ActiveZero: Mixed Domain Learning for Active Stereovision with Zero Annotation

Isabella Liu[1]   Edward Yang[1]   Jianyu Tao[1]   Rui Chen[2]   Xiaoshuai Zhang[1]

Qing Ran[3]   Zhu Liu[3]   Hao Su[1]

[1]University of California, San Diego   [2]Tsinghua University   [3]Alibaba DAMO Academy

Abstract

Traditional depth sensors generate accurate real world depth estimates that surpass even the most advanced learning approaches trained only on simulation domains. Since ground truth depth is readily available in the simulation domain but quite difficult to obtain in the real domain, we propose a method that leverages the best of both worlds. In this paper we present a new framework, ActiveZero, which is a mixed domain learning solution for active stereovision systems that requires no real world depth annotation. First, we demonstrate the transferability of our method to out-of-distribution real data by using a mixed domain learning strategy. In the simulation domain, we use a combination of supervised disparity loss and self-supervised losses on a shape primitives dataset. By contrast, in the real domain, we only use self-supervised losses on a dataset that is out-of-distribution from either training simulation data or test real data. Second, our method introduces a novel self-supervised loss called temporal IR reprojection to increase the robustness and accuracy of our reprojections in hard-to-perceive regions. Finally, we show how the method can be trained end-to-end and that each module is important for attaining the end result. Extensive qualitative and quantitative evaluations on real data demonstrate state of the art results that can even beat a commercial depth sensor. The codes of ActiveZero are available at: https://github.com/haosulab/active_zero.

1. 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 [6,17,26]. Among the different types of depth sensors available, active stereovision depth sensors (eg. Intel RealSense™D series) are the most widely adopted in both industry and academic settings due to their high spatial
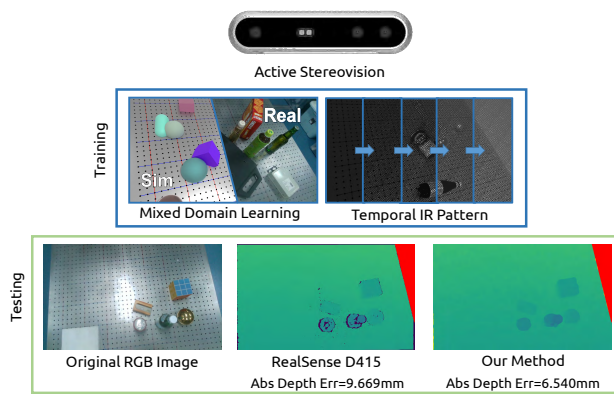

Figure 1. ActiveZero produces more accurate and complete disparity estimates on real IR stereo images for objects with complex optical characteristics (specular, transparent) than commercial depth sensors with zero real depth annotation using *mixed domain learning* by leveraging self-supervised reprojection loss on temporal IR patterns in the real domain and direct disparity supervision in the simulation domain.

resolution, high accuracy, and low cost [19]. These 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 [5].

Learning based methods can solve the aforementioned issues by generating more accurate and complete depth maps through the utilization of prior samples to understand how to correctly handle edges and uncertain pixels [2–4, 35]. However, a large scale stereo dataset with ground truth 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 [38, 39] 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 ground truth 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 [23], but the introduction of GANs makes the training process unstable [20] and the performance suboptimal.

This paper 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* (Fig. 1). 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 ground truth 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.

In addition, we observed that there are fundamental issues with performing direct image reprojection as previous works had done so we propose the use of temporal IR by periodically adjusting the brightness of the emitted IR pattern and extracting the binary pattern from the temporal image sequences. The reprojection loss on the temporal binary pattern eliminates the influence of scene texture and also the effect of illumination strength decaying with increased distance. 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.

2. Related Work

Depth sensors can be classified into four categories according to their underlying sensing principle [5]: passive stereo-vision, active stereo-vision, structured light, and time-of-flight. Each depth sensing technique has its own advantages and drawbacks. Giancola *et al.* [13] introduces the principles of different depth sensors and evaluated their metrological performance independently. Chen *et al.* [5] compared the short-range depth sensing performance of 8 commercially available 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 [29]. In this paper, 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 [12, 16, 21]. Stereo matching for depth estimation is typically done in four steps: matching cost computation, cost aggregation, optimization, and disparity refinement [31]. Zbontar and LeCun were the first to design a network for computing matching costs by utilizing a deep Siamese architecture [37]. Building on this, DispNet introduced the first end-to-end framework for predicting entire disparity maps from stereo image pairs [25]. Works such as GWCNet followed and improved on this framework by using 3D convolutions to compute better cost volumes [18]. 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 [2, 15]. However, the requirement of ground truth depth as supervision has limited the application of learning based stereo.

Self-supervised Stereo is a popular approach for stereo matching when ground truth depth is unavailable. Godard *et al.* [14] 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 [11] introduced the idea of perceptual consistency to improve reconstruction quality on regions with low texture and high color fluctuations. ActiveStereoNet [38] 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.

Domain Adaptation techniques have shown great promise in closing the gap between the simulation and real domains. Tobin *et al.* [33] 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
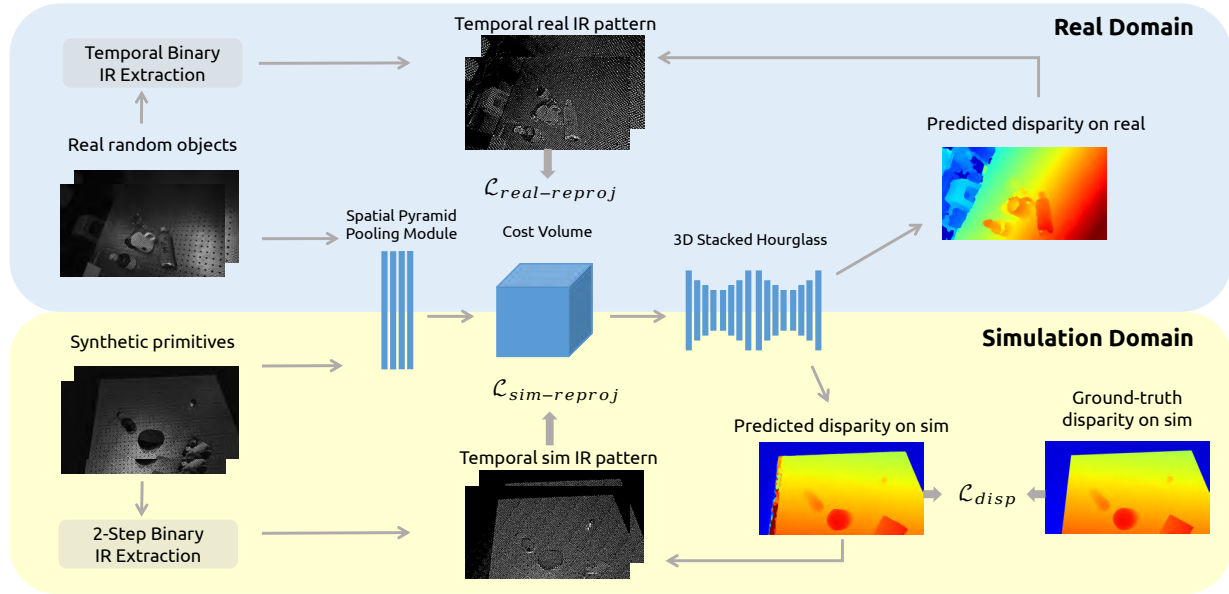
Figure 2. Architecture overview. The simulated and real stereo IR images are fed to a shared weight stereo network consisting of a CNN for noise reduction and a cost-volume-based 3D CNN for disparity prediction. The network is trained with reprojection loss on temporal binary IR pattern in the real domain, reprojection loss and disparity loss in the simulation domain as mixed domain learning.

matching their input distributions or their feature statistics [24, 32]. 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 [10]. Moreover, unsupervised losses have seen increased use for domain adaptation in tasks such as semantic segmentation and object detection [7, 30, 34].

Our work is most related to StereoGAN [23], which uses ground truth 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.

3. Method

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 $\mathbb{X}^t = \{(x_l^t, x_r^t)_i\}_{i=1}^N$, and our goal is to learn an accurate disparity estimation network F to estimate the disparity $\hat{x}_d^t = F(x_l^t, x_r^t)$. We utilize *mixed domain* data

to train the network: in real domain $\mathcal{X}$ we collect another set of real IR stereo images $\mathbb{X} = \{(x_l, x_r)_i\}_{i=1}^M$ 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 ground truth disparity annotation $\mathbb{Y} = \{(y_l, y_r, y_d)_i\}_{i=1}^K$. 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 $\mathbb{Y}$.

Figure 2 shows the framework of our proposed method. In the real domain, we propose the use of temporal binary IR reprojection loss as self-supervision (Sec. 3.1). In the simulation domain, we use the loss between predicted disparity and the ground truth disparity y_d as supervision (Sec. 3.2). The network is trained jointly using the self-supervision in real domain and supervision in simulation domain (Sec. 3.3). The stereo network architecture and other implementation details are introduced in Sec. 3.4.

3.1. Real Domain: Self-supervised Learning on IR Images

The prerequisite for computing reprojection loss of grayscale stereo images in conventional self-supervised learning methods [14, 38] 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
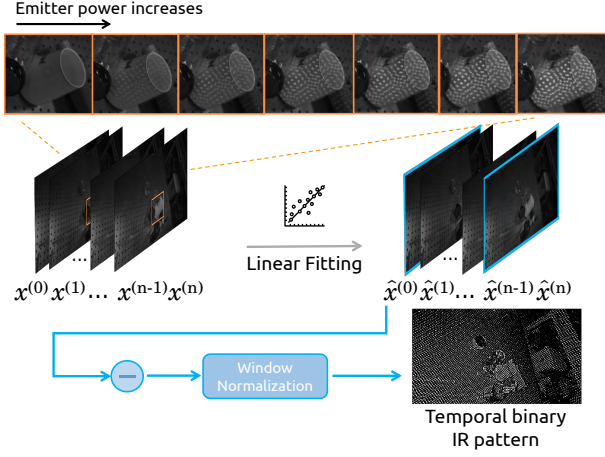
Figure 3. Temporal binary pattern extraction

active pattern from temporal IR stereo image sequences, which eliminates the adverse effect of surface reflection while maintaining the most important components of active pattern. Then, we construct the reprojection loss on this new binary pattern.

Binary Pattern Extraction From Temporal IR Images. For the real captured IR images x_l or x_r, the grayscale at pixel (u, v) is:

$$x_l(u, v) = I_l(u, v) + \alpha * e * K_l(u, v) + \epsilon \quad (1)$$

where $I_l(u, v)$ represents the environmental illumination intensity, $K_l(u, v)$ represents the binary pattern captured by the camera, α represents the reflection coefficient determined by the object surface material, texture, angle and distance, e represents the pattern emittance, and ϵ represents the camera sensor noise. For active depth sensors, we manually adjust the pattern emittance e by changing the emitter power. Therefore, as shown in Fig. 3, our pattern extraction procedure is as follows: we set e to $\{e_0, e_1, ..., e_n\}$, capture a temporal sequence of corresponding IR images $\{x^{(0)}, x^{(1)}, ..., x^{(n)}\}$, and fit $x^{(0)}, ..., x^{(n)}$ to the linear model regressed and obtain $\hat{x}^{(0)}, ..., \hat{x}^{(n)}$. We extract the binary IR pattern $K(u, v)$ from the temporal image sequence through local window normalization and binarization:

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

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

where $W(x, u, v)$ is a local window centered at pixel (u, v) in x with window size w, c is a threshold to filter out noise

and areas where the reflection coefficient is extremely small such as pure specular reflection regions. In our work, we use $n = 6$.

In Fig. 4, 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 active pattern, which is beneficial for further reprojection loss computation.

Binary Pattern Reprojection Loss. As demonstrated in traditional stereo matching and active stereo methods [1, 8, 9, 38], 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 extracted binary IR pattern (K_l, K_r) :

$$\mathcal{L}_{\text{reproj}}(K_l, K_r, \hat{x}_d) = \sum_{uv} \frac{1}{(2p+1)^2} C(u, v)$$

$$C(u, v) = \sum_{(u_p, v_p) \in P(u, v)} ||K_l(u_p, v_p) - \hat{K}_l(u_p, v_p)||^2$$

where $P(u, v)$ represents the patch centered at pixel (u, v) with patch size $(2p+1) \times (2p+1)$, $\hat{K}_l$ represents the warped right image using the predicted disparity $\hat{x}_d$.

As shown in Fig. 4, 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 binary IR pattern reaches global minima at the ground truth disparity while the losses computed on the other two patterns could be misleading for the stereo network.

3.2. 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 ground truth depth does not suffer from the aforementioned issues. However, it is costly and time-consuming to acquire ground truth depth in real world settings. Thus, we perform supervised learning only in the simulation domain.

Dataset Generation based on Ray-tracing. In the last decade, there has been significant progress in ray-tracing
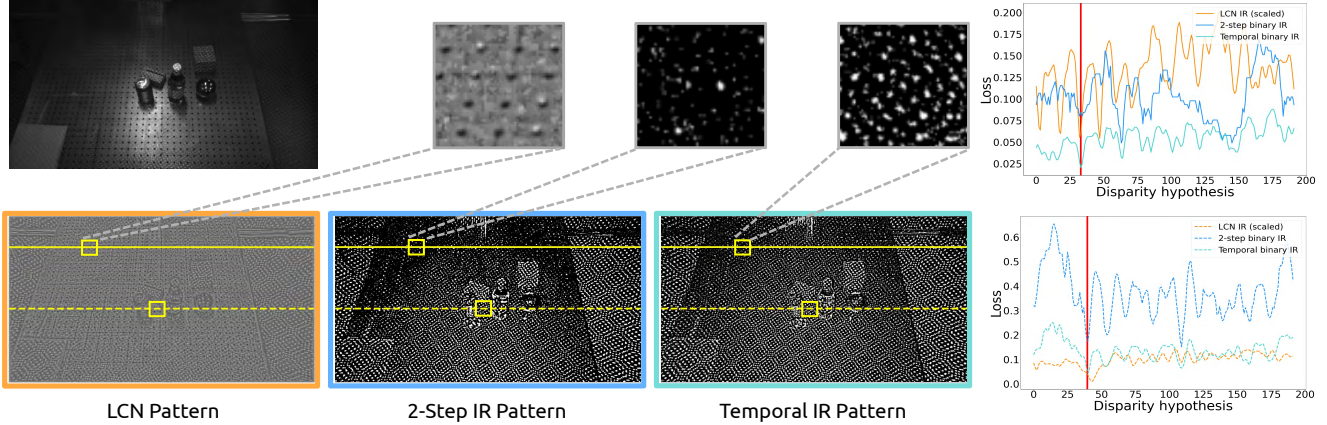
Figure 4. Comparison of extracted pattern and reprojection loss along the epipolar line. LCN pattern represents local contrast normalization [38] which consists of continuous values; 2-step IR pattern and temporal IR pattern represent the extracted binarized pattern from temporal IR image sequence using $n = 1$ and $n = 6$, respectively.

rendering techniques in terms of speed and quality. Compared with rasterization, ray-tracing rendering can accurately simulate the light transmission process on translucent and transparent objects [28]. 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.

Shape Primitives. The semantic-specific biases in CAD model datasets may mitigate the generalizability of the learned stereo network. Thus we only use base shape primitives for simulated dataset generation. We use images from tiny ImageNet [22] as object textures. The number of primitives is randomly sampled from 5 to 15. The sizes, layouts and materials are also randomly generated.

Disparity loss. Given the synthetic stereo image pair with ground-truth disparity (y_l, y_r, y_d), we follow [2] and adopt smooth L_1 loss between y_d and the predicted disparity on synthetic stereo images:

$$\mathcal{L}_{\text{disp}} = L_{1smooth}(F(y_l, y_r), y_d) \tag{3}$$

3.3. Mixed Domain Learning

Given the real stereo IR image(x_l, x_r), and the simulated stereo IR image with ground truth disparity (y_l, y_r, 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:

$$\mathcal{L}(x_l, x_r, y_l, y_r, y_d) = \lambda_r \cdot \mathcal{L}_{\text{real-reproj}}(x_l, x_r, F(x_l, x_r)) +$$
$$\lambda_s \cdot \Big[\mathcal{L}_{disp}(F(y_l, y_r), y_d) +$$
$$\mathcal{L}_{\text{sim-reproj}}(y_l, y_r, F(y_l, y_r)) \Big]$$

where λ_r and λ_s represent the weights of the real domain and the simulation domain 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 (see Sec. A.1 in supplementary material and Sec. 4.3 for empirical evidences).

3.4. Implementation Details

In the stereo matching network, we adopt PSMNet [2] as the backbone, which aggregates image features at different scales, constructs a cost volume and uses 3D CNNs to regress the disparity. The max disparity of PSMNet is set to be 192. We also use a 6-layer CNN to filter out irrelevant noise before feeding the stereo images into PSMNet. To make the model more robust, we apply color jitter and gaussian blur to the input images.

4. Experiments

4.1. Experiment details

Datasets. Figure 5 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 ground depth is required. To do so, we constructed a set of simulated scenes which are pixel-wise aligned with the real ones by precisely aligning the shapes and poses of objects and the intrinsic and extrinsic parameters of the RealSense camera. 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, 50% are textured by images from tiny imagenet [22], and the rest are set to random colors. 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 [27].

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. The objects are randomly placed on the table, and captured by the same RealSense from different viewpoints. Note that we only use the real IR stereo images to construct the temporal IR reprojection loss, and the depth images are not collected.

Training. We train the network using the Adam optimizer with the initial learning rate set to 2e-4, decaying by half every 10k iterations for a total of 40k iterations. The network is trained on 2 GPUs each with 11GB GPU memory and a batch size of 4. We use $\lambda_s = 0.01$ and $\lambda_r = 2$ for the loss weight to set the two losses to similar scales.

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.

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 (*abs depth err*)

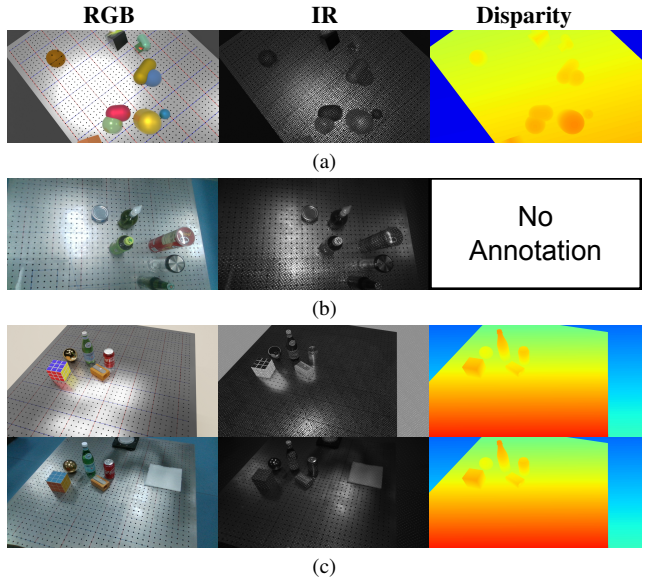| RGB | IR | Disparity |

(a)

(b)

(c)

Figure 5. 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 disparity annotation in (b).

and the percentage of depth outliers with absolute error larger than 4mm, which is denoted as *>4mm*. 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. 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.

4.2. Comparison with other Methods

For evaluation, our method is compared with other learning based methods and a decent commercial depth sensor - the RealSense D415. As shown in Tab. 1, our method outperforms other methods in all metrics.

Learning-based methods. Our method is best compared with PSMNet [2] and StereoGAN [23] and we use them as our baselines. To test vanilla PSMNet, we train it on input stereo images with and without active pattern using only the training dataset in the simulation domain and then test it directly in the real testing dataset. As shown in Tab. 1, using active pattern can improve the stereo matching accuracy across all metrics and is beneficial for eliminating the sim-real domain gap. This intuitively makes sense since active light adds pattern to textureless areas which are the most difficult to match.
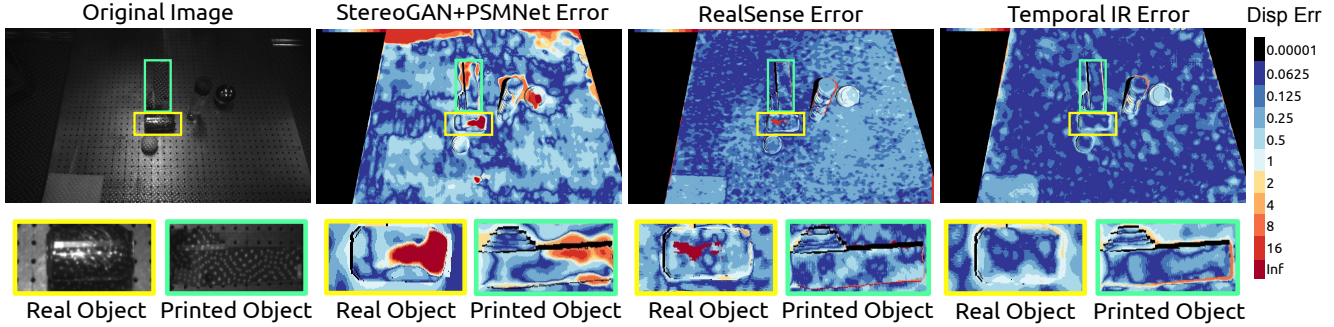
Figure 6. Comparison of the disparity error map of our method with StereoGAN and RealSense D415. Our method improves disparity accuracy on both 3D-printed objects and real objects.

Method	EPE (px) ↓	Bad 1 ↓	Abs depth err (mm) ↓			> 4mm↓		
	All	All	All	Printed	Real	All	Printed	Real
Excluding uncertain pixels								
PSMNet [2] w/o active pattern	0.664	0.187	9.218	12.600	16.467	0.478	0.686	0.836
PSMNet [2] w/ active pattern	0.476	0.077	7.135	9.174	15.570	0.504	0.591	0.800
StereoGAN [23]	5.603	0.741	44.284	36.892	42.105	0.925	0.915	0.931
StereoGAN [23] + PSMNet [2]	2.296	0.176	13.762	22.489	37.031	0.641	0.762	0.899
RealSense D415	0.392	0.032	5.817	7.851	15.826	0.565	0.612	0.817
Ours	**0.334** ± .022	**0.029** ± .001	**4.607** ± .242	**6.942** ± .376	**15.675** ± .789	**0.358** ± .036	**0.472** ± .04	**0.734** ± .028
Including uncertain pixels								
PSMNet [2] w/o active pattern	0.698	0.194	9.530	12.987	16.960	0.485	0.689	0.840
PSMNet [2] w/ active pattern	0.513	0.084	7.444	9.580	16.745	0.510	0.595	0.804
StereoGAN [23]	5.765	0.744	44.747	36.703	42.220	0.926	0.915	0.932
StereoGAN [23] + PSMNet [2]	2.472	0.185	14.318	22.818	37.753	0.645	0.764	0.902
RealSense D415	1.793	0.056	8.159	9.891	22.492	0.576	0.621	0.835
Ours	**0.420** ± .024	**0.037** ± .001	**5.027** ± .245	**7.450** ± .39	**17.430** ± .853	**0.366** ± .034	**0.479** ± .039	**0.748** ± .027

Table 1. Performance of different state of the art learning-based stereo, commercial depth sensor and our method on the real testing dataset

Furthermore, besides the original StereoGAN [23], we extend the StereoGAN architecture by using PSMNet as the disparity prediction backbone, which is denoted as **StereoGAN+PSMNet**. This improved StereoGAN uses cost volume aggregation in its stereo matching module, which makes it more powerful and comparable with our method. The results show that **StereoGAN+PSMNet** performs better than **StereoGAN** in all metrics. Although, when compared with our method, **StereoGAN+PSMNet** performs considerably worse as the absolute depth error increases from 4.377mm to 13.762mm. This is further corroborated by Fig. 6, where **StereoGAN+PSMNet** struggles to predict depth on real objects such as the metal can, 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.

Intel RealSense D415. 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 [19, 36], 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

Pattern	Abs depth err (mm) ↓	> 4mm ↓
Raw IR	32.166	0.638
LCN IR [38]	10.598	0.512
2-Step IR	4.697	0.373
Temporal IR	**4.377**	**0.335**

Table 2. Comparison of self-supervised reprojection loss on different patterns

Without sim ground-truth		
Reprojection	Abs depth err (mm) ↓	> 4mm ↓
Raw IR	43.326	0.716
Temporal IR	**4.729**	**0.367**
With sim ground-truth		
Reprojection	Abs depth err (mm) ↓	> 4mm ↓
Raw IR	32.166	0.638
Temporal IR	**4.377**	**0.335**

Table 3. Comparison of disparity supervision in the simulation domain with different self-supervised reprojection loss

Simulation Dataset	Abs depth err (mm) ↓	> 4mm ↓
Testing objects	4.388	0.347
Shape primitives	**4.377**	**0.335**

Table 4. 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

every metric. In Fig. 6, 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.

4.3. 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 Tab. 2. *Raw IR* simply computes the patch-wise Mean Squared Error (MSE) of the warped images. *LCN IR* is from ActiveStereoNet [38], 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 ground truth 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. This issue is addressed by our temporal IR patterns. By tracking the intensity difference in the temporal IR image sequence, our approach extracts a more accurate and complete IR pattern. The results prove that our reprojection on temporal IR images is superior to all other reprojection methods.

Simulation Supervision. In order to investigate the effect of simulation supervision, we implement the experiments listed on Tab. 3. Specifically, we observe a significant performance drop in the trained model after removing supervision on simulation disparity. Therefore, we can conclude that supervision on simulation domain helps the network achieve better performance.

As mentioned before, the simulation domain can help temporal IR reprojection converge closer and faster to a global minima. Then, temporal IR reprojection serves to further converge to the ground truth disparity. The results in Tab. 3 are consistent with the fact that synthetic supervision can further improve the performance.

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 4 shows the model trained on the random shape primitives dataset outperforms the model trained on the dataset that contains only shapes and textures that appear in the testing dataset, which validates the claim that greater variation of geometry, texture, and material introduced in our shape primitives dataset leads to superior generalizability of the learned stereo network.

5. Conclusion and Future Work

In this paper, we 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. One limitation of our work is that we only evaluate its effectiveness on one type of active stereovision sensor. Further study is needed to understand the extent to which our learned stereo network transfers to other out-of-distribution real datasets and types of sensors. 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] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011. 4

[2] 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. 1, 2, 5, 6, 7

[3] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019. 1

[4] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Visibility-aware point-based multi-view stereo network. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3695–3708, 2020. 1

[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, 2

[6] Joao Cunha, Eurico Pedrosa, Cristóvao Cruz, António JR Neves, and Nuno Lau. Using a depth camera for indoor robot localization and navigation. *DETI/IEETA-University of Aveiro, Portugal*, page 6, 2011. 1

[7] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018. 3

[8] Sean Ryan Fanello, Julien Valentin, Christoph Rhemann, Adarsh Kowdle, Vladimir Tankovich, Philip Davidson, and Shahram Izadi. Ultrastereo: Efficient learning-based matching for active stereo systems. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6535–6544. IEEE, 2017. 4

[9] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Found. Trends. Comput. Graph. Vis.*, 9(1–2):1–148, jun 2015. 4

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3

[11] Huachen Gao, Xiaoyu Liu, Meixia Qu, and Shijie Huang. Pdanet: Self-supervised monocular depth estimation using perceptual and data augmentation consistency. *Applied Sciences*, 11(12), 2021. 2

[12] Andreas Geiger, P Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: the kitti dataset. *The International Journal of Robotics Research*, 32:1231–1237, 09 2013. 2

[13] Silvio Giancola, Matteo Valenti, and Remo Sala. *A survey on 3D cameras: Metrological comparison of time-of-flight, structured-light and active stereoscopy technologies*. Springer, 2018. 2

[14] 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. 2, 3

[15] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[17] Minho Hwang, Daniel Seita, Brijen Thananjeyan, Jeffrey Ichnowski, Samuel Paradis, Danyal Fer, Thomas Low, and Ken Goldberg. Applying depth-sensing to automated surgical manipulation with a da vinci robot. In *2020 International Symposium on Medical Robotics (ISMR)*, pages 22–29. IEEE, 2020. 1

[18] 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

[19] 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, 7

[20] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017. 2

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 2

[22] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5, 6

[23] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li. Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12757–12766, 2020. 2, 3, 6, 7

[24] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jia-Guang Sun, Yuchen Guo, and Philip S. Yu. Transfer sparse coding for robust image representation. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 407–414, 2013. 3

[25] 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

[26] Akhilesh Kumar Mishra and Oscar Meruvia-Pastor. Robot arm manipulation using depth-sensing cameras and inverse kinematics. *2014 Oceans - St. John's*, pages 1–6, 2014. 1

[27] Steven G. Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, and Martin Stich. Optix: A general purpose ray tracing engine. *ACM Trans. Graph.*, 29(4), jul 2010. 6

[28] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 5

[29] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3642. IEEE, 2020. 2

[30] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser-Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 3

[31] DANIEL SCHARSTEIN. A taxonomy and evaluation of dense two-frame stereo correspondence. *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision, Kauai, HI, Dec, 2001*, 2001. 2

[32] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing. 3

[33] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 2

[34] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[35] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1

[36] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *European conference on computer vision*, pages 151–158. Springer, 1994. 7

[37] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. 2

[38] 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. 2, 3, 4, 5, 8

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