WiViD: Leveraging Wi-Fi and Vision for Depth Estimation via Multimodal Diffusion

Shijie Cheng
School of Software
Tsinghua University
Beijing, China
chengshijie2009@gmail.com

Yuchong Gao
School of Software
Tsinghua University
Beijing, China
gaoyc01@gmail.com

Zheng Yang
School of Software
Tsinghua University
Beijing, China
hmilyyz@gmail.com

Guoxuan Chi✉
School of Software
Tsinghua University
Beijing, China
chiguoxuan@gmail.com

Tony Xiao Han
Wireless Technology Lab
Huawei Technology Co., Ltd.
Shenzhen, China
tony.hanxiao@huawei.com

2024 20th International Conference on Mobility, Sensing and Networking (MSN) | 979-8-3315-1602-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/MSN63567.2024.00021

Abstract—**Depth estimation is crucial for numerous applications, including autonomous driving, robotic navigation and augmented reality. Existing solutions based on LiDAR and mmWave technologies are constrained by high deployment costs, while those utilizing monocular vision suffer from limited accuracy. To address these challenges, this paper proposes *WiViD*, a diffusion-based depth estimation system that leverages commercial Wi-Fi and vision. Diffusion models, with their ability to iteratively refine predictions, offer significant advantages in producing accurate and detailed estimations. We introduce a Multimodal Conditional Diffusion (MMCD) mechanism and design two encoding modules: the Complex-Valued CSI Encoder (CCE) and the Residual Image Encoder (RIE). These components fully exploit the spatio-temporal information inherent in Wi-Fi CSI and enable the effective fusion of Wi-Fi CSI and RGB image data, which results in high-precision and robust depth estimation. Experimental results in real-world scenarios demonstrate that *WiViD* outperforms state-of-the-art (SOTA) monocular methods, reducing the Absolute Relative Error (ARE) by 67.2%, highlighting the advantages of *WiViD* in terms of accuracy and reliability.**

Index Terms—**Depth estimation, Wireless sensing, Diffusion model, Multimodal fusion, Wi-Fi CSI**

I. Introduction

Depth estimation provides 3-D perception capability, which is essential for numerous applications, such as autonomous driving, robotic navigation, and augmented reality [1]. Over the past decades, significant research efforts have been devoted to this field.

Most existing depth estimation methods can be divided into two categories: active and passive. Active methods, such as mmWave Radar [2] and LiDAR [3], measure depth by projecting millimeter waves or lasers and receiving the reflected signals. While these methods are highly accurate in localization [4], they rely on expensive equipment, limiting their widespread use. Passive methods, based on a single image, are more common but less accurate. Monocular vision estimates depth using RGB images, which is a low-cost and portable method but is easily affected by lighting and texture. Recent research has attempted to combine active and passive methods, such as combining RGB with infrared measurements [5], or fusing RGB with LiDAR [6] or mmWave [7]. However, these methods have high computational complexity and do not address equipment costs. Therefore, improving the accuracy and robustness of depth estimation while maintaining cost-effectiveness remains a significant challenge.

Recently, with the rapid development of wireless communication technology, localization [8]–[10] and sensing [11]–[13] based on Wi-Fi signals has become a research focus due to its ubiquitous availability, low cost, and ease of deployment. This Wi-Fi-based sensing solution does not require expensive hardware and offers pervasiveness.

Considering the above situation, we identify an opportunity to achieve depth estimation by fusing two modalities: Wi-Fi and vision. Wi-Fi channel state information (CSI) [14], [15] can capture three-dimensional spatial information by detecting variations in signal reflections caused by objects in the environment, making them highly sensitive to changes in the signal propagation path (i.e., radial direction). In contrast, vision-based solutions excel at detecting changes in the pixel plane (i.e., tangential direction), providing detailed spatial location information with high pixel-level precision. Therefore, by fusing these two complementary modalities, a multimodal solution has great potential to enhance both the accuracy and robustness of depth estimation, providing more refined and reliable 3D perception capabilities for upper-level applications.

However, translating this intuition into a practical system faces two main challenges.

Challenge 1: How to design a high-precision Wi-Fi and vision multimodal depth estimation framework. Previous methods combining Wi-Fi and vision are primarily designed for classification tasks, such as human activity recognition. However, depth estimation, being a regression task, demands not only higher precision but also continuous and high-resolution distance data. Unlike classification, which deals with discrete labels, depth estimation requires handling continuous variables. Consequently, the limitations of traditional data fusion strategies are more pronounced in this context.

Challenge 2: How to extract precise spatial information from complex-valued Wi-Fi CSI signals. Wi-Fi CSI signals have complex domain characteristics, including amplitude and phase information. Existing real-domain processing methods are insufficient for handling the high spatio-temporal coupling features in the complex domain. Due to the multipath propagation effect, the blending of CSI signal information in time and space significantly increases the difficulty of feature extraction.

2994-3523/24/$31.00 ©2024 IEEE
DOI 10.1109/MSN63567.2024.00021

Consequently, traditional methods cannot fully utilize the rich spatio-temporal information within complex-valued CSI signals. Constructing a neural network that operates in the complex domain to extract information from CSI signals for guiding depth reconstruction has become a critical research challenge.

To tackle these challenges, we develop a system called *WiViD*, which utilizes a generative diffusion model for multimodal depth estimation. Diffusion models, known for their iterative refinement processes, excel in producing high-quality and detailed predictions. Specifically, we design two network structures in *WiViD* for processing different data modalities: the Complex-Valued CSI Encoder (CCE) and the Residual Image Encoder (RIE). The CCE is designed to encode complex-valued Wi-Fi CSI, while the RIE processes RGB image information. To effectively fuse these two data modalities, we propose the Multimodal Conditional Diffusion (MMCD) mechanism. The MMCD concatenates the features extracted by the CCE and RIE, transforming them into the control condition using linear transformations and activation functions. In the U-net [16] Denoising Module, the control condition guides the process of denoising, enabling accurate depth estimation.

We implement *WiViD* and conduct extensive experiments in real-world scenarios. We compare it with two advanced monocular visual depth estimation methods: DORN [17] and VA-DepthNet [18]. The experimental results show that *WiViD*'s ARE is only 0.022, outperforming DORN's 0.153 and VA-DepthNet's 0.059, demonstrating the outstanding performance of *WiViD*. Additionally, we conduct experiments to confirm the effectiveness of *WiViD*'s multimodal fusion mechanism.

Our contributions are summarized as follows:

- We propose *WiViD*, the first depth estimation system incorporating an innovative Wi-Fi and vision multimodal diffusion model. *WiViD* achieves excellent experimental results, representing a promising step in the exploration of depth estimation by fusing Wi-Fi and vision.
- Our proposed Complex-Valued CSI Encoder (CCE) offers unique advantages in handling spatio-temporal complex-valued signals and can be directly applied to other wireless sensing applications (e.g., human gesture recognition, fall detection) to enhance their feature extraction capabilities.
- Our proposed Multimodal Conditional Diffusion (MMCD) mechanism is a pioneering approach to Wi-Fi-vision fusion within a diffusion framework. The MMCD mechanism can be extended to other sensing modalities, establishing a new paradigm for multimodal fusion perception.

II. System Overview

WiViD is a Wi-Fi and vision depth estimation system that utilizes multimodal diffusion. As shown in Fig. 1, *WiViD* extracts features from Wi-Fi CSI signals and RGB images, encoding them as control conditions to guide the diffusion model in generating accurate depth maps.

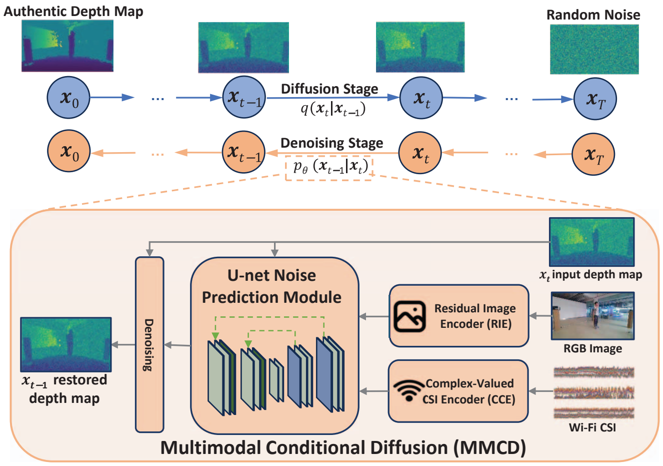


Fig. 1. System Overview of *WiViD*

Similar to DDPM [19], our diffusion model consists of two stages: the forward diffusion stage and the reverse denoising stage. During the forward diffusion stage, Gaussian noise is incrementally added to the authentic depth map, gradually transforming it into a noise distribution. In the reverse denoising stage, *WiViD* employs the U-net Noise Prediction Module to predict the noise in the $\boldsymbol{x}_t$ input depth map. *WiViD* progressively removes the predicted noise at each step, thereby reconstructing the $\boldsymbol{x}_{t-1}$ depth map until $\boldsymbol{x}_0$, the final noise-free depth map. The reverse denoising stage is divided into T steps. The training process involves both stages, whereas the prediction process only requires the reverse denoising stage.

The architecture of *WiViD* consists of three key components: the Complex-Valued CSI Encoder (CCE), the Residual Image Encoder (RIE), and the U-net Noise Prediction Module. These modules are unified under the Multimodal Conditional Diffusion (MMCD).

When performing depth estimation tasks, the algorithm samples random Gaussian noise and gradually transforms it into a depth map. The CCE and RIE extract and encode features from the input Wi-Fi CSI and RGB images into high-dimensional vectors. These vectors undergo multimodal fusion in the MMCD, forming a comprehensive representation used as a control condition during the denoising stage. This control condition, containing inherent spatial data from Wi-Fi and vision inputs, is crucial for generating the depth map. In each block of the U-net Noise Prediction Module, this control condition is added to the predicted intermediate representation, guiding the denoising process to achieve accurate depth estimation.

III. *WiViD* Design

WiViD is designed for depth estimation and proposes a deep learning strategy based on a diffusion model. It addresses two major challenges: 1) extracting and encoding features from RGB images and Wi-Fi CSI signals, particularly Wi-Fi CSI; and 2) fusing the encoded information from these two modalities to guide the denoising stage of the diffusion model for depth estimation.

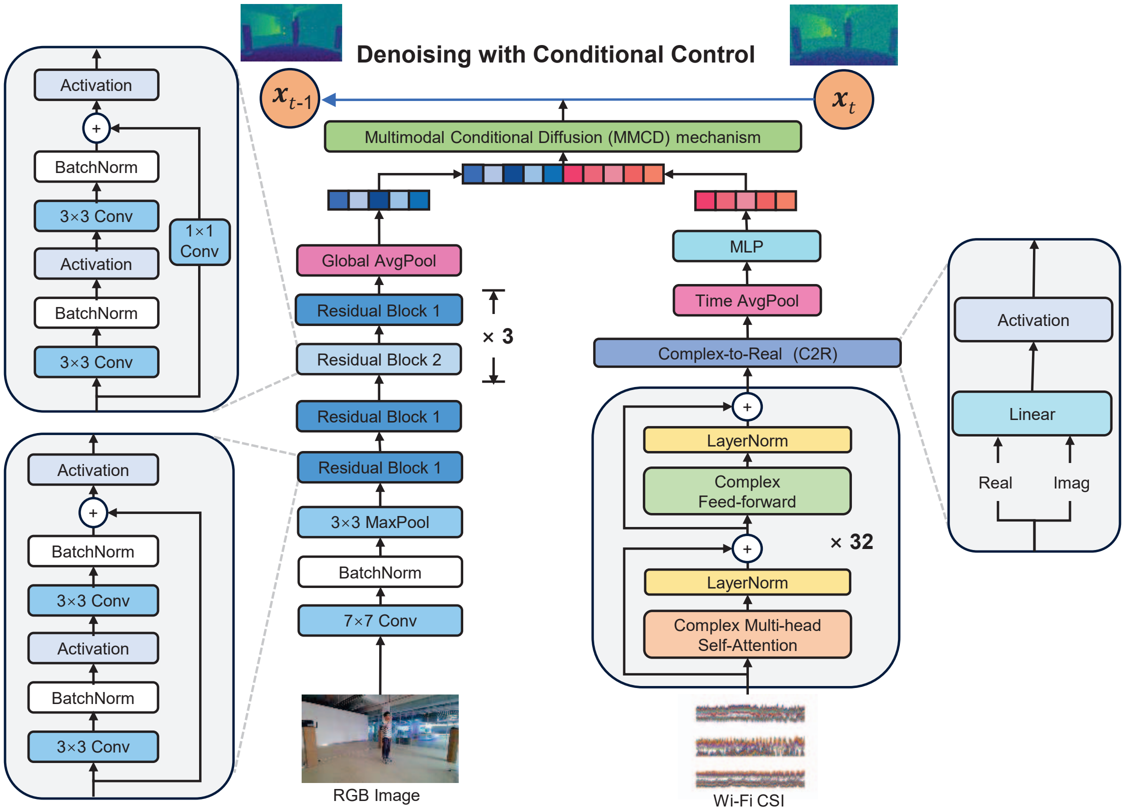


Fig. 2. Design of Denoising with Conditional Control

To address these challenges, we introduce the RGB Image Encoder (RIE) and the CSI Signal Encoder (CSE) into the diffusion model. Additionally, we employ the Multimodal Conditional Diffusion (MMCD) mechanism for multimodal fusion.

To enhance prediction accuracy in the depth estimation, *WiViD* employs the SiLU activation function.

A. Diffusion for Depth Estimation

Our diffusion model consists of two stages: the diffusion stage and the denoising stage.

In the diffusion stage, the authentic depth map $\boldsymbol{x}_0$ is gradually converted into pure noise $\boldsymbol{x}_T$ through a series of predefined noise scales over T steps. This transformation is governed by the diffusion process $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$, producing a noised depth map $\boldsymbol{x}_t$ for $t \in \{1, ..., T\}$, defined as a random process:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) := \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative product of the noise scales at each step, indicating the total noise level from step 1 to t. It is calculated as $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, where $\alpha_s = 1 - \beta_s$, and β_s represents the noise variance schedule.

The denoising stage involves training a deep neural network $\boldsymbol{\epsilon}_\theta$ to predict the noise in the noised depth map $\boldsymbol{x}_t$, using the RGB image $\boldsymbol{M}$ and Wi-Fi CSI $\boldsymbol{H}$ as the control condition. A proportion of this noise, determined by $\bar{\alpha}_t$, α_t and β_t, is then removed to obtain $\boldsymbol{x}_{t-1}$. This process is iteratively continued until the authentic depth map $\boldsymbol{x}_0$ is recovered. This procedure is described by:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{M}, \boldsymbol{H}), \boldsymbol{\sigma}_t^2\mathbf{I}\right), \tag{2}$$

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{M}, \boldsymbol{H}) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{M}, \boldsymbol{H})), \tag{3}$$

where $\boldsymbol{\sigma}_t^2$ represents the transition variance, and $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{M}, \boldsymbol{H})$ represents the process of removing predicted noise to obtain the denoised result at each step.

The training process of *WiViD* involves both a diffusion stage and a denoising stage, as illustrated in Algorithm 1. Noise is added to the authentic depth map $\boldsymbol{x}_0$, with the noise level determined by t, randomly selected from $\{1, ..., T\}$. Concurrently, the image $\boldsymbol{M}$ and CSI $\boldsymbol{H}$ are input as the control condition. Subsequently, the model $\boldsymbol{\epsilon}_\theta$ predicts the added noise, recovering the authentic depth map $\boldsymbol{x}_0$ from the noise-corrupted data. This training process optimizes the parameters of the model $\boldsymbol{\epsilon}_\theta$ to minimize the difference between the predicted noise and the actual noise added to $\boldsymbol{x}_0$.

The sampling process requires only the denoising stage, as illustrated in Algorithm 2. Gaussian noise $\boldsymbol{x}_T$ is randomly generated and input along with the corresponding image $\boldsymbol{M}$ and CSI $\boldsymbol{H}$. Using the trained model $\boldsymbol{\epsilon}_\theta$, the noise is iteratively removed, starting from step T and continuing until the authentic depth map $\boldsymbol{x}_0$ is recovered.

For depth images of *WiViD*, depth can be represented as $\mathbf{d} \in \mathbb{R}^{1\times M\times N}$, where M and N denote the height and width of the depth map. Unlike RGB images with three channels, depth maps have only one channel. The depth map is normalized to the range $[-1, 1]$ to match the Gaussian noise mean and variance in diffusion model. So that the diffusion and denoising stages are applied to these normalized depth maps.

Algorithm 1 *WiViD* Training Algorithm

```
Input: Dataset following $\boldsymbol{x} \sim q(\boldsymbol{x})$, image $\boldsymbol{M}$ and CSI $\boldsymbol{H}$
Output: Noise prediction model $\boldsymbol{\epsilon}_\theta$
repeat
    $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, $\boldsymbol{M}$ and $\boldsymbol{H}$ from dataset
    $t \sim \mathrm{Uniform}(\{1, \dots, T\})$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Take gradient descent step on
        $\nabla_\theta \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_\mathbf{0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, \boldsymbol{M}, \boldsymbol{H})\|^2$
until converged
```

Algorithm 2 *WiViD* Sampling Algorithm

```
Input: Noise prediction model $\boldsymbol{\epsilon}_\theta$, image $\boldsymbol{M}$ and CSI $\boldsymbol{H}$
Output: Depth map $\boldsymbol{x}_0$
$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ do
    $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{M}, \boldsymbol{H})\right)$
end for
return $\boldsymbol{x}_0$
```

B. Residual Image Encoder (RIE)

To fully leverage the information in RGB images, we design the RIE module using ResNet-18 [20] for feature extraction and encoding. By removing its classification layer, we obtain an intermediate representation rich in semantic content, essential for encoding. This representation encapsulates high-level image features. Using RIE for feature extraction, we encode the image $\boldsymbol{M}$ into a vector for depth estimation, $\boldsymbol{X}_\mathrm{M}$, as described below:

$$\boldsymbol{X}_\mathrm{M} = \mathrm{ResNet}(\boldsymbol{M}). \tag{4}$$

C. Complex-Valued CSI Encoder (CCE)

To effectively encode Wi-Fi CSI data and extract its spatio-temporal features, we design a Multi-Head Transformer Encoder that supports complex-valued calculation. This encoder not only preserves the authentic information of the CSI signal but also captures the interactions and dependencies between signals through the Attention mechanism.

Complex-Valued Multi-Head Attention. In this Transformer Encoder, we extend Attention to the complex domain. For given $\boldsymbol{X} \in \mathbb{C}^{L\times d_\mathrm{in}}$, firstly it will perform linear projection:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K, \quad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V, \tag{5}$$

then we get $\boldsymbol{Q} \in \mathbb{C}^{L\times d_\mathrm{Q}}$, $\boldsymbol{K} \in \mathbb{C}^{L\times d_\mathrm{K}}$ and $\boldsymbol{V} \in \mathbb{C}^{L\times d_\mathrm{V}}$, in which the multiplication follows the principles of complex multiplication, and $\boldsymbol{W}_Q \in \mathbb{C}^{d_\mathrm{in}\times d_Q}$, $\boldsymbol{W}_K \in \mathbb{C}^{d_\mathrm{in}\times d_K}$ and $\boldsymbol{W}_V \in \mathbb{C}^{d_\mathrm{in}\times d_v}$ are projection parameters. Softmax and Attention mechanism are also extended in complex domain, and can be defined as follows:

$$\mathrm{softmax}_\mathrm{complex}(z) = \frac{\mathrm{e}^{|z|}\mathrm{e}^{j\angle(z)}}{\sum_{i=1}^{n}\mathrm{e}^{|z_i|}}, \tag{6}$$

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}_\mathrm{complex}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_K}}\right)\boldsymbol{V}, \tag{7}$$

where j is the imaginary unit. Furthermore, we employ Multi-Head Attention with h heads, defined as follows:

$$\mathrm{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = (a_1, a_2, \dots, a_h)^T\boldsymbol{W}_O, \tag{8}$$

where $\boldsymbol{W}_O \in \mathbb{C}^{hd_V\times d_O}$ is the final projection parameter, and $a_i = \mathrm{Attention}(\boldsymbol{X}\boldsymbol{W}_Q^i, \boldsymbol{X}\boldsymbol{W}_K^i, \boldsymbol{X}\boldsymbol{W}_V^i)$.

Complex feed-forward module. The complex feed-forward module contains linear transformations and activation functions, which support calculations in the complex domain, and can be defined as follows:

$$\begin{aligned}\boldsymbol{w}\boldsymbol{x} + \boldsymbol{b} &= \begin{bmatrix}\Re(\boldsymbol{w}\boldsymbol{x} + \boldsymbol{b})\\ \Im(\boldsymbol{w}\boldsymbol{x} + \boldsymbol{b})\end{bmatrix}\\ &= \begin{bmatrix}w_r & -w_i\\ w_i & w_r\end{bmatrix}\begin{bmatrix}x_r\\ x_i\end{bmatrix} + \begin{bmatrix}b_r\\ b_i\end{bmatrix},\end{aligned} \tag{9}$$

$$\sigma_\mathrm{complex}(x) = \sigma(x_r) + \mathrm{j}\sigma(x_i), \tag{10}$$

where $\sigma(\cdot)$ represents the activation function.

The Complex Transformer Encoder block consists of Complex Multi-Head Attention and Complex Feed-Forward modules. The Complex Transformer Encoder is structured with 32 of these blocks.

Complex-to-Real Transformation Layer (C2R). Subsequently, we design a Complex-to-Real Transformation Layer following the Transformer Encoder. This layer converts complex features to real features. C2R is defined as follows:

$$\mathrm{C2R}(\boldsymbol{X}) = \sigma(\mathrm{Linear}_1(\Re(\boldsymbol{X})) + \mathrm{Linear}_2(\Im(\boldsymbol{X}))). \tag{11}$$

Following the Transformer Encoder and C2R, we apply an average pooling layer to the CSI data for temporal averaging. This step compresses the data and transforms dynamic CSI features into a static representation, aligning them with single-frame image features. Subsequently, an MLP further extracts and compresses the feature vector from the average pooling, resulting in a compact feature representation.

In summary, the CCE can be described as follows:

$$\boldsymbol{X}_\mathrm{H} = \mathrm{MLP}(\mathrm{AvgPool}(\mathrm{C2R}(\mathrm{CTransEnc}(\boldsymbol{H})))), \tag{12}$$

where CTransEnc represents the Complex Transformer Encoder, $\boldsymbol{H}$ is the Wi-Fi CSI matrix, and $\boldsymbol{X}_\mathbf{H}$ is the processed CSI embedding.

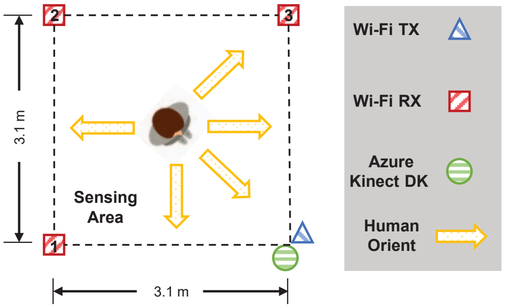


Fig. 3. Layout of Experiment Space

D. Multimodal Conditional Diffusion (MMCD)

To fuse RGB images and Wi-Fi CSI data, we integrate the Multimodal Conditional Diffusion (MMCD) mechanism into the denoising stage of the diffusion model. RGB images are encoded by the RIE to produce image embeddings, while CSI data is encoded by the CCE to produce CSI embeddings. Both encoders are trained simultaneously with the U-net.

Within U-net blocks, RGB and CSI embeddings are concatenated to form a unified feature representation, then refined through activation functions and a linear layer to produce a block-specific fusion condition embedding $\boldsymbol{X}_{\mathrm{F}}$. Each block of the U-net utilizes its own $\boldsymbol{X}_{\mathrm{F}}$ to improve feature adaptability and expressiveness. $\boldsymbol{X}_{\mathrm{F}}$ can be represented as follows:

$$\boldsymbol{X}_{\mathrm{F}} = \mathrm{Linear}(\sigma(\mathrm{RIE}(\boldsymbol{M}) \oplus \mathrm{CCE}(\boldsymbol{H}))), \quad (13)$$

where $\boldsymbol{M}$ and $\boldsymbol{H}$ are matrices of RGB image and Wi-Fi CSI.

$\boldsymbol{X}_{\mathrm{F}}$ is then added to the intermediate feature representation within the U-net blocks. For each channel of the intermediate representation, a broadcasting mechanism is used to add the corresponding values from $\boldsymbol{X}_{\mathrm{F}}$, thereby guiding noise prediction for depth estimation. The conditional control process can be described as follows:

$$\boldsymbol{N}' = \boldsymbol{N} + \boldsymbol{X}_{\mathrm{F}}, \quad (14)$$

where $\boldsymbol{N}$ and $\boldsymbol{N}'$ are intermediate representation before and after conditional control.

IV. Evaluation

A. Methodology

Implementation. We utilize MATLAB for Wi-Fi CSI data pre-processing, converting the raw CSI into tensors suitable for model training. *WiViD* is implemented in PyTorch and trained on an NVIDIA GeForce RTX 3090. During training, the number of steps T is set to 50. A cosine annealing strategy dynamically adjusts the learning rate from 1e-4 to 1e-7. The model completes 100 epochs of training with a batch size of 16.

Dataset and Data Preprocessing. For our experiments, we use a subset of the XRF55 [21] dataset, extracting RGB images and Wi-Fi CSI from a meeting room, with depth maps as the ground truth. The spatial layout is shown in Fig. 3. The dataset comprises 10,000 RGB images created by frame-sampling 400 randomly selected videos. To streamline calculations, the original 1280 × 720 RGB images and depth maps are resized to 160 × 90 pixels. Wi-Fi CSI data for each video frame is collected and concatenated from three Wi-Fi RXs to form a unified dataset. The data is then shuffled and divided into training and testing sets in a 9:1 ratio for *WiViD* training and evaluation.

Evaluation Metrics. Various evaluation metrics can measure the effectiveness of depth estimation; we select two commonly used metrics: error values and threshold accuracy. Lower error values between the predicted depth and the ground truth indicate better depth estimation, while higher threshold accuracy signifies improved performance.

Absolute Relative Error (ARE) and Relative Squared Error (SRE) describe the general distance between the predicted depth and the ground truth.

- Absolute Relative Error (ARE):

$$\mathrm{ARE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_{\mathrm{true},i} - d_{\mathrm{pred},i}|}{d_{\mathrm{true},i}}. \quad (15)$$

- Squared Relative Error (SRE):

$$\mathrm{SRE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_{\mathrm{true},i} - d_{\mathrm{pred},i}|^2}{d_{\mathrm{true},i}}. \quad (16)$$

Root Mean Square Error (RMSE) and Root Mean Square Logarithmic Error (RMSELog) are especially sensitive to local inaccuracies of depth estimation.

- Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|d_{\mathrm{true},i} - d_{\mathrm{pred},i}|^2}. \quad (17)$$

- Root Mean Square Logarithmic Error (RMSELog):

$$\mathrm{RMSELog} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\log d_{\mathrm{true},i} - \log d_{\mathrm{pred},i}|^2}. \quad (18)$$

Threshold accuracy measures the similarity between the predicted depth and the ground truth, providing a quantitative evaluation of the model's accuracy within a specific depth range. It includes three primary metrics: δ_1, δ_2 and δ_3, corresponding to different depth thresholds.

- Threshold accuracy (δ_k) is defined as the percentage of $d_{\mathrm{pred},i}$ where $\max\left(\frac{d_{\mathrm{pred},i}}{d_{\mathrm{true},i}}, \frac{d_{\mathrm{true},i}}{d_{\mathrm{pred},i}}\right) < 1.25^k$, for $k = 1, 2, 3$.

It is important to note that in the ground truth depth map, there are pixels with undefined depth. When calculating metrics, we exclude these pixels.

B. Overall Performance

To assess the performance of *WiViD*, we select two prominent monocular depth estimation algorithms as comparison benchmarks: DORN [17] and VA-DepthNet [18]. DORN transforms depth prediction into an ordinal classification problem, employing deep learning to capture relative depth order

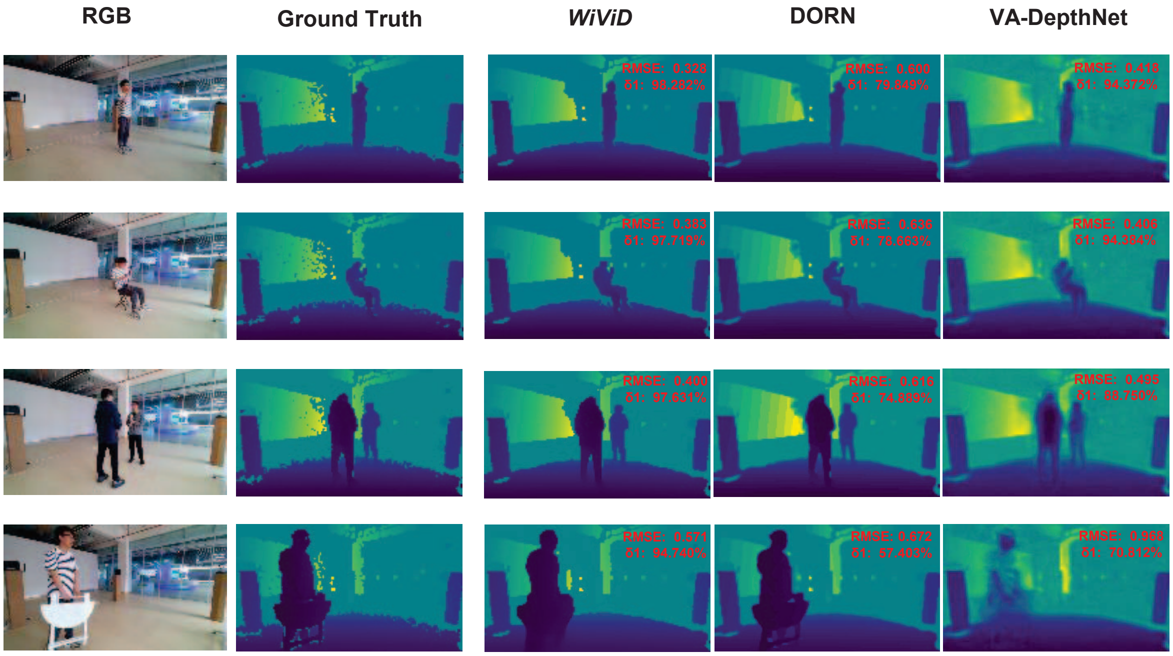
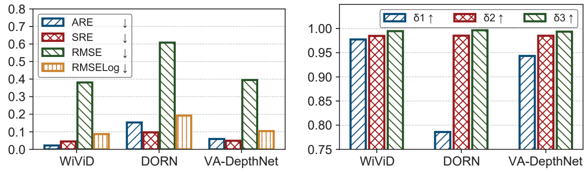
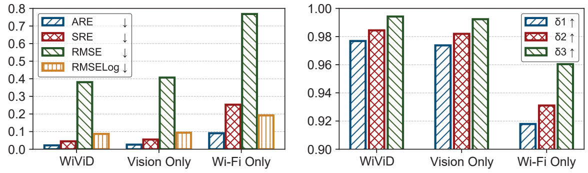


Fig. 4. Depth Estimation Samples of *WiViD* and Benchmarks



(a) Estimation Error (b) Threshold Accuracy

Fig. 5. Overall Performance Comparison with Benchmarks



(a) Estimation Error (b) Threshold Accuracy

Fig. 6. Performance Comparison between Multimodal and Single-modal

at the pixel level for precise depth mapping. VA-DepthNet incorporates first-order variational constraints and encoder-decoder network architecture. To ensure a fair comparison, we fine-tune the open-source implementations of these algorithms on our dataset.

As depicted in Fig. 5(a) and Fig. 5(b), *WiViD* outperforms existing benchmarks, particularly on the ARE metric, achieving a score of 0.022 compared to DORN's 0.153 and VA-DepthNet's 0.059. On the δ_1 metric, which represents the highest demand for depth estimation accuracy, *WiViD* reaches 0.977, significantly higher than DORN's 0.786 and VA-DepthNet's 0.943.

As depicted in Fig. 4, the advantages of *WiViD* are especially notable. While DORN accurately identifies human contours and their relative depth, it fails to provide precise depth estimation due to inaccuracies in the depth scale, resulting in a higher RMSE than *WiViD*. VA-DepthNet grasps the overall scale of depth estimation but struggles to clearly segment human actions, leading to a loss of accuracy. In contrast, *WiViD* not only accurately estimates the static depth in the scene but also precisely captures changes in depth caused by human actions, thereby achieving more accurate depth estimation.

C. Micro Benchmarks

1) Effectiveness of Multimodal Fusion

To evaluate the effectiveness of multimodal fusion in *WiViD*, we conduct an experiment comparing multimodal and single-modal methods. Specifically, we examine the performance of using RGB images alone, Wi-Fi CSI data alone, and the combination of both.

We remove one of the modalities to obtain depth estimation results for "Vision Only" and "Wi-Fi Only". As shown in Fig. 6(a) and Fig. 6(b), the multimodal fusion method in *WiViD* consistently outperforms both single-modality methods across all metrics. Specifically, the RMSE for the multimodal method, "Vision Only" and "Wi-Fi Only" are 0.381, 0.407 and 0.768, while the δ_1 values are 0.977, 0.973 and 0.918. These results indicate that the multimodal method optimizes depth estimation performance, demonstrating the effectiveness of combining vision and Wi-Fi for depth estimation.

To explore the role of fusing Wi-Fi and vision for depth estimation, we selected two sets of samples for comparative analysis, as illustrated in Fig. 7. It is evident that the "Vision

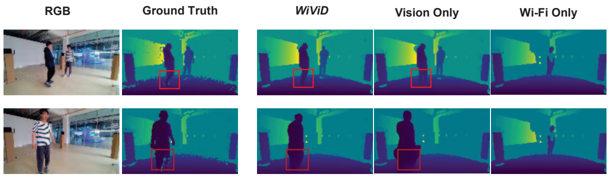


Fig. 7. Depth Estimation Samples of Multimodal and Single-modal

Only" method is less accurate in many details of depth estimation compared to the multimodal fusion method. Moreover, the "Wi-Fi Only" method struggles to capture depth changes caused by human movements.

Monocular visual depth estimation relies on extracting visual cues from a single image to infer depth but lacks direct physical depth information, limiting its accuracy. In contrast, Wi-Fi CSI, which is closely related to the geometric structure of space, provides direct physical depth information. However, when used alone, Wi-Fi CSI lacks pixel-level resolution and texture details, affecting its fine-scale depth estimation capability. Our observations indicate that fusing Wi-Fi CSI and RGB images provides more detailed and complementary information for depth estimation, thereby improving accuracy. When RGB images alone fail to provide accurate depth estimation, the integration of Wi-Fi CSI significantly enhances the system's robustness.

2) *Impact of Diffusion Steps*

TABLE I
PERFORMANCE OF *WiViD* IN THE IMPACT OF DIFFUSION STEPS

Metrics / T	**35**	**40**	**45**	**50**
ARE ↓	0.027	0.026	0.022	0.022
SRE ↓	0.058	0.050	0.046	0.045
RMSE ↓	0.404	0.395	0.380	0.381
RMSELog ↓	0.093	0.091	0.087	0.087
δ_1 ↑	0.974	0.975	0.977	0.977
δ_2 ↑	0.982	0.983	0.985	0.985
δ_3 ↑	0.993	0.994	0.994	0.995

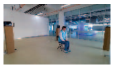
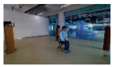
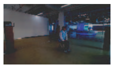
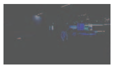
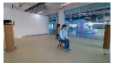
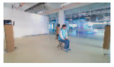
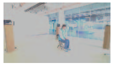
WiViD is a depth estimation method based on a diffusion model. During the denoising stage, the diffusion steps T affect both the number of denoising iterations and the amount of noise removed in each iteration. A smaller T requires more noise to be removed in a single step, and vice versa.

We conduct experiments with different T values (35, 40, 45, 50), as shown in Table I. As T increases from 35 to 50, the depth estimation performance of *WiViD* improves, but this improvement slows beyond $T = 45$. Increasing T within a certain range enhances depth estimation performance, but further increases beyond this threshold do not yield significant improvements.

D. *Robustness Experiments*

We simulate scenes with underexposure and overexposure at 25%, 50%, and 75%, as outlined in Table II. We then compare the depth estimation capabilities of the multimodal model against the "Vision Only" model across different exposure levels. Our findings show that the performance of both models degrades as texture information decreases. However, the multimodal model, which integrates vision and Wi-Fi, outperforms the "Vision Only" model on all evaluation metrics by augmenting depth information when visual texture is lacking. This superiority is most pronounced in the δ_1 and ARE metrics, as illustrated in Fig. 8 and Fig. 9. These results suggest that the enhanced robustness of the multimodal model in the presence of reduced bright or dark texture is due to the integration of Wi-Fi CSI, which provides additional depth information and reinforces the model's performance in challenging exposure conditions.

TABLE II
UNDEREXPOSURE OR OVEREXPOSURE FOR RGB IMAGE

Exposure	**0%**	**25%**	**50%**	**75%**
Underexposure				
Overexposure				

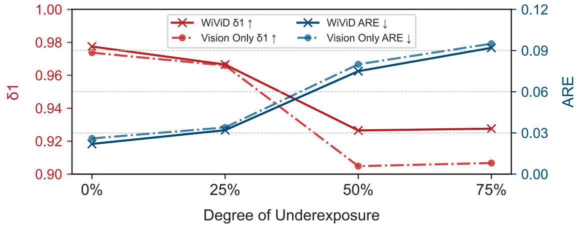


Fig. 8. δ_1 and ARE of Underexposure

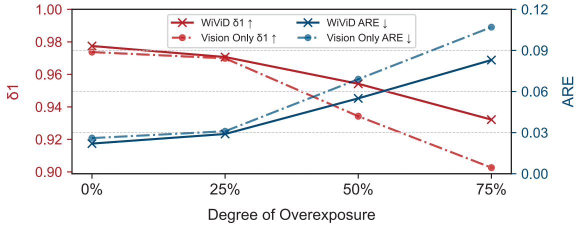


Fig. 9. δ_1 and ARE of Overexposure

V. RELATED WORK

We briefly review the related works in the following.

Diffusion Models. Diffusion models have become pivotal in deep learning and AI. Innovations such as GLIDE [22], DALL-E 2 [23], and Imagen [24] have significantly advanced text-to-image creation by enhancing image diversity and utility. Latent Diffusion Models (LDM) [25] optimize this process by operating in a low-dimensional latent space, reducing computational demands while maintaining high-quality outputs. Additionally, in the field of radio frequency, RF-Diffusion [26] is a wireless signal generation scheme utilizing a diffusion model. These advancements have expanded the capabilities and applications of diffusion models, making them integral to AI research and development.

Depth Estimation. Depth estimation is essential for various technological applications, driving the development of diverse methods to address this challenge. Convolutional Neural Networks (CNNs) are widely employed, with approaches such as those by Eigen et al. [27] utilizing multi-scale CNNs to achieve precise depth predictions. MonoDepth [28] leverages a self-supervised learning framework to estimate depth from single images, thereby eliminating the need for ground truth data and significantly enhancing efficiency. Recent advancements include monocular depth estimation using Transformer architectures, as demonstrated by MonoViT [29], which captures long-range dependencies to improve depth prediction. Collectively, these methods advance depth estimation and contribute to progress in autonomous driving, augmented reality, and robotics.

VI. Conclusion

This paper proposes *WiViD*, the first Wi-Fi and vision multimodal fusion depth estimation method. It integrates Wi-Fi and vision data through a diffusion model. The Complex-Valued CSI Encoder (CCE), Residual Image Encoder (RIE), and Multimodal Conditional Diffusion (MMCD) mechanism ensure effective fusion of Wi-Fi CSI and RGB images. Experimental results demonstrate that *WiViD* outperforms existing monocular visual depth estimation methods in real-world scenarios.

Acknowledgment

This work was supported by the Natural Science Foundation of China (NSFC) under Grant No. 62402276 and Grant No. 62372265.

References

[1] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, “Survey on deep neural networks in speech and vision systems,” *Neurocomputing*, 2020.

[2] S. Rao, “Introduction to mmwave sensing: Fmcw radars,” *Texas Instruments (TI) mmWave Training Series*, 2017.

[3] K. Park, S. Kim, and K. Sohn, “High-precision depth estimation with the 3d lidar and stereo fusion,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[4] G. Zhang, G. Chi, Y. Zhang, X. Ding, and Z. Yang, “Push the limit of millimeter-wave radar localization,” *ACM Transactions on Sensor Networks*, 2023.

[5] F. Alhwarin, A. Ferrein, and I. Scholl, “Ir stereo kinect: improving depth images by combining structured light with ir stereo,” in *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*, 2014.

[6] F. Ma, G. V. Cavalheiro, and S. Karaman, “Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[7] A. D. Singh, Y. Ba, A. Sarker, H. Zhang, A. Kadambi, S. Soatto, M. Srivastava, and A. Wong, “Depth estimation from camera image and mmwave radar point cloud,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[8] D. Li, J. Xu, Z. Yang, Y. Lu, Q. Zhang, and X. Zhang, “Train once, locate anytime for anyone: Adversarial learning based wireless localization,” in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, 2021.

[9] G. Chi, Z. Yang, J. Xu, C. Wu, J. Zhang, J. Liang, and Y. Liu, “Wi-drone: wi-fi-based 6-dof tracking for indoor drone flight control,” in *Proceedings of the 20th annual international conference on mobile systems, applications and services*, 2022.

[10] Y. Gao, G. Chi, G. Zhang, and Z. Yang, “Wi-prox: Proximity estimation of non-directly connected devices via sim2real transfer learning,” in *GLOBECOM 2023-2023 IEEE Global Communications Conference*, 2023.

[11] Z. Yang, Y. Zhang, K. Qian, and C. Wu, “{SLNet}: A spectrogram learning neural network for deep wireless sensing,” in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 2023.

[12] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, “Zero-effort cross-domain gesture recognition with wi-fi,” in *Proceedings of the 17th annual international conference on mobile systems, applications, and services*, 2019.

[13] G. Chi, G. Zhang, X. Ding, Q. Ma, Z. Yang, Z. Du, H. Xiao, and Z. Liu, “Xfall: Domain adaptive wi-fi-based fall detection with cross-modal supervision,” *IEEE Journal on Selected Areas in Communications*, 2024.

[14] Z. Yang, Z. Zhou, and Y. Liu, “From rssi to csi: Indoor localization via channel response,” *ACM Computing Surveys (CSUR)*, 2013.

[15] Z. Yang, Y. Zhang, G. Chi, and G. Zhang, “Hands-on wireless sensing with wi-fi: A tutorial,” *arXiv preprint arXiv:2206.09532*, 2022.

[16] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015.

[17] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[18] C. Liu, S. Kumar, S. Gu, R. Timofte, and L. Van Gool, “Va-depthnet: A variational approach to single image depth prediction,” *arXiv preprint arXiv:2302.06556*, 2023.

[19] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[21] F. Wang, Y. Lv, M. Zhu, H. Ding, and J. Han, “Xrf55: A radio frequency dataset for human indoor action analysis,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2024.

[22] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” *arXiv preprint arXiv:2112.10741*, 2021.

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, 2022.

[24] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in neural information processing systems*, 2022.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[26] G. Chi, Z. Yang, C. Wu, J. Xu, Y. Gao, Y. Liu, and T. X. Han, “Rf-diffusion: Radio signal generation via time-frequency diffusion,” in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024.

[27] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” *Advances in neural information processing systems*, 2014.

[28] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[29] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, “Monovit: Self-supervised monocular depth estimation with a vision transformer,” in *2022 international conference on 3D vision (3DV)*, 2022.