

Cross-View Variance Correlation in Path-Traced Stereo: A Hidden Shortcut in Synthetic Training Data

Po-Ting Lin

Abstract—Path-traced synthetic stereo data underlie a large fraction of modern disparity-estimation training pipelines. We report a previously unrecognised property of such data: while the Monte Carlo (MC) noise streams of the two cameras are statistically independent, the underlying *variance fields*—deterministic per-pixel functions of the rendering integrand—are highly correlated once aligned by the ground-truth disparity warp. Across 20 scenes rendered with Mitsuba 3, the warped Pearson correlation reaches $\rho=0.754\pm0.016$ across 20 scenes at SPP=512, and on a representative scene remains essentially invariant ($\rho=0.778\pm0.001$) over a $16\times$ range of samples per pixel. The effect is strongest in Lambertian regions ($\rho\approx0.78$) and substantially weaker in glass ($\rho\approx0.30$), as predicted by an integrand decomposition into view-independent and view-dependent components. A residual-shuffle intervention that breaks the cross-view alignment while preserving the clean image degrades the GT cost margin by 33% on non-glass and the variance-based winner-take-all accuracy on glass by $4.3\times$, confirming the structure functions as a matching cue. This signal is unique to MC-rendered data and constitutes a candidate sim-to-real shortcut whose impact on trained networks remains to be quantified.

Index Terms—Stereo matching, Monte Carlo rendering, synthetic data, sim-to-real, variance analysis, path tracing.

I. INTRODUCTION

A large fraction of modern stereo-matching networks [1]–[3] are trained, at least in part, on synthetic data generated by physically-based path tracing [4]–[6]. The attraction is practical: ground-truth disparity is free at render time, and modern path tracers produce images whose first-order statistics closely match real photographs. The implicit assumption underlying this practice is that the residual Monte Carlo (MC) noise behaves as an additive i.i.d. perturbation of an otherwise clean stereo pair; in particular, that the noise streams in the left and right views are statistically independent.

This assumption holds at the level of individual samples: the random number generators driving the two cameras are seeded independently, and the per-sample radiance estimates are uncorrelated across views by construction. However, deep stereo networks do not consume samples, they consume images, and the matching cues they learn are computed from *aggregated* pixel intensities. A natural object to consider is therefore the per-pixel variance field $\sigma^2(x, y)$ obtained from N independent renders of the same scene—a deterministic function of the rendering integrand, distinct from the noise itself.

In this letter we report that the variance fields of the two views, while constructed from independent samples, are highly

correlated once aligned by the ground-truth disparity warp (Fig. 1). Across 20 scenes rendered with Mitsuba 3 [7], the warped Pearson correlation reaches $\rho=0.754\pm0.016$, and remains essentially unchanged ($\rho=0.778\pm0.001$) over a $16\times$ range of samples per pixel. Counter-intuitively, the effect is strongest in Lambertian regions ($\rho\approx0.78$) and substantially weaker in glass ($\rho\approx0.30$). Because real binocular sensors carry independent thermal and shot-noise streams, the cross-view variance signal is unique to MC-rendered data; we argue it constitutes a shortcut signal available to stereo networks at the cost-volume level on synthetic training inputs, and a previously unrecognised contributor to the sim-to-real gap of stereo networks [8], [9].

Contributions.

- We identify and quantify the cross-view correlation of MC variance fields in path-traced stereo (Sec. II, III-B).
- We show this correlation is essentially invariant over a $16\times$ SPP range (Sec. III-C), indicating it is a deterministic property of the scene rather than an artefact of finite-sample estimation.
- We give a material-conditioned breakdown showing that the correlation is driven by view-independent integrands, and discuss its potential implication as a sim-to-real shortcut, while noting that its effect on trained networks is left to future work (Sec. III-E, IV).
- We provide causal evidence via a residual-shuffle intervention: destroying the cross-view alignment while preserving the clean image degrades the GT cost margin and the variance-based winner-take-all accuracy, confirming that the structure functions as a matching cue at the cost-volume level (Sec. III-D).

II. METHOD

A. Variance estimation

For each rectified stereo scene we render N independent images per camera using a path tracer, driven by N independent random-number-generator seeds. Let $I_L^{(n)}(x, y)$ and $I_R^{(n)}(x, y)$, $n = 1, \dots, N$, denote the resulting per-pixel radiance estimates. The left- and right-view per-pixel MC variance fields are

$$\sigma_V^2(x, y) = \frac{1}{N} \sum_{n=1}^N (I_V^{(n)}(x, y) - \bar{I}_V(x, y))^2, \quad V \in \{L, R\}, \quad (1)$$

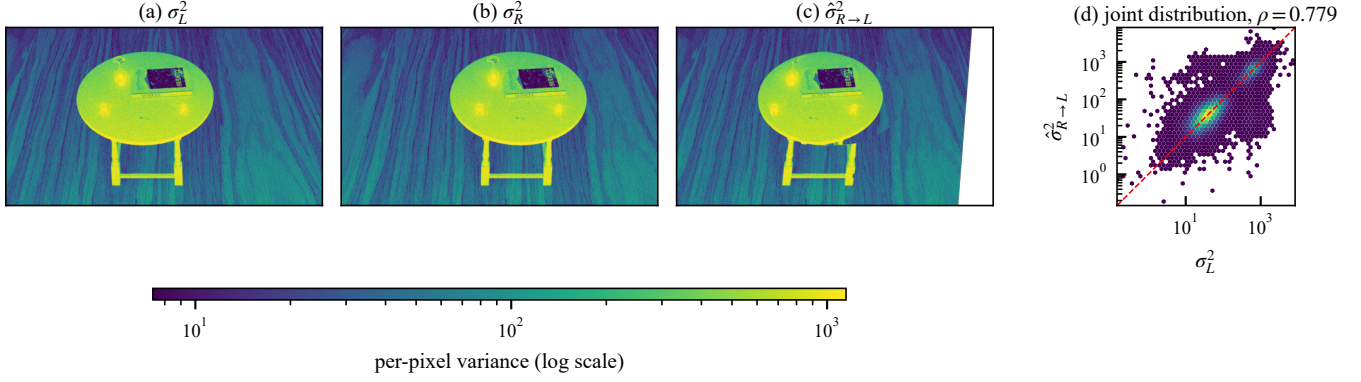


Fig. 1. Cross-view variance correlation in path-traced stereo on a representative scene. (a) and (b): per-pixel Monte Carlo variance σ_L^2 and σ_R^2 estimated from $N=30$ independent seeds at SPP=512. (c): the right-view field warped into left-view coordinates by the ground-truth disparity, $\hat{\sigma}_{R \rightarrow L}^2$. (d): joint distribution of (a) and (c) over the valid-pixel set Ω , with the diagonal $y=x$ line in red. Although the noise streams in the two views are statistically independent, the underlying variance fields are tightly aligned once the geometric transform is applied.

with \bar{I}_V the seed mean; for colour images we average (1) over RGB channels. Each σ_V^2 is a deterministic function of the rendering integrand and the scene/camera configuration; the finite-sample estimate (1) converges to that function at rate $O(1/\sqrt{N})$ as N grows [10], [11].

B. Cross-view alignment

Given the ground-truth disparity $d(x, y)$ supplied by the renderer, we warp σ_R^2 into left-view coordinates,

$$\hat{\sigma}_{R \rightarrow L}^2(x, y) = \sigma_R^2(x + d(x, y), y), \quad (2)$$

implemented by bilinear interpolation along the x axis, where $d > 0$ indicates the right-view correspondence lies to the right of the left-view pixel (i.e. the $x_R = x_L + d$ convention used throughout). We mask pixels for which $x + d(x, y)$ falls outside the right image, d is non-positive, or d is non-finite, leaving a valid-pixel set Ω . The map (2) is the same alignment used implicitly by every cost-volume-based stereo matcher [12]: any cross-view feature consumed by such a network has been brought into a common coordinate frame via this warp.

C. Correlation measure

We quantify the cross-view variance correlation by the Pearson coefficient over Ω ,

$$\rho = \text{corr}(\sigma_L^2, \hat{\sigma}_{R \rightarrow L}^2)|_{\Omega}. \quad (3)$$

Pearson is scale-invariant in both arguments, which is essential here: σ_V^2 scales as $1/\text{SPP}$ [10], so a magnitude-sensitive metric would conflate sample-budget changes with structural similarity. For the material-conditioned analysis of Sec. III-E we restrict Ω to glass pixels or to non-glass pixels using the ground-truth material mask supplied by the renderer.

III. EXPERIMENTS

A. Setup

We render 20 indoor scenes with Mitsuba 3 [7] at 1280×720 resolution and a stereo baseline of 26 mm.

TABLE I
CROSS-SCENE CORRELATION AT SPP=512 OVER 20 SCENES, WITH Ω RESTRICTED BY THE GROUND-TRUTH MATERIAL MASK.

Region	mean ρ	std	range
All pixels	0.754	0.016	0.735–0.784
Non-glass	0.779	0.011	0.758–0.797
Glass	0.301	0.102	0.140–0.473

Each scene is rendered $N=30$ times per camera with independent random-number-generator seeds, at SPP=512 by default; for the sample-budget experiment of Sec. III-C a representative scene is additionally rendered at SPP $\in \{128, 256, 512, 1024, 2048\}$. Ground-truth disparity and a per-pixel material mask (with a dedicated glass channel) are produced by the renderer. After the validity masking of Sec. II-B, each scene contributes approximately 9.2×10^5 pixels to Ω , so all correlations reported below are significant at $p < 10^{-100}$ under the standard Fisher z test.

B. Cross-scene correlation

Across the 20 scenes, the warped Pearson correlation of (3) reaches

$$\rho = 0.754 \pm 0.016 \quad (\text{range } 0.735\text{--}0.784),$$

a coefficient of variation of 2.1% that establishes the effect as a structural property of path-traced stereo rather than a per-scene anomaly. Computing Pearson without any warp gives only $\rho_{\text{no-warp}} \approx 0.36$ over the same 20 scenes: alignment via the disparity warp roughly doubles the measured correlation, confirming that the cross-view structure being measured is the same one that any cost-volume-based matcher would attempt to exploit. Material-conditioned numbers are reported in Table I and discussed in Sec. III-E.

C. SPP invariance

We test whether ρ is a residue of the finite-sample estimator (1) by sweeping the sample budget over

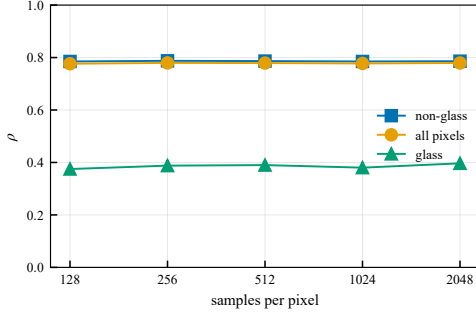


Fig. 2. Cross-view variance correlation ρ as a function of samples per pixel on a representative scene, for all valid pixels, non-glass only, and glass only. The non-glass curve varies by 0.12% across a $16\times$ range of sample budgets, ruling out a finite-sample explanation of the effect.

TABLE II
SPP SWEEP ON A REPRESENTATIVE SCENE. $\bar{\sigma}_L^2$ IS THE MEAN OVER Ω WITHIN THE INDICATED MATERIAL REGION. ACROSS THE $16\times$ SAMPLE-BUDGET RANGE, ρ IS INVARIANT TO WITHIN 0.2%.

SPP	$\bar{\sigma}_L^2 _{\text{glass}}$	$\bar{\sigma}_L^2 _{\neg\text{glass}}$	ρ_{all}	ρ_{glass}	$\rho_{\neg\text{glass}}$
128	17239	2447	0.776	0.375	0.785
256	8627	1224	0.780	0.388	0.788
512	4323	612	0.779	0.390	0.786
1024	2164	306	0.778	0.380	0.785
2048	1072	153	0.779	0.397	0.786

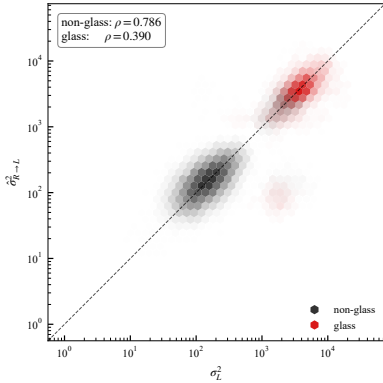


Fig. 3. Joint distribution of $(\sigma_L^2, \hat{\sigma}_{R \rightarrow L}^2)$ at SPP=512, separated by material. Non-glass pixels (gray) cluster near the diagonal at $\rho \approx 0.78$; glass pixels (red) disperse at $\rho \approx 0.30$.

$\{128, 256, 512, 1024, 2048\}$ on a representative scene. The mean per-pixel variance $\bar{\sigma}_L^2$ scales as $1/\text{SPP}$, dropping by a factor of 16 across the sweep (Table II, Fig. 2); the correlation is essentially unchanged:

$$\rho_{\text{all}} = 0.7783 \pm 0.0012, \quad \rho_{\neg\text{glass}} = 0.7860 \pm 0.0009.$$

The relative variation of $\rho_{\neg\text{glass}}$ across the $16\times$ SPP range is 0.12%, an order of magnitude smaller even than the per-scene variation of Sec. III-B. This rules out the hypothesis that the cross-view correlation is a finite-sample artefact that would vanish in the high-SPP limit: it persists into that limit.

D. Causal evidence: alignment as a matching cue

The correlation reported above is a property of the data, not yet of the matching task. To test whether the cross-view variance structure actually behaves as a disparity cue—i.e. whether removing the alignment component degrades matching at d_{GT} —we run a controlled intervention on the same 20 scenes.

a) *Decorrelation operator.*: For each seed n we decompose the right view into a clean image and a residual, $I_R^{(n)} = \bar{I}_R + \epsilon_R^{(n)}$ with \bar{I}_R the across-seed mean. We replace $\epsilon_R^{(n)}$ by a spatially block-shuffled copy and reassemble $\tilde{I}_R^{(n)} = \bar{I}_R + \Pi(\epsilon_R^{(n)})$, where Π permutes 16×16 residual blocks under a fixed seed. The clean image is preserved, so first-order intensity statistics are unchanged; only the spatial alignment between σ_L^2 and σ_R^2 is destroyed.

b) *Cost volumes.*: We build per-pixel cost volumes under the verified $x_R = x_L + d$ convention and evaluate two cost functions: a *residual SAD*,

$$C_\epsilon(x, d) = \sum_{u \in W} |\epsilon_L(x+u) - \epsilon_R(x+d+u)|$$

with patch W of size 5×5 , which isolates the MC signal from the dominant clean-image SAD; and a *variance cost*,

$$C_\sigma(x, d) = \sum_{u \in W} |\sigma_L^2(x+u) - \sigma_R^2(x+d+u)|$$

which uses the variance fields themselves as image inputs. We report three metrics: the GT margin

$$m(x) = \min_{d \neq d_{\text{GT}}} C(x, d) - C(x, d_{\text{GT}})$$

(higher is more salient); the winner-take-all accuracy

$$\Pr[\arg\min_d C - d_{\text{GT}} < 1 \text{ px}]$$

and the variance-similarity peak hit rate, the analogous accuracy of

$$\arg\max_d -|\sigma_L^2(x) - \sigma_R^2(x+d)|$$

at the pixel level.

c) *Results.*: Table III shows that destroying alignment shifts every metric in the direction predicted if the cross-view variance is a matching cue. On non-glass pixels the residual SAD margin improves by 33% (from -237 to -178); on glass pixels the variance-cost WTA accuracy increases $4.3\times$ (from 1.75% to 7.51%); and the variance peak hit rate, which uses the variance map directly as a cost, improves by 80% on glass and 15% on non-glass. These three effects are seen on the same 20 scenes that produced $\rho = 0.754 \pm 0.016$, with cross-scene standard deviations on the deltas an order of magnitude smaller than the deltas themselves. The intervention does not affect the clean image \bar{I}_R , so it isolates the contribution of the variance-alignment shortcut from ordinary intensity matching. The residual SAD column shows a direction-consistent but seemingly opposed pair of changes—the GT margin grows more negative under decorrelation while the WTA accuracy also drops; this reflects that block-shuffle inflates the overall residual cost magnitude, deepening the apparent margin while simultaneously degrading the SNR of the per-seed signal at

TABLE III

INTERVENTION OVER THE SAME 20 SCENES ($N=30$ SEEDS, PATCH 5×5 , RESIDUAL BLOCK 16×16). **NORMAL** PAIRS ARE THE RENDERED (I_L, I_R); **DECORR.** REPLACES $\epsilon_R^{(n)}$ BY ITS BLOCK-SHUFFLED COPY. MARGINS ON C_σ ARE DIMENSIONALLY LARGER THAN THOSE ON C_ϵ AS THE INPUTS ARE VARIANCE FIELDS, NOT INTENSITIES; WHAT MATTERS IS THE WITHIN-ROW CONTRAST.

Cost / Metric	Region	Normal	Decorr.
<i>Residual SAD C_ϵ</i>			
GT margin	non-glass	-178 ± 4	-237 ± 7
	glass	-589 ± 52	-236 ± 14
WTA accuracy	non-glass	$1.97\% \pm 0.03$	$1.83\% \pm 0.03$
	glass	$1.48\% \pm 0.17$	$1.95\% \pm 0.15$
<i>Variance cost C_σ</i>			
GT margin	non-glass	-3156 ± 100	-6744 ± 440
	glass	-19270 ± 3200	-8695 ± 510
WTA accuracy	non-glass	$1.57\% \pm 0.05$	$1.77\% \pm 0.04$
	glass	$7.51\% \pm 2.1$	$1.75\% \pm 0.6$
<i>Variance peak hit rate</i>			
σ -argmax hit	non-glass	$2.15\% \pm 0.04$	$1.87\% \pm 0.04$
	glass	$3.36\% \pm 0.38$	$1.87\% \pm 0.33$

d_{GT} , which is what the WTA metric reads. Two material-conditioned regularities are worth noting: the residual SAD is most informative on non-glass, where $\rho_{\text{glass}}=0.78$ provides a strong per-seed alignment; the variance cost is most informative on glass, where the absolute variance level is $7\times$ larger. The latter resolves an apparent puzzle: although ρ_{glass} is only 0.30, the $7\times$ -larger absolute magnitude makes $|\sigma_L^2(x) - \sigma_R^2(x+d)|$ a high-amplitude function of d whose minimum at d_{GT} is still distinguishable, so *moderate* alignment of a *large* signal can dominate winner-take-all matching even when the per-seed residual itself is poorly aligned. The two costs together verify the cue across both regimes.

E. Material breakdown

Restricting Ω to glass and non-glass pixels separately (Table I, Fig. 3) reveals the most surprising aspect of the phenomenon:

$$\rho_{\text{glass}} = 0.301 \pm 0.102, \quad \rho_{\text{non-glass}} = 0.779 \pm 0.011.$$

Non-glass regions are more than twice as cross-view correlated as glass regions, and an order of magnitude more stable across scenes (relative variation 1.4% versus 33.9%). Yet glass pixels dominate the absolute noise level: at SPP=512 we measure $\bar{\sigma}_L^2|_{\text{glass}} \approx 4.3 \times 10^3$ versus $\bar{\sigma}_L^2|_{\text{non-glass}} \approx 6.1 \times 10^2$, a $7\times$ ratio. The cross-view structure is therefore *inversely* related to the magnitude of the variance: pixels with the most MC noise carry the least cross-view alignment. The physical mechanism behind this inversion is the subject of Sec. IV.

IV. DISCUSSION

Variance integrand decomposition: The radiance integrand at a surface point P viewed from direction ω admits the decomposition

$$f(P; \omega) = f_{\text{ind}}(P) + f_{\text{dep}}(P, \omega), \quad (4)$$

where f_{ind} collects view-independent contributions—direct shadowing from area lights, indirect illumination, caustic

projection onto diffuse surfaces, and colour bleeding—and f_{dep} collects view-dependent contributions from Fresnel reflectance, specular and glossy lobes, and refraction [10], [13]. Under MC integration the per-pixel variance field inherits the same split. After warping the right view to left coordinates by ground-truth disparity, both pixels sample the same P ; the warped variance fields therefore agree to the extent that f_{ind} dominates and disagree to the extent that f_{dep} does.

Why Lambertian outranks glass: This decomposition predicts the observed ordering $\rho_{\text{Lamb}} \gg \rho_{\text{glass}}$. In Lambertian regions the variance is governed almost entirely by f_{ind} —ambient occlusion edges, indirect-bounce structure, and caustic spots projected from glass elsewhere in the scene—none of which depend on the viewing direction. The warped fields then align down to the noise floor of the finite-seed estimator. Specular and refractive materials behave oppositely: their variance is driven by Fresnel-modulated reflection and refraction-path geometry, both of which differ between left and right views even after correct geometric alignment. The intuition that “complex transparent materials carry more cross-view structure” inverts the actual ordering.

A cue available in synthetic data, absent in real sensors: The intervention of Sec. III-D confirms the signal is exploitable at the cost-volume level [1]–[3], while real binocular captures, with independent thermal and shot-noise streams, carry $\rho \approx 0$. Whether a trained network in fact draws on this cue—and how strongly it contributes to the sim-to-real gap of any specific architecture [8], [9]—is left to future work; the cue’s invariance to sample budget (Sec. III-C) implies higher SPP alone would not remove it.

Limitations: Our experiments use a single renderer; the mechanism is generic to MC path tracing, but cross-renderer confirmation is left for future work. The warp assumes rectified stereo with known ground-truth disparity, which hold by construction for synthetic data. The intervention of Sec. III-D establishes that the cross-view variance structure behaves as a usable matching cue at the cost-volume level, but does not quantify how strongly a particular trained stereo network draws on this cue versus on intensity matching: cost-level evidence constrains what the data *makes available* to a matcher, not what a fully optimised network ends up using in practice. Designing mitigations—e.g. variance equalisation or seed-coupled rendering—and measuring their effect on trained networks is the natural next step.

V. CONCLUSION

Path-traced synthetic stereo carries a near-deterministic cross-view structure that real binocular sensors lack: although the MC noise streams are independent, the per-pixel variance fields of the two views agree to $\rho \approx 0.78$ in Lambertian regions and persist unchanged over a $16\times$ sample-budget range. The signal is dominated by view-independent integrand contributions, and is therefore most pronounced precisely where the variance itself is smallest. A controlled intervention confirms it is exploitable as a matching cue at the cost-volume level. Whether trained stereo networks in fact rely on this cue, and how to neutralise it without sacrificing data utility, are open questions that we hope this characterisation makes tractable.

REFERENCES

- [1] J.-R. Chang and Y.-S. Chen, “Pyramid stereo matching network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5410–5418.
- [2] L. Lipson, Z. Teed, and J. Deng, “RAFT-Stereo: Multilevel recurrent field transforms for stereo matching,” in *Proc. Int. Conf. 3D Vis. (3DV)*, 2021, pp. 218–227.
- [3] G. Xu, X. Wang, X. Ding, and X. Yang, “Iterative geometry encoding volume for stereo matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 21 919–21 928.
- [4] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, “A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4040–4048.
- [5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 611–625.
- [6] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, “Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10 912–10 922.
- [7] W. Jakob, S. Speierer, N. Roussel, M. Nimier-David, D. Vicini, T. Zeltner, B. Nicolet, M. Crespo, V. Leroy, and Z. Zhang, “Mitsuba 3 renderer,” 2022, <https://www.mitsuba-renderer.org>.
- [8] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. Di Stefano, “Real-time self-adaptive deep stereo,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 195–204.
- [9] J. Watson, O. Mac Aodha, D. Turmukhambetov, G. J. Brostow, and M. Firman, “Learning stereo from single images,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 722–740.
- [10] E. Veach, “Robust monte carlo methods for light transport simulation,” Ph.D. dissertation, Stanford University, Dec. 1997.
- [11] M. Zwicker, W. Jarosz, J. Lehtinen, B. Moon, R. Ramamoorthi, F. Rouselle, P. Sen, C. Soler, and S.-E. Yoon, “Recent advances in adaptive sampling and reconstruction for Monte Carlo rendering,” *Comput. Graph. Forum*, vol. 34, no. 2, pp. 667–681, 2015.
- [12] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, 2002.
- [13] M. Pharr, W. Jakob, and G. Humphreys, *Physically Based Rendering: From Theory to Implementation*, 4th ed. MIT Press, 2023.