

HybridGS: Decoupling Transients and Statics with 2D and 3D Gaussian Splatting

Supplementary Material

1. Positioning of Our Work

In the community, there are currently two predominant approaches for tackling the challenge of novel view synthesis in wild images with different complexities. We distinguish these approaches according to the primary datasets they utilize and provide a detailed comparison in Tab. 1.

NeRF On-the-go [7] processes casually captured images that lack inter-frame continuity, with the goal of eliminating the interference from transient objects to reconstruct statics.

Photo Tourism [10] gathers photo collections from the web, resulting in completely unconstrained conditions with more complex lighting variations and increased foreground interference. It focuses more on integrating appearance embedding to model the photometric changes in the scene.

In summary, our method belongs to the first category and aims to decompose transients and statics from casually captured images in scenes with minimal illumination changes. Experiments have demonstrated our state-of-the-art results on two widely used benchmark datasets, such as NeRF On-the-go [7] and RobustNeRF [8]. Handling varying lighting conditions will be our future work as discussed in the paper.

2. More Discussions

2.1. 2D Gaussians

The fitting capability of 2D Gaussians is inherited from 3D Gaussians. Given \mathbf{J} , the Jacobian of the affine projective transformation, and \mathbf{W} , the viewing transformation, the 3D Gaussians can be projected to 2D image plane and blended through a fast, differentiable α -blending process to render 2D images following the Eq. 3 and Eq. 4. Therefore, the 2D Gaussians can be viewed as the projection of 3D Gaussians.

During training, the warm-up allows 3DGS to establish an initial model of the entire scene. It is noteworthy that intuitively, the residuals between the results of 3DGS rendering and the ground-truth would potentially model transients. However, the 3DGS itself is constrained only by RGB loss, therefore, the transient objects from different viewpoints are eventually fitted into the 3D Gaussians, leading to less effective fitting of static scenes with vanilla 3DGS. We address this issue by incorporating additional 2D Gaussians. During the iterative training stage, the 2DGS learns the residuals per view, focusing more on the unique elements of each image. The output soft mask or matting can effectively direct 3DGS to concentrate on areas with smaller residuals, which represent the common and shared elements of the scene. In the final joint training stage, we perform a deep

Table 1. Comparison of NeRF On-the-go and Photo Tourism.

	NeRF On-the-go [7]	Photo Tourism [10]
Data source	Casually captured photos	Web photos in the wild
Photometric	Similar lighting	Varying lighting over time
Scene	Indoor and outdoor scenes	Mostly outdoor scenes
Evaluation	Statics	Statics with their illumination
Related Works	[5, 7, 9] & Our HybridGS	[2–4, 6, 11, 12]

Table 2. We conduct ablation studies on a pure static scene *Garden* from the MipNeRF 360 dataset [1], and *Corner*, scene that includes both transients and statics, from the NeRF On-the-go dataset [7], to explore the potential influence of our designs.

Method Settings	Garden			Corner		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
3DGS	29.323	0.924	0.050	20.148	0.686	0.202
+ Multi-view	29.572	0.925	0.053	21.758	0.769	0.147
+ Multi-view + 2DGS	29.512	0.924	0.054	25.020	0.847	0.077

Table 3. Comparison of different number of views in *Corner* in multi-view supervision. The best results are highlighted in **bold**.

Metric	K=1	K=2	K=4	K=8
PSNR	24.30	24.78	25.03	24.87
SSIM	0.820	0.839	0.847	0.842
LPIPS	0.123	0.081	0.079	0.077

integration of 2D and 3D Gaussians for fine-tuning. Therefore, in our method, 3D Gaussians tend to learn the elements that are consistent across different viewpoints, which we define as statics. Meanwhile, 2D Gaussians capture image-specific information, such as dynamic objects and occlusions, referred to as transients.

2.2. Multi-view Supervision

We have further investigated the performance of multi-view 3DGS in different scenarios in Tab. 2. To be specific, we select the static scene *Garden* from the MipNeRF 360 [1] dataset and the dynamic scene *Corner* from the NeRF On-the-go [7] dataset. The results indicate that in static scenes, employing multi-view supervision has minimal impact on achieving the best results for novel view synthesis. However, during training, it is noted that 3DGS tends to over-fit the training views, leading to a gradual decline in both visual quality and metric performance for novel views. In contrast, multi-view 3DGS demonstrates more stable convergence and effectively reduces the over-fitting problem.

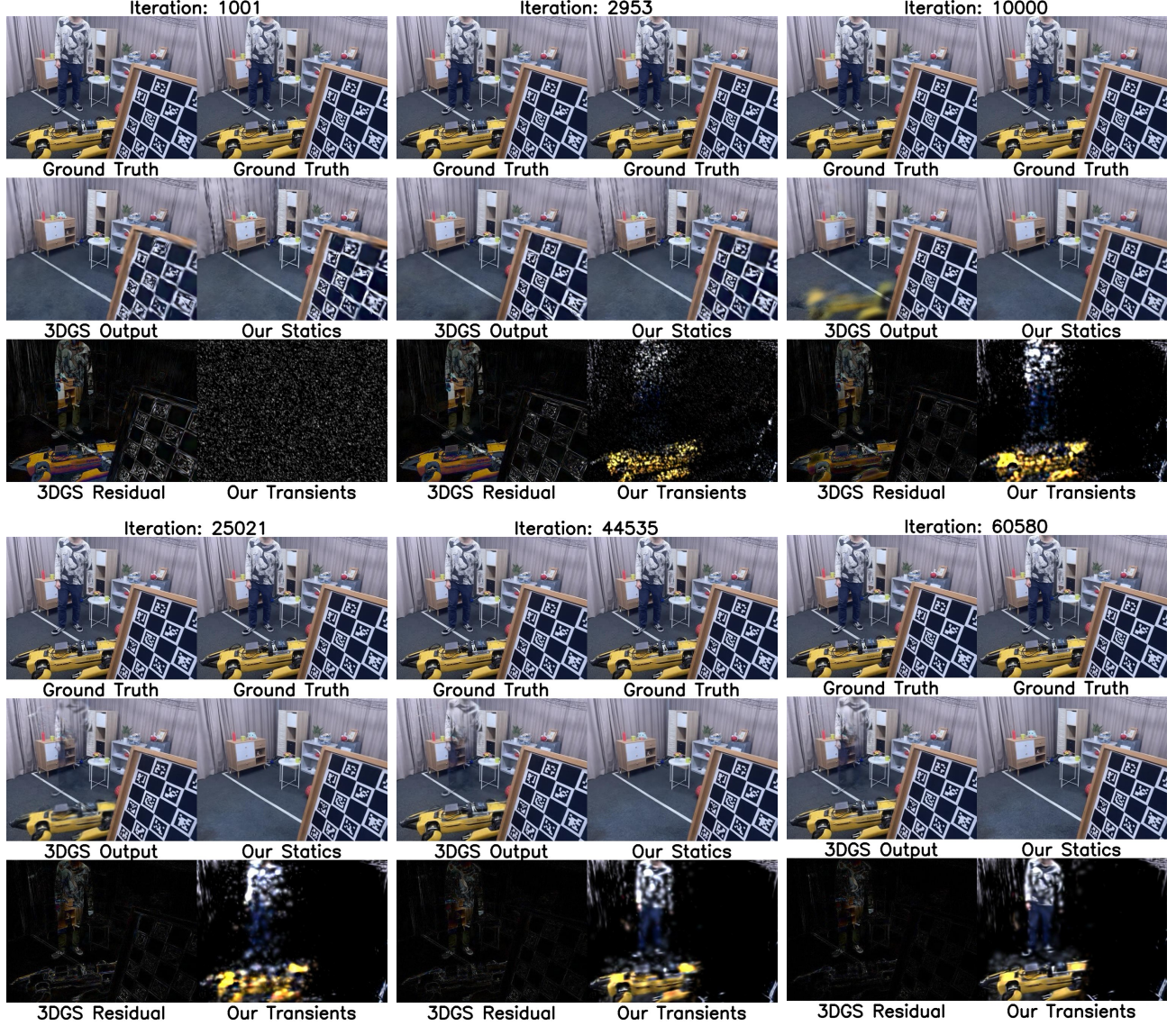


Figure 1. **Qualitative results compared to 3DGS at each randomly selected training step.** For HybridGS, the training steps for each stage are: Warm-up: 0~1,010, Iterative Training: 1,010~40,400, Joint Training: 40,400~60,600. Notably, this scene contains 101 images. Referring to Fig. 3, before the warm-up step (the top left), during the initial rough training phase, both our and 3DGS results are somewhat blurry. However, since we adopt a multi-view strategy, the occluded parts in our results are slightly clearer. As training progresses, by the 2953rd iteration (the top central), 3DGS reaches its optimum. Nevertheless, at this point, the background in the transients remains quite blurry for 3DGS, whereas our approach has already transitioned into the iterative training phase, allowing us to model static elements more accurately. Moving forward, we maintain stable training (top right and bottom left), largely due to our introduction of 2D Gaussians to decouple transients from statics. This effectively prevents over-fitting to transients that 3DGS begins to experience, leading to diminishing rendering quality. By the 44535th iteration (the bottom central), during the joint optimization phase, our results reach their optimum. The bottom right shows the results at the end of the training process.

In dynamic scenes, obviously, 3DGS is prone to over-fitting, which adversely affects performance in novel view synthesis. On the other hand, multi-view 3DGS benefits from mutual supervision in areas visible to multiple views, significantly improving the visual results.

Unlike standard multi-batch 3DGS, we focus on multi-

view-visible regions for better efficiency and effectiveness. Here, $K = 1$ means training transients on a single image without considering co-visible areas. An ablation study shows that using $K = 4$ views achieves the best performance across metrics.



Figure 2. Visualization on 6 remaining scenes of NeRF On-the-go [7] dataset.

2.3. Comparison with in-the-wild methods

As illustrated in the Sec 1 In-the-wild methods focus on modeling photometric variations in Photo Tourism differs from our goal of removing randomly captured transients, evaluated on NeRF On-the-go and RobustNeRF. We compared with SOTA in-the-wild approaches on NeRF On-the-go dataset using open-source codes with default settings. Our method outperforms methods such as WildGaussians [5], demonstrating its effectiveness.

3. More Implementation Details

3.1. Training and Storage

For the training of 3D Gaussians, we perform the densification of 3D Gaussians during the warm-up stage. Then, in the subsequent stages, we maintain a constant number of existing 3D Gaussians and focus solely on optimizing their parameters. For 2D Gaussians, we maintain a constant number 10,000 per image throughout the entire training process without any densification. During the iterative training process, while optimizing 3DGS with 2DGS held fixed, we binarize the uncertainty mask obtained from 2DGS into 0s and 1s using a threshold value of $\epsilon = 0.1$.

Table 4. **Comparison with other state-of-the-art in-the-wild methods.** The best results are highlighted in **bold**. * indicates that the results are reproduced from the official implementation.

Dataset(ratio)	On-the-go low.(5%)			On-the-go medium.(17%)			On-the-go high.(26%)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF-W [6]	17.63	0.451	0.518	18.88	0.620	0.397	14.69	0.366	0.648
Gaussian in-the-wild* [12]	20.32	0.601	0.256	22.46	0.769	0.146	22.20	0.685	0.246
WildGaussians [5]	20.62	0.657	0.235	22.80	0.811	0.092	23.02	0.770	0.172
Ours	21.42	0.684	0.206	23.50	0.827	0.092	23.05	0.768	0.114

We used an early version of Taming-3DGS (24.09) for baseline, which speeds up training but does not optimize storage. Our method is more efficient in both storage and training time. Specifically, baseline 3DGS requires many small Gaussians to overfit view-dependent transients, taking over 150 epochs on densification and resulting in longer optimization time and higher memory usage. In contrast, our method represents the static scene with a densification process for only 10 epochs, while decomposing transient parts into 2D Gaussians per image, with minimal additional storage and reduced optimization time (Ours 10.8 mins vs Baseline 12.1 mins). Compared to 3DGS, our method use 2D Gaussians to model transients instead of forcing 3D Gaussians to fit them, significantly reducing both storage and computational requirements, showcasing its superior efficiency in both training and testing phases.

3.2. Datasets

We follow the same training/testing split and resolution settings as the official rules in NeRF On-the-go [7] and RobustNeRF [8]. Specifically, for the NeRF On-the-go dataset, we downsample images from most scenes by $8\times$ to 504×378 . Note that *Arcdetriomphe* and *Patio* are downsampled by $4\times$ to 480×270 . For the RobustNeRF dataset, all scenes are downsampled by $8\times$, with *Android* and *Statue* resized to 503×377 , and *Crab* and *Yoda* to 431×431 .

4. More Visualization Results

4.1. Training Process

To better demonstrate the changes during our training process, we select `IMG_7195.JPG` of *Corner* from NeRF On-the-go dataset as example, visualizing the statics and transients during different training stages and comparing them with vanilla 3DGS in Fig. 1. As training iterations increase, 3DGS tends to gradually integrate transient elements into the static components, rendering the residuals being almost incapable of capturing transient contents. In contrast, our HybridGS effectively distinguishes transients from statics over time, leading to consistent improvements in Fig. 3.

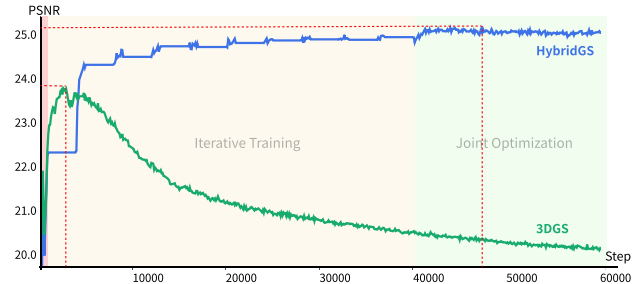


Figure 3. **The PSNR on testing set of Corner during training process.** Training steps for each stage of our HybridGS: Warm-up : 0~1,010, Iterative Training : 1,010~40,400, Joint Training : 40,400~60,600. We use the same data as in Fig. 1. Note that, 3DGS reaches its optimum at step 2953, but as training continues, 3DGS tends to overfit transients in dynamic scenes, leading to gradual decline in performance. In comparison, our method is able to train steadily. This directly validates the statements in Sec. 2 and 4.1.

4.2. More Scenes

In addition to providing metrics and results on the 6 commonly used scenes of NeRF On-the-go [7] dataset, we also present the visualization results on the remaining scenes as shown in Fig. 2. These complex scenes include some variations in lighting and shadows. We find that in addition to removing dynamic objects, our statics can also eliminate elements lacking specific semantics, such as shadows of pedestrians (in *Drone* and *Train Station*) and cars (in *Arcdetriomphe* and *Train*). This separation of non-semantic transients illustrates that our method is fundamentally a versatile, low-level and semantics-free scene decomposition approach, effectively highlighting its generality and robustness.

4.3. More Datasets

We apply our method on Photo Tourism [10] dataset, which consists of unconstrained photo collections with photometric variations. As shown in Fig. 4, we have some intriguing and reasonable observations. First, the statics generated using 3D Gaussians are rendered under an average light condition derived from the training images, similar to the dif-

fuse lighting on an overcast day. Additionally, we discover that besides modeling dynamic objects, 2D Gaussians also capture photometric differences in our transients, since the illumination difference is indeed a per-image characteristic. This finding perfectly aligns with the perspective we presented in Sec. 2 that transients can capture unique aspects of each image, broadening the scope for future research to further isolate photometric information from 2D Gaussians.

5. Limitations

HybridGS may classify sparsely occurring static elements as transients, especially in sparse scenes. Similarly, issues can occasionally arise with illumination and shadows lacking multi-view consistency. A potential solution is to decompose the input into reflectance and shadow layers, and focus on intrinsic colors, laying the foundation for applying 3DGS in more complex scenes.

References

- [1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 1
- [2] Jiahao Chen, Yipeng Qin, Lingjie Liu, Jiangbo Lu, and Guanbin Li. Nerf-hugs: Improved neural radiance fields in non-static scenes using heuristics-guided segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19436–19446, 2024. 1
- [3] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Feng Ying, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12933–12942, 2021.
- [4] Hiba Dahmani. Swag: Splatting in the wild images with appearance-conditioned gaussians. In *ECCV*, 2024. 1
- [5] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. 2024. 1, 3, 4
- [6] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7206–7215, 2020. 1, 4
- [7] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 3, 4
- [8] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J. Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20626–20636, 2023. 1, 4
- [9] Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David J. Fleet, and Andrea Tagliasacchi. SpotLessSplats: Ignoring distractors in 3d gaussian splatting. *arXiv:2406.20055*, 2024. 1
- [10] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH Conference Proceedings*, pages 835–846, New York, NY, USA, 2006. ACM Press. 1, 4, 6
- [11] Yuze Wang, Junyi Wang, and Yue Qi. We-gs: An in-the-wild efficient 3d gaussian representation for unconstrained photo collections, 2024. 1
- [12] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. *ECCV*, 2024. 1, 4

