Deep Stereo Image Inpainting

Minseo (Sonia) Kim
Stanford University
kminseo@stanford.edu

Jasmine Cheng
Stanford University
chengjz@stanford.edu

Abstract

Immersive VR/AR needs high-quality, consistent stereo images. Single-view inpainting methods ignore stereo geometry, causing visual discomfort. Therefore, stereo consistency in inpainting is essential for realistic and comfortable user experiences. Recent deep learning advances can model relationships between stereo pairs. We adapt the Parallax Attention approach [3] for stereo image inpainting, training from scratch on the 48k pairs of stereo images from Flickr1024 dataset to achieve perceptually coherent results.

1. Introduction

Recent advances in virtual and augmented reality technologies have created an urgent need for high-quality stereo image processing capabilities. While immersive VR/AR experiences depend on consistent stereo imagery to maintain user comfort and realism, traditional single-view inpainting methods fail to account for the geometric relationships between stereo image pairs. This oversight leads to visual inconsistencies that can cause discomfort and break immersion. The challenge of maintaining stereo consistency during inpainting—filling in missing or corrupted regions of images—is therefore critical for delivering realistic and comfortable user experiences in immersive environments.

To address this challenge, we propose a novel approach that adapts the Parallax Attention Mechanism (PAM) from PASSRnet [3] for stereo image inpainting. Unlike existing methods such as SICNet [2], which relies heavily on additional disparity networks and struggles with misaligned inputs, our approach leverages PAM's global receptive field along epipolar lines to effectively aggregate features across stereo pairs. We modify the original architecture to output both left and right inpainted images simultaneously, ensuring bidirectional consistency through a comprehensive loss function that incorporates inpainting loss, left-right consistency, PAM map consistency, and cycle consistency constraints.

Our model is trained from scratch on the Flickr1024 [4] dataset, comprising 48,000 stereo image pairs, to learn the complex relationships between corresponding views. This extensive training enables the network to produce perceptually coherent results that maintain geometric consistency across the stereo pair. By combining the flexibility of PAM with multiple consistency losses, our approach achieves improved stereo inpainting performance, resulting in reduced disparity between the left and right views compared to vanilla PAM and ultimately contributing to more comfortable and realistic immersive visual experiences.

2. Related Work

SICNet SICNet [2] introduces an X-shaped convolutional neural network designed to jointly inpaint both views of a stereo pair. It is trained with a combination of reconstruction loss, adversarial loss, and stereo consistency loss to enforce geometrically coherent inpainting across the left and right images. While SICNet performs well in regular stereo setups, it assumes well-aligned stereo pairs and relies heavily on an auxiliary disparity estimation network. This results in high training costs and limits its robustness to misalignment or noisy depth priors.

PAM PASSRnet [3] addresses the task of stereo image super-resolution through a parallax-attention module (PAM), which leverages a global receptive field along epipolar lines to fuse complementary features across views. It is capable of handling large disparity variations and benefits from diverse loss functions to preserve structure. However, the original PASSRnet only enhances the left image and is not designed for completing missing regions in both views.

Our Approach. To leverage the strengths of PASSRnet in stereo correspondence, we adapt it for stereo inpainting by retrieving the reconstructed right view from batchwise multiplication of parallax attention matrix and the reconstructed left view. This hybrid design avoids the need of ground truth disparity via PASSRnet's powerful stereo attention mechanism, aiming for coherent and structure-preserving stereo inpainting.

3. Method

Our proposed approach leverages stereo inpainting utilizing parallax-attention mechanisms and a comprehensive composite loss function to ensure high-quality image completion and cross-view consistency. Given a masked stereo image pair, the framework, shown in Figure 2, generates two parallax-attention maps (PAM) to capture correspondence between the left and right images. The PAM facilitates accurate reconstruction of missing regions by allowing information transfer between views.

3.1. Parallax-Attention Mechanism

The parallax-attention mechanism (PAM) captures stereo correspondence by attending to matching pixels along epipolar lines. Given a stereo image pair with size $H \times W$, PAM produces attention maps of size $H \times W \times W$, where each slice $h \in H$ encodes dependencies between corresponding rows of the left and right images. When no disparity exists, the attention maps resemble stacked H identity matrices, indicating direct pixel-to-pixel alignment. For regions with disparity, attention shifts to off-diagonal positions that reflect the spatial offset between corresponding features. Examples of parallax attention are shown in Figure 1.

Moreover, PAM can infer occlusion: regions without valid matches are indicated by rows or columns lacking active attention weights. This allows PAM to robustly align features even under large disparity variations and partial occlusion.

3.2. Loss Function

The total training objective combines inpainting error, stereo photometric consistency, PAM smoothness, and cycle consistency. The loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{inpaint}} + \lambda \underbrace{\left(\mathcal{L}_{\text{photometric}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{cycle}}\right)}_{\text{stereo loss}} \quad (1)$$

where $\mathcal{L}_{\text{inpaint}}$ is an ℓ_2 loss between the predicted inpainted image and the ground truth, encouraging accurate restoration of missing content.

The stereo loss components ($\mathcal{L}_{\text{photometric}}$, $\mathcal{L}_{\text{smooth}}$, and $\mathcal{L}_{\text{cycle}}$) regularize the solution by leveraging the inherent geometric and photometric relationships between stereo pairs.

Photometric loss ($\mathcal{L}_{\text{photometric}}$): Enforces left-right consistency by minimizing the difference between a view and its reconstruction from the counterpart via the PAM.

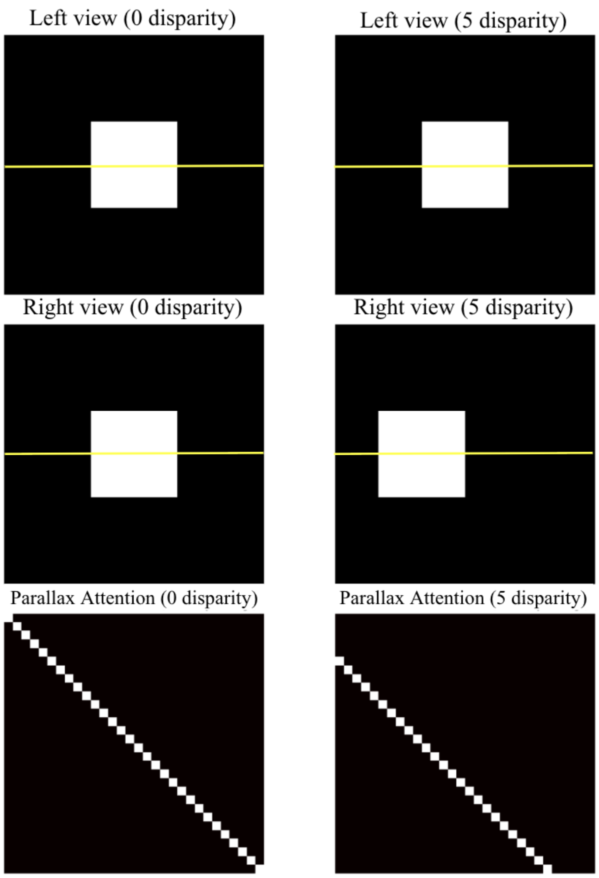

Figure 1. **Example of parallax attention.** The left column has no disparity and the right column has 5 pixel of disparity. The yellow line is the horizontal epipolar line which the parallax attention map is generated upon.

$$\begin{aligned}\mathcal{L}_{\text{photometric}} = &\sum_{p \in \mathcal{V}_{l\to r}} \left\| \mathbf{I}_l^L(p) - \left(\mathbf{M}_{r\to l} \otimes \mathbf{I}_r^L\right)(p) \right\|_1 \\ &+ \sum_{p \in \mathcal{V}_{r\to l}} \left\| \mathbf{I}_r^L(p) - \left(\mathbf{M}_{l\to r} \otimes \mathbf{I}_l^L\right)(p) \right\|_1\end{aligned} \quad (2)$$

Smoothness loss ($\mathcal{L}_{\text{smooth}}$): Encourages spatial smoothness in the PAMs by penalizing abrupt variations along spatial dimensions, promoting coherent correspondence estimation.

$$\begin{aligned}\mathcal{L}_{\text{smooth}} = \sum_{\mathbf{M}} \sum_{i,j,k} \Big(&\|\mathbf{M}(i,j,k) - \mathbf{M}(i+1,j,k)\|_1 \\ &+ \|\mathbf{M}(i,j,k) - \mathbf{M}(i,j+1,k+1)\|_1 \Big)\end{aligned} \quad (3)$$

Cycle consistency loss ($\mathcal{L}_{\text{cycle}}$): Ensures that mapping a pixel from one view to the other and back via PAM should return the original pixel, thus enforcing reliable cross-view matching.

$$\mathcal{L}_{\text{cycle}} = \sum_{p \in \mathcal{V}_{l\rightarrow r}} \|\mathbf{M}_{l\rightarrow r\rightarrow l}(p) - I(p)\|_1 + \sum_{p \in \mathcal{V}_{r\rightarrow l}} \|\mathbf{M}_{r\rightarrow l\rightarrow r}(p) - I(p)\|_1 \tag{4}$$

In practice, each masked stereo input generates left-to-right and right-to-left PAMs. These mappings enable reconstruction of one view from the other by aggregating information according to the parallax correspondences. For instance, the right inpainted image is derived from the left inpainted image using the left-to-right PAM (5), where $\otimes$ indicates batch-wise matrix multiplication. This process ensures that the underlying geometry and texture continuity are preserved across both images, resulting in plausible and consistent inpainted outputs.

$$\mathbf{I}_r^{inpaint} = \mathbf{M}_{l\rightarrow r} \otimes \mathbf{I}_l^{inpaint} \tag{5}$$

4. Results

4.1. Experimental Setup

The experimental setup involved training our model on 48,000 stereo image pairs from the Flickr1024 dataset. We perform data preprocessing to standardize input size and promote efficient learning. To prepare the dataset for stereo image inpainting, we extract overlapping patches of size 120×360 pixels from stereo pairs in the Flickr1024 dataset using a sliding window with a stride of 80 pixels. To simulate occlusions, each extracted patch pair is then corrupted with a fixed central rectangular mask, which removes 25% of the area by zeroing out a centrally located region. This simulates missing content in both views while preserving the geometric alignment. For each patch, we store the original left and right images, the masked versions, and the corresponding binary mask for supervised training.

Evaluation was performed on the KITTI 2012 [1] dataset, with the same data preprocessing to obtain 120×360 pixel patches to ensure consistency with the training conditions.

To generate the baseline results, we train two independent models—one for the left view and one for the right view—using standard image inpainting without leveraging stereo information. Each model receives a single masked image as input and predicts the reconstructed version of the same view. For consistency with the stereo setting, we replicate the masked image and feed it into both branches of the stereo inpainting architecture, effectively disabling any cross-view interaction, which means that the PAM module should only output identity matrix. When training the baseline model, we only use the inpaint MSE loss without the stereo loss. In this way, the left-view model is trained and tested solely on the masked left image, and likewise for the right-view model. This setup serves as a control that isolates the contribution of stereo supervision by removing parallax attention and stereo correspondence learning from the pipeline.

The total training time was approximately 9 hours, while testing took about 2 minutes, allowing us to comprehensively assess the model's stereo inpainting performance under diverse and realistic conditions.

4.2. Qualitative Evaluation

The qualitative results in Figure 3 demonstrate the importance of incorporating stereo consistency loss into the training objective. Without the stereo loss, the model focuses solely on reconstructing the left view, and the right view is generated through the learned parallax attention mechanism. However, because the model does not receive any explicit supervision on stereo alignment, it fails to learn accurate correspondence between the two views. As a result, the reconstructed right view in the "Recon. w/o Stereo Loss" column is often heavily blurred or distorted in the masked regions.

By introducing stereo consistency loss during training, the model is encouraged not only to reconstruct the left view but also to learn a meaningful stereo mapping via the parallax attention mechanism. This results in more coherent and aligned completions across both views. Interestingly, we note that the right-view output with stereo loss is worse than the baseline, despite using the PAM module. We hypothesize that this is due to the masked regions lacking valid stereo cues, which means that the mask has 0 disparity. This confuses the attention mechanism during training.

4.3. Quantitative Evaluation

Method	**Left PSNR**	**Right PSNR**	**Disparity RMSE**
Baseline	33.27	33.31	0.8071
w/o stereo loss	33.25	14.31	7.6461
w/ stereo loss	33.06	25.12	1.0043

Table 1. **Quantitative Results.** We show the average PSNR over 800 validation KITTI image patches and RMSE to measure the disparity map consistency.

We quantitatively evaluate the performance of our model variants using PSNR for the left and right reconstructed views and RMSE of the disparity maps of the reconstructed

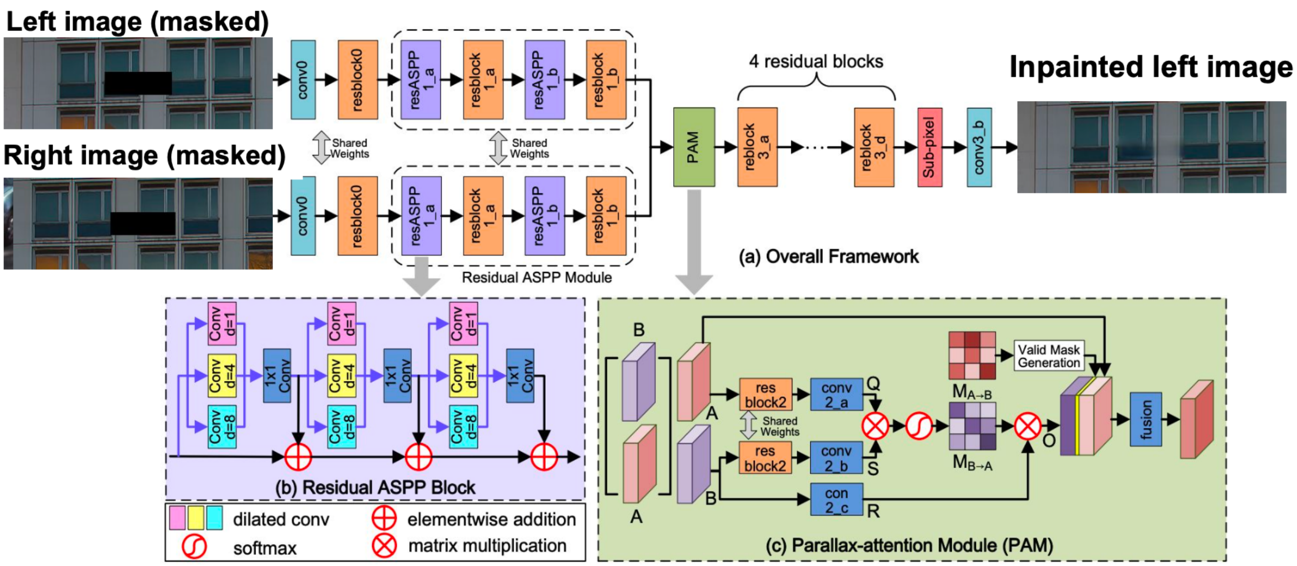

Figure 2. **Overview of the PAM model architecture for inpainting task.**

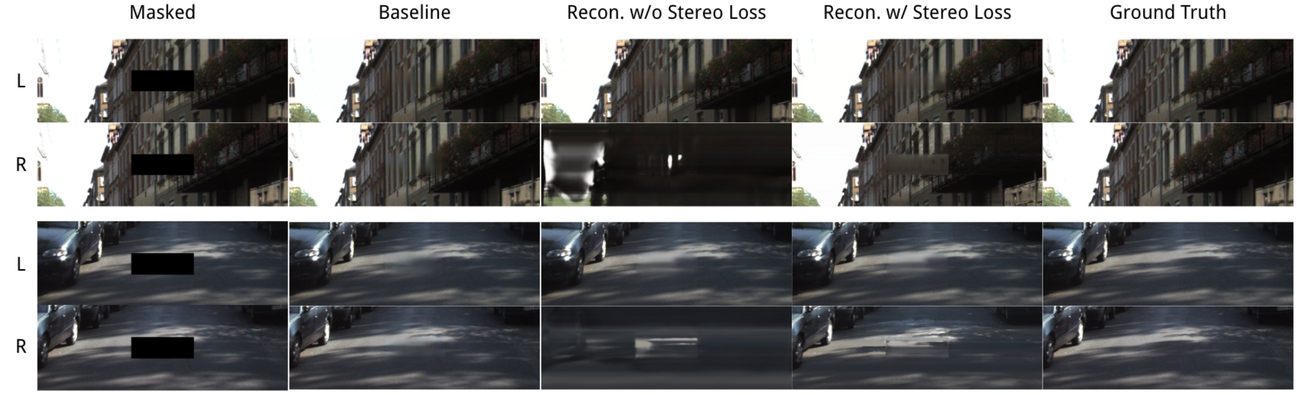

Figure 3. **Qualitative comparison of stereo inpainting results.** Each group shows the left (L) and right (R) stereo images with masked inputs (first column), baseline reconstruction (second column), reconstruction without stereo consistency loss (third column), and our final model trained with stereo consistency loss (fourth column). The inclusion of stereo loss leads to more coherent and geometrically consistent completions across both views, especially in structured regions such as building facades and car contours.

views. As shown in Table 1, the baseline achieves balanced PSNR across both views, serving as a reference point. Without stereo loss, the model achieves a similar left-view PSNR (33.25 dB) but suffers a severe drop in right-view quality (14.31 dB), indicating that the parallax attention mechanism alone fails to generalize stereo correspondence in the absence of explicit supervision. Introducing stereo loss improves the right-view PSNR to 25.12 dB, confirming that stereo consistency guidance significantly enhances reconstruction quality. However, the disparity RMSE increases slightly compared to the baseline (1.0043 vs. 0.8071), and visualization of the disparity map error, shown in Figure 4, reveals that much of this error is concentrated in the masked regions. This suggests that, while stereo loss helps enforce geometric alignment, occluded or ambiguous regions remain challenging for disparity estimation and can confuse the parallax attention module when supervision is sparse or noisy. This will not only degrade the reconstruction of the left view but also the attention.

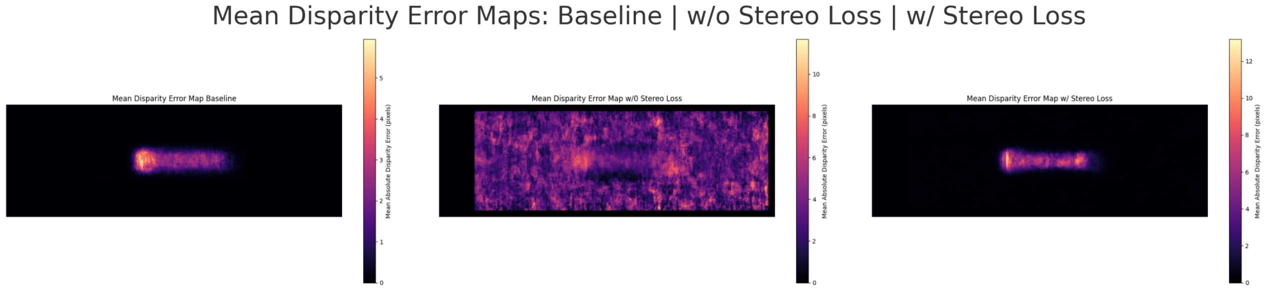

Figure 4. **Mean Disparity Error Maps across validation patches.** Each map visualizes the mean absolute disparity error over 800 validation patches for (left) the baseline model, (center) the model without stereo loss, and (right) the model with stereo loss.

5. Conclusion

We presented a novel stereo image inpainting framework that leverages the Parallax Attention Mechanism (PAM) to enforce cross-view consistency. By adapting the PASSRnet architecture and introducing a multi-term stereo loss, our model effectively completes missing regions in both stereo views while maintaining geometric alignment. Experimental results on KITTI2012 datasets demonstrate that incorporating stereo loss significantly improves right-view reconstruction quality and reduces disparity error, validating the importance of enforcing stereo constraints.

While our model does not outperform the baseline in all quantitative metrics—particularly in left-view PSNR and overall disparity RMSE—it provides valuable insights into the trade-offs between reconstruction fidelity and stereo consistency. Our findings highlight the importance of balanced loss weighting and the challenges of learning reliable disparity in occluded regions. These insights can inform future work on stereo inpainting and other multi-view generation tasks, where geometric coherence is critical for downstream applications in immersive VR/AR environments.

5.1. Limitations and Future Work

Several limitations, such as blurriness of the right inpainted image and suboptimal disparity RMSE, in our current approach may explain these performance gaps:

- **Feature Representation Mismatch:** The stereo loss may force the network to compromise detail reconstruction in favor of stereo consistency, resulting in blurrier outputs, particularly in the right view.
- **Imbalanced Learning Objectives:** The weight balancing between reconstruction quality and stereo consistency appears suboptimal, causing the model to underperform in one aspect while attempting to satisfy the other.
- **Disparity Estimation Challenges:** Our method relies on accurate disparity estimation, which remains challenging in regions with occlusions, textureless surfaces, and reflective materials.

To address these limitations, we propose several promising directions for future research:

- **Adaptive Stereo Loss Weighting:** Implementing a spatially-adaptive weighting mechanism that applies stronger stereo constraints in regions with reliable disparity estimates while relaxing constraints in ambiguous areas.
- **Attention-based Cross-view Fusion:** Developing specialized attention mechanisms that can selectively transfer information between views without compromising detail preservation, similar to approaches in recent stereo video inpainting work [5].
- **X-shaped Architecture:** Designing an X-shaped encoder-decoder architecture with two symmetric decoding branches—one for the left and one for the right view—allowing the model to reconstruct both images simultaneously. Cross-view features are fused through a shared bottleneck or attention layers, enabling the network to jointly reason about occlusions, correspondences, and structural consistency across views. This setup avoids the imbalance of using one view as the supervisory target and promotes symmetrical learning for stereo inpainting.

References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] W. Ma, M. Zheng, W. Ma, S. Xu, and X. Zhang. Learning across views for stereo image completion. *IET Computer Vision*, 05 2020.

[3] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo. Learning parallax attention for stereo image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *International Conference on Computer Vision Workshops*, pages 3852–3857, Oct 2019.

[5] Z. Wu, C. Sun, H. Xuan, and Y. Yan. Deep stereo video inpainting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5702, 2023.