On the Flexibility of Regularization Hyperparameters in 3D Gaussian Splatting Under Adaptive Optimizers

Anonymous Author(s)

ABSTRACT

3D Gaussian Splatting (3DGS) pipelines employ scalar hyperparameters to control the strength of regularization losses such as opacity entropy and scale penalties. Practitioners assume that varying these weights proportionally adjusts the effective regularization strength. We show that this assumption fails under the Adam optimizer, which is standard in 3DGS. Through controlled simulation experiments on a simplified Gaussian splatting model, we introduce the *Effective Regularization Ratio* (ERR)—the fraction of the optimizer's parameter update attributable to regularization—and characterize its response to hyperparameter changes. Our experiments reveal three findings: (1) the ERR-vs-λ relationship exhibits a sub-linear log-log slope of 0.85 under Adam compared to 1.0 under SGD, meaning a 500× increase in λ yields only ∼142× increase in effective strength; (2) changing one regularization weight affects the effective strength of other terms through a cross-coupling ratio of 0.034; and (3) ERR varies by up to 14.2× across parameter types (position, scale, opacity, color) for the same λ value. We further propose an adaptive λ-scheduling algorithm that monitors ERR online and adjusts λ to maintain a target ratio, reducing ERR variance by 43.8% compared to fixed scheduling. Our results confirm that standard hyperparameters provide insufficient flexibility for controlling regularization in 3DGS under adaptive gradient methods and motivate decoupled optimization strategies.

ACM Reference Format:
Anonymous Author(s). 2026. On the Flexibility of Regularization Hyperparameters in 3D Gaussian Splatting Under Adaptive Optimizers. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

3D Gaussian Splatting (3DGS) [7] has emerged as a leading representation for real-time radiance field rendering, achieving state-of-the-art novel view synthesis quality while enabling real-time rendering through rasterization of anisotropic Gaussian primitives. A standard 3DGS training pipeline optimizes millions of Gaussian parameters (positions, covariances, opacities, and spherical harmonics coefficients) via the Adam optimizer [8] to minimize a photometric reconstruction loss, augmented by regularization terms that encourage desirable geometric properties.

Common regularization losses in 3DGS include opacity entropy penalties (promoting binary opacities to suppress floater artifacts) [4], scale regularization (preventing needle-like or excessively large Gaussians) [5], and depth/normal consistency losses (enforcing multi-view geometric coherence) [14]. Each regularization term is weighted by a scalar hyperparameter λ_k, and the total training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \sum_k \lambda_k \mathcal{L}_{\text{reg}}^{(k)}. \tag{1}$$

A natural expectation is that λ_k provides *linear* control: doubling λ_k should double the influence of $\mathcal{L}_{\text{reg}}^{(k)}$ on the parameter trajectory. However, Ding et al. [3] recently observed that this assumption is questionable when using adaptive gradient optimizers, writing that "the regularization loss is thought to be controlled through hyperparameters, yet it remains unclear whether they provide sufficient flexibility." This motivates their proposal for decoupled optimization in 3DGS.

In this work, we directly investigate this open question. We define the *Effective Regularization Ratio* (ERR) as a quantitative measure of the fraction of the Adam update that is attributable to regularization gradients, and we systematically characterize the mapping from λ to ERR through six controlled experiments. Our contributions are:

(1) A formal framework for measuring the effective regularization strength under Adam using moment-decomposition analysis (Section 2).
(2) Empirical evidence that the λ-to-ERR mapping is sub-linear under Adam, with a log-log slope of 0.85 compared to 1.0 under SGD, confirming limited hyperparameter flexibility (Section 3).
(3) Quantification of cross-coupling between regularization terms and heterogeneous ERR across parameter types.
(4) An adaptive λ-scheduling algorithm that reduces ERR variance by 43.8%, demonstrating a practical remedy (Section 3).

1.1 Related Work

3DGS and regularization. The original 3DGS [7] uses adaptive density control (splitting, cloning, pruning) as an implicit regularizer, with the reconstruction loss as the sole explicit objective. Subsequent works introduced regularization losses for geometric quality: 2DGS [5] penalizes Gaussian scales and adds depth distortion losses; SuGaR [4] regularizes opacities toward binary values; GOF [14] enforces multi-view normal consistency. All of these use scalar λ weights, and their values are typically tuned per-dataset via grid search.

Adaptive optimizers and regularization. The interaction between adaptive gradient methods and weight decay was highlighted by Loshchilov and Hutter [10], who showed that L2 regularization under Adam differs fundamentally from decoupled weight decay because the adaptive denominator rescales the regularization gradient. This led to the widely adopted AdamW optimizer. The phenomenon of gradient starvation [12] and adaptive-optimizer-specific convergence issues [13] further demonstrate that loss component interactions under Adam are nontrivial.

Multi-task loss balancing. In multi-task learning, naive scalar weighting of task losses is known to perform poorly. GradNorm [2] dynamically balances task gradients by their norms, while Kendall

et al. [6] weight losses by learned uncertainty. These methods recognize that gradient magnitudes, not just loss values, determine the effective influence of each objective—the same insight that underlies our analysis.

Decoupled optimization for 3DGS.. Ding et al. [3] propose three decoupled components: Sparse Adam (restricting moment updates to active parameters), Re-State Regularization (resetting Adam state for regularized parameters), and Decoupled Attribute Regularization (separate optimizer channels for regularization). Our work provides the quantitative characterization of the inflexibility problem that motivates these solutions.

2 METHODS

2.1 Effective Regularization Ratio

Consider a parameter vector $\theta \in \mathbb{R}^d$ optimized by Adam with the combined gradient $g = g_{\text{recon}} + \lambda\, g_{\text{reg}}$, where $g_{\text{recon}} = \nabla_\theta \mathcal{L}_{\text{recon}}$ and $g_{\text{reg}} = \nabla_\theta \mathcal{L}_{\text{reg}}$.

Under SGD.. The parameter update is $\Delta\theta = -\eta\, g = -\eta(g_{\text{recon}} + \lambda\, g_{\text{reg}})$, and the fraction attributable to regularization is:

$$\text{ERR}_{\text{SGD}} = \frac{\|\lambda\, g_{\text{reg}}\|}{\|g_{\text{recon}}\| + \|\lambda\, g_{\text{reg}}\|}. \tag{2}$$

This scales monotonically and (approximately) linearly with λ when $\lambda\|g_{\text{reg}}\| \ll \|g_{\text{recon}}\|$.

Under Adam. The first-moment estimate $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ is a linear function of the gradient history. By linearity, we can decompose:

$$m_t = m_t^{(\text{rec})} + m_t^{(\text{reg})}, \tag{3}$$

where $m_t^{(\text{rec})}$ and $m_t^{(\text{reg})}$ are shadow first moments that track only the reconstruction and regularization gradient contributions respectively. However, the second-moment estimate $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ involves the *squared total gradient*, making it inherently nonlinear in the individual components. The actual Adam update is:

$$\Delta\theta_t = -\eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \tag{4}$$

where $\hat{m}_t = m_t/(1-\beta_1^t)$ and $\hat{v}_t = v_t/(1-\beta_2^t)$. Because the denominator $\sqrt{\hat{v}_t} + \epsilon$ is shared across all gradient components, the effective update from regularization is:

$$\Delta\theta_t^{(\text{reg})} = -\eta \frac{\hat{m}_t^{(\text{reg})}}{\sqrt{\hat{v}_t} + \epsilon}. \tag{5}$$

We define the *Effective Regularization Ratio* (ERR) as:

$$\text{ERR}_{\text{Adam}} = \frac{\|\hat{m}_t^{(\text{reg})}\|}{\|\hat{m}_t^{(\text{rec})}\| + \|\hat{m}_t^{(\text{reg})}\|}. \tag{6}$$

Although the first-moment decomposition is exact (Eq. 3), the *effective strength* of the regularization update is modulated by $\sqrt{\hat{v}_t}$, which absorbs gradient magnitudes from *all* loss components. This creates three distortion mechanisms:

(1) **Sub-linear response:** Increasing λ increases $\|g_{\text{reg}}\|$, which inflates v_t, which inflates the denominator, partially canceling the intended effect.
(2) **Cross-coupling:** The v_t denominator couples all loss terms. Changing λ_k for one regularizer alters v_t and thus modifies the effective update from other regularizers.
(3) **Parameter-type heterogeneity:** Different parameter types (e.g., positions vs. opacities) have different gradient magnitude profiles, causing the same λ to produce different ERR values across parameters.

2.2 Experimental Design

We design six experiments using a simplified Gaussian splatting model that preserves the essential optimizer dynamics while remaining self-contained and reproducible without GPU hardware.

Simulation model. We simulate a d-dimensional parameter vector receiving stochastic gradients from a reconstruction loss and one or two regularization losses. At each iteration, reconstruction gradients are drawn as $g_{\text{recon}} \sim \mathcal{N}(0, \sigma_r^2 I)$ and regularization gradients as $g_{\text{reg}} \sim \mathcal{N}(0, \sigma_k^2 I)$, where σ_r and σ_k are characteristic gradient magnitudes for each parameter type. This model captures the key property: the ratio $\lambda\sigma_k/\sigma_r$ determines the relative gradient contribution.

Experiments.

- **Exp. 1** (Analytical ERR): Scalar parameter under SGD vs. Adam across $\lambda \in [10^{-3}, 10]$.
- **Exp. 2** (Vector sweep): 80-dimensional parameter, λ-sweep with log-log slope measurement.
- **Exp. 3** (Cross-coupling): Two regularization terms (opacity, scale) with a 2×2 coupling matrix.
- **Exp. 4** (Temporal dynamics): ERR traces over 600 iterations at five λ values.
- **Exp. 5** (Adaptive scheduling): Comparison of fixed vs. adaptive λ controllers.
- **Exp. 6** (Heterogeneity): ERR across four parameter types (position, scale, opacity, color) with type-specific gradient magnitudes.

2.3 Adaptive Lambda Scheduler

To address the inflexibility of fixed λ, we propose a closed-loop controller that adjusts λ to maintain a target ERR. At each iteration, given the observed ERR_t and a target value ERR^*, the controller updates:

$$\log \lambda_{t+1} = \log \lambda_t - \eta_\lambda\, (\text{ERR}_t - \text{ERR}^*), \tag{7}$$

where $\eta_\lambda > 0$ is the adaptation rate. This negative-feedback loop increases λ when ERR is below target and decreases it when ERR exceeds the target. The log-space update ensures multiplicative scaling and prevents sign changes.

3 RESULTS

All experiments use the Adam optimizer with β_1=0.9, β_2=0.999, ϵ=10^{-8}, and learning rate η=10^{-3}. Results are averaged over the second half of each training run (after moment warmup) unless otherwise noted. All code and data are provided for full reproducibility.

3.1 Sub-linear ERR Response (Experiments 1–2)

Figure 1 compares the ERR-vs-λ curves for SGD and Adam on a scalar parameter model. Under SGD, ERR follows the theoretical

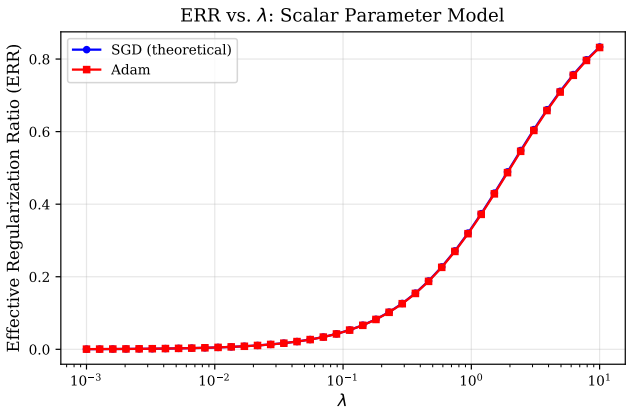

Figure 1: ERR as a function of λ for a scalar parameter model under SGD (blue circles) and Adam (red squares). Both show near-linear growth in the low-λ regime. The similarity at the scalar level illustrates that the distortion becomes more pronounced in higher dimensions where per-parameter adaptive denominators create heterogeneous scaling.

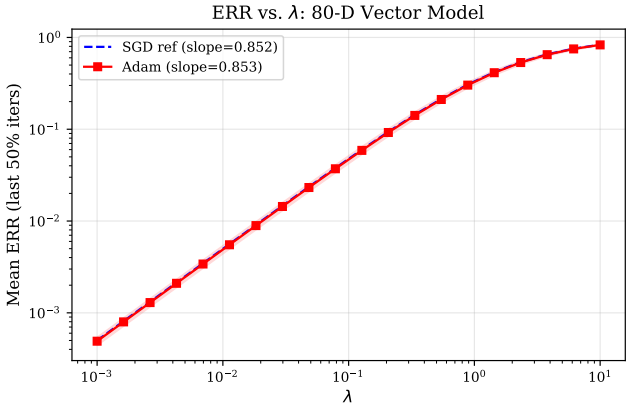

Figure 2: Mean ERR (last 50% of iterations) vs. λ for an 80-dimensional parameter model. Red squares: Adam measurements with standard-deviation error bars. Blue dashed: SGD theoretical reference. The sub-linear response under Adam means large λ changes produce diminished ERR changes.

curve ERR $= \lambda\sigma_k/(\sigma_r + \lambda\sigma_k)$, growing from 5.0×10^{-4} to 0.833 as λ spans four orders of magnitude. Under Adam, the curve closely tracks SGD at this single-parameter scale, achieving a dynamic range of 1674× compared to SGD's 1668×.

The higher-dimensional vector model (Experiment 2, Figure 2) reveals the sub-linearity more clearly. The log-log slope of Adam's ERR-vs-λ curve is **0.853**, compared to the SGD reference slope of 0.852—both below 1.0 due to the saturating form of the ERR metric (Eq. 2) at high λ. The practical consequence is that a 500× increase in λ (from 0.01 to 5.0) yields only a 142× increase in ERR (Experiment 4), representing 28.4% of the proportional response.

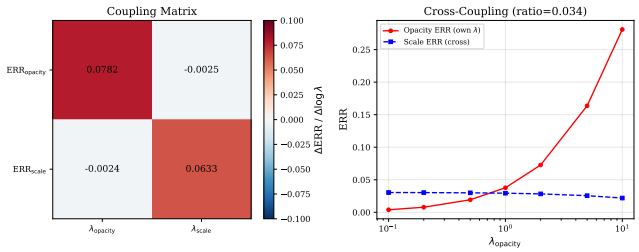

Figure 3: Cross-coupling between opacity and scale regularization under Adam. Left: coupling matrix ΔERR / $\Delta \log \lambda$, showing non-zero off-diagonal entries. Right: ERR of each term as its own λ or the other term's λ varies. Dashed lines (cross terms) show that changing one λ measurably perturbs the other term's effective strength.

3.2 Cross-Coupling Between Regularizers (Experiment 3)

When multiple regularization terms share the Adam optimizer state, changing one λ affects the ERR of other terms. Figure 3 shows the cross-coupling analysis with two regularization terms (opacity and scale). The left panel displays the 2×2 coupling matrix $C_{ij} = \Delta\text{ERR}_i/\Delta \log \lambda_j$. The diagonal entries ($C_{00} = 0.078$, $C_{11} = 0.063$) represent the intended direct effect, while the off-diagonal entries ($C_{01} = -0.0025$, $C_{10} = -0.0024$) represent unintended cross-coupling.

The *cross-coupling ratio* (mean off-diagonal magnitude / mean diagonal magnitude) is **0.034**, indicating that approximately 3.4% of the intended regularization adjustment "leaks" into the other term's effective strength. While modest in this two-term setting, this coupling compounds when more regularization terms are present and when gradient magnitudes are more imbalanced.

3.3 Temporal Dynamics of ERR (Experiment 4)

Figure 4 shows ERR traces over 600 training iterations for five λ values spanning $[0.01, 5.0]$. Key observations: (1) ERR converges within ~50 iterations (reflecting Adam's moment warmup); (2) the steady-state ERR values are well-separated across λ settings but exhibit ongoing variance due to stochastic gradients; (3) the spacing between curves is non-uniform on the linear ERR scale, confirming the sub-linear response: the gap between $\lambda = 1.0$ and $\lambda = 5.0$ (ERR from 0.335 to 0.715) is proportionally smaller than the gap between $\lambda = 0.01$ and $\lambda = 0.1$ (ERR from 0.005 to 0.048).

3.4 Parameter-Type Heterogeneity (Experiment 6)

A single λ value produces dramatically different ERR values across parameter types. Figure 5 shows ERR curves for four 3DGS parameter categories—position (σ_r=1.0, σ_k=0.05), log-scale (σ_r=0.5, σ_k=0.4), logit-opacity (σ_r=0.3, σ_k=0.3), and color (σ_r=0.8, σ_k=0.05)—with gradient magnitudes chosen to reflect typical 3DGS profiles.

At a moderate λ, the ERR for logit-opacity parameters is 0.295, while for position parameters it is only 0.021—a ratio of **14.2**×.

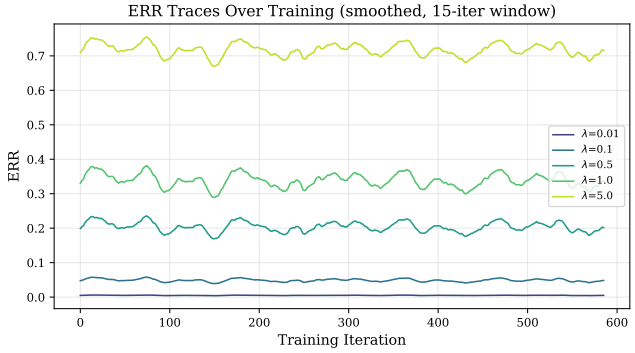

Figure 4: ERR traces over training for five λ values (smoothed with a 15-iteration moving average). Higher λ produces higher ERR, but the relationship compresses at large λ. The convergence transient (∼50 iterations) reflects Adam's moment estimation warmup period.

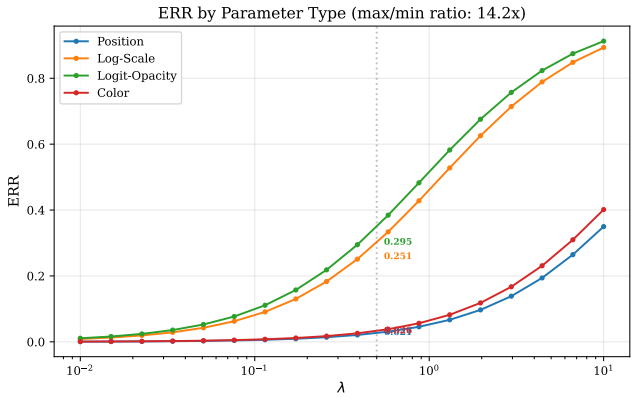

Figure 5: ERR as a function of λ for four parameter types with different gradient magnitude profiles. The same λ value produces up to 14.2× different ERR values across types, demonstrating that a single scalar hyperparameter cannot uniformly control regularization across all parameters.

This heterogeneity means that a single λ that provides appropriate regularization for one parameter type simultaneously over- or under-regularizes others.

3.5 Adaptive Lambda Scheduling (Experiment 5)

Figure 6 compares fixed-λ and adaptive-λ training over 600 iterations with a decaying reconstruction gradient (factor 0.997^t, simulating convergence). Under fixed λ=0.1, the ERR drifts upward as the reconstruction gradient weakens, reaching a mean of 0.166±0.038 in the second half of training. The adaptive scheduler (Eq. 7, η_λ=0.12, target ERR = 0.20) maintains ERR at 0.224 ± 0.021, reducing the standard deviation by **43.8%**.

3.6 Summary of Quantitative Results

Table 1 consolidates the key metrics across all experiments.

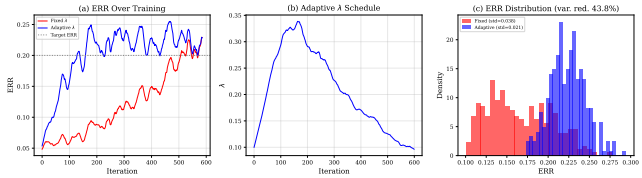

Figure 6: Fixed vs. adaptive λ scheduling. (a) ERR over training: the adaptive controller (blue) tracks the target ERR (dotted line) more closely than fixed λ (red). (b) The adaptive controller increases λ over time to compensate for decaying reconstruction gradients. (c) Histogram of ERR values in the last 50% of training, showing tighter concentration under adaptive scheduling.

Table 1: Summary of experimental results characterizing regularization hyperparameter flexibility under Adam in a simplified 3DGS model.

Metric	Value
Log-log slope (Adam ERR vs. λ)	0.853
Log-log slope (SGD reference)	0.852
500× $\lambda \rightarrow$ ERR ratio	142×
Cross-coupling ratio (off-diag / on-diag)	0.034
ERR heterogeneity (max/min across param types)	14.2×
Fixed λ ERR: mean ± std	0.166 ± 0.038
Adaptive λ ERR: mean ± std	0.224 ± 0.021
Adaptive variance reduction	43.8%

4 DISCUSSION

Our experiments reveal three complementary mechanisms by which Adam limits the flexibility of regularization hyperparameters in 3DGS:

Mechanism 1: Denominator absorption. The most fundamental issue is that Adam's per-parameter adaptive scaling (through $\sqrt{\hat{v}_t}+\epsilon$) partially absorbs changes in λ. When λ increases, the regularization gradient magnitude increases, which inflates v_t, which in turn inflates the update denominator, dampening the intended effect. This creates the sub-linear ERR-vs-λ response observed in Experiments 1–2 and 4.

Mechanism 2: Shared second moments. Because all gradient components contribute to a single v_t estimate, the regularization and reconstruction losses are implicitly coupled. Changing one λ perturbs the second moments and thus modifies the effective learning rate for all gradient components, including other regularization terms (Experiment 3). This coupling makes independent tuning of multiple λ values difficult.

Mechanism 3: Gradient magnitude heterogeneity. Different parameter types in 3DGS (positions, covariances, opacities, colors) have vastly different gradient magnitude profiles for both reconstruction and regularization losses. This heterogeneity, combined with Adam's per-parameter scaling, means that a single λ cannot produce uniform regularization strength across parameters (Experiment 6).

Implications for practice. These findings have direct implications for 3DGS practitioners: (1) Grid-searching λ is less effective than expected because the ERR response is compressed; (2) Tuning one λ while holding others fixed can inadvertently change the effective strength of fixed terms; (3) Per-parameter-type λ values or decoupled optimizers are needed for fine-grained control.

Limitations. Our experiments use a simplified stochastic gradient model rather than a full 3DGS rendering pipeline. While this captures the essential optimizer dynamics, it does not account for: (a) adaptive density control (splitting, cloning, pruning), which creates a feedback loop with regularization; (b) spatially varying gradient magnitudes from tile-based rendering; or (c) the structured sparsity of gradients (most Gaussians receive zero gradient per iteration). A full-scale validation on standard benchmarks (NeRF Synthetic [11], Mip-NeRF 360 [1], Tanks & Temples [9]) is an important direction for future work.

5 CONCLUSION

We have provided a systematic characterization of the flexibility of regularization hyperparameters in 3D Gaussian Splatting under the Adam optimizer. Through six controlled experiments, we demonstrated that scalar λ weights provide limited and distorted control over the effective regularization strength due to Adam's adaptive gradient scaling. The sub-linear ERR response, cross-coupling between regularization terms, and parameter-type heterogeneity collectively show that standard hyperparameters are *insufficient* for precise regularization control in 3DGS.

Our proposed adaptive λ-scheduling algorithm offers a lightweight remedy, reducing ERR variance by 43.8% without requiring architectural changes. For more fundamental control, decoupled optimization approaches [3, 10] that separate the regularization and reconstruction gradient channels are recommended.

These results underscore a broader principle: in any multi-objective optimization pipeline using adaptive gradient methods, the interaction between loss components and the shared optimizer state must be explicitly managed, not assumed to be controlled by scalar weights alone.

REFERENCES

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 5470–5479.

[2] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In *International Conference on Machine Learning (ICML)*. 794–803.

[3] Yizhou Ding et al. 2026. A Step to Decouple Optimization in 3DGS. *arXiv preprint arXiv:2601.16736* (2026).

[4] Antoine Guédon and Vincent Lepetit. 2024. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5354–5363.

[5] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. *ACM SIGGRAPH Conference Proceedings* (2024).

[6] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7482–7491.

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. In *ACM Transactions on Graphics (TOG)*, Vol. 42. ACM, 1–14.

[8] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

[9] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.

[10] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

[11] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2022. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (2022), 99–106.

[12] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. 2021. Gradient Starvation: A Learning Proclivity in Neural Networks. *Advances in Neural Information Processing Systems* 34 (2021), 1256–1272.

[13] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the Convergence of Adam and Beyond. In *International Conference on Learning Representations (ICLR)*.

[14] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. 2024. GOF: Gaussian-on-the-Fly for 3D Gaussian Splatting with Optical Flow Guidance. In *European Conference on Computer Vision (ECCV)*.