

# Iterative Geometry-Aware Cross Guidance Network for Stereo Image Inpainting

Ang Li<sup>1</sup>, Shanshan Zhao<sup>2</sup>, Zhang Qingjie<sup>1</sup> and Qihong Ke<sup>3\*</sup>

<sup>1</sup>Aviation University of Air Force

<sup>2</sup>JD Explore Academy, JD.COM

<sup>3</sup>The University of Melbourne

angli.cs@outlook.com, sshan.zhao00@gmail.com, nudtzhang@hotmail.com,  
qihong.ke@unimelb.edu.au

## Abstract

Currently, single image inpainting has achieved promising results based on deep convolutional neural networks. However, inpainting on stereo images with missing regions has not been explored thoroughly, which is also a significant but different problem. One crucial requirement for stereo image inpainting is stereo consistency. To achieve it, we propose an Iterative Geometry-Aware Cross Guidance Network (IGGNet). The IGGNet contains two key ingredients, i.e., a Geometry-Aware Attention (GAA) module and an Iterative Cross Guidance (ICG) strategy. The GAA module relies on the epipolar geometry cues and learns the geometry-aware guidance from one view to another, which is beneficial to make the corresponding regions in two views consistent. However, learning guidance from co-existing missing regions is challenging. To address this issue, the ICG strategy is proposed, which can alternately narrow down the missing regions of the two views in an iterative manner. Experimental results demonstrate that our proposed network outperforms the latest stereo image inpainting model and state-of-the-art single image inpainting models.

## 1 Introduction

Image inpainting aims to synthesize alternative contents for the missing regions of an image while simultaneously preserving the overall structure. It has a wide range of applications such as photo recovery and object removal. In recent years, relying on the powerful representation capabilities, deep learning based methods [Liu *et al.*, 2018; Yu *et al.*, 2019; Li *et al.*, 2020; Zeng *et al.*, 2021] have been dominating for image inpainting and are able to generate promising results. Apart from single images, stereo images are also important media due to the use of head-mounted devices (e.g., AR/VR glasses), dual-lens smart phones, and stereo cameras. As a result, in some scenarios such as object removal and content editing, we are also required to recover the missing regions in

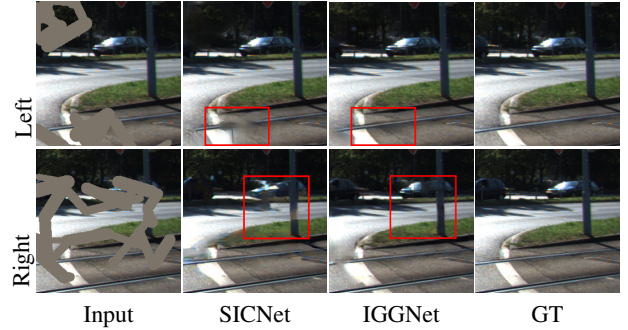


Figure 1: An example of visual comparison with state-of-the-art stereo image inpainting model SICNet [Ma *et al.*, 2020] on a stereo pair (which contains two images, i.e., left view and right view) from the KITTI2015 dataset [Mayer *et al.*, 2016]. GT is the ground truth image. As shown here, the input stereo pair can have different missing regions in each view. The proposed IGGNet is effective to incorporate geometry correlation and generate stereo consistent results.

the stereo images. However, inpainting stereo images has not been explored thoroughly yet.

A straightforward solution of stereo image inpainting is to directly apply single image inpainting methods (developed for individual images) over the left view and the right view separately. However, this may lead to severe stereo inconsistency, since inpainting an individual image only considers the undamaged spatial statistics in that image, i.e., only one view of the stereo images, but ignores the geometric relationship between the two views. Here, stereo inconsistency refers to that the inpainted stereo pair, i.e., the left and the right output images, has different structures or textures in the corresponding region. To alleviate this issue, Ma *et al.* [2020] recently propose SICNet for stereo image inpainting, which associates the two views via a feature map concatenation operation. Although simple, the concatenation cannot fully explore the stereo cues, and thus the stereo inconsistency is not addressed well. We give an example in Figure 1.

To achieve better stereo consistency, we consider the epipolar geometry existing in the stereo pair. In particular, the epipolar geometry refers to that most of contents in one view can be found in another view. Therefore, this cue can motivate us to associate the features of potential corresponding regions in the stereo pair. In this way, if a region is missing

\*Corresponding author.

in one view but exists in another, this cue can help us recover the region through querying related contents from the other view. Additionally, this cue is also helpful to make the structure and texture of corresponding regions consistent. However, for the co-existing region, *i.e.*, a region missing in both views, it is still challenging to recover the content using the epipolar-geometry-based feature association since it is difficult to model the geometric correspondence under this situation.

Motivated by the above observations and analysis, we propose a stereo inpainting network named **Iterative Geometry-Aware Cross Guidance Network (IGGNet)**. This network performs inpainting on the stereo images through exploring and integrating the stereo geometry in an iterative manner. IGGNet consists of two key ingredients, *i.e.*, the *Geometry-Aware Attention (GAA)* module and the *Iterative Cross Guidance (ICG)* strategy. Relying on the epipolar geometry cues, the GAA module associates the two views through explicitly learning attention maps. In such manner, it is capable of adaptively capturing the useful information from the reference view to guide the restoration of the target view without the requirement of disparity information. The ICG strategy is implemented to alternately narrow down the missing regions in the two views based on a confidence-driven mask updating policy. It is beneficial to alleviate the impact of co-existing missing regions and thus obtain better visual consistency in the inpainted results. More specifically, in a certain iteration, we take one view as the target view (the view to be inpainted) and the other one as the reference (guidance). Then, given the inpainted target view obtained in this iteration and the reference view, we swap their roles (reference and target), apply the updated masks to them, and repeat such cross-guidance process for the following iterations. By processing the two views using the geometry-aware cross guidance in an iterative manner, our model is able to yield high-quality stereo image inpainting results with better geometry consistency.

In summary, our main contributions are:

- We propose a novel end-to-end network for stereo image inpainting, which models the association between the two views through learning geometry-aware guidance and crossly performs inpainting in an iterative manner.
- We introduce a *Geometry-Aware Attention (GAA)* module that incorporates the correlation between the two views of a stereo pair based on the epipolar geometry. The GAA module relies on the epipolar geometry cue and benefits generating stereo consistent results.
- We adopt an *Iterative Cross Guidance (ICG)* strategy that exploits the geometry-aware guidance to alternately shrink the missing regions of the two views in an iterative manner. In this way, we can deal with co-existing missing regions better in the stereo pair.
- Experiments on real data demonstrate that our proposed IGGNet outperforms state-of-the-art single image inpainting methods and stereo image inpainting methods.

## 2 Related Work

### 2.1 Image Inpainting

Different from conventional image inpainting approaches, deep learning-based image inpainting models can generate more visually plausible details or fill large missing regions with new contents that never exist in the input image [Pathak *et al.*, 2016; Iizuka *et al.*, 2017; Yu *et al.*, 2018; Liu *et al.*, 2018; Li *et al.*, 2019b; Yu *et al.*, 2019; Liu *et al.*, 2019; Li *et al.*, 2019a; Li *et al.*, 2020; Yi *et al.*, 2020; Zeng *et al.*, 2021]. An advantage of these models is to adaptively synthesize visually plausible contents for different semantics. Pathak *et al.* [2016] first introduce the *Context Encoder (CE)* model where a convolutional encoder-decoder network is trained with the combination of an adversarial loss [Goodfellow *et al.*, 2014] and a reconstruction loss. Yu *et al.* [2019] present DeepFillv2 to provide a learnable dynamic mask-updating mechanism for free-form image inpainting. Zeng *et al.* [2021] propose CR-Fill with a learnable contextual reconstruction loss. Since these single image inpainting models only take one view as input and ignore the epipolar geometry correlation when dealing with stereo images, they are unable to generate stereo consistent results.

### 2.2 Stereo Image Inpainting

Several conventional methods [Wang *et al.*, 2008; Hervieu *et al.*, 2010; Morse *et al.*, 2012] have been proposed to address stereo image inpainting. However, these methods fail to generate meaningful structures when facing complex semantic scenes in the missing regions, which is also a shared limitation for conventional single image and video inpainting methods. Chen *et al.* [2019] design the first CNN-based model for stereo image inpainting based on [Pathak *et al.*, 2016]. However, this method can only deal with square holes in the centre. Ma *et al.* [2020] propose the SICNet with a cross-fusion design. However, the fusion module simply concatenates the two views' feature maps and it is unable to effectively explore the stereo cues.

## 3 Proposed Method

In this section, we first present the overall network architecture. Then, we describe the GAA module and the ICG strategy in more details, respectively. Finally, the training target functions are introduced.

### 3.1 Network Overview

Given a rectified stereo image pair  $(X_{left}, X_{right})$  annotated with binary masks  $(M_{left}, M_{right})$  which indicate the missing region, the goal of our network is to generate the output pair  $(\hat{Y}_{left}, \hat{Y}_{right})$ . The output pair should not only have visually plausible restored contents but also maintain the geometry consistency with the ground truth pair  $(Y_{left}, Y_{right})$ . To achieve this goal, we restore the two images in an iterative manner with the proposed **Iterative Cross Guidance (ICG)** strategy using the deep network containing a **Geometry-Aware Attention (GAA)** module.

Specifically, for the first iteration, we take the left view  $X_{left}$  as the target (the view to be inpainted) and the right

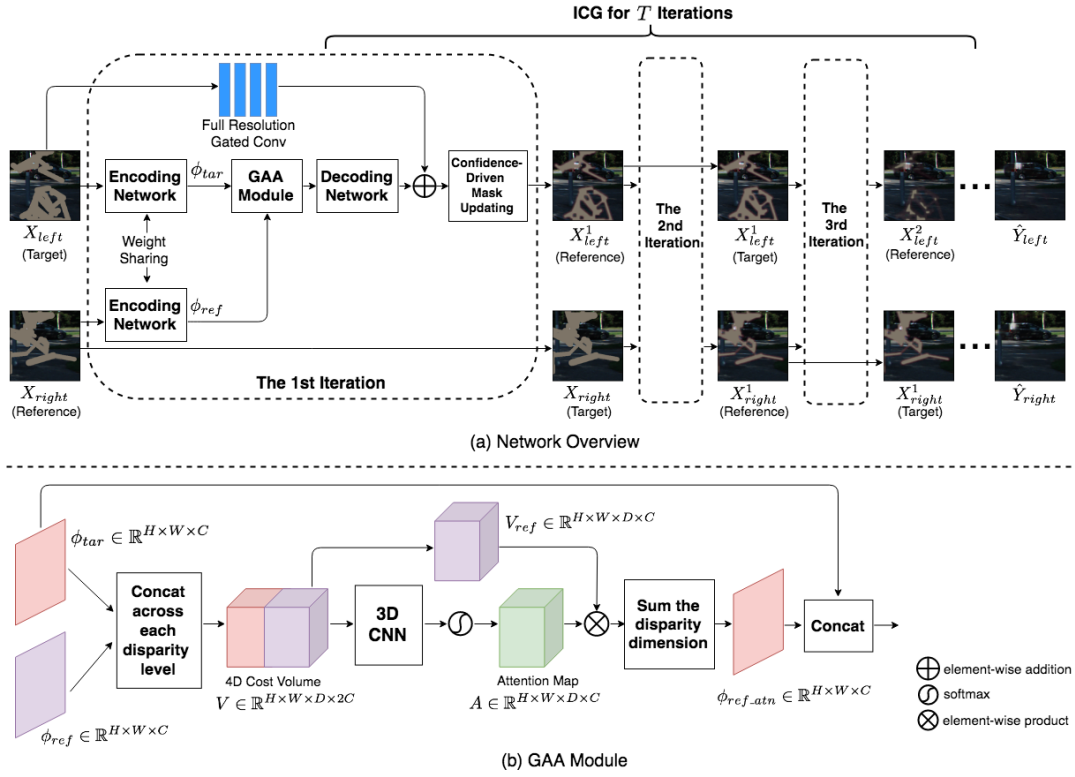


Figure 2: Our proposed IGGNet network contains two key parts, a Geometry-aware Attention (GAA) module and an Iterative Cross Guidance (ICG) strategy. The Geometry-aware Attention (GAA) module explores the geometry correlation of the stereo pair and performs attention-driven feature aggregation. The Iterative Cross Guidance (ICG) strategy exploits the guidance through a confidence-driven mask updating module to alternately shrink the missing regions of the two views.

view  $X_{right}$  as the reference (guidance). As shown in Figure 2 (a), the IGGNet includes two parallel branches, *i.e.*, an encoder-decoder branch and a full resolution branch. We apply Gated convolution [Yu *et al.*, 2019] as the basic convolution operation in both branches. For the encoder-decoder branch, both the target and reference images are fed into the encoding network, which generates the corresponding feature maps  $\phi_{tar}$  and  $\phi_{ref}$ . The two feature maps are then processed by the GAA module, which explores the geometry correspondence between them and incorporates most relevant context from  $\phi_{ref}$  to guide the inpainting of  $\phi_{tar}$ . (More details are explained in Section 3.2). Then, the decoding network decodes the aggregated feature maps (output of the GAA module) to a restored target image. In addition, the full resolution branch consists of four consecutive full-resolution Gated convolution [Yu *et al.*, 2019] layers, which are beneficial to the maintenance of high-resolution representations [Guo *et al.*, 2019]. The outputs of two parallel branches are then fused through the element-wise addition operation.

After the first iteration, we obtain the inpainted left image  $X_{left}^1$  with narrowed missing regions, as shown in Figure 2 (a). Next, we swap the roles of the two images, *i.e.*, the inpainted left image  $X_{left}^1$  as the reference and the original right images  $X_{right}$  as the target, and then feed them to the second iteration, which yields the inpainted right image  $X_{right}^1$ . In the following iterations, we continue to swap the roles of tar-

get and reference, *e.g.*,  $X_{left}^1$  (target) and  $X_{right}^1$  (reference) as input for the third iteration, and repeat the cross guidance process over  $T$  iterations in total to obtain the final restored images  $\hat{Y}_{left}$  and  $\hat{Y}_{right}$ . In the following, we detail the two main involved components.

### 3.2 Geometry-Aware Attention Module

In this part, we introduce the GAA module by taking the example of the left image as the target and the right as the reference. Since the missing content of the left image might occur in the right one, it is crucial to exploit the guidance from the right image to inpaint the left. To achieve this, we can warp the right image according to the disparity map [Liang *et al.*, 2018], which indicates the displacement between the corresponding pixels of the two views, and then directly exploit the warped image as the guidance. However, the disparity information is often unavailable, which can be obtained by traditional SGM algorithms [Hirschmuller, 2005] or deep learning based approaches [Liang *et al.*, 2018]. Both of them require extra computational and time costs. Additionally, if there are missing regions or noises in the disparity map, we also need to process the disparity map firstly. To dispense with this needless step, we examine the epipolar geometry existing in the stereo images. In specific, for a rectified stereo image pair, there only exists the horizontal displacement between the corresponded pixels in the two views. As a result, for a pixel in

one view, we can model the relationship between it and its potential correspondences through learning an attention map upon the 4D cost-volume created from the feature maps of the two views. As a result, our model is able to adaptively capture the useful information from the reference to guide the restoration of the target without the requirement of disparity information. Next, we provide more details.

As shown in Figure 2 (b), the GAA module first builds a cost volume upon the encoded feature maps  $\phi_{tar} \in \mathbb{R}^{H \times W \times C}$  and  $\phi_{ref} \in \mathbb{R}^{H \times W \times C}$ . Specifically,  $\phi_{ref}$  is shifted by different disparity levels ( $\{0, 1, \dots, D-1\}$ ) along the horizontal direction and then concatenated with  $\phi_{tar}$ , which is repeated  $D$  times to generate a 4D volume  $V \in \mathbb{R}^{H \times W \times D \times 2C}$ . Here  $D$  denotes the predefined maximum disparity value. In order to aggregate the feature information along the disparity dimension as well as the channel dimension, we first apply consecutive 3D convolutional operations on the channel dimension followed by a Softmax operation on the disparity dimension to get a 4D attention map  $A \in \mathbb{R}^{H \times W \times D \times C}$ . We then perform an attention-driven feature aggregation to update the features of the target adaptively. Specifically, we apply element-wise product on  $A$  and  $V_{ref} \in \mathbb{R}^{H \times W \times D \times C}$ , where  $V_{ref}$  represents the corresponding part of the reference view in  $V$ , and sum the elements along the disparity dimension (the third dimension) to obtain the attended reference feature map  $\phi_{ref\_atn} \in \mathbb{R}^{H \times W \times C}$ :

$$\phi_{ref\_atn} = \text{sum}(A \cdot V_{ref}, \text{dim} = 3), \quad (1)$$

where  $\cdot$  denotes the element-wise product.

Finally, we concatenate  $\phi_{tar}$  and  $\phi_{ref\_atn}$  on the channel dimension as the output of GAA module.

### 3.3 Iterative Cross Guidance

The co-existing missing regions in the stereo image pair make the inpainting task extremely challenging. When the model uses the reference view as guidance to inpaint the target view, missing regions in the reference view inevitably causes inaccurate attention computation and feature aggregation. Therefore, we propose an iterative cross guidance inpainting strategy. The whole inpainting procedure is composed of a number  $T$  of inpainting iterations, in which the two views alternately serve as the target view and reference view, respectively. Each view is inpainted by  $T/2$  times ( $T$  is supposed to be an even number). In each iteration, we shrink the missing region in the target view with a confidence-driven mask updating module. Note that we use the same IGGNet in all the iterations.

**Confidence-Driven Mask Updating** We observe that in the failure cases of existing approaches, despite the artifacts, there often exist sub-regions with plausible predictions. By trusting the plausible part and treating the remaining part as a new mask, we can run the network again and again. In this way, the mask becomes progressively smaller and the model can produce better results. Inspired by this observation, we exploit a confidence-driven mask updating module to keep the inpainted region with higher confidence in the target image but ignore those with large prediction error. To realize the mask confidence prediction, we utilize the Gated Convolution

[Yu *et al.*, 2019] in both the full-resolution and the encoder-decoder branches instead of the standard 2D convolution. Specifically, we modify the last Gated convolution layer of the two branches by setting a confidence threshold on the generated three-channel soft mask  $S \in [0, 1]^{H \times W \times 3}$ , and obtain the updated one-channel binary mask  $M \in \{0, 1\}^{H \times W \times 1}$ .

$$M(i, j) = \begin{cases} 1, & \text{if } \max_{d=0,1,2} (S(i, j, d)) > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

We represent this process as a function  $g$ , i.e.,  $M = g(S)$ .

Algorithm 1 illustrates the procedure of the ICG strategy.

---

#### Algorithm 1 Iterative Cross Guidance

---

**Require:** number of iterations  $T$ , left input image  $X_{left}$ , right input image  $X_{right}$ , full resolution branch Branch\_f, encoder-decoder branch Branch\_ed

**Ensure:** restored left image  $\hat{Y}_{left}$ , restored right image  $\hat{Y}_{right}$

```

1:  $X_{tar}^0 = X_{left}, X_{ref}^0 = X_{right}$ 
2: for  $t$  in  $1, \dots, T$  do
3:    $R_f, S_f = \text{Branch\_f}(X_{tar}^{t-1})$ 
4:    $R_{ed}, S_{ed} = \text{Branch\_ed}(X_{tar}^{t-1}, X_{ref}^{t-1})$ 
5:    $M_f = g(S_f)$  (Eq. 2)
6:    $M_{ed} = g(S_{ed})$  (Eq. 2)
7:    $M_t = M_f \cdot M_{ed}$ 
8:    $R^t = \frac{1}{2}(M_t \cdot (R_f + R_{ed}))$ 
9:    $X_{tar}^t = X_{ref}^{t-1}$ 
10:   $X_{ref}^t = R^t$ 
11:  if  $t \% 2 == 1$  then
12:     $X_{left}^{(t+1)/2} = R^t$ 
13:  else
14:     $X_{right}^{t/2} = R^t$ 
15:  end if
16: end for
17:  $\hat{Y}_{left} = X_{left}^{T/2}, \hat{Y}_{right} = X_{right}^{T/2}$ 

```

---

### 3.4 Loss Function

We train our IGGNet following a typical GAN [Goodfellow *et al.*, 2014] procedure. The loss function used for training is:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{adv} \mathcal{L}_{adv}. \quad (3)$$

Here,  $\mathcal{L}_{rec}$  and  $\mathcal{L}_{adv}$  denote reconstruction loss and adversarial loss respectively. The balancing weight  $\lambda_{adv}$  is empirically set to 0.01. The formulation of  $\mathcal{L}_{rec}$  and  $\mathcal{L}_{adv}$  are as follows:

$$\mathcal{L}_{rec} = \sum_t^{T/2} \|(X_{left}^t - Y_{left}) + (X_{right}^t - Y_{right})\|_1, \quad (4)$$

$$\begin{aligned} \mathcal{L}_{adv} = & \sum_t^{T/2} (\log(\text{Dis}(Y_{left})) + \log(1 - \text{Dis}(X_{left}^t)) \\ & + \log(\text{Dis}(Y_{right})) + \log(1 - \text{Dis}(X_{right}^t))), \end{aligned} \quad (5)$$

where  $\text{Dis}$  denotes the discriminator of GAN. We also adopt spectral normalization [Miyato *et al.*, 2018] to stabilize the convergence of the discriminator. Here the generator of the GAN is IGGNet. Details of network structures are provided in the supplementary material.



## 4 Experiments

We evaluate our method and make comparisons against state-of-the-art single image and stereo image inpainting models qualitatively and quantitatively. A comprehensive ablation study is also conducted on our proposed model.

KITTI2015									
Model	mask 0-20%			mask 20-40%			mask 40-60%		
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID
DeepFillv2	27.15	0.919	6.42	21.91	0.844	12.37	17.68	0.694	26.56
CR-Fill	27.84	0.927	5.86	22.45	0.863	11.60	17.89	0.711	25.18
SICNet	28.18	0.935	5.47	22.74	0.852	10.48	18.04	0.706	23.65
IGGNet (w/o ICG)	28.45	0.932	5.18	23.37	0.860	9.91	18.81	0.714	22.47
IGGNet	<b>29.31</b>	<b>0.941</b>	<b>4.69</b>	<b>24.20</b>	<b>0.881</b>	<b>8.07</b>	<b>19.58</b>	<b>0.723</b>	<b>20.24</b>

MPI-Sintel									
Model	mask 0-20%			mask 20-40%			mask 40-60%		
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID
DeepFillv2	28.40	0.918	5.51	23.26	0.838	13.64	18.11	0.730	25.54
CR-Fill	29.07	0.935	4.61	24.03	0.884	10.47	18.29	0.746	23.72
SICNet	29.87	0.944	3.97	24.21	0.876	8.83	18.65	0.743	22.40
IGGNet (w/o ICG)	30.25	0.947	4.08	24.72	0.885	8.47	19.50	0.754	21.64
IGGNet	<b>31.02</b>	<b>0.951</b>	<b>3.75</b>	<b>25.56</b>	<b>0.903</b>	<b>7.29</b>	<b>20.81</b>	<b>0.768</b>	<b>19.82</b>

Table 1: Quantitative comparisons on KITTI2015 and MPI-Sintel datasets under three different mask ratio settings regarding 3 metrics: PSNR (higher is better), SSIM (higher is better) and FID (lower is better). Here, “w/o” means “without”. The best results are in **bold**.

**Datasets.** We train and evaluate our model on two well-known stereo image datasets: **KITTI2015** [Mayer *et al.*, 2016] and **MPI-Sintel** [Butler *et al.*, 2012]. For the KITTI2015 dataset, we use the original training and testing splits. For the MPI-Sintel dataset, we select 16 video directories for training, and the other 7 for testing. All training images are randomly cropped to  $256 \times 256$  for every epoch. Each testing image in the KITTI2015 dataset is horizontally split to 4 images and resize them to  $256 \times 256$ , while that in the MPI-Sintel dataset is directly resized to  $256 \times 256$ . It is worth noting that we do not apply any pre-processing or post-processing operation. In our experiments, we define the maximum image disparity to be 192. Since we apply GAA module on the  $1/4$  scale level,  $D$  is set as 48.

**Mask Settings.** To simulate diverse scenarios, we use the irregular mask dataset [Yu *et al.*, 2019] that contains masks with arbitrary shapes and random locations. For each stereo image pair, we apply different masks on the two views, respectively. We use three different mask ratios (mask area relative to the entire image): 0 – 20%, 20 – 40% and 40 – 60%.

**Baseline Models.** We compare our model with two state-of-the-art single image inpainting models **DeepFillv2** [Yu *et al.*, 2019] and **CR-Fill** [Zeng *et al.*, 2021], and the latest stereo image inpainting model **SICNet** [Ma *et al.*, 2020]. For fair comparison, we train and test the baseline models and our model with the same dataset (KITTI2015 or MPI-Sintel). For DeepFillv2 and CR-Fill, we use their released code but not use their open-source pretrained weights. For SICNet, we reproduce it according to the descriptions in its paper.

### 4.1 Quantitative Results

We consider three metrics for evaluation: 1) *Peak Signal-to-Noise Ratio* (PSNR, measures image distortion), 2) *Structural Similarity* (SSIM, measures structure similarity) and 3) the *Fréchet Inception Distance* (FID, measures perceptual quality) [Heusel *et al.*, 2017]. As shown in Table 1, even without

ICG strategy, our model still achieves superior results over the baseline models, which indicates that GAA can effectively aggregate stereo correspondence into the inpainting process. When utilizing ICG strategy, our model has a more obvious lead, especially when dealing with larger masks (which contain more co-existing missing regions). This verifies that ICG can alleviate the problem of co-existing missing regions.

### 4.2 Qualitative Results

To further evaluate the visual quality of the inpainted stereo images, we show three examples from the KITTI2015 dataset and three examples from the MPI-Sintel dataset in Figure 3. Object-removal masks or irregular masks are applied on the ground truth images to obtain the corresponding input images. It is obvious that both DeepFillv2 and CR-Fill struggle to maintain the stereo consistency of the inpainted views, which is a common problem for single-image inpainting models. SICNet utilizes feature-fusion strategy to incorporate the geometry correlation during inpainting. However, the fusion module directly concatenates the two views’ feature maps, which fails to effectively explore the stereo cues. In contrast, our proposed IGGNet can generate not only vivid textures but also stereo consistent contents.

### 4.3 Ablation Study

We investigate the effectiveness of Geometry-Aware Attention (GAA) and Iterative Cross Guidance (ICG), respectively.

#### Effectiveness of GAA

As described in Sec. 3.2, the purpose of GAA is to effectively explore and integrate the geometry correspondence from  $\phi_{ref}$  to guide the inpainting of  $\phi_{tar}$ . We study the performance of our full model with two variants of GAA: 1) “**concat**”, which directly concatenates  $\phi_{ref}$  and  $\phi_{tar}$  without constructing the 4D cost volume and attention-driven aggregation. 2) “**max**”, which performs the 4D cost volume regression by directly choosing the maximum value on the disparity dimension and takes it as the input for the decoding network. The performance drop of these two variants as shown in Table 2 justifies the effectiveness of the GAA module. As further shown in Figure 4, the model with GAA (concat) suffers from inaccurate feature alignment due to ignoring geometry correlation, thus produces distorted contents. GAA (max) apparently improves the results, but it still brings negative impacts due to the simple feature aggregation. In contrast, using full GAA leads to better results with geometry consistent details.

	PSNR	SSIM	FID
concat	25.17	0.863	10.73
max	25.68	0.875	9.04
GAA	<b>26.42</b>	<b>0.897</b>	<b>7.86</b>

Table 2: Ablation study on the GAA module over the KITTI2015 dataset with mask ratio 0 – 60% and iteration number  $T = 6$ .

#### Effectiveness of ICG

The ICG strategy aims to alternately use one view as the reference to guide the inpainting of the other view in an iterative manner. We inspect the impact of the number of iterations  $T$ .

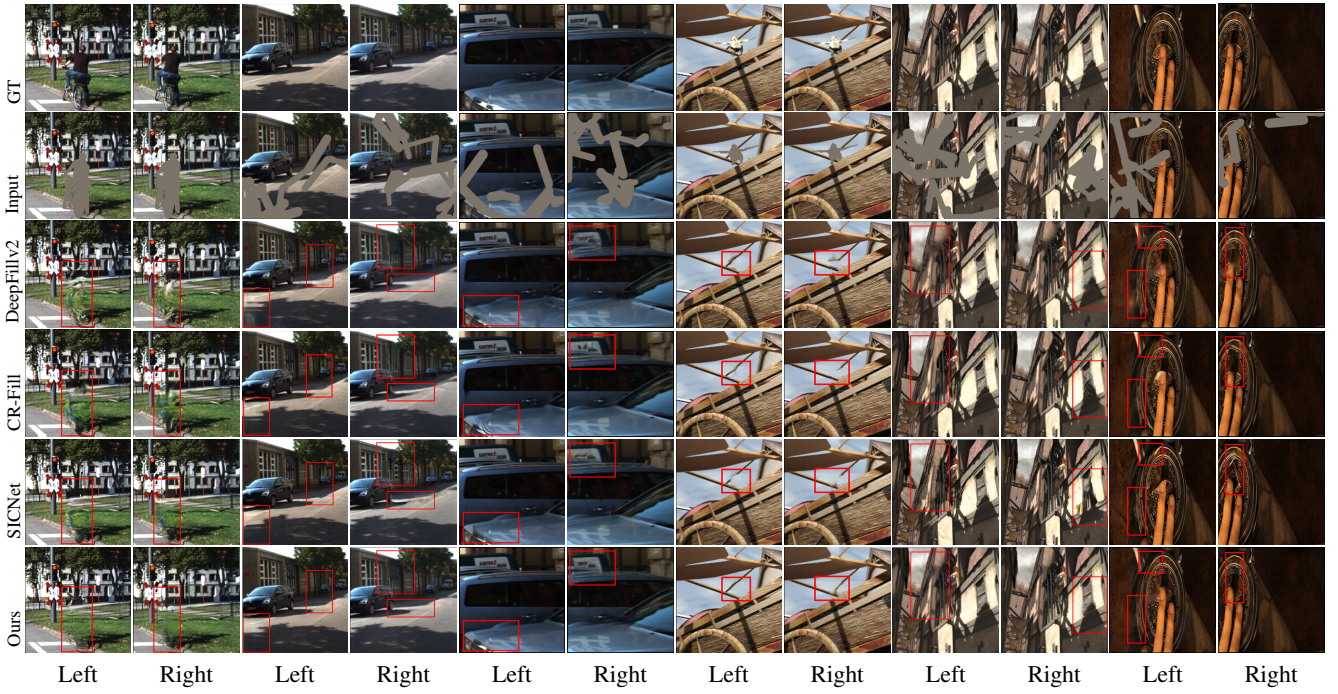


Figure 3: Qualitative comparison of our proposed model with baseline models on KITTI2015 and MPI-Sintel datasets. Better viewed at zoom level 400%.

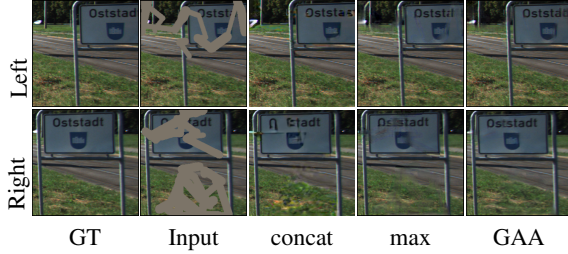


Figure 4: Ablation study on the GAA module. Better viewed at zoom level 400%.

Specifically, we trained our model with  $T = 4, 6, 8, 10$ . The evaluation results over the KITTI2015 dataset are shown in Table 3. The best results are obtained when  $T = 6$ . A small  $T$  (i.e., a small number of iterations) is insufficient to alleviate the co-existing missing region problem, leading to model performance degradation. In contrast, a large  $T$  shows no further performance improvement with the increase in computation time, because it increases the depth and complexity of the whole procedure and makes model convergence more difficult and even overfitting. An example is further shown in Figure 5. When  $T = 4$ , the inpainting result suffers from visual artifacts and blurriness. When  $T = 6$  and  $T = 10$ , our model yields visual plausible results.

## 5 Conclusion

In this work, we studied stereo image inpainting, attempting at inpainting the missing regions of the left and right image and maintaining their stereo consistency. To achieve this, we

	$T = 4$	$T = 6$	$T = 8$	$T = 10$
PSNR	25.98	<b>26.42</b>	26.35	26.17
SSIM	0.865	<b>0.879</b>	0.869	0.876
FID	8.40	<b>7.86</b>	7.93	8.22

Table 3: Ablation study on the  $T$  value of ICG over the KITTI2015 dataset with mask ratio 0 – 60%.

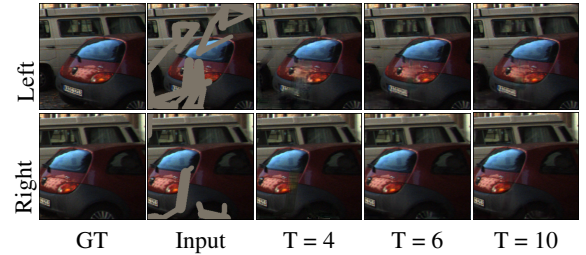


Figure 5: Ablation study on different number of iterations  $T$  of the ICG strategy. Better viewed at zoom level 400%.

first proposed a GAA module, which aims to associate the two images’ features through learning an attention map by considering the epipolar geometry. Additionally, to address the co-existing missing regions, we further proposed an ICG strategy, which alternately inpaints one of two views. The whole network is named IGGNet. Taking advantage of the proposed modules, IGGNet outperforms the latest stereo image inpainting method and state-of-the-art single image inpainting methods.

## References

- [Butler *et al.*, 2012] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision*, 2012.
- [Chen *et al.*, 2019] Shen Chen, Wei Ma, and Yue Qin. Cnn-based stereoscopic image inpainting. In *International Conference on Image and Graphics*, 2019.
- [Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2014.
- [Guo *et al.*, 2019] Zongyu Guo, Z. Chen, Tao Yu, Jianxian Chen, and S. Liu. Progressive image inpainting with full-resolution residual network. *Proceedings of the ACM International Conference on Multimedia*, 2019.
- [Hervieu *et al.*, 2010] Alexandre Hervieu, Nicolas Papadakis, Aurélie Bugeau, Pau Gargallo, and Vicent Caselles. Stereoscopic image inpainting: distinct depth maps and images inpainting. In *International Conference on Pattern Recognition*, 2010.
- [Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2017.
- [Hirschmuller, 2005] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [Iizuka *et al.*, 2017] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 2017.
- [Li *et al.*, 2019a] Ang Li, Jianzhong Qi, Rui Zhang, Xingjun Ma, and Kotagiri Ramamohanarao. Generative image inpainting with submanifold alignment. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.
- [Li *et al.*, 2019b] Ang Li, Jianzhong Qi, Rui Zhang, and Kotagiri Ramamohanarao. Boosted gan with semantically interpretable information for image inpainting. In *Proceedings of the International Joint Conference on Neural Networks*, 2019.
- [Li *et al.*, 2020] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [Liang *et al.*, 2018] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [Liu *et al.*, 2018] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision*, 2018.
- [Liu *et al.*, 2019] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE international Conference on Computer Vision*, 2019.
- [Ma *et al.*, 2020] Wei Ma, Mana Zheng, Wenguang Ma, Shibiao Xu, and Xiaopeng Zhang. Learning across views for stereo image completion. *IET Computer Vision*, 14:482–492, 2020.
- [Mayer *et al.*, 2016] 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*, 2016.
- [Miyato *et al.*, 2018] Takeru Miyato, T. Kataoka, Masanori Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *ArXiv*, abs/1802.05957, 2018.
- [Morse *et al.*, 2012] Bryan Morse, Joel Howard, Scott Cohen, and Brian Price. Patchmatch-based content completion of stereo image pairs. In *International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, 2012.
- [Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [Wang *et al.*, 2008] Liang Wang, Hailin Jin, Ruigang Yang, and Minglun Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [Yi *et al.*, 2020] Zili Yi, Qiang Tang, Shekoofeh Azizi, Dae-sik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [Yu *et al.*, 2018] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [Yu *et al.*, 2019] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.
- [Zeng *et al.*, 2021] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M. Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.