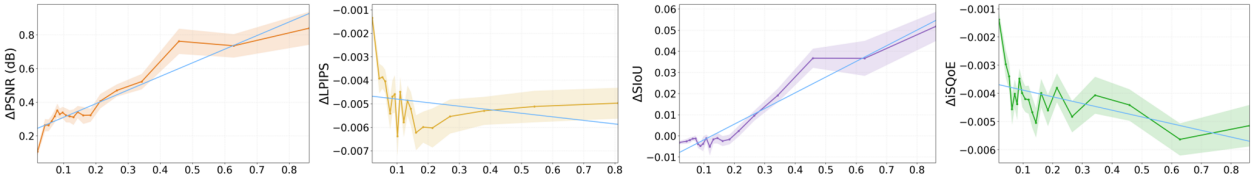
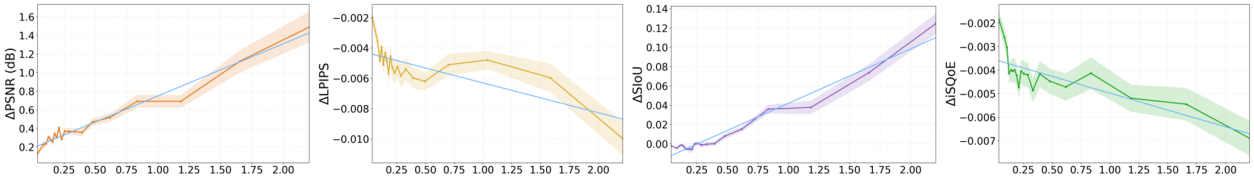
XPaintNet: An eXtreme Lightweight Framework for Stereoscopic Conversion without Inpainting Network

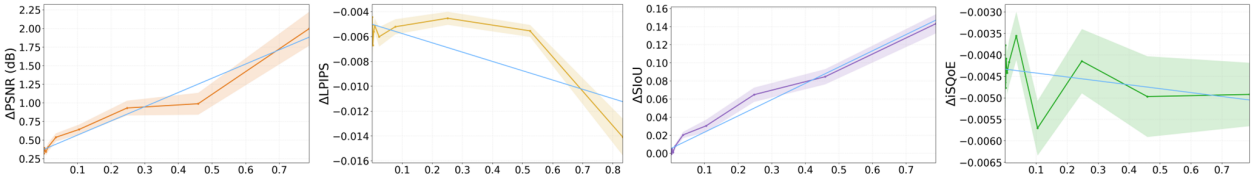
Supplementary Material



(a) The improvement of each metric according to Root Mean Square (RMS) of vertical component ($|v|$)



(b) The improvement of each metric according to the 95th percentile of vertical component ($|v|$)



(c) The improvement of each metric according to fraction of pixels with vertical component ($|v| > \tau, \tau = 1$)

Figure 8. Analysis of metric improvement (ΔPSNR↑, ΔLPIPS↓, ΔSIoU↑, ΔiSQoE↓) as the vertical component distribution varies (RMS of $|v|$, 95th percentile of $|v|$, and fraction of pixels with $|v|$>1 px).

A. Analysis about residual vertical component

In this section, we further analyze the benefit of using the vertical component (v) alongside the horizontal component (u) in stereoscopic conversion, relative to horizontal-only warping. For analysis, we compute three statistics: (1) RMS of the $|v|$, (2) 95th percentile of $|v|$, and (3) the fraction of pixels with $|v| > 1$ px. Fig. 8 shows each frame binned by statistical value (percentile). Overall trends are summarized by the global least squares slope and Spearman's rank correlation coefficient (ρ). Statistics are computed per frame across the entire dataset. These predictors collectively capture complementary aspects-global magnitude (RMS), tail severity (95th), and spatial support (area fraction).

RMS of the vertical component. RMS reflects the average energy of frame-level vertical motion. As RMS increases, ΔPSNR and ΔSIoU monotonically increase, while ΔLPIPS and ΔiSQoE decrease (improve), indicating gains in precision and edge alignment.

95th percentile of the vertical component. The 95th percentile emphasizes rare but large deviations that typically occur near depth discontinuities and occlusions, isolating worst-case misalignment. Improvements are steepest with this predictor: ΔPSNR and ΔSIoU increase and ΔLPIPS/ΔiSQoE decrease more sharply than with RMS.

Fraction of pixels with $|v| > 1$ px. This metric captures how widely meaningful vertical motion is distributed across the image, a practical indicator of where horizontal-only warps begin to blur or tear. As this fraction increases, all metrics improve. Perceptual metrics exhibit slightly higher variance in low-motion regions but follow the same monotonic relationship.

Residual vertical parallax, caused by mis-rectification, lens/zoom asymmetry, stabilization, or cropping, generally appear more frequently and are larger near depth boundaries. Under horizontal-only warping, these regions exhibit vertical misalignment and tearing. Allowing v directly corrects these errors and reduces hole formation during warping. This alleviates the burden of filling holes via inpainting or other methods. Positive slopes and significant ρ across all three predictors indicate that the benefit of using the vertical component increases with the frequency and magnitude of vertical motion, supporting Secs. 2.3 and 3.1.

B. Forward-from-Backward Approximation & Stability

We propose the Bi-Warp and Fusion approach, which estimates flow in one direction and derives the opposite direction through approximation. This method offers three advantages. First, by estimating flow in one direction and approximating the opposite direction, it reduces cross-direction inconsistencies that occur when estimating both directions independently. This mitigates pixel-level discrepancies between left and right views. Second, the approximated forward flow can preserve the direction and magnitude of the directly estimated forward flow with minimal loss. Finally, since only one direction is predicted, it is more efficient in terms of parameters, computational complexity, and latency. To demonstrate these advantages, we perform a more detailed analysis in this section.

We first evaluate the accuracy of the forward-from-backward approximation. For evaluation, we generate bidirectional pseudo ground truth flow, using well-trained flow estimate network [29]. To verify accuracy of the approximate flow, we conduct the analysis between the approximated forward flow (obtained by approximate backward ground truth flow) and the forward ground truth flow. For detailed evaluation, we report four complementary metrics: pixel-wise EPE, normalized-EPE (n-EPE) scaled by local flow magnitude, angular agreement (θ), and the fraction of pixels with EPE ≤ 1 and ≤ 2.

	Benchmarks						
	Mono2Stereo						Inria3D
	animation	complex	indoor	outdoor	simple	avg	
EPE	0.66	0.67	0.78	0.58	0.97	0.72	0.43
n-EPE	0.17	0.20	0.21	0.23	0.32	0.22	0.11
θ	0.93	0.93	0.93	0.91	0.88	0.92	0.97
EPE$\leq$1px	0.88	0.89	0.88	0.90	0.82	0.88	0.94
EPE$\leq$2px	0.93	0.94	0.93	0.94	0.90	0.93	0.97

Table 5. Accuracy of the forward-from-backward approximation.

Tab. 5 summarizes the results for the Mono2Stereo subset and Inria3D. The mean EPE is 0.72, indicating an error rate that does not exceed 1 pixel on average, while the θ is 0.92, preserving the directionality of the flow. Furthermore, the fraction of EPE within $1/2$ px is 0.88 and 0.93. Moreover, a higher degree of correspondence is observed in Inria3D. These results demonstrate that the approximation preserves both the direction and scale of the forward flow.

Second, to show that the approximation reduces cross-direction inconsistencies, we compare the warped images rather than the flows. For comparison, we calculate the error between backward warped image and forward warped image using the directly estimated forward flow. And the error between backward warped image and forward warped image using the approximated forward flow. To evaluate consistency specifically with the backward warped frame, we excluded hole regions from both the forward warped and approximated forward warped frames using a shared mask, and computed the ℓ_2 distance over the shared valid region. As shown in Tab. 6, approximation can preserve the consistencies of the cross-direction.

	Benchmarks						
	Mono2Stereo						inria3D
	animation	complex	indoor	outdoor	simple	avg	
Bidirection	21.49	19.27	16.62	19.26	9.59	17.35	29.75
Approximation	21.25	19.07	16.44	19.05	9.50	17.17	29.51

Table 6. Image-space disagreement between backward warped and forward warped images.

	Benchmarks					
	Mono2Stereo			Inria3D		
	LPIPS (↓)	SIoU (↑)	iSQoE (↓)	LPIPS (↓)	SIoU (↑)	iSQoE (↓)
Bidirection	0.125	0.284	0.639	0.267	0.175	0.556
Approximation	0.099	0.292	0.631	0.139	0.271	0.537

Table 7. Results comparison between direct estimated bi-direction flow and forward-from-backward approximation.

We further validate the effectiveness of the forward-from-backward approximation in our real-time network, XPaintNet. For the bi-flow baseline, we modify each Lite-MonoFlow estimator to predict both directions by expanding its output from 3 to 5 channels (forward, backward flow, and mask). Furthermore, to compare the effects of approximation, comparisons were conducted excluding the proposed loss function $L_{\text{bi-perc}}$, while all other training details remained identical. As shown in Tab. 7, directly estimating bi-direction flow can cause misalignment between the two directions. Therefore, employing an approximation that can mitigate this issue yields better performance.

	Efficiency metrics		
	MACs (G)	Params (M)	Latency (ms)
Bi-flow	15.31	1.50	9.80
Approximation	15.04	1.48	9.17

Table 8. Efficiency comparison between direct estimated bi-direction flow and forward-from-backward approximation.

Moreover, as shown in Tab. 8, approximation is superior in terms of complexity and latency because directly estimating both directions requires more channels than approximation.

In summary, (i) F and $\hat{F}$ agree strongly in flow space across diverse scenes, (ii) The warping result by approximation reduces pixel disagreement relative to a separately estimated forward flow, and (iii) the approximation-based approach improves quality and efficiency. These findings justify the use of forward-from-backward approximation in our Bi-Warp and Fusion approach.

Resolution	Desktop	Edge Device GPU	Mobile Device		
	RTX 3090 GPU	Jetson AGX Orin 64GB	Exynos 2400	Snapdragon 8 Gen 3	
			GPU	GPU	NPU
720p	250.04	122.32	82.64	85.47	94.34
2K	109.05	62.68	25.01	34.97	35.59
4K	29.10	15.67	6.29	10.66	10.14

Table 9. **XPaintNet throughput (FPS↑) across resolutions on a desktop GPU and memory-constrained edge devices, all in FP32.** Here, edge devices denote mobile accelerators (NPUs/GPUs on Exynos/Snapdragon) and Jetson edge GPUs. Measurements are end-to-end (include pre/post-processing) and averaged after warm-up.

C. Effectiveness of Bi-Warp compare with Inpainting

Networks	Benchmarks					
	Mono2Stereo			Inria3D		
	LPIPS (↓)	SIoU (↑)	iSQoE (↓)	SIoU (↑)	LPIPS (↓)	iSQoE (↓)
FuseFormer [15]	0.367	0.200	0.759	0.506	0.205	0.795
E2FGVI [13]	0.323	0.201	0.759	0.463	0.209	0.794
MI-GAN [22]	0.285	0.237	0.762	0.462	0.239	0.788
ProPainter [41]	0.234	0.214	0.764	0.422	0.220	0.794
StrDiffusion [14]	0.241	0.221	0.761	0.427	0.232	0.793
StereoCrafter [40]	0.310	0.216	0.767	0.442	0.223	0.797
Bi-Warp (Ours)	0.098	0.269	0.752	0.384	0.246	0.773

Table 10. Quantitative results comparison on benchmark datasets.

In Sec. 2.2, we compared our approach with existing inpainting networks to demonstrate the efficiency of Bi-Warp and Fusion. To further specify the effectiveness of our proposed method, we compare reconstruction quality only within the disoccluded region where holes occur via forward warping, rather than across the entire frame. For comparison, we masked the entire image using the mask commonly employed by both the inpainting network and our method. Furthermore, to demonstrate that our method preserves geometric consistency better than inpainting in stereoscopic conversion, we report perceptual metrics including LPIPS, SIoU, and iSQoE.

As shown in Tab. 10, Bi-Warp tends to yield lower LPIPS and higher SIoU than inpainting baselines on Mono2Stereo, indicating less perceptual distortion and more consistent, well-aligned disparities. It also achieves lower iSQoE, suggesting better stereo viewing comfort. On Inria3D, the same tendency is observed across all metrics.

Inpainting-based methods can produce visually plausible results, but they may break geometric consistency, which can degrade stereoscopic comfort. By contrast, Bi-Warp fills the hole regions by retaining valid backward samples via grid sampling and fusing them with forward constraints. This reduces reliance on hallucinated content and better preserves geometry at depth boundaries, yielding a cleaner appearance and a more comfortable stereoscopic experience. Taken together, these observations suggest that Bi-Warp is a practical alternative to inpainting, achieving comparable or superior quality without an explicit inpainting stage.

D. Latency on Memory Constrained Devices

We substantiate the claim that our XPaintNet is designed to operate on memory-constrained edge devices by measuring end-to-end inference latency on actual edge-aware devices across common input resolutions (720p, 2K, and 4K). Table Tab. 9 reports throughput in frames per second (FPS, higher is better), the corresponding latency in milliseconds is 1000/FPS. All results were obtained under FP32. To evaluate throughput on edge devices, we use mobile devices (NPUs/GPUs on Exynos and Snapdragon systems) and an edge-device GPU (Jetson). Concretely, the desktop system uses an RTX 3090 with a CUDA-based runtime, Jetson AGX Orin 64GB uses a TensorRT GPU backend, Exynos 2400 uses the mobile GPU runtime, and Snapdragon 8 Gen 3 is evaluated on both the GPU and the NPU backends. For mobile devices, the FP32 checkpoints were converted and packaged for on-device execution using LiteRT[1]. Measurements are end-to-end with identical pre-processing and post-processing across all devices. After an initial warm-up phase, the network was run for 50 iterations and the average was reported. Mobile results were collected using the AI-Benchmark application[2], with its invocation aligned to our runtime configuration.

Using 30 FPS as the real-time threshold, all platforms achieve real-time at 720p. At 2K, the desktop GPU and Jetson are real-time, Snapdragon 8 Gen 3 meets real-time on both GPU and NPU, while Exynos 2400 GPU is slightly below. At 4K, none of the edge or mobile platforms achieve the 30 FPS real-time threshold, and the desktop and Jetson configurations also remain below strict real time. Nevertheless, further speedups are plausible through optimizations such as mixed-precision or integer inference (FP16/INT8) via post-training quantization or quantization-aware training, kernel tuning and scheduling, hardware-aligned layout packing, offline compilation with caching, and tighter memory/buffer reuse. These optimizations could reduce latency toward real-time 4K, and a systematic study of the accuracy–latency–memory trade-offs is left to future work.

[1] https://ai.google.dev/edge
[2] https://ai-benchmark.com/

Figure 9. Anaglyph results comparison between state-of-the-arts StereoScopic Conversion Networks and Our XPaintNet.

E. Qualitative Comparison with Anaglyph Image

We visualized the anaglyph images generated by fusing the left view image with the right view image converted by network to compare stereo comfort. Comparisons are conducted on the Mono2Stereo [36] and Inria3D [23] benchmarks. And we compare Deep3D [30], which is efficient in model size and latency, and Mono2Stereo [36], which demonstrated the best performance among existing stereo image conversion networks.

As shown in Fig. 9, Deep3D generates a slightly blurry result because it generates the right view based on a shift-weighted composite layer. And Mono2Stereo appears satisfactory when viewed independently. However, when combined with the left view to anaglyph, geometric inconsistencies caused by the right view synthesis become apparent. These inconsistencies are particularly noticeable in facial regions. In contrast, our proposed XPaintNet generates the right view by fusing backward warped samples with forward constraints via Bi-Warp fusion. This preserves left–right geometric consistency in the anaglyphs and fills disocclusions with minimal ghosting, reducing reliance on hallucinated content.

Overall, the anaglyph results supports that *XPaintNet* maintains left-right geometric consistency while preserving texture fidelity, reducing binocular conflict and providing a more comfortable stereoscopic experience.