision. In *Proc. of the IEEE/CVF Int. Conf. on Comput. Vision Workshops*, pages 0–0, 2019. 1

[5] Herng-Hua Chang, Audrey H Zhuang, Daniel J Valentino, and Woei-Chyn Chu. Performance measure characterization for evaluating neuroimage segmentation algorithms. *Neuroimage*, 47(1):122–135, 2009. 5, 6

[6] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2624–2632, 2019. 1

[7] Zeyu Cheng, Yi Zhang, and Chengkai Tang. Swindepth: Using transformers and multi-scale fusion for monocular-based depth estimation. *IEEE Sensors Journal*, 21(23):26912–26920, 2021. 2

[8] Weiqin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, David Suter, and Alireza Bab-Hadiashar. Semantic guided long range stereo depth estimation for safer autonomous vehicle applications. *IEEE Trans. on Intell. Transp. Syst.*, 23(10):18916–18926, 2022. 1

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Inf. Process. Syst.*, 27, 2014. 2, 5, 6, 7

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proc. of the IEEE Conf. on Comput. Vision and Pattern Recognit.*, pages 2002–2011, 2018. 7

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The Int. Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 5

[12] Ali Hatamizadeh, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. *arXiv preprint arXiv:2206.09959*, 2022. 4

[13] Varsha Hedau, Derek Hoiem, and David Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *Eur. Conf. on Comput. Vision*, pages 224–237. Springer, 2010. 1

[14] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11130–11140, 2021. 1, 2, 5, 7, 8

[15] Yasunori Ishii and Takayoshi Yamashita. Cutdepth: Edge-aware data augmentation in depth estimation. *arXiv preprint arXiv:2107.07684*, 2021. 2, 5, 7, 8

[16] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2

[17] Doyeon Kim, Woonghyun Ga, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 2, 5, 7, 8

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

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 4th Int. Conf. on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2

[20] 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. 7

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

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[23] Michele Mancini, Gabriele Costante, Paolo Valigi, and Thomas A Ciarfuglia. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. In *2016 IEEE/RSJ Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 4296–4303. IEEE, 2016. 1

[24] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proc. of the IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 9685–9694, 2021. 1

[25] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021. 1, 2, 7, 8

[26] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proc. of the IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 1610–1621, 2022. 7

[27] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21477–21487, 2023. 1, 2, 7

[28] Jiamin Ping, Yue Liu, and Dongdong Weng. Comparison in depth perception between virtual reality and augmented

reality systems. In *2019 IEEE Conf. on Virtual Reality and 3D User Interfaces (VR)*, pages 1124–1125. IEEE, 2019. 1

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. of the IEEE/CVF Int. Conf. on Comput. Vision*, pages 12179–12188, 2021. 1, 7

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

[31] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. on Comput. Vision*, pages 746–760. Springer, 2012. 5

[32] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022. 2, 6

[33] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset, 2019. 2

[34] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 2

[35] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013. 2

[36] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[37] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 1, 2, 7

[38] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proc. of the IEEE/CVF Int. Conf. on Comput. Vision*, pages 16269–16279, 2021. 1, 2, 7

[39] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proc. of the IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 3916–3925, 2022. 1, 2, 7

[40] Min Yue, Guangyuan Fu, Ming Wu, and Hongqiao Wang. Semi-supervised monocular depth estimation based on semantic supervision. *Journal of Intelligent & Robotic Systems*, 100(2):455–463, 2020. 1, 2, 7

[41] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 1, 2, 7, 8

[42] Yan Zhang, XiaoPeng Fan, and DeBin Zhao. Semisupervised learning-based depth estimation with semantic inference guidance. *Science China Technological Sciences*, 65(5):1098–1106, 2022. 1, 2, 7

[43] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 5, 6