# Context-Enhanced Stereo Transformer

Weiyu Guo[1,2(✉)], Zhaoshuo Li[3(✉)], Yongkui Yang[1(✉)], Zheng Wang[1(✉)], Russell H. Taylor[3], Mathias Unberath[3], Alan Yuille[3], and Yingwei Li[3(✉)]

[1] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Johns Hopkins University, Baltimore, USA
{wy.guo,yk.yang,zheng.wang}@siat.ac.cn, {zli122, yingwei.li}@jhu.edu

**Abstract.** Stereo depth estimation is of great interest for computer vision research. However, existing methods struggles to generalize and predict reliably in hazardous regions, such as large uniform regions. To overcome these limitations, we propose Context Enhanced Path (CEP). CEP improves the generalization and robustness against common failure cases in existing solutions by capturing the long-range global information. We construct our stereo depth estimation model, Context Enhanced Stereo Transformer (CEST), by plugging CEP into the state-of-the-art stereo depth estimation method Stereo Transformer. CEST is examined on distinct public datasets, such as Scene Flow, Middlebury-2014, KITTI-2015, and MPI-Sintel. We find CEST outperforms prior approaches by a large margin. For example, in the zero-shot synthetic-to-real setting, CEST outperforms the best competing approaches on Middlebury-2014 dataset by 11%. Our extensive experiments demonstrate that the long-range information is critical for stereo matching task and CEP successfully captures such information[†].

**Keywords:** Stereo depth estimation, transformer, context extraction

## 1 Introduction

Stereo depth estimation is a critical task in computer vision that has been widely used in various fields, such as robotics [27], autonomous driving [24], and 3D scene reconstruction [29]. Recent developments in learning-based stereo disparity estimation algorithms generally use using techniques restricted to local information for matching the feature patterns between the left and right images. For example, prior works [2,8,36] construct a cost volume with pre-defined disparity range and use 3D convolutions to process the cost volume, limiting themselves to the receptive field of convolution kernel. Xu *et al.*[32] proposed to instead process the cost volume using 2D convolutions, however, facing the same challenge. Recently, approaches that attempt to capture more global information have been proposed. For example, STTR [17] and RAFT-Stereo [19] computes

[†]Code available at: github.com/guoweiyu/Context-Enhanced-Stereo-Transformer

attention or correlation between all pixels of the left and right images on the same epipolar lines. However, they all fail to take advantage of cross-epipolar line information, which is a critical component of global information processing. Thus, as shown in Figure 1, these methods cannot address hazardous regions like textureless, large uniform regions, specularity, and transparency [15,37], which are particularly challenging for stereo algorithms to produce reliable estimates. The features of left and right frames in these regions are often similar or misleading, which makes the feature matching ambiguous [37]. If disparities of these regions cannot be reliably predicted, downstream applications, such as 3D object detection [28], may be severely impacted due to missing or wrong predictions. Therefore, in this paper, we seek to answer this critical question: how to guide the stereo models properly handle those hazardous regions.

To address this question, we hypothesize that the long-range contextual information help to improve the predictions on hazardous regions. For example, as shown in Figure 1 (a), previous work performs unreliably in large white wall. However, if we could use the global information (*e.g.*, orientation, edge information) of the house, the prediction can be improved. Such global context information in theory will inform the model about the geometry on a global scale and guide the model to resolve the ambiguity in prediction. To this end, we proposed a plug-in module, called Context Enhanced Path (CEP), which helps stereo matching models to better understand the global structure of the hazardous regions. Compared to existing methods, CEP offers the following three unique advantages: (1) strong generalization ability, compared with previous methods [2,17], CEP shows strong results on unseened real-world data even if only training on synthetic data; (2) robustness against hazardous, thanks to modeling the long-range contextual information.(3) generic, unlike [9,14,34], our method serves as a plug-in that can be potentially applied to most of stereo matching methods. We construct our stereo depth estimation model based on CEP, namely Context Enhanced Stereo Transformer (CEST). We have examined CEST on several popular and diverse datasets, such as, Middlebury-2014[26], KITTI-2015 [24], and MPI sintel [1]. Our extensive experiments demonstrate that (1) the long-range information is critical for stereo depth estimation, (2) CEST attains strong generalization ability, and (3) more importantly, CEST can better handle hazardous regions, such as texturelessness and disparity jumps (shown in Figure 1 and Table 3). This result is attributed to our simple yet powerful observation: using long-range contextual information to better understand the global structure of the image can significantly help stereo depth estimation especially for those hazardous area. This result suggests that modeling long-range context information is critical for building a robust and generalizable stereo depth estimation algorithm.

To summarize, our contributions are 3-fold: (1) we found global contextual information is critical for stereo depth estimation; (2) we design a plug-in module, Context Enhanced Path (CEP), for generic stereo depth estimation models; (3) we integrate our plug-in module and build a stereo matching model named Context Enhanced Stereo Transformer (CEST), which achieves the state-of-the-

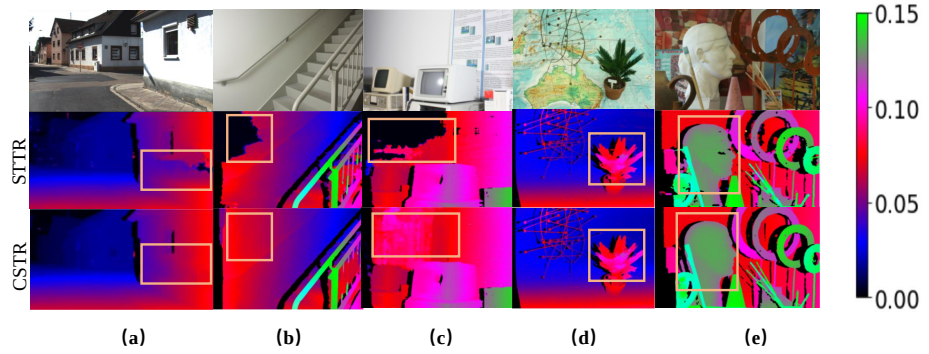


**Fig. 1.** Sample visualizations of hazardous regions taken from KITTI-2015 and Middlebury-2014 datasets. First row is the input left images. Second row is the disparity predicted by Stereo Transformer (STTR) [17]. Third row is the disparity predicted by our proposed Context Enhanced Stereo Transformer (CSTR). The color map shown on the right is based on the disparity value relative to the image width.


**Fig. 2.** Examples of hazardous regions including: (a) Texturelessness: the wall and the ceiling in the room a (b) Specularity: the screen of a TV (c) Transparency: the sliding door (d) Disparity jumps: objects such as bamboos, fences and plants give frequent disparity discontinuities. Images are from Zhang *et al*. [37].

art generalisation results on several popular datasets, including Middlebury-2014-2014[26], KITTI-2015 [24], and MPI-sintel [1].

## 2 Related Work

**Rectified stereo depth estimation** obtains per-pixel depth from the left and right frames provided by the binocular camera. It has a wide range of applications in robotics, autonomous driving, scene understanding, 3D modeling, *etc*. In contrast to the success of deep learning in many high-level vision problems, low-level deep learning algorithms for vision tasks are still in their early stages [15]. In the field of stereo depth estimation, many works aim to improve a single step of the classical pipeline by replacing it with a deep learning module [15,38], where the quality of cost volume directly determines the accuracy of the disparity map. Chen *et al*. proposed Deep Embed to learn a cost function from

different windows by processing multi-patches at different resolutions [4]. After cost volume computation, cost aggregation is essential for gathering large context information from the huge cost volume. One of the most popular cost aggregation techniques is Semiglobal Matching (SGM) [12]. A global energy function related to the disparity map is set to minimize this energy function to solve the optimal disparity of each pixel. The raw disparity map should be refined by a post-processing algorithm.

Although there are still several remaining challenges, recently, end-to-end deep learning begin to be used in binocular stereo depth estimation and dominate dense disparity estimation in several well-known benchmarks. In order to keep memory feasible and inference speed manageable, many researchers adopt 2D convolution-based methods. These architectures always contain a self-design layer namely correlation layer in charge of computing correlation scores between left and right features. Mayer *et al.* proposed an encoder-decoder architecture based on U-net named DispNet [23]. Some researchers adopt 3D convolutions in stereo matching which take a 4D tensor (disparity range, height, weight, feature) as the input and directly process a matching volume-like representation. Chang *et al.* proposed Pyramidal Stereo Matching network (PSMNet) to integrate Spatial Pyramidal Pooling layers (SPP) in the feature extractor [2]. However, these methods lead to large computational costs, such as huge memory cost and low inference speed. Besides, the disparity range of the conventional methods are limited, preventing them to be used in many cases when the scenes are close to the camera. Recently, Li *et al.* use a sequence-to-sequence perspective to replace cost volume construction with dense pixel matching [17]. Lipson *et al.* unify stereo and optical flow approaches and utilize GRU to iteratively generate the final disparity map [19]. Others [18,16] exploit auxiliary information for detph estimation. However, stereo depth estimation is still limited by difficulties like textureless surfaces, disparity jumps, and occlusions.

**Hazardous Regions** Most of stereo algorithms rely on the following basic assumptions [37]: (1) well-textured local surface for feature extraction without large homogeneous regions; (2) single image layer assumption with only Lambertian surface; (3) the disparity varies slowly and smoothly in space without sudden jumps. However, as shown in Figure 2, these assumptions can easily be broken in many real world scenarios. For example, textureless regions like large wall are commonly seen and specular surfaces will create multiple image layers. Furthermore, disparity jumps can break the local smoothness assumption. The aforementioned regions are called hazardous regions [35]. In this work, we specifically study these commonly seen yet challenging scenarios for more robust stereo depth estimation.

**Efficient Attention** Attention has a good ability to capture correspondence between two sequences and solves the problem that RNN cannot be calculated in parallel[30]. There are many successful applications that adopt attention to encode long-range sequences [3]. Recently, attention has been applied to extract non-local features in computer vision and led to SOTA performance for many vision tasks [7]. However, it is computational expensive when the input of attention

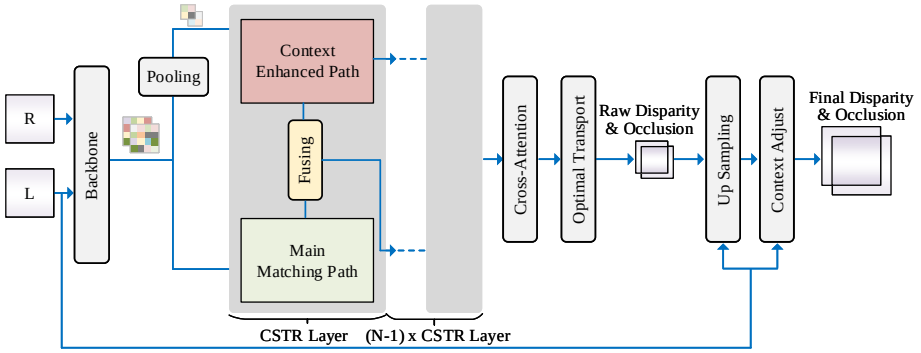


**Fig. 3.** CSTR consists of two main components:(1) Context Enhanced Path that extracts long-range context information in low resolution feature. (2) Main Matching Path that use Axial-Attention to enhance context and Cross-Attention to compute raw disparity. Then a learnable Up Sampling block up restore the original scale of disparity and Context Adjustment block refines the disparity with context information across epipolar lines conditioned on the left image.

module is large. In order to reduce its complexity, efficient attention approaches have been proposed. Yang *et al.*incorporate coarse-grained global attention and fine-grained local attentions depending on the distance to the token [33] .

**Axial Attention** Wang *et al.* factorize 2D self-attention into two 1D self-attentions to propose Axial-Attention [31]. In this paper, we adopt Axial-Attention to enhance context of feature before pixel matching. Most previous works proposed efficient attention by adding various local constraints. However, these constraints always sacrifice the global context and limit the attention's receptive field.

To ensure both efficient computation and global context, Wang et al. employ two Axial-Attention layers consecutively for the height-axis and width-axis, respectively[31]. A width-axis attention layer can be described as:

$$y_i = \sum_{j \in N_{(W*1)}(i)} S(q_i^T k_i + q_i^T r_{j-i})(v_j) \tag{1}$$

where $N_{w*1}(i)$ is the $w * h$ scale 1D region around $i$ stands for relative position encoding, and $q, k, v, S$ denote query, key, value, soft-max, respectively. In practice, $w * h$ is much smaller than the full feature shape.

Compared with local constraints attention, width-Axial-Attention computes the attention line by line with weight sharing. $W$ is equal to the width of input. Height-axis attention is the same as width-Axial-Attention besides computing the attention column by column.

Furthermore, positional information is critical for pixel matching, especially in large textureless regions. Due to shift-invariance in an image, we adopt relative position encoding to add data-only-dependent spatial information. A classical

attention mechanism with relative position encoding can be described as follows.

$$\begin{aligned} a_{i,j} = x_i W_q W_k^T x_j^T + x_i W_q W_k^T p_j^T \\ + p_i W_q W_k^T x_j^T + p_i W_q W_k^T p_j^T \end{aligned} \tag{2}$$

In Equation (2), the four terms for addition represent content-content, content-position, position-content, position-position, respectively. However, disparity computation mainly depends on the image content. To remove redundancy and ensure efficiency, we delete the last term in Equation (2) and the equation becomes:

$$a_{i,j} = x_i W_q W_k^T x_j^T + x_i W_q W_k^T p_j^T + p_i W_q W_k^T x_j^T \tag{3}$$

In the field of NLP, a similar design is adopted in DeBERT[11] and it is found that most tasks only require relative position information.

## 3 Context Enhanced Path

We propose a plug-in module, Context Enhanced Path (CEP), that provides additional context information to help stereo matching model to better understand the global structure of the input images. The goal of CEP is to maintain the context features for left and right images, and provide the context features to the Main Matching Path as additional complementary information. The detailed structure of the CEP is shown in Figure 4. As a layer-by-layer module, CEP first obtains the context feature from the previous CEP layer. Then, the Axial-Attention layer and the Cross-Attention layer are applied to further process the context features. The processed context features are served as the complementary information used for fusing with the Main Matching Path. Finally, we generate the context features as the input of the next CEP layer with 3 different strategies ($M_1$, $M_2$, $M_3$). From $M_1$ to $M_3$, the enhancement of the context information extraction increases sequentially. In the $M_1$, we only use low-level features to extract context information. Specifically, the features output by backbone only go through one layer of Axial-Attention and one layer of Cross-Attention before fusing with the main matching path. Compared to $M_1$, $M_2$ extract higher-level context information. In the $M_2$, the features output by backbone go through $L$ layers of Axial-Attention and one layer of Cross-Attention before being fused to the $L$-th layer of the main matching path. In the $M_3$, the features output by backbone go through $L$ layers of Axial-Attention and $L$ layers of Cross-Attention before the fusion.

## 4 Context Enhanced Stereo Transformer

Based on our proposed Context Enhanced Path, we further propose a transformer-based stereo depth estimation model, Context Enhanced Stereo Transformer (CSTR). We will first introduce the architecture of CSTR, and then introduce each component in detail.

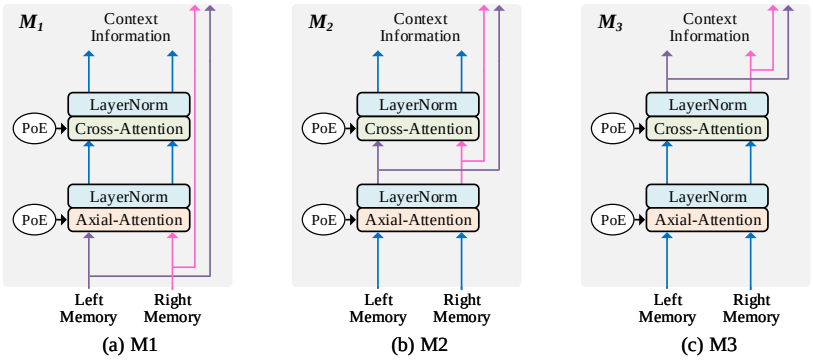


**Fig. 4.** Three different design choices for Context Enhanced Path (CEP). All strategies are composed of Axial-Attention and Cross-Attention, but the feature fed to next layer is different.

### 4.1 Pipeline

The architecture of CSTR is shown in Figure 3. The whole pipeline is mainly following the architecture of STTR [17] but the context information is enhanced. Given the pair of left (L) and right (R) input images, a convolution-based backbone is used to extract the left and the right features separately. The pair of left and right features then processed by several CSTR layers to obtain the disparity map with a coarse-to-fine manner. In each CSTR layer, there are 3 critical modules (Context Enhanced Path, Main Matching Path, and the path fusion module) that helps to incorporate the context information for generating better disparity map. The Context Enhanced Path is discussed in Section 3, the other two modules, Main Matching Path and the path fusion module, will be explained in detail in the rest part of this section. Finally, we apply several post-processing modules (*e.g.*, optimal transport layer, upsampling layer, and context adjust layer) to obtain the final disparity.

### 4.2 Main Matching Path

Main Matching Path is similar to the Transformer module from STTR [17], which includes a self-attention module followed by a Cross-Attention module as shown in Figure 6 (a). The self-attention module is used to aggregate the information in the same image, while the Cross-Attention module is used to compute the similarity of pixels from the different images. Note that the self-attention module only computes attention between pixels along the *same* epipolar line in the same image, leading to difficulty to collection contextual information from other epipolar lines.

To help the model gather more context information, as shown in Figure 6 (b), we replace the original self-attention layer to an Axial-Attention layer, including

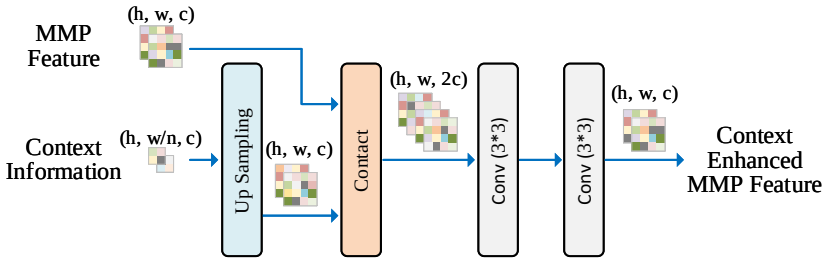


**Fig. 5.** Path Fusion Module. (h, w,c) represent height, width and the number of feature channel. (1)The MMP feature is first concatenated with the upsampled context feature. (2) Two convolution layers are applied to the concatenated feature to aggregate the context information to the Main Matching Path(MMP) features. (3) We use the fused feature as the input of the next MMP module

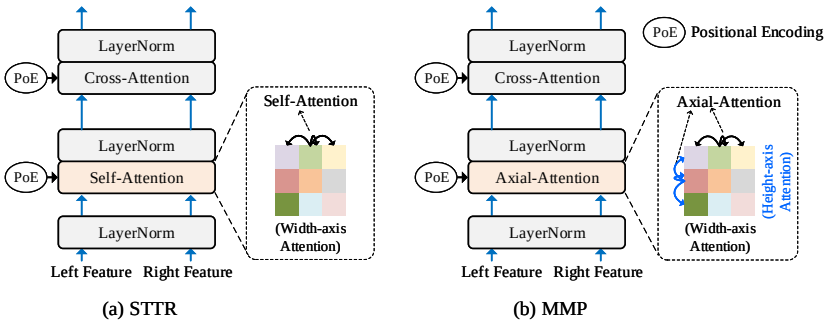


**Fig. 6.** (a)Overview of the Matching Path in STTR. (b) Overview of the Main Matching Path with alternating Axial-Attention and Cross-Attention in CSTR.

a horizontal Axial-Attention module and a vertical Axial-Attention module, to collect the context information from both horizontal and vertical axials.

### 4.3 Path Fusion Module

The path fusion module aims to fuse the context feature from the Context Enhanced Path (CEP) to the main matching features in the Main Matching Path (MMP). This will keep the main matching path capturing long-range context from low-resolution features. The architecture is shown in Figure 5. Specifically, the MMP feature is first concatenated with the upsampled context feature. Then, two convolution layers are applied to the concatenated feature to aggregate the context information to the main features. Finally, we use the fused feature as the input of the next main matching path module.

### 4.4 Other Important Modules

This section discuss other important modules in our pipeline, including Attention Mask, Optimal Transport, Raw Disparity and Occlusion Computation. Other details are illustrated in Section 5.2.

**Attention Mask and Optimal Transport** We further compress the pixels' matching space based on the following two observations.

First, when a point in the physical world is imaged by a binocular camera, the imaging position in the left image will be more to the left than the imaging position in the right image. Let us denote the $P_L$, $P_R$ as the imaging point of a real point in the left and right image. Then the following formula always holds:

$$P_L - P_R <= 0 \tag{4}$$

Therefore, the point at $P_L$ in the left image should just match the candidate point at $P > P_L$ in the right image.

Second, every pixel in the left image can only match one pixel in the right image which is called uniqueness constraint. We adopt entropy-regularized optimal transport [5] to implement such constraints in a soft way. Entropy-regularized optimal transport is proposed to improve the network performance in a similar task of semantic correspondence matching [21]. In the following section, we denote the optimal transport assignment matrix as $T$ which contains a correlation score of pixels in two images.

**Raw Disparity and Occlusion Computation** In order to improve the model robustness in multi-modal distributions, we use a small number of candidate disparity in a local region rather than use all candidate disparity. First, we compute raw disparity by finding the location($S_h$) of the highest correlation score. Then,a 3 px window $N_{3x3}(S_h)$ is built around $S_h$ in matrix $T$ to regress raw disparity. $t$ is used to represent correlation score in $N_{3x3}(S_h)$. The raw disparity regression can be described as:

$$\sum_{i \in N_{3*3}(S_h)} t_i = 1, i \in N_{3*3}(S_h) \tag{5}$$

$$\overline{t_i} = \frac{t_i}{\sum_{i \in N_{3*3}(S_h)}}, i \in N_{3*3}(S_h) \tag{6}$$

$$\overline{d_{raw}}(S_h) = \sum_{i \in N_{3*3}(S_h)} d_i \overline{t_i} \tag{7}$$

where $\overline{d_{raw}}$ represents regressed raw disparity and $d_i$ denotes the raw disparity in $N_{3*3}(S_h)$. Occlusion probability($p_{occ}(S_h)$) can be interpreted as the probability that one pixel has no matching pixel in another image. Thus it can be described as:

$$p_{occ}(S_h) = 1 - \sum_{i \in N_{3*3}(S_h)} t_i \tag{8}$$


**Fig. 7.** Results on KITTI-2015, Middlebury-2014, MPI Sintel in zero-shot synthetic-to-real setting. Black represent occlusion. The color map is the image width $\times$ 0.2 and is shown on the right which used to visualize disparity.

## 5 Experiments

### 5.1 Experimental Settings

**Datasets.** We evaluate CSTR on four popular but diverse datasets: Scene Flow [23], KITTI-2015 [24], Middlebury [26], and MPI Sintel [1]. These datasets contain random objects, real street scene, indoor scene, and realistic artifacts, respectively. Scene Flow is a synthetic dataset of random object with many subset. We use FlyingThings3D subset with 21818 training samples (960×540) in the experiment. KITTI-2015 contains stereo videos of road scenes from a calibrated pair of cameras mounted on a car with 200 training samples (1242×375). MPI Sintel contains sufficiently realistic scenes including natural image degradations such as fog and motion blur with 1064 training samples (1024×436).

**Evaluation Metric.** We use both EPE (end-point-error) and 3 px Error (percentage of EPE $> 3$) as evaluation metrics. we use Intersection over Union (IOU) to evaluate occlusion estimation. In the rest of this Section, we report the results for the non-occluded regions.

### 5.2 Implementation Details

CSTR is implemented in Pytorch [25] and is trained using one Tesla A100 GPU. During training, we use the AdamW[22] optimizer with weight decay of 1e-4. We pre-train on Scene Flow for 17 epochs using a fixed learning rate of 1e-4 for the CSTR layer and backbone, and 2e-4 for the context adjustment layer.

**Feature Extractor** In order to efficiently extract both global and local context information, we adopt an hourglass-shaped feature extractor composed of encoding and decoding paths. The encoding path is based on spatial pyramid pooling [2] modules and residual blocks [10] while the decoding path consists of dens-blocks [13], transposed convolution layer, a final average pooling layer for generating multi-scale features. The scale of feature map output by transposed convolution layer is at 1/4 resolution as the input image. For an input


**Fig. 8.** KITTI-2015's ground truth is missing part of occlusion. However CSTR can accurately give this part of the missing occlusion. First row are left image and ground truth. Second row are right image and our predicted disparity.

like ($H$,$W$), we generate a multi-scale feature($H$,$W/2K$) with repeating average pooling of the width dimension, where $K$ is the down sample rate.

**Supervision** Motivated by Relative Response loss $L_{rr}$[20], we split assignment matrix $T$ to matched pixel sets $M$ and unmatched pixel sets $U$. The loss can be described as:

$$t_i^* = LinerInterp(T_i, p_i - d_{gt,i}) \tag{9}$$

$$L_{rr} = \frac{1}{N_M} \sum_{i \in M} -log(t_i^*) + \frac{1}{N_U} \sum_{i \in M} -log(t_{i,\Phi}) \tag{10}$$

where $t_i$ stands for $i$-th matching probability and $d_{gt,i}$ represents $i$-th ground truth disparity. To accelerate the convergence of the model, we adopt smooth L1 [6] on both raw and final disparities. Furthermore, we use a binary-entropy loss to supervise the occlusion map. The total loss $L$ is computed as:

$$\begin{aligned} L = w_1 L_{rr,raw} + w_2 L_{d1,raw} + \\ w_3 L_{d1,final} + w_4 L_{be,final} \end{aligned} \tag{11}$$

where $L_{rr,raw}$, $L_{d1,raw}$, $L_{d1,final}$, $L_{be,final}$ represent Relative Response loss on raw disparity, L1 loss on raw disparity, L1 loss on final disparity,binary-entropy loss on final occlusion, respectively.

**Hyperparameters.** In our experiments, we use 6 CSTR layers with feature of 128 channels. We use multi-head attention with 4 heads. The resolution of feature in MMP is set to 1/4 of full resolution. Sinkhorn algorithm is run for 10 iterations [5].

**Baselines.** In this work, we compare CSTR with prior work based on different learning-based stereo depth paradigms: **PSMNet** [2] is a 3D convolution-based model consists two main modules — spatial pyramid pooling and 3D CNN; **AANet**[32] is a correlation-based model which is proposed to replace 3D convolutions to realize fast inference speed while ensure comparable accuracy;

**Table 1.** Generalization experiment. The models are only trained on Scene Flow without fine-tuning on MPI Sintel, KITTI-2015, Middlebury-2014 dataset. **Bold** is the best result.

| | Middlebury 2014(varies) | | | MPI Sintel† (1024 * 436) | | | KITTI-2015 (1242 * 375) | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3px Error ↓ | EPE ↓ | Occ IOU ↑ | 3px Error ↓ | EPE ↓ | Occ IOU ↑ | 3px Error ↓ | EPE ↓ | Occ IOU ↑ |
| AANet | 6.29 | 2.24 | Null | 9.57 | **1.71** | Null | 7.06 | 1.31 | Null |
| PSMNet | 7.93 | 3.70 | Null | 10.24 | 2.02 | Null | 7.43 | 1.39 | Null |
| GwcNet-g | 5.83 | 1.32 | Null | 6.60 | 1.95 | Null | 6.75 | 1.59 | Null |
| RAFT-Stereo | 7.57 | 1.21 | Null | 13.02 | 17.36 | Null | **5.68** | **1.10** | Null |
| STTR | 6.19 | 2.33 | 0.95 | 5.75 | 3.01 | 0.86 | 6.74 | 1.50 | 0.98 |
| CSTR (Ours) | **5.16** | **1.16** | **0.95** | **5.51** | 2.58 | **0.92** | 5.78 | 1.43 | **0.98** |

**GwcNet-g**[8] is a correlation and 3D convolution hybrid approach which constructs the cost volume by group-wise correlation; **STTR**[17] is a transformer-based model which revisits the problem from a sequence-to-sequence correspondence perspective to replace cost volume construction; **RAFT-stereo**[19] is a state-of-the-art recurrent model on Middlebury-2014 and Scene Flow datasets using iterative refinement to compute disparity.

### 5.3 Zero-Shot Generalization

We compare the zero-shot generalization ability between our proposed CSTR and previous popular stereo depth estimation methods. Specifically, the models are trained on the SceneFlow synthetic dataset, and then test on real data such as KITTI-2015 (real outdoor scene), Middlebury-2014 (real indoor scene), and MPI Sintel (Synthesized complex game scenes).

The results are shown in Table 1. Our model CSTR is better than our baseline method STTR [17] on all datasets and on all different metrics. For example, compared with the STTR baseline, the 3px error on Middlebury dataset is improved from 6.19 to 5.16. These improvement shows that the design of context extraction of our network facilitates generalization.

Besides, our model achieves the best results on both Middlebury 2014 and MPI Sintel datasets compared with previous methods. The quantitative results on KITTI-2015 dataset is not as good as RAFT-Stereo. Compared with RAFT-Stereo, the 3px error is dropped from 5.68 to 5.78. However, by visualizing and comparing the ground-truth label and the output of CSTR, we observe that our predicted results are even more precise than the grounding truth in the occlusion areas. See Figure 8 for more details.

### 5.4 Ablations

In Table 2, we provide quantitative results for the effects of the Axial-Attention and three different context extraction strategies. All ablated models are trained on FlyingThings of Scene Flow. Below we describe each of the experiments in more detail.

**Table 2.** Ablation generalization experiments.The model only trained on Scene Flow without fine-tune. Following prior work, we validate on the Scene Flow test set. STTR: Stereo Transformer; MMP: Main Matching Path with Aixal-Attention; $M_1$,$M_1$,$M_1$ are three different Context Enhanced Path.

| Experiment | | | | | Scene Flow | | | Middlebury-2014 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| STTR | MMP | $M_1$ | $M_2$ | $M_3$ | 3px Err | EPE | IOU | 3px Err | EPE | IOU |
| ✓ | | | | | 1.54 | 0.50 | 0.97 | 6.93 | 2.24 | 0.95 |
| ✓ | ✓ | | | | 1.28 | 0.43 | 0.98 | 5.55 | 2.03 | 0.95 |
| ✓ | ✓ | ✓ | | | **1.18** | 0.42 | 0.98 | 5.47 | 1.44 | 0.95 |
| ✓ | ✓ | | ✓ | | 1.20 | 0.42 | 0.98 | 5.38 | 1.60 | 0.95 |
| ✓ | ✓ | | | ✓ | 1.20 | **0.42** | **0.98** | **5.13** | **1.16** | **0.95** |

**Main Matching Path** Main Matching Path adopt Axial-Attention which factorizing 2D self-attention into two 1D self-attentions rather than the stand-alone self attention. This allows performing attention in a larger region to extract context information with acceptable computation cost. As shown in Table 2, comparing with STTR which adopt 1D attention on epipolar, Main Matching Path have better EPE and 3px Err. Especially, it reduce EPE by 17% and reduce 3px Err by 14% on Scene Flow. This improvement shows that the global information extracted by Axial-Attention is benefit to stereo matching.

**Three Context Enhanced Path Strategies** We design three different context enhanced strategies($M_1$,$M_2$,$M_3$) that extract the context from low resolution features to enhanced the Main Matching Path. $M_1$,$M_2$,$M_3$ improve the result of EPE and 3px Error on Scene Flow, especially, $M_3$ achieves an EPE 7% reduction. As it has been approved, these context enhanced strategies of CEP further impove the stereo matching performance. The result on Scene FLow of three strategies are verly similar, we will further compare their real-world generalization performance and robustness in following Section.

**Real World Generalization Experiment** We evaluate the generalization performance of MMP and three CEP on Middlebury-2014. The model only trained on FlyingThings of Scene Flow. As listed in Table 2, the MMP significantly outperform the baseline setting STTR which only has 1D self-attentions epipolar. The Axial-Attention for global information extraction reduce the EPE and the 3px Err by 20% and 9% on Middlebury-2014. This shows the global information is critical for model's generalization.

All three context enhanced strategies outperform the MMP. It shows that the context information provided by CET facilitates generalization. $M_3$ achieves the best results on both EPE and 3px Err. For example, compared with MMP, $M_3$ reduce the 3px error by 7% and even reduce the EPE by 42%. This proves that the design of our CEP is important for enhancing generalization performance. Finally, compared with STTR, the best setting of CSTR with Axial-Attention and CEP of $M_3$ reduce the 3px error by 26% and even reduce the EPE by 48%.

**Table 3.** EPE results of Ablation and Rubostness experiments.The model only trained on Scene Flow without fine-tune. Hazardous Data is a dataset that label the hazardous regions in KITTI-2015[37]; SPL: Specularity ; TEL:Texturelessness; TRS:Transparency.

| Experiment | | | | | Hazardous Data | | | |
|---|---|---|---|---|---|---|---|---|
| STTR | MMP | $M_1$ | $M_2$ | $M_3$ | SPL | TEL | TRS | AVG |
| ✓ | | | | | 5.43 | 10.42 | 7.03 | 7.63 |
| ✓ | ✓ | | | | 4.98 | 8.59 | 6.8 | 6.79 |
| ✓ | ✓ | ✓ | | | 4.75 | 10.74 | 7.64 | 7.71 |
| ✓ | ✓ | | ✓ | | **3.68** | 11.28 | 7.01 | 7.32 |
| ✓ | ✓ | | | ✓ | 4.54 | **8.21** | **6.01** | **6.25** |

**Robustness Against Hazardous regions** The images regions like textureless-ness, transparency, specularity are likely to cause the failure of an algorithm, namely hazardous regions[35]. Zhang *et al.* [37] lable the hazardous regions in KITTI-2015 and we use it to provide quantitative results for the effects of the MMP, three CEP strategies summarized in Table 3. Using Axial-Attention instead of stand alone self-attention can effectively improve average EPE of harzardous regions in with 11%, especially on Textureless regions with 17%. Using $M_1$ or $M_2$, which memory feature just past one Cross-Attention layer, lead to a decrease in average EPE. However, $M_3$ which are used in our final CSTR, can bring additional 8 % improvement in average EPE compared with MMP. This may be because $M_3$ uses the same number of Cross-Attention layers as MMP, which is beneficial for MMP to better integrate context information.

## 6 Conclusions

Current stereo depth estimation models usually fail to handle the hazardous regions. In this paper, we found using global context information mitigate this issue. Therefore, we proposed a plug-in module, Context Enhanced Path. Based on CEP, we then built a stereo depth estimation model, Context Enhanced Stereo Transformer. According to our experimental results, our method achieves strong cross dataset generalization ability, handles hazardous regions robustly, and provides accurate occlusion prediction.

**Acknowledgments** This paper is supported by Key-Area Research and Development Program of Guangdong Province (Grant No. 2019B010155003), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2020B1515120044, 2020A1515110495), Johns Hopkins University internal funds, ONR award N00014-21-1-2812, and NIH award K08DC019708.

## References


1. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. pp. 611–625. Springer (2012)
2. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5418 (2018)
3. Chaudhari, S., Mithal, V., Polatkan, G., Ramanath, R.: An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology (TIST) **12**(5), 1–32 (2021)
4. Chen, Z., Sun, X., Wang, L., Yu, Y., Huang, C.: A deep visual correspondence embedding model for stereo matching costs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 972–980 (2015)
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
6. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
7. Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.: Attention mechanisms in computer vision: A survey. arXiv preprint arXiv:2111.07624 (2021)
8. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3273–3282 (2019)
9. Hartmann, W., Galliani, S., Havlena, M., Gool, L.V., Schindler, K.: Learned multi-patch similarity. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. IEEE (2016)
11. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv (2020)
12. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence **30**(2), 328–341 (2007)
13. Huang, G., Liu, Z., Laurens, V., Weinberger, K.Q.: Densely connected convolutional networks. IEEE Computer Society (2016)
14. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. IEEE (2018)
15. Laga, H., Jospin, L.V., Boussaid, F., Bennamoun, M.: A survey on deep learning techniques for stereo-based depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
16. Li, Z., Drenkow, N., Ding, H., Ding, A.S., Lu, A., Creighton, F.X., Taylor, R.H., Unberath, M.: On the sins of image synthesis loss for self-supervised depth estimation. arXiv preprint arXiv:2109.06163 (2021)
17. Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M.: Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: 2021 IEEE International Conference on Computer Vision (ICCV) (2021)
18. Li, Z., Ye, W., Wang, D., Creighton, F.X., Taylor, R.H., Venkatesh, G., Unberath, M.: Temporally consistent online depth estimation in dynamic scenes. arXiv preprint arXiv:2111.09337 (2021)

19. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021)
20. Liu, X., Zheng, Y., Killeen, B., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Extremely dense point correspondences using a learned feature descriptor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4847–4856 (2020)
21. Liu, Y., Zhu, L., Yamada, M., Yang, Y.: Semantic correspondence as an optimal transport problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4463–4472 (2020)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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. pp. 4040–4048 (2016)
24. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3061–3070 (2015)
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
26. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German conference on pattern recognition. pp. 31–42. Springer (2014)
27. Schmid, K., Tomic, T., Ruess, F., Hirschmüller, H., Suppa, M.: Stereo vision based indoor/outdoor navigation for flying robots. In: 2013 IEEE/RSJ international conference on intelligent robots and systems. pp. 3955–3962. IEEE (2013)
28. Sun, J., Chen, L., Xie, Y., Zhang, S., Jiang, Q., Zhou, X., Bao, H.: Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10548–10557 (2020)
29. Tomono, M.: Robust 3d slam with a stereo camera based on an edge-point icp algorithm. In: 2009 IEEE International Conference on Robotics and Automation. pp. 4306–4311. IEEE (2009)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12349 LNCS**, 108–126 (2020)
32. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1959–1968 (2020)
33. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641 (2021)

34. Yao, C., Jia, Y., Di, H., Li, P., Wu, Y.: A decomposition model for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6091–6100 (2021)
35. Zendel, O., Murschitz, M., Humenberger, M., Herzner, W.: Cv-hazop: Introducing test data validation for computer vision. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2066–2074 (2015)
36. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 185–194 (2019)
37. Zhang, Y., Qiu, W., Chen, Q., Hu, X., Yuille, A.: Unrealstereo: Controlling hazardous factors to analyze stereo vision. In: 2018 International Conference on 3D Vision (3DV). pp. 228–237. IEEE (2018)
38. Zhao, C., Sun, Q., Zhang, C., Tang, Y., Qian, F.: Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences **63**(9), 1612–1627 (2020)