# GS-ASM: 2DGS-Supervised Active Stereo Matching

Zhengling Wu[1,2] Rongfeng Lu[1,3,†] Quan Chen[4] Longjian Zeng[1] Ming Lu[1] Yaoqi Sun[3]
Yahong Chen[3] Baofeng Ji[5] Chenggang Yan[1]

[1]Hangzhou Dianzi University [2]Shandong University [3]Lishui University [4]Jiaxing University [5]Henan University of Science and Technology

zhenglingwu2001@gmail.com rongfeng-lu@hdu.edu.cn

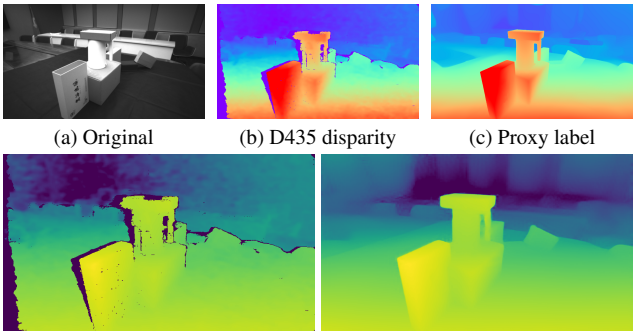
## Abstract

*Due to the lack of ground truth, existing methods of active stereo matching generally employ fully self-supervised learning to produce precise depth estimates. Although they can achieve promising results, their performance still has a noticeable gap compared with supervised models. To fill this gap, we propose a novel framework that synthesizes proxy labels to enable supervised training of deep active stereo networks without requiring any ground-truth depth. To expand the training data and generate disparity proxy labels, we develop an active 2D Gaussian Splatting (2DGS)-based synthesis method that explicitly models the scene geometry and the projected active pattern. Furthermore, to balance the varying contributions of different supervisions during training, we design a hybrid supervision regularization strategy that dynamically adjusts the loss weights to achieve stable optimization. We also contribute a real-world dataset captured by a handheld RealSense camera, along with our active 2DGS model, which facilitates future research on active stereo matching. Extensive experiments with multiple backbone networks demonstrate that our method achieves state-of-the-art performance on active stereo matching task. The code and dataset will be publicly released.*

## 1. Introduction

Depth acquisition is crucial for many computer vision tasks, including embodied intelligence [37], autonomous driving [44], 3D reconstruction [8], and mixed reality [56]. There are various ways to obtain depth information, but active stereo vision depth cameras (e.g., the Intel RealSense™ series) are the most widely used in practice.

Although monocular depth estimation methods [50, 51, 55] have achieved remarkable progress, these methods can only predict relative depth rather than absolute depth. This limitation restricts their use in real-world scenarios that require precise depth information. Stereo matching methods [3, 43] can estimate absolute depth by exploiting visual relationships between two color images. Yet, they often fail in scenes with low textures, repetitive patterns, or non-Lambertian materials, which are common in the real world. Active depth sensors, such as Light Detection and Ranging (LiDAR), provide accurate and reliable depth measurements. Nevertheless, their point cloud density is much lower than the resolution of modern cameras, and high-density LiDAR systems are costly [25].



(a) Original (b) D435 disparity (c) Proxy label

(d) D435 depth (e) Ours predict

Figure 1. The comparison of disparity and depth from commercial depth sensors and our GS-ASM.

Active stereo vision depth cameras project infrared (IR) patterns onto real-world scenes using an IR pattern emitter. Two IR cameras then capture the projected patterns to perform stereo matching and estimate the depth map of the scene. This approach combines the robustness and accuracy of active sensors with the low cost and high resolution of cameras. As a result, such sensors have been widely used in both academic research and industrial applications [18]. However, since these cameras rely on classical stereo matching algorithms, they often suffer from issues such as over-smoothing and edge information loss [5].

Although learning-based methods can generate more accurate and complete depth maps by training on large depth data, these approaches typically rely on a vast number of ground-truth depth maps for supervised learning. However, such specialized datasets are still scarce in the field of active stereo matching. In addition, collecting active IR stereo

†Corresponding author.

matching datasets in real-world environments is extremely costly and time-consuming, especially when dense and accurate depth labels are required.

To address this challenge, some researchers have adopted a self-supervised learning paradigm [45], where the reprojection error between the left and right views serves as the supervisory signal. However, due to the instability of this loss function, the network often fails to achieve satisfactory results. Compared with supervised learning, the predictions produced by this approach usually lack clarity, detail, and accuracy. Other researchers have used Blender to create simulated active IR scenes and obtain ground-truth depth maps [21, 30, 42], enabling the network to generate clearer, more detailed, and more accurate predictions. Nevertheless, the domain gap between simulated and real-world environments remains a significant problem. ActiveZero [27] proposes a mixed-domain strategy that combines the two paradigms above to reduce the domain gap. Although it performs well on synthetic data, its performance degrades significantly in real-world scenarios. This limitation continues to hinder active stereo matching in practical applications. Unlike previous methods, our method constructs proxy labels directly from real-world scenes to enable supervised training in real environments. This greatly improves the model's real-world performance. As shown in Figure 1, (c) illustrates the proxy labels synthesized by our method, while (d) and (e) show that our results outperform those captured by the commercial active stereo depth camera D435.

We first build a new large-scale active stereo matching dataset using a RealSense D435 depth camera, which contains over 100,000 images across various real-world scenes. For each scene, we capture multi-view active IR stereo pairs under different object arrangements and viewpoints. To obtain high-quality binary IR patterns, seven images with varying IR emission intensities are captured at each viewpoint. We then propose an active 2D Gaussian Splatting [15] (2DGS)-base synthesis method, which reconstructs a 3D scene from multi-view images captured in real-world environments. In the reconstructed scene, we generate disparity proxy labels for active stereo pairs. Moreover, its novel-view rendering capability allows us to expand the dataset by synthesizing additional active stereo pairs and their corresponding disparity proxy labels. Using these synthesized proxy labels, we perform supervised training of active stereo networks in real-world settings. To dynamically balance the contributions of supervised and self-supervised learning during training, we design a hybrid supervision regularization strategy that adaptively adjusts the loss weights. Experimental results show that our method outperforms state-of-the-art deep stereo matching approaches and leading commercial depth sensors in both quantitative metrics and qualitative comparisons. Extensive ablation studies further confirm the effectiveness and robustness of our proposed approach.

In summary, the main contributions are as follows:

1. We capture a large-scale IR stereo dataset in real-world scenarios using a handheld RealSense camera, along with our active 2DGS model, addressing the lack of open-source datasets in this field.
2. We propose an active 2DGS-based synthesis framework that renders IR images with active projection patterns and generates high-quality disparity proxy labels, as well as additional novel-view training data, enabling supervised learning for active stereo matching in real-world scenes.
3. We propose a hybrid supervision regularization strategy to balance the contributions of proxy-supervised and self-supervised learning during training, leading to more stable and effective optimization.
4. Our method achieve state-of-the-art performance, demonstrating the highest accuracy on simulation data and delivering the best visual quality on real-world data.

## 2. Related Work

**Depth Stereo Matching.** In the past, stereo matching relied heavily on handcrafted algorithms [29, 31, 34, 35, 54]. However, in recent years, the development of deep learning has led to groundbreaking advancements. With the emergence of DispNet [30], end-to-end learning quickly surpasses previous methods, establishing new performance benchmarks [4, 6, 16, 24, 36, 41, 49, 57, 58]. Recently, techniques inspired by RAFT [38] or utilizing transformers [13, 23] have emerged, achieving significant success on public benchmarks. However, their performance is largely dependent on the accuracy of data labels. In the domain of active stereo matching, there is no large-scale and accurate dataset available for training.

**Self-Supervised Stereo.** Self-supervision through left-right consistency is a common strategy for stereo matching in scenarios with uncertain ground truth depth. This method involves reconstructing the right view from the left view and its predicted disparity map. The reconstruction loss then acts as the training supervision signal. Other approaches use photometric losses from a single stereo image pair [12, 39, 40, 60]. While these self-supervised methods are practical, they typically perform best within specific domains. Despite progress in pseudo-label guided self-supervision, they often show limited generalizability [1, 53], leading to results that may not match the quality achieved with labeled supervision.

**Synthetic Data for Stereo.** Using synthetic data in stereo matching is promising for improving model performance. However, the domain shift remains a challenge, limiting model effectiveness on real-world data unseen during training. To tackle this, recent researches introduces

novel methods. He et al. [14] propose a semi-synthetic approach that integrates realistic textures into synthetic data, enabling efficient dataset generation. It effectively reduces the gap between synthetic and real data, showing enhanced performance on limited-data benchmarks like Middlebury [2], KITTI [10], and ETH3D [33], and outperforming traditional synthetic datasets. Meanwhile, [22, 48] develop a Synthetic-to-Real Domain Adaptation (SDA) network. It aims to minimize differences between synthetic and real data by using edge cues for domain adaptation and combining features with a Spatial Feature Transform (SFT) layer. By focusing on synthetic data training, this method outperforms several existing domain adaptation techniques, boosting the generalization ability of stereo matching CNNs.

**Gaussian Splatting.** In recent years, Neural Radiance Fields (NeRF) [47] have emerged as a pivotal approach for novel view synthesis through neural rendering. By implicitly encoding scenes using multilayer perceptrons, NeRF achieves impressive photorealism but faces challenges in computational efficiency and dynamic scene modeling. This has spurred the development of 3D Gaussian Splatting (3DGS) [17] as an efficient alternative. Representing scenes explicitly with anisotropic 3D Gaussians, 3DGS enables both high-quality view synthesis and real-time performance. The framework has since been improved and extended to dynamic scene reconstruction, material editing, and animatable avatars, demonstrating broader applicability beyond static view synthesis [20, 26, 28, 46, 52].

## 3. Method

Figure 2 illustrates the overall framework of our GS-ASM. We first briefly introduce the background of 2DGS. Then, we present the detailed implementation of our method, including the active 2DGS construction, generation of depth proxy labels, active stereo novel view synthesis, active stereo network, and hybrid supervision regularization.

### 3.1. Preliminaries: Gaussian Splatting

In 3DGS, scenes are represented by a set of 3D Gaussian primitives and rendered into images via a splatting-based rasterization technique, each primitive is explicitly parameterized by a 3D covariance matrix $\Sigma_{3D}$ and the center position $\mathbf{x}_k$ of the $k$-th Gaussian, with its mathematical expression presented as follows:

$$G(\mathbf{x}) = \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{x_k})^\top \Sigma_{3D}^{-1}(\mathbf{x}-\mathbf{x_k})] \quad (1)$$

Here, $\mathbf{x}$ denotes the current position, and the covariance matrix $\Sigma_{3D}$ can be decomposed into a scaling matrix $\mathbf{S}$ and a rotation matrix $\mathbf{R}$, following the relationship $\Sigma_{3D} = \mathbf{RSS}^\top\mathbf{R}^\top$.

To render 3D Gaussians onto 2D planes, they are projected from 3D space to camera space. The projected 2D covariance matrix $\Sigma_{3D}^{'}$ is derived as:

$$\Sigma_{3D}^{'} = JW\Sigma_{3D}W^\top J^\top \quad (2)$$

where $W$ is the world-to-camera transformation matix and $J$ is the Jacobian of the affine approximation of the projective transformation.

During the rendering process, all Gaussians are sorted by their depth values from near to far. The final color $\mathbf{c}$ at pixel $\mathbf{x}$ is computed through volumetric alpha blending in this depth-ordered sequence:

$$\mathbf{c}(\mathbf{x}) = \sum_{i\in\mathcal{N}} \omega_i \mathbf{c}_i, \quad (3)$$

where $\omega_i = T_i\alpha_i G_i(\mathbf{x})$ denotes the weight of the $i$-th Gaussian, $T_i = \prod_{j=1}^{i-1}(1-\alpha_j G_j(\mathbf{x}))$ represents its accumulated transmittance.

While 3DGS enables efficient rendering, its volumetric representation often suffers from multi-view geometric inconsistency, limiting surface reconstruction quality. To address this, 2DGS [15] replaces 3D Gaussians with planar 2D disks by reducing the 3D covariance matrix to its 2D form, which involves removing the third row and column of the matrix. Rather than splatting in image space, 2DGS computes ray-disk intersections in local tangent space, avoiding splat degeneration at grazing angles. The rasterization process then sorts these 2D Gaussians by depth and blends them front-to-back using alpha compositing.

### 3.2. Active Neural Rendering for Proxy Label

Due to the lack of open-source datasets in active stereo matching, we construct training and testing data using the D435 depth sensor. However, collecting ground truth disparity for target views in real-world scenarios is challenging. To address this, we propose an active neural rendering approach that leverages 2DGS to generate high-quality disparity maps as synthetic proxy labels.

The IR imaging process in active stereo systems involves both ambient lighting and projected patterns. The grayscale intensity at pixel $\mathbf{x}(u,v)$ is modeled as:

$$x_l(u,v) = I_l(u,v) + \alpha \cdot e \cdot K_l(u,v) + \epsilon, \quad e \geq 0 \quad (4)$$

where $I_l(u,v)$ represents ambient light intensity, $K_l(u,v)$ denotes the projected binary pattern, $\alpha$ is the surface reflectance coefficient, $e$ is the pattern emissivity, and $\epsilon$ accounts for sensor noise.

Since the projected binary patterns violate the assumptions of volumetric rendering, we use IR images captured with zero emitter power ($e = 0$) to train the 2DGS model. This ensures the rendering process accurately captures scene geometry without pattern interference.

Once the 2DGS training is finished, the depth map is generated. The depth $z(\mathbf{x})$ at each pixel $\mathbf{x} = (u,v)$ is

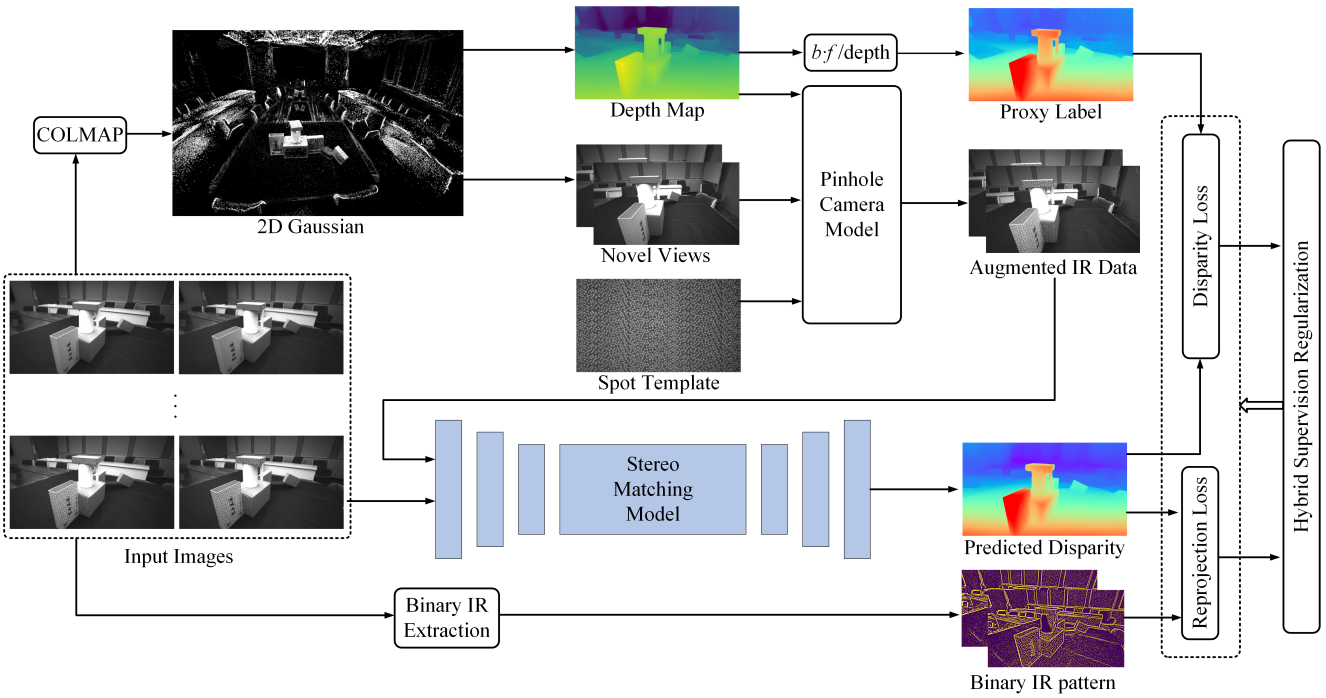



Figure 2. The pipeline of GS-ASM. The 2DGS model renders IR images in no-projection mode to generate disparity proxy labels from depth maps for proxy-supervision. Its rendered novel views (with depth maps) undergo the Pinhole Camera Model for data augmentation. The active stereo matching network utilizes IR images and binary projection patterns for reprojection, employing joint training through disparity supervision and reprojection loss.

computed as the weighted sum of the depths of all overlapping 2D Gaussians, using the same alpha-blending weights $\omega_i$ as in color rendering. Afterward, 2DGS utilizes TSDF fusion[7] for mesh extraction with depth, the disparity $d(\mathbf{x})$ can be extracted from the rendered depth, which corresponds to the depth of the rendered image:

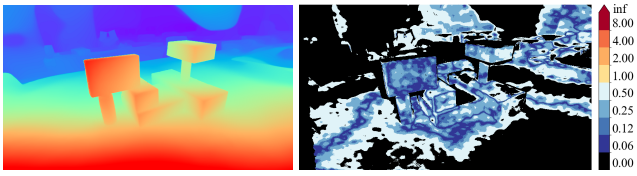
$$z(\mathbf{x}) = \frac{\sum_i \omega_i z_i}{\sum_i \omega_i} \quad d(\mathbf{x}) = \frac{b \cdot f}{z(\mathbf{x})} \tag{5}$$

where $f$ represents the focal length estimated by COLMAP [32], and $b$ is the baseline of the depth camera. As shown in Figure 3, after converting to disparity, the disparity values align closely with the results from the RealSense D435 sensor, with an error margin within 1 pixel.

This active 2DGS pipeline enables the generation of accurate disparity proxy labels from multi-view IR images, providing essential supervision for training our stereo matching network.

### 3.3. Active Stereo Novel View Synthesis

2DGS introduces precise geometric constraints by confining Gaussian primitives to 2D planes, thereby preserving local geometric consistency. Building on this advantage, our work explores 2DGS for data augmentation in binocular vision tasks, leveraging its novel view synthesis capability defined in Equation 3 and Equation 5.




(a) Proxy label (b) Disp error

Figure 3. Comparison of proxy label with D435 sensor parallax, error range within 1 pixel.

To maintain rendering quality while increasing viewpoint diversity, we apply a constrained perturbation strategy to the COLMAP-estimated camera poses. This generates novel views with limited deviation from original perspectives, enhancing detail coverage while preserving fidelity. For stereo pairs, the baseline $b$ is derived from relative poses in the COLMAP coordinate system, ensuring geometric and metric consistency with the original reconstruction.

For active IR pattern synthesis, we project spot template onto novel views using the pinhole camera model with COLMAP-derived intrinsics and 2DGS-generated depth maps. The 3D coordinate $\mathbf{P}\text{cam} = (X, Y, Z)^\top$ for pixel $\mathbf{x} = (u, v)$ is back-projected as:

$$\mathbf{P}_{\text{cam}} = z(\mathbf{x}) \cdot \mathbf{K}^{-1} \cdot [u, v, 1]^\top \tag{6}$$

The spot template is then scaled according to depth values $z(\mathbf{x})$ and blended with the synthesized views. This integra-

tion ensures geometric consistency between active illumination and novel viewpoints, completing our data augmentation pipeline.

The entire process yields geometrically consistent stereo image pairs with corresponding active IR patterns and, importantly, accurate proxy labels derived from 2DGS rendering, providing rich training data with reliable ground truth annotations.

### 3.4. Active Stereo Network

Traditional stereo matching relies on self-supervised reprojection between stereo pairs, leveraging geometric consistency [11, 59] but often failing in textureless regions due to limitations of photometric consistency.

In contrast, active stereo networks utilize active light sources (e.g., IR projectors) to project specific patterns onto the scene, providing additional features in textureless regions. This enables the model to learn not only geometric consistency but also to optimize reprojection loss by learning the shape and distribution of projected patterns, thereby improving matching accuracy.

**Reprojection loss based on projection pattern.** To extract patterns from the temporal image sequence, the method first compares the original IR images $x^{(0)}, x^{(1)}, \ldots, x^{(n)}$ with the estimated images obtained through linear regression $\hat{x}^{(0)}, \hat{x}^{(1)}, \ldots, \hat{x}^{(n)}$. The differences between these images are then subjected to local window normalization to enhance key features, followed by binarization using the following formula [27].

$$K(u,v) = \begin{cases} 1 & ||\hat{x}^{(n)}(u,v) - \hat{x}^{(0)}(u,v)|| > \delta(u,v) + c \\ 0 & otherwise \end{cases}$$

$$\delta(u,v) = \frac{1}{w^2} \sum ||W(\hat{x}^{(n)}, u, v) - W(\hat{x}^{(0)}, u, v)|| \quad (7)$$

where $W(x,u,v)$ represents a local window centered at pixel $(u,v)$ within image $x$, with window size $w$. The parameter $c$ is a threshold for suppressing noise and small regions. In real data, $n$ ranges from 0 to 6, while in synthetic data, $n$ takes values of 0 and 1.

We construct the reprojection loss on the extracted binary IR patterns $K_l, K_r$:

$$\mathcal{L}_{\text{self}}(K_l, K_r, \hat{I}_c^d) = \left\| K_l(u_p, v_p) - \hat{K}_r(u_p, v_p) \right\|^2 \quad (8)$$

As shown in Figure 4, the binary IR pattern eliminates the influence of object textures and ambient lighting, preserving only the projected pattern for robust stereo matching.

**Disparity loss.** For stereo image pairs of real and synthetic data $(y_l, y_r, y_d)$, we follow [4] and employ a smooth $L_1$ loss between the ground truth disparity $y_d$ and the predicted disparity:

$$\mathcal{L}_{\text{disp}} = L_{1\text{smooth}}(F(y_l, y_r), y_d) \quad (9)$$

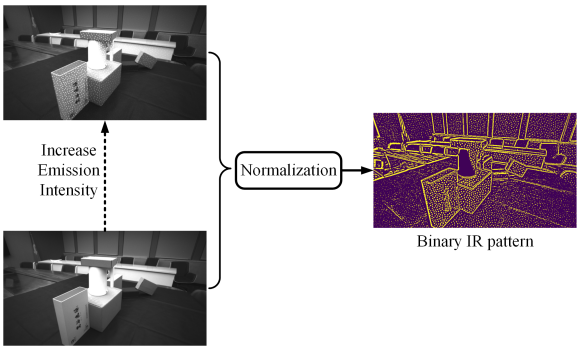



Figure 4. The illustration of Binary IR pattern extraction.

where $F(\cdot)$ represents stereo matching network.

### 3.5. Hybrid Supervision Regularization

To mitigate loss oscillations arising from noisy proxy supervision in real-world data, we propose an adaptive dynamic weighting mechanism for mixed-domain training. This method balances the proxy-supervised and self-supervised losses to enable stable and efficient convergence.

Our mixed-domain loss function integrates proxy-supervised and self-supervised terms from both real-world and synthetic data, with adaptive weights that evolve during training:

$$\begin{aligned} \mathcal{L}(x_l, x_r, y_d) = \mu(t) \cdot [\mathcal{L}_{\text{real-disp}} + \mathcal{L}_{\text{sim-disp}}] \\ + \lambda(t) \cdot [\mathcal{L}_{\text{real-self}} + \mathcal{L}_{\text{sim-self}}] \end{aligned} \quad (10)$$

where $\mathcal{L}_{\text{self}}$ (Equation 8) and $\mathcal{L}_{\text{disp}}$ (Equation 9) denote self-supervised and proxy-supervised losses, with subscripts indicating real/simulated domains. The weights $\mu(t)$ and $\lambda(t)$ are adjusted based on the optimization progress of their corresponding loss terms, prioritizing those with rising losses and de-emphasizing those that are converging. Here, $t$ represents the iteration step during model training.

**Disparity loss.** We initialize $\mu(0) = 0.01$ and $\lambda(0) = 2$ for balanced optimization across loss terms.

**Normalization.** To prevent numerical instability, weights are normalized to maintain meaningful proportions, $\mu(t)$ and $\lambda(t)$ are normalized using softmax to preserve relative proportions.

**Adaptive Adjustment.** The weights are dynamically adjusted in inverse proportion to their loss changes between iterations to balance the optimization. Formally:

$$\begin{aligned} \hat{\mu}(t+1) &= \hat{\mu}(t) \cdot \left[ 1 + \alpha \cdot \left( \frac{\mathcal{L}_{\text{disp-total}}(t)}{\mathcal{L}_{\text{disp-total}}(t-1)} - 1 \right) \right] \\ \hat{\lambda}(t+1) &= \hat{\lambda}(t) \cdot \left[ 1 + \alpha \cdot \left( \frac{\mathcal{L}_{\text{self-total}}(t)}{\mathcal{L}_{\text{self-total}}(t-1)} - 1 \right) \right] \end{aligned} \quad (11)$$

Here, $\alpha = 0.1$ is the update rate hyperparameter, $\mathcal{L}_{\text{disp-total}} = \mathcal{L}_{\text{real-disp}} + \mathcal{L}_{\text{sim-disp}}$ and $\mathcal{L}_{\text{self-total}} = \mathcal{L}_{\text{real-self}} + \mathcal{L}_{\text{sim-self}}$ denote total proxy-supervised and self-supervised

losses, respectively. To avoid extreme values, raw weights are clamped: $\hat{\mu}, \hat{\lambda} \in [10^{-3}, 10]$.

# 4. Experiments

## 4.1. Dataset

We collect data in indoor environments using the D435 depth camera, with a captured image resolution of 848×480. For each scene, the collected data include color images, depth images, and six corresponding left and right IR views at different intensity levels per frame. We use COLMAP [32] to estimate the intrinsic and extrinsic parameters of the camera for each scene. Additionally, we construct active 2DGS to render more novel-view stereo IR images for data augmentation. Meanwhile, we generate proxy labels for each pair of stereo IR images to enable supervised learning. Furthermore, following the DREDS framework [9], we render 20,000 synthetic training images and 3,000 synthetic test images using Blender, including IR images with and without light patterns, as well as depth maps. By aligning the intrinsic and extrinsic parameters of the RealSense camera, we construct a synthetic dataset that allows us to quantitatively compare our method with the depth camera on the same test set. The training and test sets are split at the scene level.

## 4.2. Implementation Details

**Deep stereo training.** We use the PSMNet [4], RAFT [38], and StereoNet [19] models as evaluation subjects. To ensure a fair comparison, our method applies the same data augmentation operations as the baseline method. Specifically, the brightness and contrast of the images are uniformly scaled within the ranges of 0.4 to 1.4 and 0.8 to 1.2, respectively. Gaussian blur is added, with a Gaussian kernel size of 9×9 and standard deviation uniformly selected between 0.1 and 2. During the experiments, we set the batch size to 4 and apply random cropping of images to a size of 256×512. All models are trained on both real and synthetic data simultaneously.

**Evaluation metrics.** Our evaluation protocol involves calculating both the End-point-error (EPE), defined as the mean absolute disparity error, and the proportion of pixels for which the disparity error does not exceed the thresholds of 1, 3, and 5 pixels compared to the actual depth values. These criteria serve as robust measures of our model's precision and the effectiveness of our disparity estimation techniques under diverse experimental settings. All experiments are conducted on one 4090 NVIDIA GPU.

## 4.3. Comparison with other Methods

To comprehensively evaluate our method, we compare it with other learning-based approaches and a well-regarded commercial depth sensor, the RealSense D435. We conduct both quantitative and qualitative evaluations on the Blender synthetic dataset, and qualitative comparisons on our captured real-world scenes. The results show that our method consistently outperforms all other approaches in both synthetic and real-world scenarios.

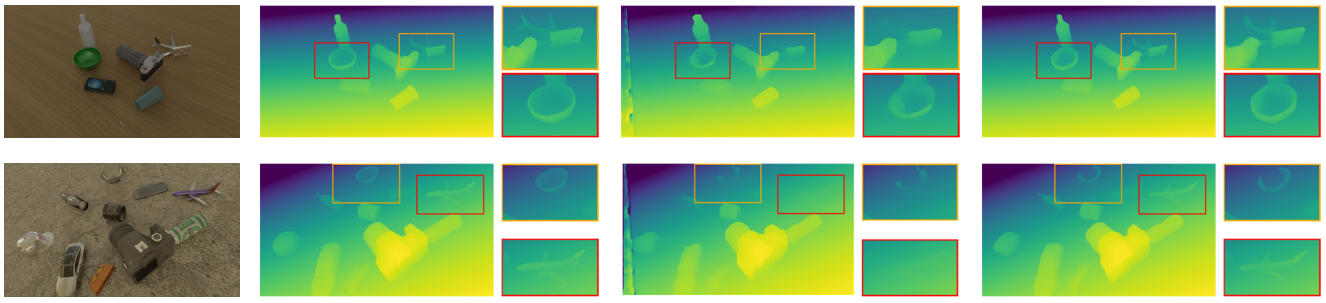
| **Model** | **Method** | EPE(px)↓ | 1px↑ | 3px↑ | 5px↑ |
|---|---|---|---|---|---|
| PSMNet | D435 | 0.5488 | 0.9032 | 0.9811 | 0.9911 |
| | Baseline | 0.4300 | 0.9446 | 0.9829 | 0.9910 |
| | **Ours** | **0.2613** | **0.9597** | **0.9897** | **0.9955** |
| RAFT | D435 | 0.5428 | 0.9513 | 0.9820 | 0.9889 |
| | Baseline | 0.4715 | 0.9306 | 0.9719 | 0.9856 |
| | **Ours** | **0.1279** | **0.9744** | **0.9925** | **0.9955** |
| StereoNet | D435 | 0.7190 | 0.8340 | 0.9752 | 0.9890 |
| | Baseline | 0.5551 | 0.9206 | 0.9691 | 0.9826 |
| | **Ours** | **0.4124** | **0.9223** | **0.9839** | **0.9936** |

Table 1. Quantitative Evaluation of disparity estimation on simulated test datasets.

**Comparison with Learning-based Methods.** We adopt ActiveZero [27] as our baseline framework, which employs a hybrid-domain training strategy—self-supervised learning in real-world domains and supervised learning on the Blender synthetic dataset. For a fair comparison, our method and the baseline are trained with identical hybrid-domain inputs. To further compare our method with a commercial sensor–based learning method, we use the depth maps captured by the RealSense D435 as supervision and train the model under a combined supervised and self-supervised setting. We evaluate different stereo learning models, including PSMNet [4], StereoNet [19], and RAFT [38], to validate the generality of our framework.

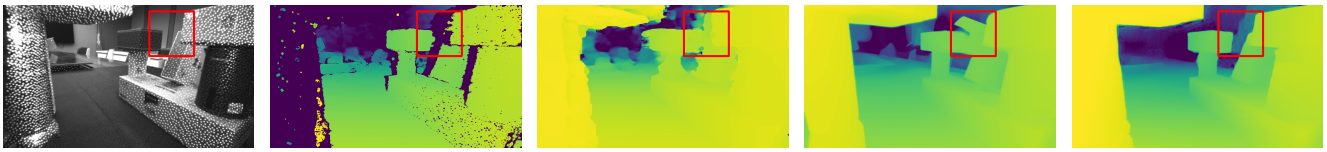
Quantitative results on the synthetic test set are presented in Table 1. Regardless of the underlying network architecture, our method consistently outperforms existing approaches across all evaluation metrics. Figure 5 provides qualitative comparisons of depth estimation in synthetic scenes. While ActiveZero produces generally plausible predictions, it exhibits noticeable artifacts near the left image boundaries. In contrast, our method preserves fine-grained geometric details, especially for distant objects where conventional methods typically suffer performance degradation. This demonstrates the strong capability of our model in handling complex synthetic scenarios. Figure 7 visualizes the pixel-wise disparity errors between the estimated maps and ground truth, where our method achieves the lowest error rates among all evaluated approaches.

Figure 6 shows the depth estimation results of different active stereo methods on real-world scenes using RAFT



(a) Color (b) GT (c) ActiveZero (d) Ours

Figure 5. Depth estimation comparison on simulated test datasets.



(a) IR reference (b) D435 disparity (c) ActiveZero (d) Proxy labels (e) Ours

Figure 6. Depth estimation comparison on real test datasets.

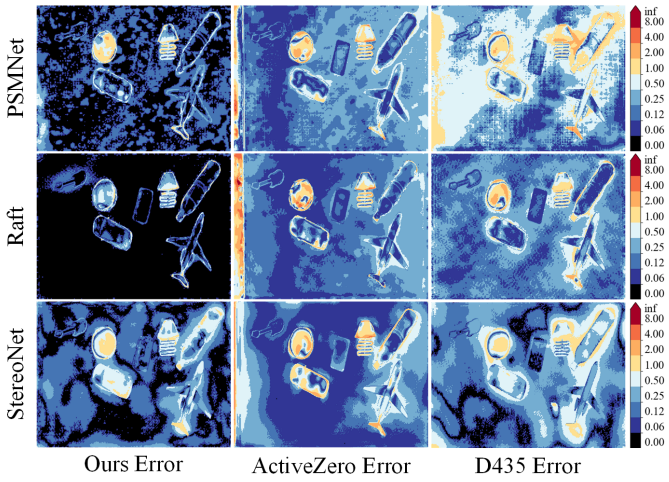
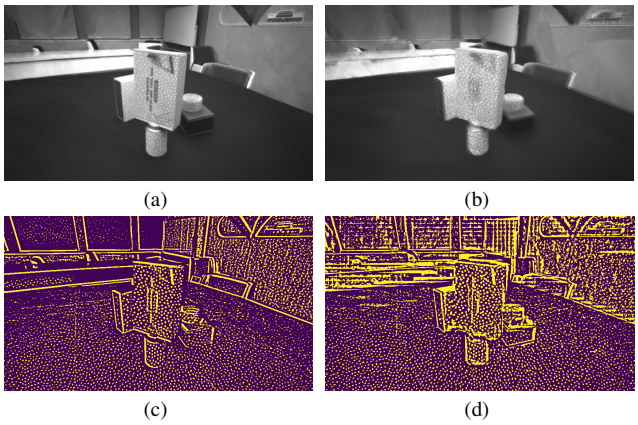



Figure 7. Pixel-wise disparity error comparison of our method, ActiveZero, and D435 on three backbone models.

as the backbone. Although ActiveZero performs well in synthetic environments (as shown in Figure 5), it generalizes poorly to real-world data—producing even worse results than those captured by the commercial D435 sensor. In particular, as highlighted by the red circles, ActiveZero exhibits severe depth estimation failures in many regions. In contrast, our method significantly improves real-world depth estimation and remains robust even in challenging regions where the generated proxy labels (Figure 6d) may contain inaccuracies, as shown in Figure 6e. These results demonstrate the strong practical potential and robustness of our method for real-world active stereo applications.

**Comparison with Intel RealSense D435.** Although commercial depth sensors such as the RealSense D435 are widely used in both industrial and academic fields, they often suffer from severe measurement noise and significant depth loss around object boundaries. We conduct qualitative comparisons with the D435 sensor, as shown in Figure 6. Our method achieves a substantial improvement in both accuracy and visual quality—it not only suppresses noise effectively but also recovers fine-grained disparity details, particularly along object edges.



(a) (b) (c) (d)

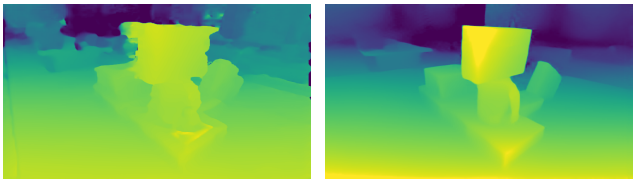
Figure 8. (a) and (b) show the rendering results at intensity levels 0 and 6, respectively. (c) and (d) present the binary extraction results of the corresponding IR images.

| IR | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| $I^0$ | 33.23 | 0.9683 | 0.0762 |
| $I^1$ | 32.73 | 0.8959 | 0.1655 |
| $I^2$ | 32.42 | 0.8522 | 0.2454 |
| $I^3$ | 32.20 | 0.8222 | 0.3001 |
| $I^4$ | 31.99 | 0.7958 | 0.3442 |
| $I^5$ | 31.86 | 0.7706 | 0.3749 |
| $I^6$ | 31.75 | 0.7533 | 0.3936 |

Table 2. The reconstruction effect of 2DGS under different IR intensities.

| Method | Proxy Labels | Novel Viewpoints | Regularization | EPE(px)↓ | 1px↑ | 3px↑ | 5px↑ |
|---|---|---|---|---|---|---|---|
| Baseline | – | – | – | 0.4715 | 0.9306 | 0.9719 | 0.9856 |
| Ours | ✓ | – | – | 0.3482 | 0.9526 | 0.9880 | 0.9942 |
| | ✓ | ✓ | – | 0.2736 | 0.9647 | 0.9885 | 0.9945 |
| | ✓ | ✓ | ✓ | **0.1279** | **0.9744** | **0.9925** | **0.9955** |

Table 3. Ablation Study on RAFT Stereo Matching Network: Effects of key components on disparity estimation performance



(a) Baseline (b) Baseline w/ Proxy Labels

Figure 9. Ablation Study: Depth estimation details of Baseline vs. Baseline with Proxy Labels in simulated domains.

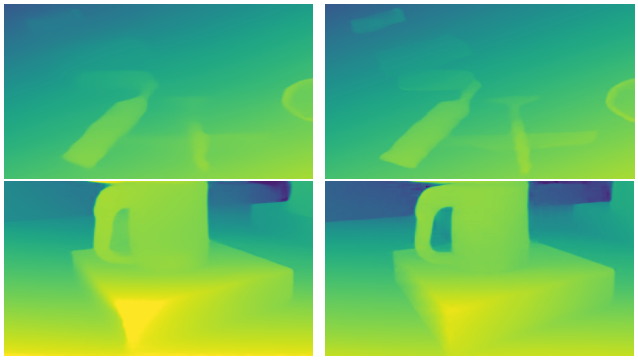
### 4.4. Ablation Study

In this section, we validate the effectiveness of each design selection through ablation experiments.

**Construction of Active 2DGS.** 2DGS rendering was applied to all IR intensity levels, with reconstruction outcomes compared in Table 2. A consistent decline in quality was observed as IR intensity increased. Binary extraction was performed on all rendered IR images, showing that intensity level 0 produced the best results, with its pattern successfully capturing detailed textures. However, as IR intensity increased, the clarity and detail in binary IR images diminished, as shown in the Figure 8.

**Component Effectiveness.** We conduct ablation studies to evaluate three key components: proxy labels, novel viewpoint data augmentation, and hybrid supervision regularization. As evidenced in Table 3, each component contributes to progressive performance gains.

As shown in Figure 9, compared with the baseline method, incorporating the proxy labels generated by our active 2DGS for supervised training significantly improves disparity estimation in real-world scenes, especially around object boundaries. This demonstrates that the proxy labels generated from real-world active scenes effectively enhance the generalization ability of ActiveZero in real environments. Furthermore, Figure 10 shows that our proposed hybrid supervision regularization further improves disparity estimation for fine geometric details in real-world scenes. It also enhances the accuracy of distant object reconstruction in synthetic environments, validating the effectiveness of our hybrid training strategy. The complete framework, integrating all three components, achieves optimal performance across all evaluation metrics. The adaptive weighting strategy successfully balances multiple supervision sources, resulting in more accurate and robust disparity estimation.



(a) Ours w/o Regularization (b) Ours

Figure 10. Ablation Study: Depth estimation details of Our method without Regularization vs. Complete our method in simulated and real domains.

## 5. Conclusion

We present a large-scale active stereo dataset to address the lack of public resources in this field. Leveraging real-world active stereo images, we construct an active 2DGS model to generate disparity proxy labels for supervised training, without relying on any ground-truth input. The active 2DGS also enables novel-view data augmentation, expanding the diversity of training samples. Furthermore, we introduce a hybrid supervision regularization strategy to dynamically balance the self-supervised and proxy-supervised losses during training, enhancing fine details and distant object accuracy. Through a mixed-domain training framework, our method surpasses state-of-the-art stereo matching networks and commercial depth sensors in both quantitative and qualitative evaluations.

**Limitations.** Active stereo networks face challenges due to reliance on active light sources, which are vulnerable to interference from ambient light, causing noise and data acquisition failures in outdoor scenes. Their dependence on specific hardware and controlled environments also limits generalization across diverse scenarios.

**Future Directions.** Future research can integrate complementary technologies like LiDAR or ToF. Multimodal data fusion offers potential to enhance generalization and robustness under complex lighting conditions.

## Acknowledgements

This research was supported by the Joint Fund of the Zhejiang Provincial National Science Foundation of China (No.LLSSZ25F030002). This work was also supported by the National Key Research and Development Program of China under Grant (2023YFB4502800), Zhejiang Provincial Key Laboratory of Low Altitude Ubiquitous Networking Technology, Hangzhou Dianzi University, Hangzhou, 310018, China

## References

[1] Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 614–632. Springer, 2020. 2

[2] Simon Baker, Daniel Scharstein, James P Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011. 3

[3] Luca Bartolomei, Matteo Poggi, Fabio Tosi, Andrea Conti, and Stefano Mattoccia. Active stereo without pattern projector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18470–18482, 2023. 1

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2, 5, 6

[5] Rui Chen, Jing Xu, and Song Zhang. Comparative study on 3d optical sensors for short range applications. *Optics and lasers in engineering*, 149:106763, 2022. 1

[6] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in neural information processing systems*, 33:22158–22169, 2020. 2

[7] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 4

[8] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4): 1, 2017. 1

[9] Qiyu Dai, Jiyao Zhang, Qiwei Li, Tianhao Wu, Hao Dong, Ziyuan Liu, Ping Tan, and He Wang. Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In *European Conference on Computer Vision (ECCV)*, 2022. 6

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 5

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 2

[13] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *European Conference on Computer Vision*, pages 263–279. Springer, 2022. 2

[14] Ju He, Enyu Zhou, Liusheng Sun, Fei Lei, Chenyang Liu, and Wenxiu Sun. Semi-synthesis: A fast way to produce effective datasets for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2884–2893, 2021. 3

[15] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 2, 3

[16] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[18] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–10, 2017. 1

[19] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European conference on computer vision (ECCV)*, pages 573–590, 2018. 6

[20] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 505–515, 2024. 3

[21] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 2

[22] Xing Li, Yangyu Fan, Zhibo Rao, Guoyun Lv, and Shiya Liu. Synthetic-to-real domain adaptation joint spatial feature transform for stereo matching. *IEEE Signal Processing Letters*, 29:60–64, 2021. 3

[23] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 2

[24] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820, 2018. 2

[25] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10233–10240. IEEE, 2020. 1

[26] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 3

[27] Isabella Liu, Edward Yang, Jianyu Tao, Rui Chen, Xiaoshuai Zhang, Qing Ran, Zhu Liu, and Hao Su. Activezero: Mixed domain learning for active stereovision with zero annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13033–13042, 2022. 2, 5, 6

[28] Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. Thermal-gaussian: Thermal 3d gaussian splatting. *arXiv preprint arXiv:2409.07200*, 2024. 3

[29] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016. 2

[30] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2

[31] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 2

[32] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 4, 6

[33] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 3

[34] Akihito Seki and Marc Pollefeys. Patch based confidence prediction for dense disparity map. In *BMVC*, page 4, 2016. 2

[35] Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 231–240, 2017. 2

[36] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 2

[37] Fuchun Sun, Runfa Chen, Tianying Ji, Yu Luo, Huaidong Zhou, and Huaping Liu. A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives. *CAAI Artificial Intelligence Research*, 3:9150042, 2024. 1

[38] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2, 6

[39] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2019. 2

[40] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 195–204, 2019. 2

[41] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8942–8952, 2021. 2

[42] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 2

[43] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 1

[44] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019. 1

[45] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8071–8081, 2019. 2

[46] Zhicong Wu, Hongbin Xu, Gang Xu, Ping Nie, Zhixin Yan, Jinkai Zheng, Liangqiong Qu, Ming Li, and Liqiang Nie. Textsplat: Text-guided semantic fusion for generalizable

gaussian splatting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8478–8487, 2025. 3

[47] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, pages 641–676. Wiley Online Library, 2022. 3

[48] Chenggang Yan, Tong Teng, Yutao Liu, Yongbing Zhang, Haoqian Wang, and Xiangyang Ji. Precise no-reference image quality evaluation based on distortion identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s):1–21, 2021. 3

[49] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 2

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 1

[51] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1

[52] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024. 3

[53] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised cross-media hashing learning via knowledge graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022. 2

[54] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(65):1–32, 2016. 2

[55] Longjian Zeng, Zunjie Zhu, Rongfeng Lu, Ming Lu, Bolun Zheng, Chenggang Yan, and Anke Xue. Depthdark: Robust monocular depth estimation for low-light environments. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11239–11248, 2025. 1

[56] Chenyang Zhang, Tiansu Chen, Eric Shaffer, and Elahe Soltanaghai. Focusflow: 3d gaze-depth interaction in virtual reality leveraging active visual depth manipulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024. 1

[57] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 185–194, 2019. 2

[58] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 420–439. Springer, 2020. 2

[59] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–801, 2018. 5

[60] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017. 2