ActiveZero++: Mixed Domain Learning Stereo and Confidence-Based Depth Completion With Zero Annotation

Rui Chen, Isabella Liu, Edward Yang, Jianyu Tao, Xiaoshuai Zhang, Qing Ran, Zhu Liu, *Senior Member, IEEE*, Jing Xu, *Member, IEEE*, and Hao Su

***Abstract*—Learning-based stereo methods usually require a large scale dataset with depth, however obtaining accurate depth in the real domain is difficult, but groundtruth depth is readily available in the simulation domain. In this article we propose a new framework, ActiveZero++, which is a mixed domain learning solution for active stereovision systems that requires no real world depth annotation. In the simulation domain, we use a combination of supervised disparity loss and self-supervised loss on a shape primitives dataset. By contrast, in the real domain, we only use self-supervised loss on a dataset that is out-of-distribution from either training simulation data or test real data. To improve the robustness and accuracy of our reprojection loss in hard-to-perceive regions, our method introduces a novel self-supervised loss called temporal IR reprojection. Further, we propose the confidence-based depth completion module, which uses the confidence from the stereo network to identify and improve erroneous areas in depth prediction through depth-normal consistency. Extensive qualitative and quantitative evaluations on real-world data demonstrate state-of-the-art results that can even outperform a commercial depth sensor. Furthermore, our method can significantly narrow the Sim2Real domain gap of depth maps for state-of-the-art learning based 6D pose estimation algorithms.**

***Index Terms*—Depth completion, learning-based stereo, sensors, Sim2Real.**

I. Introduction

DEPTH sensors can provide 3D geometry information about a target scene, which is critical in various robotic applications, including mapping, navigation, and object manipulation [1], [2], [3]. Among the different types of depth sensors commercially available, active stereovision depth sensors (eg. Intel RealSenseD series [4]) are the most widely adopted in both industry and academic settings due to their high spatial resolution, high accuracy, and low cost [5]. Active stereovision sensors are composed of an infrared (IR) pattern emitter and two IR cameras with the IR pattern projected onto the target scene to facilitate stereo matching. However, since these sensors use classical stereo algorithms, they suffer from common stereo matching issues such as over smoothing, edge fattening and holes for specular and transparent objects so they are non-ideal for robotic applications which require high precision and completeness [6].

Learning based methods can solve the aforementioned issues by generating more accurate and complete depth maps through more robust feature extraction and the utilization of shape priors [7], [8], [9], [10]. However, a large scale stereo dataset with groundtruth depth is required to train these learning based methods, which is costly and time-consuming to collect in the real world. Therefore, one way to alleviate this problem is to use self-supervised learning. Self-supervised stereo methods [11], [12] use reprojection or other related losses between binocular images as supervision, but the fluctuation of these losses prohibit the network from reaching a meaningful optima. Another approach is to use simulation data where groundtruth depth is readily available. However due to the domain gap between the simulation and real world, networks trained on only simulation data cannot be reliably transferred to the real domain. Domain adaptation methods have been proposed to overcome the Sim2Real problem [13], but the introduction of Generative Adversarial Networks (GANs) makes the training process unstable [14] and the performance suboptimal.

This article first proposes an end-to-end learning stereo method that combines the advantages of self-supervised learning in the real domain and supervised learning in the simulation domain which we call *mixed domain learning*. This strategy significantly boosts the stereo network performance while also stabilizing and speeding up the optimization process. Specifically, by only needing to train on shape primitives in the simulation domain with groundtruth depth as supervision and an unrelated set of scenes in the real domain with reprojection as self-supervision, we are able to achieve comparable performance on completely out-of-distribution objects in the real domain as though we were directly training on those objects. Moreover, in order to improve the robustness of the reprojection loss to

Manuscript received 20 July 2022; revised 5 August 2023; accepted 10 August 2023. Date of publication 15 August 2023; date of current version 3 November 2023. Recommended for acceptance by J. Kwon. *(Corresponding author: Jing Xu.)*

Rui Chen and Jing Xu are with the State Key Laboratory of Tribology, Beijing Key Laboratory of Precision/Ultra-Precision Manufacturing Equipment Control, Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China (e-mail: ruichenthu14@gmail.com; jingxu@tsinghua.edu.cn).

Isabella Liu, Edward Yang, Jianyu Tao, Xiaoshuai Zhang, and Hao Su are with the Department of Computer Science and Engineering, University of California, La Jolla, CA 92093 USA (e-mail: lal005@ucsd.edu; edwardyang016@gmail.com; jit124@ucsd.edu; xiz040@eng.ucsd.edu; haosu@eng.ucsd.edu).

Qing Ran is with Alibaba DAMO Academy, Hangzhou 311121, China (e-mail: ranqing@zju.edu.cn).

Zhu Liu is with Amazon, Seattle, WA 98109 USA (e-mail: zhu.liu@ieee.org).

Digital Object Identifier 10.1109/TPAMI.2023.3305399

environmental illumination and scene texture, we propose the use of temporal IR by progressively adjusting the brightness of the emitted IR pattern and extracting the binary pattern from the temporal image sequences.

Although the mixed domain learning stereo method can accurately reconstruct real diffused and specular objects, its performance on transparent and translucent objects remains limited because the light transport on these objects in the real world is complex for high-fidelity simulation. In contrast to depth prediction, normal estimation is more locally constrained and easier since light's reflection and refraction are determined by the the incident ray's direction, the local surface geometry (normal, curvature), and the material properties of the surface. Specifically for active stereovision sensors, the position of the light source relative to the camera is constant for all images and independent of the scene. Therefore, the reflection and refraction of the actively projected pattern is directly affected by the local geometry of the surface. This makes it easier to use deep networks to infer the normal from image appearances (e.g. distorted patterns, direct specular highlights) than inferring the depth directly.

Recently, researchers have studied using learning-based method to predict the normal map and complete the depth measurement of transparent objects [15]. It must first learn to segment the entire transparent objects from the RGB image and is only trained and tested on fully transparent glass objects. However, for objects of multiple materials, only some regions have large depth errors and it is difficult to annotate these regions manually. We surprisingly find that the depth prediction error is highly related to the confidence map from the stereo network, so that we can extract the erroneous regions without the segmentation module or any manual annotation. We further develop a mixed domain learning stereo normal estimation module and improve the precision and completeness of transparent and translucent objects measurement through depth-normal consistency. Experimental results demonstrate that our method is able to outperform state-of-the-art learning-based stereo methods and commercial depth sensors, and ablation studies verify the effectiveness of each module in our work.

This article is an extension of our previous CVPR work [16]. There are three main additional contributions in this work:

(1) We design the confidence-based depth completion module which consists of confidence mask extraction, a mixed domain learning stereo normal estimation module and global depth-normal consistency optimization.

(2) We modify the domain randomization technique to suit the active stereovision synthetic dataset generation, including randomized IR pattern transformation, density, and environmental illumination.

(3) We evaluate the effectiveness of the proposed method on closing the depth Sim2Real domain gap: We train three state-of-the-art 6D pose estimation methods on the synthetic dataset generated by our method and directly test their performance on real depth maps by our method. It shows significant improvement when compared to using commercial depth sensors.

The rest of this article is organized as follows. Related works are reviewed in Section II. The proposed mixed domain learning stereo framework is described in Section III. The confidence-based depth completion method is introduced in Section IV. Experimental results of the stereo network and depth completion in the real world are presented in Section V. Finally, our work and its limitation are concluded in Section VI.

II. Related Work

Optical Depth Sensors: can be classified into four categories according to their underlying sensing principle [6]: passive stereovision, active stereovision, structured light, and time-of-flight. Each depth sensing technique has its own advantages and drawbacks. Giancola et al. [17] introduces the principles of different depth sensors and evaluated their metrological performance independently. Chen et al. [6] compared the short-range depth sensing performance of 8 commercially available optical depth sensors for different illumination settings and objects and found that active stereovision sensors and structured light sensors have similar performance to each other and better performance than the other two kinds of sensors. Furthermore, depth sensor performance varies among different objects with these sensors performing especially poorly on objects with complex optical characteristics [15]. In this article, we focus on improving the visual and numerical performance of active stereovision depth sensors, but the framework can also be applied for structured light sensors.

Learning Based Stereo: has become much more prevalent with large-scale benchmarks and higher computational ability [18], [19], [20]. Stereo matching for depth estimation is typically done in four steps: matching cost computation, cost aggregation, optimization, and disparity refinement [21]. Zbontar and LeCun were the first to design a network for computing matching costs by utilizing a deep Siamese architecture [22]. Building on this, DispNet introduced the first end-to-end framework for predicting entire disparity maps from stereo image pairs [23]. Works such as GWCNet followed and improved this framework by using 3D convolutions to compute better cost volumes [24]. Recent works have improved performance even further by utilizing multi-scale context aggregation to estimate depth at different resolutions in order to leverage global image-level information [7], [25]. However, the requirement of groundtruth depth as supervision has limited the application of learning based stereo.

Self-supervised Stereo: is a popular approach for stereo matching when groundtruth depth is unavailable. Godard et al. [26] explored the use of left-right consistency in a rectified stereo image pair for self-supervision. They reconstruct the right view based on the given left view and its predicted disparity map and then use the reconstruction loss as a supervision for training. PDANet [27] introduced the idea of perceptual consistency to improve reconstruction quality on regions with low texture and high color fluctuations. ActiveStereoNet [11] used local-contrast-normalized (LCN) reprojection loss on IR images as self-supervision to train a stereo network. However, this reprojection loss fluctuates along the epipolar line and is heavily influenced by occlusion and viewpoint variance. Not only that, LCN loss also suffers in areas where camera noise and

environmental illumination dominate the projected IR pattern since it only uses the IR image with projected pattern. Our method addresses these concerns using temporal IR reprojection loss by way of actively adjusting the brightness of the emitted IR pattern which is more robust to camera noise and environmental illumination.

Learning Based Active Stereo: approaches improve the depth estimation accuracy for texture-less areas by actively projecting a pattern or patterns onto the scene. Riegler et al. [28] proposed to use CNN to estimate the depth from the monocular image with pattern and train the CNN by using photometric and geometric losses. Johari et al. [29] improved the depth accuracy by using the estimated optical flow from video frames to guide the network training. Schreiberhuber et al. [30] proposed to combine a CNN-based backbone with a neural decision tree with regressors at the leaf nodes to improved the depth estimation accuracy. Li et al. [31] proposed a single-shot system consisting of a RGB camera and a projector to simultaneously acquire scene depth and spectral reflectance. These methods estimate depth by applying stereo matching to the image captured by a single camera and the reference pattern, and have a simpler and more cost-effective setup than ours. However, they suffer from reduced accuracy in regions where the pattern is not visible due to specular or transparent surfaces. Xu et al. [32] built a stereo system which consists of an RGB camera, an IR camera and an IR speckle projector. They first compute the initial depth map by matching the IR camera image and the IR pattern. Then a stereo network refines the initial depth using the IR camera image and the RGB camera image as input. Because the RGB camera receives lights across a wide spectrum, the accuracy of depth estimation is influenced by the environmental illumination. Warbug et al. [33] proposed a depth completion network using 3D landmarks from the visual-inertia SLAM as the sparse depth supervision. However, this method cannot be used for single-frame depth estimation, as it requires a sparse depth map from visual-inertial SLAM as input during inference.

Domain Adaptation: techniques have shown great promise in closing the gap between the simulation and real domains. Tobin et al. [34] proposed using domain randomization through randomizing rendering in the simulator to train a robust model that would interpret the real domain as just another variation of the simulation domain. Previous works have also tried aligning the source and target domains by matching their input distributions or their feature statistics [35], [36]. GraftNet [37] transforms the image feature pretrained on ImageNet [38] into a broad-spectrum and task-oriented feature to improve the stereo network's domain generalizability. Other works have attempted to learn domain-invariant representations by augmenting the input based on certain criterion set forth in the task and approach itself [39]. Moreover, unsupervised losses have seen increased use for domain adaptation in tasks such as semantic segmentation and object detection [40], [41], [42].

Our work is most related to StereoGAN [13], which uses groundtruth depth maps in the simulated domain and reprojection loss in the real domain along with unsupervised GAN losses in order to close the domain gap between simulation and real images. Our work differs from theirs in three key ways: 1) we utilize IR images with actively projected patterns for stereo matching instead of passive RGB images, which leads to a smaller sim2real gap and better transferability; 2) we use the proposed temporal IR reprojection loss as self-supervision which is more effective in correlating local matching features; 3) we train using only shape primitives and random real objects that are out-of-distribution from test time data.

Depth Completion: refers to the task of filling the depth information of incomplete depth maps. Recently learning based methods have shown great progress on this task [15], [43], [44], [45], [46]. Zhang et al. [15] used CNNs to predict transparent object masks, boundaries and surface normals from RGB images, and completed the depth through optimization. Tang et al. [44] used GAN where the generator predicts the missing depth values and the discriminator helps the training process. Xu et al. [45] generate the complete depth maps by first generating a point cloud distribution, then projecting the generated point cloud to a depth map and completing the depth map using 2D CNNs. However, these works all require a given or predicted transparent object mask. For objects of multiple materials, only some regions have large depth errors and other regions can be measured accurately. Ignoring all the depth information of the objects is sub-optimal. From our experimental results, we surprisingly find that the depth prediction error is highly related to the confidence map from the stereo network. Therefore, we only remove the depth prediction of low-confidence regions and perform the depth completion through normal-depth consistency, leading to better performance.

III. Mixed Domain Learning Stereo Network

In this section, we introduce *mixed domain learning:* for active stereovision. We first define the task setup: in real domain $\mathcal{X}$, we have a target set of real IR stereo images with projected pattern $\mathbb{X}^t = \{(\mathbf{X}_l^t, \mathbf{X}_r^t)_i\}_{i=1}^N$, and our goal is to learn an accurate disparity estimation network F to estimate the disparity $\hat{\mathbf{X}}_d^t = F(\mathbf{X}_l^t, \mathbf{X}_r^t)$. We utilize *mixed domain* data to train the network: in real domain $\mathcal{X}$ we collect another set of real IR stereo images $\mathbb{X} = \{(\mathbf{X}_l, \mathbf{X}_r)_i\}_{i=1}^M$ without disparity annotation. To be clear, the objects appearing in $\mathbb{X}$ are different from the ones in $\mathbb{X}^t$. In simulation domain $\mathcal{Y}$, we generate a set of synthetic IR stereo images with groundtruth disparity annotation $\mathbb{Y} = \{(\mathbf{Y}_l, \mathbf{Y}_r, \mathbf{Y}_d)_i\}_{i=1}^K$. In order to guarantee the generalizability of the trained network to unseen objects, we only use shape primitives (sphere, cube, capsule) with different scales, textures and materials to generate $\mathbb{Y}$.

Fig. 1 shows the framework of our proposed method. In the real domain, we propose the use of temporal binary IR reprojection loss as self-supervision (Section III-A). In the simulation domain, we use the loss between predicted disparity and the groundtruth disparity y_d as supervision (Section III-B). The network is trained jointly using the self-supervision in real domain and supervision in simulation domain (Section III-C). The stereo network architecture and other implementation details are introduced in Section III-D.

A. Real Domain: Self-Supervised Learning on IR Images

Conventional self-supervised learning methods [11], [26] rely on grayscale images for reprojection loss computation, which are

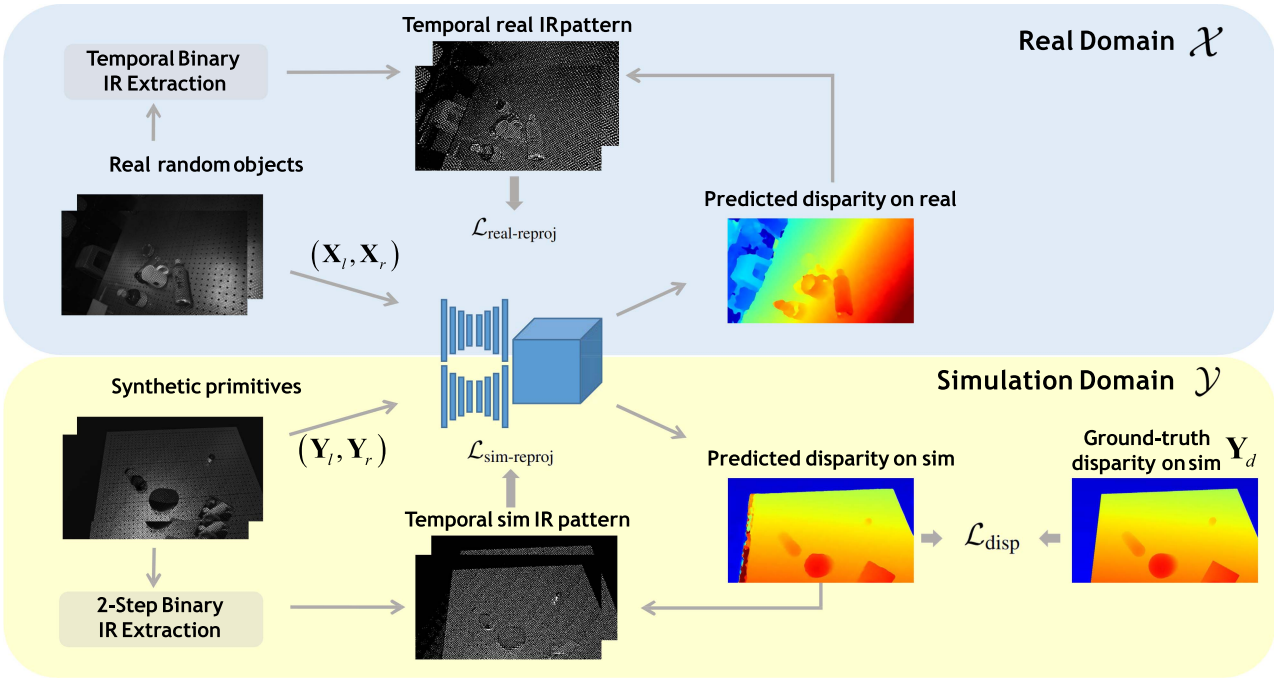


Fig. 1. Mixed domain learning stereo network architecture overview. The simulated and real stereo IR images are fed to a shared weight stereo network consisting of a CNN for noise reduction and a cost-volume-based 3D CNN for disparity prediction. The network is trained with reprojection loss on temporal binary IR pattern in the real domain, reprojection loss and disparity loss in the simulation domain as mixed domain learning.

susceptible to various factors that influence the pixel intensities, including the environmental illumination, the actively projected pattern, the material property of the object, and the viewpoint. The prerequisite for these methods is that the object surface is Lambertian diffused where the reflection intensity is invariant to the viewpoint, which is usually not satisfied in real world. Therefore, we propose to extract the binary projected active pattern from temporal IR stereo image sequences, which eliminates the adverse effect of these factors while maintaining the most important components of active pattern. Then, we construct the reprojection loss on this new binary pattern.

1) Binary Pattern Extraction From Temporal IR Images: For the real captured IR images $\mathbf{X}_l$ and $\mathbf{X}_r$, the grayscale at pixel (u, v) is

$$\begin{aligned}\mathbf{X}_l(u,v) &= \mathbf{I}_l(u,v) + e * \mathbf{R}_l(u,v) * \mathbf{K}_l(u,v) + \varepsilon_l \\ \mathbf{X}_r(u,v) &= \mathbf{I}_r(u,v) + e * \mathbf{R}_r(u,v) * \mathbf{K}_r(u,v) + \varepsilon_r \end{aligned} \quad (1)$$

where $\mathbf{I}_l, \mathbf{I}_r$ represent the environmental illumination intensity, $\mathbf{K}_l, \mathbf{K}_r$ represent the binary pattern captured by the camera, $\mathbf{R}_l, \mathbf{R}_r$ represent the reflection coefficient determined by the object surface material, texture, angle and distance, e represents the pattern emittance, and $\varepsilon_l, \varepsilon_r$ represent the camera sensor noise. To simplify the notation, we denote the variable for both cameras without the subscript l or r.

For active stereovision depth sensors, we manually adjust the pattern emittance e by changing the emitter power. Therefore, as shown in Fig. 2, our pattern extraction procedure is as follows: we set e to $\{e_0, e_1, \ldots, e_n\}$, capture a temporal sequence of corresponding IR images $\{\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}\}$, and fit the sequence to the linear model, regressed and obtain $\{\widetilde{\mathbf{X}}^{(0)}, \ldots, \widetilde{\mathbf{X}}^{(n)}\}$. Because the environmental illumination can be assumed to be constant through the short capture process, we can use linear fitting to eliminate its effect. We extract the binary IR pattern $\mathbf{K}$ from the temporal image sequence through local window normalization and binarization

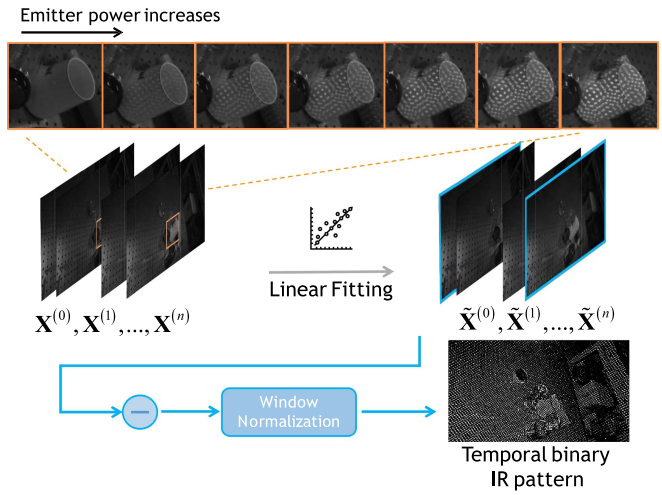


Fig. 2. Temporal binary pattern extraction.

$$\mathbf{K}(u,v) = \begin{cases} 1 & \text{if} \|\widetilde{\mathbf{X}}^{(n)}(u,v) - \widetilde{\mathbf{X}}(u,v)\| > \delta + \eta \\ 0 & \text{otherwise} \end{cases}$$

$$\delta = \frac{1}{w^2} \sum \|\mathbf{W}(\widetilde{\mathbf{X}}^{(n)}, u, v) - \mathbf{W}(\widetilde{\mathbf{X}}^{(0)}, u, v)\| \quad (2)$$

where $\mathbf{W}(\mathbf{X}, u, v)$ is a local window centered at pixel (u, v) in $\mathbf{X}$ with window size w, η is a threshold to filter out noise and areas

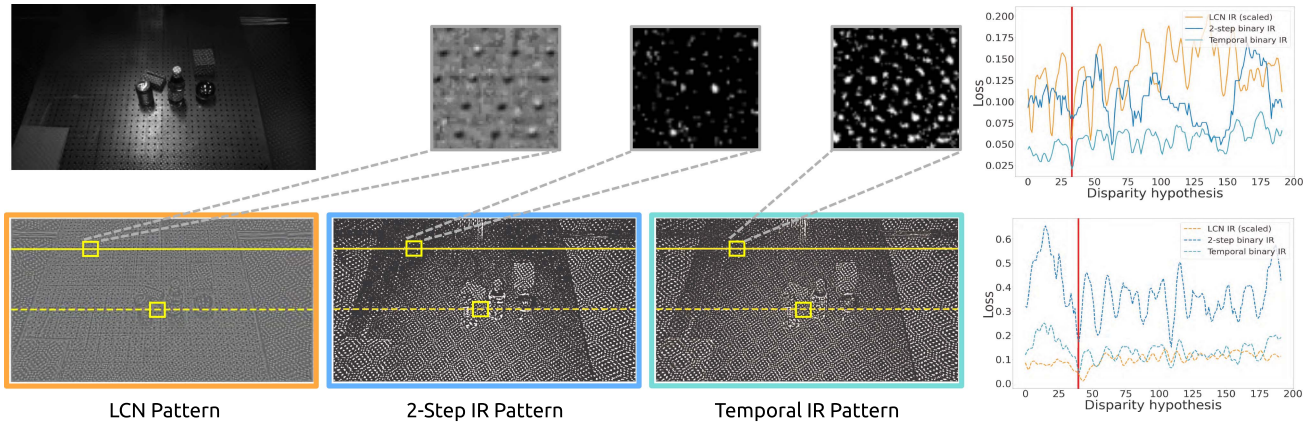


Fig. 3. Comparison of extracted pattern and reprojection loss along the epipolar line. LCN pattern represents local contrast normalization [11] which consists of continuous values; 2-step IR pattern and temporal IR pattern represent the extracted binarized pattern from temporal IR image sequence using $n = 1$ and $n = 6$, respectively.

where the reflection coefficient is extremely small such as pure specular reflection regions. In our work, we use $n = 6$. For non-Lambertian surfaces, although the intensity or grayscale of the projected pattern changes with the viewpoint, the relationship of pixel intensities of "on" pixels and "off" pixels (i.e. whether the pixel is illuminated by the pattern) in the local neighborhood remains constant. Thus, the binary IR pattern computed by the proposed method is able to achieve viewpoint invariance.

In Fig. 3, we compare the pattern extracted by different methods. By utilizing the temporal image sequence, our method is able to extract the pattern accurately and completely even in distant areas where the SNR (signal noise ratio) is low. The local normalization and binarization window filters out camera sensor noise and environmental illumination while retaining the projected pattern, which is beneficial for further reprojection loss computation.

2) Binary Pattern Reprojection Loss: As demonstrated in traditional stereo matching and active stereo methods [11], [47], [48], [49], patch-wise reprojection losses are smoother and more accurate than pixel-wise losses and are beneficial for matching. Therefore, we construct the patch-wise reprojection loss on the binary IR pattern $(\mathbf{K}_l, \mathbf{K}_r)$ computed in (2) and the disparity prediction $\mathbf{X}_d$

$$\mathcal{L}_{\text{reproj}}\left(\mathbf{K}_l, \mathbf{K}_r, \hat{\mathbf{X}}_d\right) = \sum_{u,v} \frac{1}{(2p+1)^2} \mathbf{C}(u,v)$$

$$\mathbf{C}(u,v) = \sum_{(u_p,v_p)\in P(u,v)} ||\mathbf{K}_l(u_p, v_p) - \mathbf{K}_r(u_p - \hat{\mathbf{X}}_d(u_p, v_p), v_p)||^2 \quad (3)$$

where $P(u,v)$ represents the patch centered at pixel (u,v) with patch size $(2p+1) \times (2p+1)$.

As shown in Fig. 3, since the temporal binary IR pattern eliminates the influence of object texture and environmental illumination and only retains the projected pattern, the reprojection loss computed on the temporal binary IR pattern reaches global minima at the groundtruth disparity while the losses computed on the other two patterns could be misleading for the stereo network training.

B. Simulation Domain: Supervised Learning on Shape Primitives

Although the proposed temporal IR reprojection loss can be used as the sole loss for stereo network training, it still has some limitations: the binary IR pattern cannot be extracted accurately for translucent and transparent objects and there are local minima in the loss with respect to the disparity hypotheses. On the other hand, traditional supervised learning with groundtruth depth does not suffer from the aforementioned issues. However, it is costly and time-consuming to acquire groundtruth depth in real world settings. Thus, we perform supervised learning only in the simulation domain.

1) Dataset Generation Based on Ray-Tracing: In the last decade, there has been significant progress in ray-tracing rendering techniques in terms of speed and quality. Compared with rasterization, ray-tracing rendering can simulate the light transmission process on translucent and transparent objects [50] more accurately. Therefore, we use ray-tracing rendering to generate the simulated training dataset: we first build a cone lighting with mask to imitate the pattern emitter in the real active stereovision depth sensor, and then construct two cameras similar to stereo cameras in the real setting. The relative position between cameras and lighting are set using parameters from real sensors. We also add dim ambient light in the simulation environment to imitate the filtered environmental light in the real setting.

2) Shape Primitives: In order to make the trained stereo network generalizable to varied unseen scenarios, we only use base shape primitives, including cube, sphere and capsule, for simulated dataset generation. We use images from tiny ImageNet [51] as object textures and transmissivity masks. The number of primitives is randomly sampled from 25 to 50. The sizes, layouts and materials are also randomly generated.

3) Domain Randomization: In our pervious CVPR work, a standard online data augmentation pipeline is employed, which

applies color jitter and Gaussian blur to the rendered input images on-the-fly during training. However, these augmentations are global operations that do not account for the image variations caused by changes in lighting conditions, such as shadows and specular highlights of the projected pattern. Therefore, we enhance our data augmentation by using domain randomization techniques during the image rendering stage. We modify the domain randomization technique to suit our application scenario. It includes random rotation and scaling of the IR pattern and random placement and intensity of light sources. The additional data randomization increases the rendered image diversity and improves the stereo network's Sim2Real generalizability.

4) Disparity Loss: Given the synthetic stereo image pair with ground-truth disparity $(\mathbf{Y}_l, \mathbf{Y}_r, \mathbf{Y}_d)$, we follow [7] and adopt smooth L_1 loss between $\mathbf{Y}_d$ and the predicted disparity on synthetic stereo images

$$\mathcal{L}_{\text{disp}} = L_{1\text{smooth}}(F(\mathbf{Y}_l, \mathbf{Y}_r), \mathbf{Y}_d) \tag{4}$$

where

$$L_{1\text{smooth}}(x, y) = \begin{cases} 0.5(x-y)^2 & \text{if } |x-y| < 1 \\ |x-y| - 0.5 & \text{otherwise} \end{cases} \tag{5}$$

C. Mixed Domain Learning

Given the real stereo IR image$(\mathbf{X}_l, \mathbf{X}_r)$, and the simulated stereo IR image with groundtruth disparity $(\mathbf{Y}_l, \mathbf{Y}_r, \mathbf{Y}_d)$, we train the stereo network $F(\cdot, \cdot)$ by combining the reprojection loss in the real domain and the disparity loss along with reprojection loss in the simulation domain

$$\begin{aligned} &\mathcal{L}\left(\mathbf{X}_l, \mathbf{X}_r, \mathbf{Y}_l, \mathbf{Y}_r, \mathbf{Y}_d\right) \\ &\quad = \lambda_r \cdot \mathcal{L}_{\text{real-reproj}}\left(\mathbf{X}_l, \mathbf{X}_r, F(\mathbf{X}_l, \mathbf{X}_r)\right) \\ &\qquad + \lambda_s \cdot \mathcal{L}_{\text{sim-reproj}}\left(\mathbf{Y}_l, \mathbf{Y}_r, F(\mathbf{Y}_l, \mathbf{Y}_r)\right) \\ &\qquad + \lambda_{gt} \cdot \mathcal{L}_{\text{disp}}(F(\mathbf{Y}_l, \mathbf{Y}_r), \mathbf{Y}_d) \end{aligned} \tag{6}$$

where λ_r, λ_s and λ_{gt} represent the weights of the real domain reprojection loss, the simulation domain reprojection loss, and the simulation domain groundtruth disparity loss, respectively.

The loss terms on real domain guarantee transferrability to unseen real data. However, we find that it is quite hard to train the network using these terms alone, due to noise in the self-supervision signals. Interestingly enough, after adding the supervised loss terms in simulation domain on primitive shapes, the behavior of loss minimization is much more tame: not only does the network converge faster, but also the final solution has better quality.

D. Implementation Details

For the stereo network, we adopt PSMNet [7] as the backbone, which aggregates image features at different scales, constructs a cost volume and uses 3D CNNs to regress the disparity. We make a few changes to the original version of PSMNet to improve the performance: the disparity range is set to be from 12 to 96 and the number of disparity hypotheses is set to be 192 so that it can achieve sub-pixel accuracy; we adopt dilated convolutions [52] to enlarge the receptive fields.

IV. Confidence-Based Depth Completion

Since the light transport on translucent and transparent objects are too complex for realistic rendering in the simulation domain and the temporal IR pattern extraction procedure is invalid for such objects in the real domain, the mixed domain learning stereo network's performance on these objects is inferior to its performance on Lambertian objects. In [15], the depth of transparent objects is completed through depth-normal consistency, where the transparent objects are first segmented from the RGB image, the normal map is estimated from the RGB image, and the depth is completed by global depth-normal consistency optimization. However, the segmentation module requires a large-scale dataset with manual annotation, and its performance on unseen objects is limited. Moreover, it cannot handle objects composed of heterogeneous materials, which are prevalent in real life and may exhibit large depth errors only in certain regions. We adopt the pipeline and make a few critical modifications to make it suitable for learning-based active stereovision. The pipeline overview is shown in Fig. 4.

A. Confidence Map

In cost-volume-based stereo networks [7], [8], [53], 3D CNNs are usually used to regularize the feature cost volume into a probability volume, and the depth map is predicted as the weighted sum along the depth direction. As shown in Fig. 5, only a portion of the transparent object contains erroneous depth prediction while other parts are reconstructed accurately. For these erroneous pixels, the probability distributions are not concentrated. Following [8], the confidence map $\mathbf{H}$ is computed as the sum over the probability of the $2k$ nearest disparity hypotheses to the disparity estimation. We set $k = 2$. In [8], because they only focus on the reconstruction of diffused objects, the confidence map is only used as background filtering for multi-view depth fusion. In this work, we propose to use the confidence map to remove inaccurate regions and complete these regions using depth-normal consistency. By doing so, we avoid the requirement of transparent object mask annotation and are able to generalize to objects of different materials. Moreover, it can retain the portions of transparent objects that are measured accurately, improving the reconstruction quality.

B. Mixed Domain Stereo Normal Estimation

Because the depth completion is based on depth-normal consistency, the reconstruction quality is mainly determined by the normal estimation accuracy. Existing depth completion works mainly estimate the normal map from the single view RGB image [15], [43]. However, monocular normal estimation is an ill-posed problem, leading to inferior generalizability. In this article, we develop a cost-volume-based stereo normal estimation network $N(\cdot, \cdot)$. As shown in Fig. 6, the network architecture is similar to PSMNet [7], except that we output 3 values for each voxel instead of 1. Following [54], we add the output along the depth dimension and normalize the sum to generate the normal estimation. The stereo normal estimation can be considered as expectation of predictions conditioned on the disparity hypotheses.

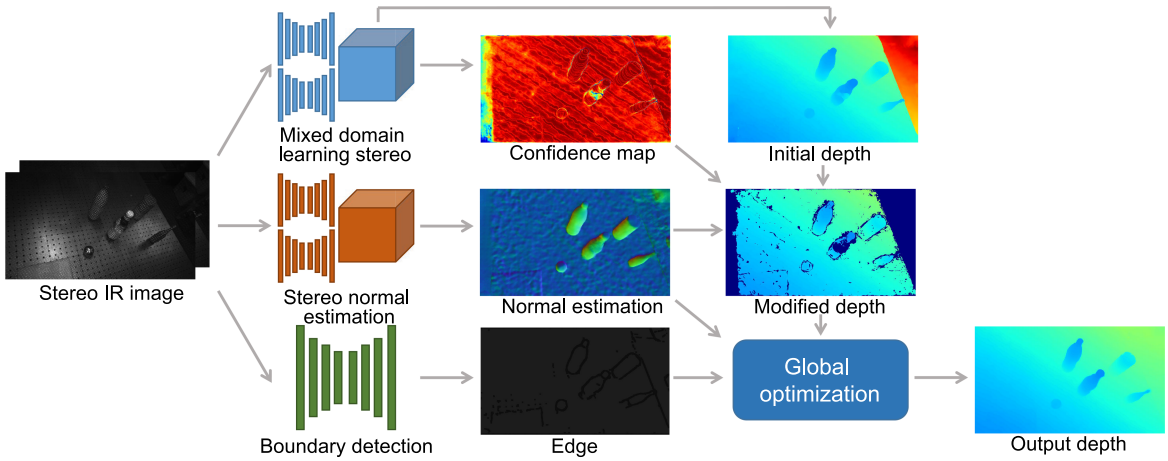


Fig. 4. Overview of the confidence-based depth completion pipeline. The input is a pair of stereo IR images in the real world. We first use three networks to infer the confidence map, initial depth, normal estimation, and edges. Then, we identify uncertain and erroneous regions and remove them. Finally, we use global depth-normal consistency optimization to generate the completed depth map.

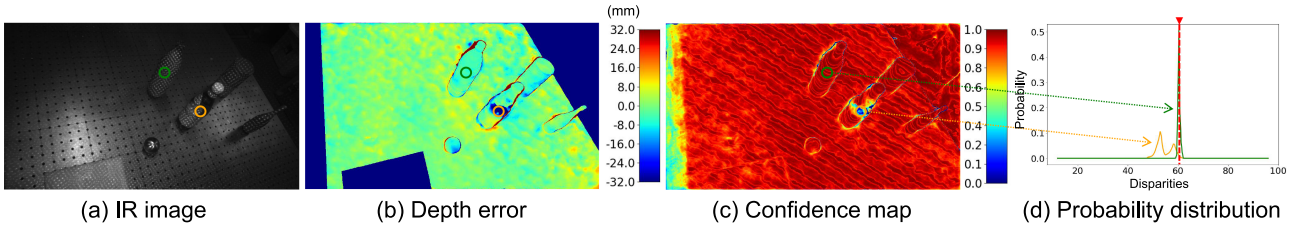


Fig. 5. Illustration of real IR image, depth error, confidence map, and probability distributions. The red dot line represents the groundtruth disparity for the two pixels.

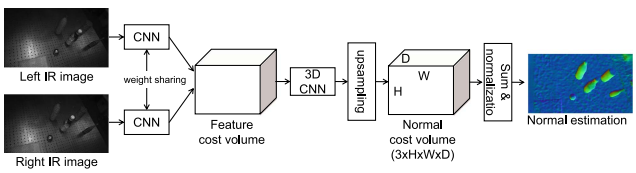


Fig. 6. Stereo normal estimation network architecture.

In order to take the advantage of both simulation domain and real domain, we use mixed domain learning for the stereo normal estimation network training. Since we do not have the groundtruth normal for the real domain data beforehand, we first train the mixed domain learning stereo network F, use F to generate the estimated depth for $\mathbb{X}$, compute the normal map using covariance analysis, and train the stereo normal estimation network on $\mathbb{X}$ and $\mathbb{Y}$ jointly. Because F may generate inaccurate depth for challenging regions in $\mathbb{X}$, we use the confidence-aware normal loss

$$\begin{aligned}\mathcal{L}_{\text{normal}} = \lambda_{nr} \cdot \mathbf{H}^c \cdot \mathcal{L}_{\cos}\left(N\left(\mathbf{X}_l, \mathbf{X}_r\right)\right) \\ + \lambda_{ns} \cdot \mathcal{L}_{\cos}\left(N\left(\mathbf{Y}_l, \mathbf{Y}_r\right)\right)\end{aligned} \tag{7}$$

where $\mathcal{L}_{\cos}$ is the cosine similarity loss, λ_{nr} and λ_{ns} represent the weights for the real domain and the simulation domain, $\mathbf{H}$ is the confidence map generated as described in Section IV-A, the confidence coefficient c is the exponent of $\mathbf{H}$ and is a non-negative integer.

C. Boundary Detection

The boundary, where the depth is discontinuous, is required for the following global depth-normal consistency optimization. Different from disparity and normal estimation, we only use the left IR image for boundary detection. The network consists of a sequence of multi-scale CNNs. The network is trained only in the simulation domain, where the groundtruth boundary mask can be generated from the simulated depth using edge detection. We find it achieves good performance on real IR images. We use a weighted binary cross-entropy loss to balance the positive and negative pixels.

D. Global Depth-Normal Consistency Optimization

Given the real IR stereo images $(\mathbf{X}_l^t, \mathbf{X}_r^t)$ in the target set, we generate the depth prediction $F(\mathbf{X}_l^t, \mathbf{X}_r^t)$ with the confidence map $\mathbf{H}$, the stereo normal estimation $N(\mathbf{X}_l^t, \mathbf{X}_r^t)$, and the boundary detection $\mathbf{G}$. And we perform the confidence-based depth completion using the global depth-normal consistency optimization proposed in [43].

We first compute the normal map $\widehat{\mathbf{N}}$ from the depth prediction $F(\mathbf{X}_l^t, \mathbf{X}_r^t)$. Then we compute the uncertain mask $\mathbf{U}$

$$\mathbf{U}(\mathbf{p}) = \begin{cases} 1 & \text{if } \mathbf{H}(\mathbf{p}) < \mu_c \\ & \text{or } || < \widehat{\mathbf{N}}(\mathbf{p}), N(\mathbf{X}_l^t, \mathbf{X}_r^t)(\mathbf{p}) > || < \cos\mu_n \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

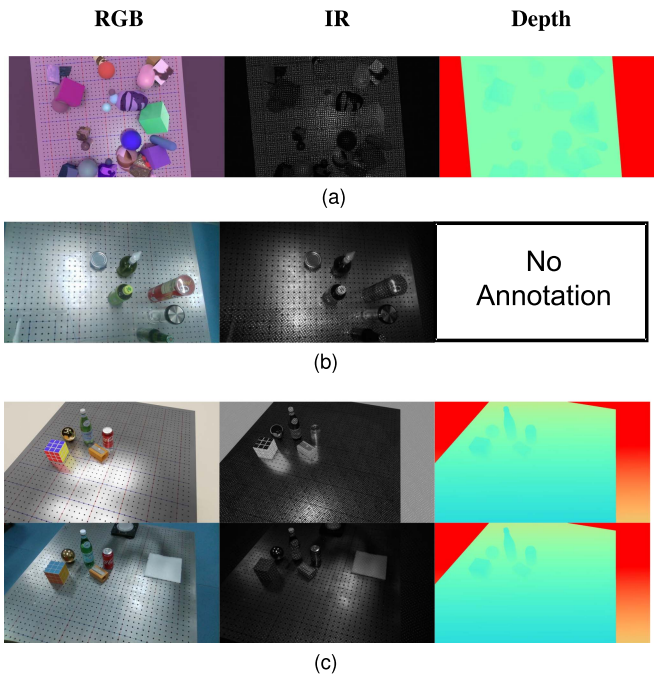


Fig. 7. Example images from our dataset. (a) the simulation training dataset of random shape primitives; (b) the real training dataset of random objects different from testing; (c) the sim2real aligned testing dataset, including specular surfaces such as metals and translucent bodies such as liquids. Note: we don't rely on any annotation for real scenes which is why we have no depth annotation in (b).

where $\mathbf{p}$ is the image pixel coordinate, μ_c is the confidence threshold, μ_n is the normal difference threshold. We remove the depth and normal of uncertain pixels and generate a dense depth map $\widehat{\mathbf{D}}$ by solving the optimization problem

$$\widehat{\mathbf{D}} = \underset{\mathbf{D}}{\operatorname{argmin}} \lambda_D E_D + \lambda_S E_S + \lambda_N E_N$$

$$E_D = \sum_{\{\mathbf{p}|\mathbf{U}(\mathbf{p})=0\}} ||\widehat{\mathbf{D}}(\mathbf{p}) - F(\mathbf{X}_l^t, \mathbf{X}_r^t)(\mathbf{p})||^2$$

$$E_S = \sum_{\mathbf{p}} \sum_{\mathbf{q}\in Neigh(\mathbf{p})} ||\mathbf{D}(\mathbf{q}) - \mathbf{D}(\mathbf{p})||^2$$

$$E_N = \sum_{\mathbf{p}} \sum_{\mathbf{q}\in Neigh(\mathbf{p})} NE(\mathbf{p}, \mathbf{q})$$

$$NE(\mathbf{p}, \mathbf{q}) = || < v(\mathbf{p}, \mathbf{q}), N(\mathbf{X}_l^t, \mathbf{X}_r^t)(\mathbf{p}) > \cdot \mathbf{G}(\mathbf{p})||^2 \quad (9)$$

where $\mathbf{p}, \mathbf{q}$ are image pixel coordinates, E_D measures the difference between $\mathbf{D}$ and $F(\mathbf{X}_l^t, \mathbf{X}_r^t)$, E_S measures the depth smoothness, and E_N measures the consistency between the depth map and the normal estimation. Finally, we fuse $\widehat{\mathbf{D}}$ and $F(\mathbf{X}_l^t, \mathbf{X}_r^t)$ using the uncertain mask $\mathbf{U}$ to generate the final depth prediction $\mathbf{D}^*$

$$\mathbf{D}^* = \mathbf{U} \cdot \widehat{\mathbf{D}} + (\mathbf{1} - \mathbf{U}) \cdot F(\mathbf{X}_l^t, \mathbf{X}_r^t) \quad (10)$$

V. Experiments

A. Dataset

Fig. 7 shows example images from the three datasets in our work. For the testing dataset, we used an Intel RealSense D415 as the active stereovision depth sensor. All the real RGB and IR images are captured using the RealSense camera. In order to quantitatively evaluate the performance of the camera, the complete and accurate groundtruth depth is required. To do so, we constructed a set of simulated scenes which are precisely aligned with the real ones. The construction process is as follows:

1) retrieve the intrinsic matrices of the depth sensor from its firmware;
2) calibrate the relative transformations among the depth sensor, the robot, and the table;
3) use the transformations and intrinsic matrices to construct and maintain a consistent simulation environment across all scenes;
4) employ physical simulation to generate physically plausible object layouts;
5) render synthetic images based on the object layouts;
6) generate mixed images in real time by overlaying rendered images on captured images by the depth sensor;
7) align real object poses with simulated ones using the mixed image as online feedback.

We measure the real-world object's pose using a motion capture system. It shows that the object pose alignment error is below 2 mm, 1.5°, and the depth error is less than 2.5 mm, which is sufficient for depth prediction evaluation.

To evaluate the influence of object material on depth estimation performance, we include two categories of objects: 3D-printed objects and real objects. The 3D-printed objects are printed using color plaster powder, and are considered Lambertian diffused, while the real objects' material are complex (*specular, translucent, transparent*) and difficult for active stereovision depth sensors. Overall, the testing dataset consists of 504 stereo images of 24 different scenes.

For the training dataset in the simulation domain, we rendered 20,000 stereo IR images with ground-truth disparity annotation using random shape primitives, including spheres, cubes and capsules. 10% of the primitives are set to be transparent with a random binary mask, 50% are textured by images from tiny imagenet [51], and the rest are set to random colors. In order to make the scene more complicated, the primitives can overlap with each other and are not strictly attached to the Table Therefore, they can either overlap with the table or float above the table. The IR pattern used in the simulation is obtained as follows: we position an Intel RealSense D415 in front of a white planar surface and obtain the image captured by the left IR camera; then we extract the binary IR pattern from the captured image using adaptive thresholding in OpenCV. Customized domain randomization as described in Section III-B3 is applied during data generation. For the ray-tracing rendering, the number of samples per pixel is 128 and the max bounces is set to 8. The rendered IR images are post-processed by the NVIDIA OptiX denoiser [55].

For the training dataset in the real domain, we collected 1,047 real stereo IR images of random objects which are different from the testing dataset. To preserve its generalizability, the optical properties of the objects are diversely selected. The objects are randomly placed on the table, and captured by the same RealSense from different viewpoints. We only use the real IR

TABLE I
PERFORMANCE OF LEARNING-BASED STEREO, COMMERCIAL DEPTH SENSOR AND OUR METHOD ON THE REAL TESTING DATASET

Excluding uncertain pixels									
Training Data	Method	EPE (px) ↓	Bad 1 ↓	ADE (mm) ↓			> 8mm↓		
		All	All	All	Printed	Real	All	Printed	Real
Sim	PSMNet	2.249	0.523	18.00	8.80	10.96	0.604	0.364	0.471
	PSMNet + DR	1.231	0.278	11.12	8.59	11.44	0.370	0.358	0.491
	PSMNet + AP + DR	0.370	0.053	5.08	5.72	**9.83**	0.395	0.180	**0.444**
Sim+Real	PSMNet + StereoGAN [13]	1.293	0.238	11.46	12.04	17.44	0.517	0.517	0.743
Sim+Real(ImageNet)	GraftNet [37]	0.712	0.183	9.24	7.18	11.86	0.364	0.300	0.557
-	RealSense D415	0.348	0.029	4.99	6.33	12.73	0.170	0.251	0.577
Sim+Real	Ours	**0.249**	**0.024**	**3.46**	**5.26**	10.29	**0.074**	**0.162**	0.472
Sim+Real	Ours + DC	0.247	0.023	3.45	5.48	10.75	0.070	0.174	0.490
Including uncertain pixels									
Training Data	Method	EPE (px)↓	Bad 1↓	ADE (mm)↓			> 8mm ↓		
		All	All	All	Printed	Real	All	Printed	Real
Sim	PSMNet	2.257	0.523	18.00	9.12	11.34	0.605	0.375	0.483
	PSMNet + DR	1.253	0.283	11.22	8.91	11.84	0.375	0.369	0.504
	PSMNet + AP + DR	0.402	0.060	5.26	6.14	**10.44**	0.178	0.197	**0.467**
Sim+Real	PSMNet + StereoGAN [13]	1.376	0.250	11.74	12.47	18.04	0.517	0.517	0.743
Sim+Real(ImageNet)	GraftNet [37]	0.752	0.191	9.42	7.55	12.42	0.370	0.314	0.577
-	RealSense D415	0.572	0.058	5.78	7.21	14.47	0.195	0.278	0.620
Sim+Real	Ours	**0.282**	**0.031**	**3.67**	**5.64**	10.90	**0.083**	**0.177**	0.494
Sim+Real	Ours +DC	0.272	0.028	3.61	5.89	11.38	0.078	0.190	0.514

AP: active pattern. DR: domain randomization. DC: depth completion.

stereo images to construct the temporal IR reprojection loss, and the depth images are not collected. Note that the temporal IR image sequences are only collected offline for the training dataset in the real domain. It takes about 0.7 s to capture one image sequence for the temporal IR pattern computation. For inference, we employ only a single-frame image acquired by each of the left and right IR cameras, and the temporal IR pattern is NOT required.

B. Mixed Domain Learning Stereo Network

In this section, we first compare the proposed mixed domain learning stereo network with other learning-based methods and a decent commercial depth sensor - the RealSense D415. Note that the confidence-based depth completion is not applied in this section.

1) Training Details: We train the network using the Adam optimizer with the initial learning rate set to 2e–4, decaying by half every 10 k iterations for a total of 60 k iterations. The network is trained on 2 GPUs each with 11 GB GPU memory and a batch size of 4.

For fair comparison, data augmentation is applied to both our method and baseline methods. Specifically, brightness and contrast is uniformly scaled by a value between 0.4 to 1.4, and 0.8 to 1.2 respectively. For Gaussian blur, kernel size is fixed to 9×9 and the standard deviation is selected uniformly between 0.1 to 2.

2) Evaluation Metrics: Several common stereo estimation metrics are used to evaluate the proposed method. End-point-error (*EPE*) is the mean absolute disparity error. *Bad 1* is the percentage of pixels with disparity errors larger than 1 pixel. By converting disparity to depth, we also measure the average absolute depth error (*ADE*) and the percentage of depth outliers with absolute error larger than 8 mm, which is denoted as >8 mm. To evaluate the performance of our model on objects of different materials, these depth metrics are measured separately on two kinds of objects in the testing dataset using object masks. In order to reduce the impact of outliers on the average, we clip the disparity and depth errors at 8 pixels and 32 mm respectively, which makes the numbers in the tables different from our CVPR version. Since the RealSense camera outputs a value of zero at areas with high depth uncertainty, metrics are computed in terms of excluding and including uncertain pixels so that the evaluation is in the same completeness level.

3) Comparison Results: Table I shows the comparison result. We first build three baseline methods: PSMNet, PSMNet+DR, PSMNet+AP+DR. The three methods use the same network architecture as our method, but we only train them on the simulation data and test them on the real data directly. By using the customized domain randomization, the performances on most metrics are improved significantly, which demonstrates the effectiveness of changing the lighting condition during simulation data generation. By adding the active pattern, the performances are further improved, showing that the active pattern can help stereo matching and eliminate the Sim2Real domain gap. This intuitively makes sense since active light adds pattern to textureless areas which are the most difficult to match. Our method is slightly worse than PSMNet+AP+DR on the real objects. The probable reason is that the temporal IR reprojection loss may be invalid for challenging regions in the real domain. On the other hand, it also validates the importance of simulation data for learning-based stereo where the groundtruth depth is available for shapes of different textures and materials.

We extend the StereoGAN [13] architecture by using PSMNet as the disparity prediction backbone, which is denoted as PSMNet+StereoGAN. This improved StereoGAN uses cost volume aggregation in its stereo matching module, which makes it more powerful and comparable with our method. When compared with our method, PSMNet+StereoGAN performs considerably worse as the absolute depth error increases from 3.67 mm

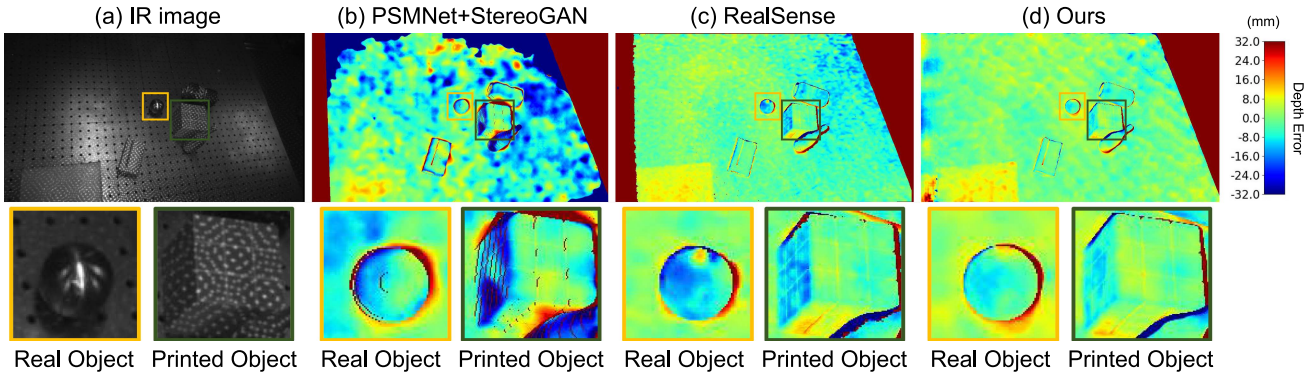


Fig. 8. Comparison of the disparity error map of our method with PSMNet+StereoGAN and RealSense D415. Our method improves disparity accuracy on both 3D-printed objects and real objects.

to 11.74 mm. This is further corroborated by Fig. 8, where PSMNet+StereoGAN struggles to predict depth on real objects such as the metal ball, which is a specular surface. On the other hand, our mixed domain learning method has improved accuracy on these types of objects. This large performance improvement can be attributed to direct supervision in the simulation domain of primitives with random shapes and materials, a well-shaped temporal IR reprojection which accurately locates the correct correspondences, and a more robust pipeline overall since it doesn't use the GAN module. To adapt GraftNet [37] to our dataset, we graft the feature extractor pretrained on ImageNet to the synthetic dataset. GraftNet surpasses PSMNet + StereoGAN and even achieves better results than RealSense for transparent objects. However, the large domain gap between the real IR images and the ImageNet images limits the performance of GraftNet.

To the best of our knowledge, we are the first work to be quantitatively compared with commercial products. The Intel RealSense D415 uses a traditional CENSUS-based stereo matching method [4], [56], which has high computation efficiency but will leave uncertain pixels without depth values. Therefore, we report our results on the same completeness levels as RealSense and demonstrate that our method outperforms RealSense in every metric. In Fig. 8, because the active pattern is not reflected, RealSense is unable to accurately predict pixels in specular areas, while our method is able to match those pixels well. In addition, for 3D-printed objects, our model also demonstrates lower depth error.

4) Ablation Study: In this section, we validate the effectiveness of each component and design choice through ablation experiments.

Reprojection loss: We compare the network's performance when doing reprojection on different patterns which is shown in Table II. Raw IR simply computes the patch-wise Mean Squared Error (MSE) of the warped raw IR images. LCN IR is from ActiveStereoNet [11], which uses an LCN module to alleviate the condition where two matched pixels have large residuals due to the distance from the camera and the physical properties of the surface. For the sake of fairness, we add synthetic groundtruth depth supervision to all of the experiments above. The Raw IR reprojection has the worst result because it doesn't take into account the different intensities of IR light of two matched pixels. While LCN IR helps address this issue, it employs reprojection on the continuous local normalized grayscale IR image, which is still affected by environmental illumination and object texture. To tackle this issue, we proposed a reprojection loss on 2-Step IR patterns which shows better performance since the binary pattern eliminates the small residual of two matched pixels. Lastly, since the SNR is low for pixels that are far away from the camera, 2-Step IR cannot properly extract the active light pattern in distant areas. Temporal IR patterns addresses this issue by tracking the intensity difference in the temporal image sequence, which enables it to extract a more accurate and complete IR pattern. The results prove that our reprojection on temporal IR images is superior to all other reprojection methods.

TABLE II
COMPARISON OF REPROJECTION LOSS ON DIFFERENT PATTERNS IN THE REAL DOMAIN

Pattern	ADE (mm) ↓	> 8mm ↓
Raw IR	4.30	0.122
LCN IR [11]	3.95	0.096
2-Step IR	3.90	0.092
Temporal IR	**3.67**	**0.083**

TABLE III
COMPARISON OF DISPARITY SUPERVISION IN THE SIMULATION DOMAIN WITH DIFFERENT SELF-SUPERVISED REPROJECTION LOSS

Reprojection	Sim GT	ADE (mm) ↓	> 8mm ↓
Raw IR	✓	**4.30**	**0.122**
	×	15.73	0.507
Temporal IR	✓	**3.67**	**0.083**
	×	4.79	0.142

Simulation Supervision: In order to investigate the effect of simulation supervision on disparity, we implement the experiments listed on Table III. We observe a significant performance drop after removing supervision on simulation disparity when using the Raw IR reprojection loss in the real domain. But the performance drop is small when using the Temporal IR reprojection loss. Therefore, we can conclude that supervision on simulation domain helps the network achieve better performance, especially when the self-supervision in the real domain is inferior.

TABLE IV
ABLATION STUDY ON LOSS WEIGHTS

λ_r	λ_s	ADE (mm) ↓	$>$ 8mm ↓
1.0	0.0	3.84	0.089
2.0	0.0	3.67	0.083
5.0	0.0	3.84	0.081
2.0	0.01	3.68	0.081
2.0	0.1	**3.61**	**0.079**
2.0	1.0	3.71	0.082
2.0	5.0	3.87	0.088
2.0	10.0	4.07	0.098

$\lambda_{gt} = 1$ for all the experiments.

TABLE V
PERFORMANCE OF NETWORK TRAINED ON DIFFERENT SIMULATION DATASETS, 'TESTING OBJECTS' CONSISTS OF ONLY OBJECTS IN THE TESTING DATASET, 'SHAPE PRIMITIVES' CONSISTS OF SHAPE PRIMITIVES OF DIFFERENT SIZE, TEXTURE AND MATERIAL

Simulation Dataset	ADE (mm) ↓	$>$ 8mm ↓
Testing objects	3.64	0.075
Shape primitives	3.67	0.083

TABLE VI
ABLATION STUDY OF DOMAIN RANDOMIZATION (DR)

DR	ADE (mm) ↓	$>$ 8mm ↓
✓	**3.67**	**0.083**
×	4.08	0.101

TABLE VII
ABLATION STUDY ON DIFFERENT BACKBONES

Backbone	DR	Pattern	ADE (mm) ↓	$>$ 8mm ↓
CFNet [57]	×	-	27.13	0.920
	✓	-	7.51	0.262
	✓	LCN IR	6.33	0.210
	✓	Temporal IR	**6.09**	**0.193**
RAFT-Stereo [58]	×	-	7.52	0.241
	✓	-	5.39	0.160
	✓	LCN IR	5.34	0.170
	✓	Temporal IR	**4.78**	**0.143**

DR: domain randomization.

TABLE VIII
PERFORMANCE OF NETWORK TRAINED ON SIMULATION DATASETS WITH DIFFERENT IR PATTERN

IR Pattern in Simulation	ADE (mm) ↓	$>$ 8mm ↓
D415	**3.67**	**0.083**
D435	3.86	0.086

'D415' is the same pattern as the real test data, while 'D435' is different from the real test data.

TABLE IX
PERFORMANCE OF NETWORK TRAINED ON DIFFERENT NUMBER OF SHAPE PRIMITIVES

No. of Shape Primitives	ADE (mm) ↓	$>$ 8mm ↓
5-15	3.98	0.096
25-50	**3.67**	**0.083**

Loss weights: Table IV shows the effects of loss weights on the stereo network's performance. When the weight of the real domain reprojection loss λ_r increases, the percentage of outliers decreases but the average depth error rises. This could be attributed to the reprojection enhancing the smoothness of the disparity prediction. When the weight of the simulation domain reprojection loss λ_s varies from 0.0 to 10.0, the performance first improves then declines. This may be due to over-fitting on the simulation domain.

Generalization: In order to evaluate the generalizability of the learned stereo network trained on the simulated dataset consisting of shape primitives, we construct another simulated dataset using the same objects as in the testing dataset. Table V shows the model trained on the random shape primitives dataset can achieve comparable performance with the model trained on the dataset that contains only shapes and textures that appear in the testing dataset. It validates the claim that the stereo network can achieve good generalizability to unseen objects, because it learns to find correspondences in stereo images rather than "memorizing" shape priors.

Customized domain randomization: Table VI compares the performance of mixed domain learning stereo network trained with and without customized simulation domain randomization. Compared with the online data augmentation used in our CVPR paper, the additional data randomization is able to increase the variation of rendered images and improve the stereo network's Sim2Real generalizability. And because the self-supervision in the real domain are also used for the two models, the improvement from domain randomization is less significant than only using the simulation data as shown in Table I.

Different backbones: Besides PSMNet [7], we further evaluate the effectiveness of the proposed mixed domain learning framework and temporal IR reprojection on two state-of-the-art stereo networks: CFNet [57] and RAFT-Stereo [58]. Table VII shows the results of the ablation study, which validates that our proposed method is able to improve the performance for different backbones.

Different IR pattern in synthetic dataset: To evaluate how the IR pattern in simulation affects the model's performance, we extract the IR pattern from RealSense D435, which differs from the one of D415, and generate another synthetic dataset. Table VIII compares the model's performance trained on simulation datasets with different IR pattern. The results indicate that using an IR pattern in simulation dataset generation that differs from the one used in real inference causes only a minor decrease in performance.

Number of shape primitives: In our previous CVPR work, we set the number of shape primitives in the synthetic primitives to 5-15. Here we change the number to 25-50, which enlarges the variation of the scene geometry and disparity maps. Table IX shows the comparison result. It demonstrates that increasing the number of shape primitives can enhance the stereo network's ability to generalize to different scenarios.

Disparity range: In our previous CVPR work, we follow PSMNet [7] and set the disparity range to 192. Here we reduce the disparity range to 12-96 with the same number of disparity hypotheses. The resolution of disparity hypotheses is thus refined to sub-pixel level. Table X shows the comparison result. It demonstrates that reducing the disparity range with the

TABLE X
PERFORMANCE OF NETWORK USING DIFFERENT DISPARITY RANGE

Disparity Range	ADE (mm) ↓	> 8mm ↓
1-192	3.95	0.097
12-96	**3.67**	**0.083**

TABLE XI
COMPARISON OF NORMAL ESTIMATION ERROR

Method	Overall (rad)↓	Printed (rad)↓	Real (rad)↓
ClearGrasp [15]	**0.169**	0.436	0.425
Ours	0.188	**0.316**	**0.386**

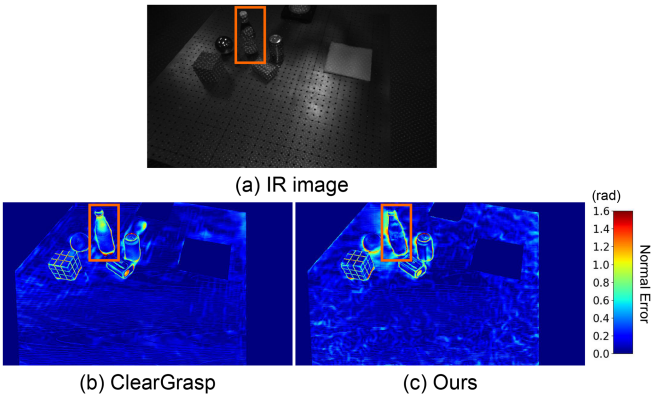


Fig. 9. Visualization of normal estimation results.

same number of hypotheses can improve the stereo network's prediction accuracy.

C. Confidence-Based Depth Completion

1) Implementation Details: Because the training dataset in the real domain $\mathbb{X}$ does not contain the groundtruth depth or normal, we use the trained stereo network to generate the estimated depth, normal map, along with confidence map. Then we train the stereo network, the stereo normal estimation network and the boundary detection network from scratch jointly on $\mathbb{X}$ and $\mathbb{Y}$. We use the Adam optimizer with the initial learning rate set to 2e-4, decaying by half every 10 k iterations for a total of 60 k iterations. The network is trained on 2 GPUs each with 11 GB GPU memory and a batch size of 4. We set $\lambda_{nr} = 1$, $\lambda_{ns} = 1$, $c = 2$ for (7). The confidence threshold is set as 0.8, and the normal threshold is set as 20°.

2) Comparison Results: We first compare the performance of the cost-volume-based stereo normal estimation network with the 2D network in ClearGrasp [15]. For fair comparison, we use the IR image as the input of the 2D network instead of RGB image in the original ClearGrasp work. Table XI and Fig. 9 show the comparison result. While ClearGrasp has better overall performance, the advantage mainly comes from the planar table surface. Our method achieves better performance on objects, especially real objects, demonstrating that our model has better generalizability and higher robustness to object materials.

We then compare the depth predictions of stereo network + depth completion and stereo network alone. As shown in Table I, both approaches yield similar outcomes. It is worth noting that the introduction of depth completion may lead to additional errors due to the fact that the object boundary detection network employed is solely trained on synthetic data. Consequently, some object boundaries may be missed, resulting in oversmoothing in the global optimization.

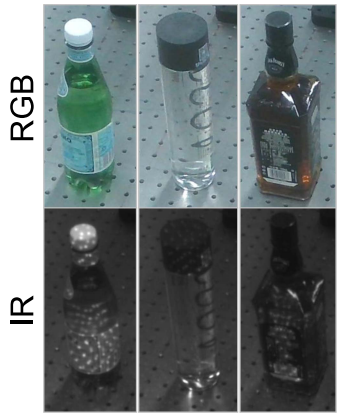


Fig. 10. Transparent and translucent objects for depth completion evaluation.

TABLE XII
COMPARISON OF REAL DEPTH SENSOR, STEREO NETWORK AND DEPTH COMPLETION ON REAL TRANSPARENT AND TRANSLUCENT OBJECTS

Method	ADE (mm) ↓	> 8mm ↓	Norm Err (rad)↓
RealSense	17.24	0.681	0.968
Stereo	12.99	**0.539**	0.607
Segmentation-based	25.04	0.883	0.573
Confidence-based	**12.95**	0.545	**0.531**

We further compare the proposed confidence-based depth completion with the segmentation-based depth completion in ClearGrasp [15]. Because the public dataset in ClearGrasp does not contain the stereo IR images or the confidence map, we choose three transparent and translucent objects (Fig. 10) in the real test dataset to evaluate the performance. The testing dataset consists of 160 stereo images of 10 different scenes. For the segmentation-based depth completion, we use the groundtruth object mask as the segmentation, and the normal estimation and boundary detection are the same as our method. Table XII and Fig. 11 show the completion results. For the transparent and translucent objects, the depth from the RealSense is incomplete because the projected pattern is refracted and cannot be captured by the cameras. Our mixed domain learning stereo network can generate more complete and accurate result for these objects, but still have large depth noise in some regions. While the segmentation-based depth completion is able to reconstruct the complete shape, the depth error is larger than our method, due to the fact that the region of object whose depth is accurately predicted by the stereo network is not used for the depth completion and it leads to larger cumulative error. Our method takes advantage of the confidence map from the stereo network and only completes the depth of uncertain regions, achieving better reconstruction quality. Compared to the original depth from the mixed domain learning stereo network, the accuracy of the completed depth is close, but the normal accuracy is remarkably

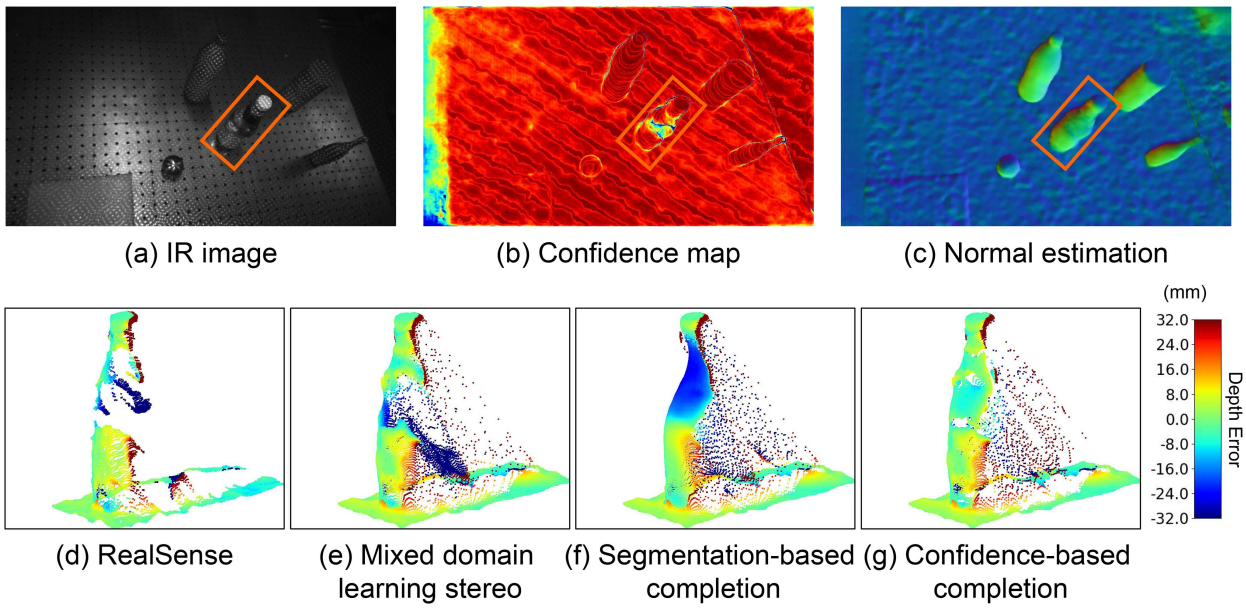


Fig. 11. Illustration of confidence-based depth completion results.

TABLE XIII
ABLATION STUDY OF STEREO NORMAL ESTIMATION

Confidence c	Sim	Real	Overall (rad)	Printed (rad)	Real (rad)
2	✓	✓	**0.184**	**0.315**	**0.385**
1	✓	✓	0.188	0.316	0.386
0	✓	✓	0.196	0.318	0.387
2	×	✓	0.192	0.333	0.401
2	✓	×	0.199	0.366	0.435
0	×	✓	0.194	0.332	0.405

improved, meaning that our confidence-based depth completion can generate more smooth and noiseless depth, which is beneficial for downstream tasks, such as grasp detection.

3) Ablation Study: In this section, we first validate the effectiveness of mixed domain learning and the the confidence-aware loss for stereo normal estimation through ablation experiments. Table XIII shows the experimental results. Comparing the performances of the confidence coefficient $c = 0, 1, 2$, we can find that it achieves higher accuracy when using the confidence-aware normal loss in the real domain, because the normal supervision in uncertain regions could be misleading for training. For the mixed domain learning, both the simulation data and the real data are critical for the model training: the simulation data contain the accurate groundtruth normal and the real data helps the network extract better features from real IR images.

We further study the impact of the confidence threshold μ_c and the normal threshold μ_n for the confidence-based depth completion. As shown in Table XV, the confidence threshold affects the accuracy of the optimization results. A low μ_c retains more pixels with low confidence and large depth error, which increases the depth and normal error after global optimization. Conversely, a high μ_c may exclude too many pixels, which causes an accumulative error in global optimization. The normal threshold μ_n has a minor impact on the optimization results. This may be because the pixels with incorrect normals are also low-confidence pixels and are already excluded by μ_c.

4) Computation Time: In this section, we evaluate the computation time of the depth completion module on a computer with a CPU @ 2.45 GHz and a NVIDIA RTX 4090 GPU. Since it is a global optimization algorithm running on CPU, the time is highly determined by the resolution of the input image. Table XVI shows the time for different resolutions.

D. Application: Sim2Real 6D Pose Estimation

In this section, we show the value of the proposed method in potential applications. The reason we choose 6D pose estimation is that accurate 6D pose estimation is useful for various downstream applications, such as robot manipulation and augmented reality, and it is difficult to annotate the 6D poses in real images. In [59], the authors use classical stereo matching algorithm to generate simulated depth maps on rendered stereo images, train 6D pose estimation models on the simulation data, and directly test the trained models on real depth maps from the RealSense D415. In this work, we use the same simulation scenes as [59] for fair comparison. We first use the trained stereo network as in Section III to generate simulated depth maps on rendered stereo images, train 6D pose estimation models on the simulation data. Then we use the same stereo network to generate depth maps from real stereo IR images, and test the trained pose estimation models. As shown in Table XIV, our method can significantly improve the Sim2Real performance on most metrics for all the three pose estimation algorithms. For real objects which are challenging for the commercial depth sensor, the pose estimation algorithms trained and tested by our method can output more accurate results (Fig. 12).

E. Failure Cases

In this section, we show some typical failure cases. Since the depth completion relies on detected edges for optimization, it may result in over-smoothing of the depth between the

TABLE XIV
COMPARISON OF SIM2REAL 6D POSE ESTIMATION IN THE REAL DOMAIN

Pose Algo.	Depth	Overall(%)		Real(%)		Printed(%)	
		10°,10mm	20°,20mm	10°,10mm	20°,20mm	10°,10mm	20°,20mm
PVN3D	RealSense D415 [59]	61.42	80.03	41.77	71.67	78.59	**88.26**
	Ours	68.85	80.99	55.60	78.53	**80.85**	83.29
Frustum	RealSense D415 [59]	42.70	74.60	38.10	73.61	47.00	76.87
	Ours	58.57	82.66	56.14	84.93	60.17	80.57
SegICP	RealSense D415 [59]	65.71	79.16	49.26	72.84	79.75	85.29
	Ours	**73.29**	**84.88**	**64.65**	**86.84**	80.81	82.69

The best result of all three pose estimation algorithms is marked as **BOLD**, and the best result of each algorithm is marked as underlined.

TABLE XV
ABLATION STUDY ON CONFIDENCE THRESHOLD μ_c AND NORMAL THRESHOLD μ_n

μ_c	$\mu_n(^\circ)$	ADE (mm) ↓	$>$ 8mm ↓	Norm Err (rad)↓
0.5	20.0	12.85	0.536	0.574
0.6	20.0	12.79	0.537	0.562
0.7	20.0	12.74	0.537	0.545
0.8	20.0	12.95	0.545	0.531
0.9	20.0	14.94	0.594	0.534
0.8	10.0	13.19	0.548	0.532
0.8	30.0	12.88	0.543	0.532
0.8	40.0	12.84	0.542	0.532

TABLE XVI
COMPUTATION TIME OF THE DEPTH COMPLETION MODULE FOR DIFFERENT RESOLUTIONS

Resolution	960 × 540	480 × 270	320 × 180
Time (s)	12.20	2.09	0.82

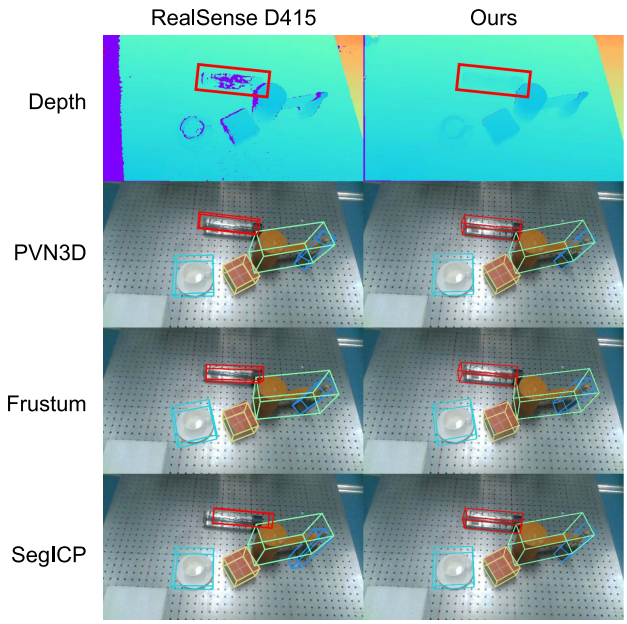


Fig. 12. Visualization of Sim2Real 6D pose estimation results.

foreground and background objects as shown in Fig. 13(d), if the edge is not detected successfully. In translucent and dark regions as shown in Fig. 13(e), the active pattern cannot be reflected back to the camera, leading to insufficient information and a significant domain gap between synthetic images and

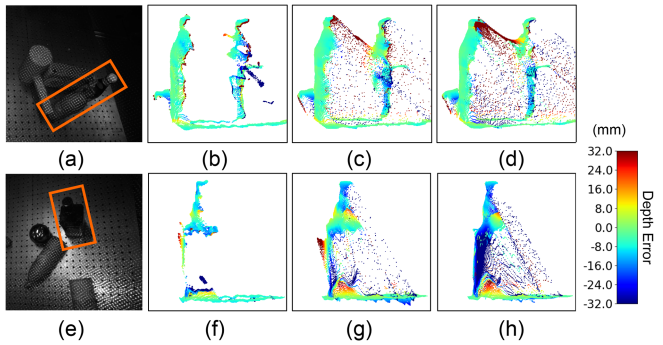


Fig. 13. Illustration of typical failure cases, (a), (e) IR images, (b), (f) results of RealSense, (c), (g) results of mixed domain learning stereo, (d), (h) results of confidence-based completion.

real captured images. While our method may generate produce greater errors in these regions, the results remain more complete and accurate than those generated by RealSense.

VI. CONCLUSION

In this article, we first propose a novel end-to-end training framework, *mixed domain learning*, for learning-based active stereo that surpasses commercial depth sensors and state-of-the-art methods in the real world without any real depth annotation. We further improve the depth completeness by confidence-based depth completion. Additionally, our method is applicable to Sim2Real 6D pose estimation to greatly narrow the domain gap of depth maps. One limitation of our work is that we only evaluate our method's effectiveness in the robotics working environment and its generalizability to completely unseen real scenarios may be limited. Further study is needed to improve the generalizability of the method, such as by expanding the diversity and scale of the synthetic and real-world datasets used for mixed domain learning and improving the capacity of the stereo network. Additionally, in order for this framework to be useable in real applications, we would need to investigate how to accelerate network inference to achieve real-time depth predictions.

REFERENCES

[1] A. K. Mishra and O. Meruvia-Pastor, "Robot arm manipulation using depth-sensing cameras and inverse kinematics," in *Proc. Oceans -St. John's*, 2014, pp. 1–6.
[2] M. Hwang et al., "Applying depth-sensing to automated surgical manipulation with a da Vinci robot," in *Proc. Int. Symp. Med. Robot.*, 2020, pp. 22–29.

[3] J. Cunha, E. Pedrosa, C. Cruz, A. J. Neves, and N. Lau, "Using a depth camera for indoor robot localization and navigation," DETI/IEETA-University of Aveiro, Portugal, p. 6, 2011.

[4] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1–10.

[5] S.-H. Baek and F. Heide, "Polka lines: Learning structured illumination and reconstruction for active stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5757–5767.

[6] R. Chen, J. Xu, and S. Zhang, "Comparative study on 3D optical sensors for short range applications," *Opt. Lasers Eng.*, vol. 149, 2022, Art. no. 106763. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0143816621002335

[7] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.

[8] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 767–783.

[9] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1538–1547.

[10] R. Chen, S. Han, J. Xu, and H. Su, "Visibility-aware point-based multi-view stereo network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3695–3708, Oct. 2021.

[11] Y. Zhang et al., "Activestereonet: End-to-end self-supervised learning for active stereo systems," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 784–801.

[12] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," 2017, *arXiv: 1709.00930*.

[13] R. Liu, C. Yang, W. Sun, X. Wang, and H. Li, "StereoGAN: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12757–12766.

[14] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv: 1705.07215*.

[15] S. Sajjan et al., "Clear grasp: 3D shape estimation of transparent objects for manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3634–3642.

[16] I. Liu et al., "Activezero: Mixed domain learning for active stereovision with zero annotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13033–13042.

[17] S. Giancola, M. Valenti, and R. Sala, *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Berlin, Germany: Springer, 2018.

[18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Robot. Res.*, vol. 32, pp. 1231–1237, Sep. 2013.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2012, pp. 1106–1114.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[21] D. Scharstein, "A taxonomy and evaluation of dense two-frame stereo correspondence," in *Proc. IEEE Workshop Stereo Multi-Baseline Vis.*, Kauai, HI, 2001, pp. 131–140. [Online]. Available: https://ci.nii.ac.jp/naid/10011856636/en/

[22] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1592–1599. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298767

[23] N. Mayer et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4040–4048.

[24] A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 66–75.

[25] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2492–2501.

[26] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 270–279.

[27] H. Gao, X. Liu, M. Qu, and S. Huang, "PDANet: Self-supervised monocular depth estimation using perceptual and data augmentation consistency," *Appl. Sci.*, vol. 11, no. 12, p. 5383, 2021. [Online]. Available: https://www.mdpi.com/2076--3417/11/12/5383

[28] G. Riegler, Y. Liao, S. Donne, V. Koltun, and A. Geiger, "Connecting the dots: Learning representations for active monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7624–7633.

[29] M. M. Johari, C. Carta, and F. Fleuret, "Depthinspace: Exploitation and fusion of multiple video frames for structured-light depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6039–6048.

[30] S. Schreiberhuber, J.-B. Weibel, T. Patten, and M. Vincze, "GigaDepth: Learning depth from structured light with branching neural networks," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Springer, 2022, pp. 214–229.

[31] C. Li, Y. Monno, and M. Okutomi, "Deep hyperspectral-depth reconstruction using single color-dot projection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19770–19779.

[32] Y. Xu, X. Yang, Y. Yu, W. Jia, Z. Chu, and Y. Guo, "Depth estimation by combining binocular stereo and monocular structured-light," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1746–1755.

[33] F. Warburg, D. Hernandez-Juarez, J. Tarrio, A. Vakhitov, U. Bonde, and P. F. Alcantarilla, "Self-supervised depth completion for active stereo," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 3475–3482, Apr. 2022.

[34] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.

[35] M. Long, G. Ding, J. Wang, J.-G. Sun, Y. Guo, and P. S. Yu, "Transfer sparse coding for robust image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 407–414.

[36] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Springer International Publishing, 2016, pp. 443–450.

[37] B. Liu, H. Yu, and G. Qi, "GraftNet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13012–13021.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[39] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2015, pp. 1180–1189.

[40] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.

[41] S. Sankaranarayanan, Y. Balaji, A. Jain, S.-N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3752–3761.

[42] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 994–1003.

[43] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 175–185.

[44] Y. Tang, J. Chen, Z. Yang, Z. Lin, Q. Li, and W. Liu, "DepthGrasp: Depth completion of transparent objects using self-attentive adversarial network with spectral residual for grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 5710–5716.

[45] H. Xu, Y. R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, "Seeing glass: Joint point-cloud and depth completion for transparent objects," in *Proc. Conf. Robot Learn.*, 2022, pp. 827–838.

[46] H. Fang, H.-S. Fang, S. Xu, and C. Lu, "TransCG: A large-scale real-world dataset for transparent object depth completion and a grasping baseline," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 7383–7390, Jul. 2022.

[47] Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial," *Found. Trends, Comput. Graph. Vis.*, vol. 9, no. 1/2, pp. 1–148, Jun. 2015. [Online]. Available: https://doi.org/10.1561/0600000052

[48] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch stereo-stereo matching with slanted support windows," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.

[49] S. R. Fanello et al., "Ultrastereo: Efficient learning-based matching for active stereo systems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6535–6544.

[50] M. Pharr, W. Jakob, and G. Humphreys, *Physically Based Rendering: From Theory to Implementation*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[51] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," *Convolutional Neural Netw. Vis. Recognit.*, vol. 7, no. 7, p. 3, 2015.

[52] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2016:main.html
[53] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2524–2534.
[54] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2186–2196.
[55] S. G. Parker et al., "OptiX: A general purpose ray tracing engine," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 66:1–66:13, Jul. 2010. [Online]. Available: https://doi.org/10.1145/1778765.1778803
[56] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 1994, pp. 151–158.
[57] Z. Shen, Y. Dai, and Z. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13906–13915.
[58] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 218–227.
[59] X. Zhang et al., "Close the optical sensing domain gap by physics-grounded active stereo sensor simulation," *IEEE Trans. Robot.*, vol. 39, no. 3, pp. 2429–2447, Jun. 2023.

Rui Chen received the BE degree in mechanical engineering, and the PhD degree in mechatronical engineering from Tsinghua University, Beijing, China, in 2020 and 2014, respectively. He is currently a research Assistant Professor with the Department of Mechanical Engineering, Tsinghua University. His research interests include three-dimensional computer vision and robot learning.

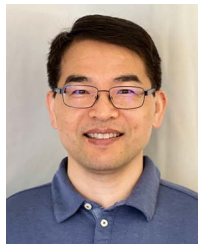
Isabella Liu received the BS degree in computer science and statistics from the University of Illinois, Urbana-Champaign. She is currently working toward the PhD degree with the University of California, San Diego. Her research interests include 3D reconstruction and efficient rendering.

Edward Yang received the BS degree in computer science from the University of California- San Diego, in 2022, and is currently working toward the MS degree with Yale University. His research interests include 3D computer vision and reinforcement learning.

Jianyu Tao received the master's degree of electrical computer engineering from the University of California, San Diego. He joined Prof. Hao Su's Lab from November 2020 to present. His research interests include 3D depth estimation and reconstruction.

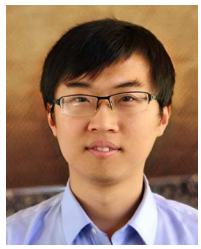
Xiaoshuai Zhang received the BS degree from Peking University, Beijing, China, in 2019. He is currently working toward the PhD degree in computer science with the University of California, San Diego. His current research interests include 3D scene reconstruction and understanding, computer graphics, and simulation.

Qing Ran received the PhD degree in computer engineering from Zhejiang University, Hangzhou, China, in 2019. She is currently working as a senior algorithm engineer with the Alibaba DAMO Academy. Her research interests include three-dimensional reconstruction and image-based rendering.



Zhu Liu (Senior Member, IEEE) received the BS and MS degrees in electronic engineering from Tsinghua University, and the PhD degree in electrical and computer engineering from New York University. He is a machine learning science manager with Amazon. He was a senior staff algorithm engineer with Alibaba DAMO and a principal inventive scientist with AT&T Labs - Research. His research interests include video content analytics, computer vision and machine learning. He co-authored a book on video search engines and has published more than 75 technical papers. He holds 222 granted US patents and won the AT&T Science & Technology Medal, in 2017. He was an adjunct professor of Columbia University and New York University. He is an associate editor of *IEEE Transactions on Multimedia*, and he has served on the organizing committees of many IEEE conferences.

Jing Xu (Member, IEEE) received the PhD degree in mechanical engineering from Tsinghua University, Beijing, China, in 2008. He was a postdoctoral researcher with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing. He is currently an associate professor with the Department of Mechanical Engineering, Tsinghua University. His research interests include vision-guided manufacturing, image processing, and intelligent robotics.



Hao Su has been in UC San Diego as assistant professor of computer science and engineering since July 2017. He is affiliated with the Contextual Robotics Institute and Center for Visual Computing. He served on the program committee of multiple conferences and workshops on computer vision, computer graphics, and machine learning. He is the area chair of ICCV'19, CVPR'19, senior program chair of AAAI'19, IPC of Pacific Graphics'18, program chair of 3DV'17, publication chair of 3DV'16, and chair of various workshops with CVPR, ECCV, and ICCV. He is also invited as keynote speakers at workshops and tutorials in NIPS, 3DV, CVPR, RSS, ICRA, S3PM, etc.