

LiteVGGT: Boosting Vanilla VGGT via Geometry-aware Cached Token Merging

Supplementary Material

1. More Implementation Details

Fine-tuning Details. Following VGGT’s official fine-tuning pipeline, we fine-tune only the aggregator together with the camera head and depth head, starting from the pretrained VGGT checkpoints. We use a diverse mixture of synthetic and real-world datasets, including Co3Dv2[9], BlendMVS[13], DL3DV[6], MegaDepth[5], WildRGB[12], ScanNet++[14], HyperSim[10], Mapillary[7], Replica[11], MVS-Synth[3], Virtual KITTI[1], Aria Synthetic Environments, and Aria Digital Twin[8]. We sample 4–48 images per batch and train for 20K iterations on 8 H20 GPUs (approximately 3 days). To stabilize training under token merging, we adopt a composite learning rate schedule: a 5% linear warm-up from 1×10^{-6} to 4×10^{-5} , followed by cosine decay to 7×10^{-7} over the remaining iterations.

FP8 Quantization Details. We apply FP8 mixed precision via NVIDIA’s Transformer Engine directly during inference. To balance efficiency and accuracy, we quantize only the modules inside the aggregator (including the DINOv2 encoder). For example, each `nn.LayerNorm` + two-layer MLP block is replaced with its FP8-enabled `te.LayerNormMLP` counterpart, while the prediction heads remain in `bfloat16` since quantizing them leads to a noticeable drop in accuracy. We adopt an FP8 (E4M3) delayed-scaling recipe with an 80-step amax history and max-based amax computation.

2. Additional Ablation Studies

Caching merge indices. Leveraging the stability of inter-layer token similarity, we cache merge indices and evaluate how different recomputation intervals affect the accuracy–efficiency trade-off.

Interval (layers)	Total Recompute	CD ↓	Acc ↓	Comp ↓	Overall ↓	Time (s) ↓
1	24	0.421	0.691	0.800	0.745	265.1
2	12	0.440	0.701	0.778	0.741	230.5
3	8	0.452	0.721	0.788	0.754	214.1
6	4	0.467	0.746	0.775	0.761	202.0
24	1	0.555	3.761	5.342	4.552	193.1

Table 1. **Ablation on merge-indices recomputation intervals.** Quantitative results of point cloud reconstruction on the Scannet-50[2] and DTU[4] dataset.

As shown in Table 1, enlarging the interval from every 1 layer to every 6 layers preserves nearly the same point cloud reconstruction quality while substantially reducing infer-

ence latency, further validating the stability of token similarity across adjacent layers. In contrast, using an excessively large interval (i.e., computing only at the first layer) causes a clear drop in reconstruction accuracy. Based on this trade-off, we adopt the 6-layer interval (4 total computations) as our preferred setting.

Geometry-aware Token Merging. We further evaluate our Geometry-aware Token Merging against standard similarity-based merging. As shown in Table 2, our method better preserves crucial geometric information and achieves better performance on both datasets. Meanwhile, the latency remains almost unchanged, highlighting the efficiency of our approach.

Method	CD ↓	Acc ↓	Comp ↓	Overall ↓	Time (s) ↓
VGGT*	0.485	0.508	0.561	0.534	1275.0
VGGT*+Naive Token Merging	0.442	0.824	0.655	0.739	202.0
VGGT*+GA Token Merging (Ours)	0.402	0.789	0.601	0.696	202.4

Table 2. **Ablation on Geometry-aware Token Merging.** Quantitative results of point cloud reconstruction on the Scannet-50[2] and DTU[4] dataset.

3. Robotic Grasping Experiments

To further verify that LiteVGGT remains practical even with its 10× reconstruction speed-up, we conduct robotic grasping experiments. As shown in Fig. 1, LiteVGGT reconstructs two real-world scenes, and the resulting point clouds are used for robotic arm grasping. Despite minor reconstruction deviations, the accuracy is sufficient for end-side grasp execution, demonstrating the practical reliability of LiteVGGT.

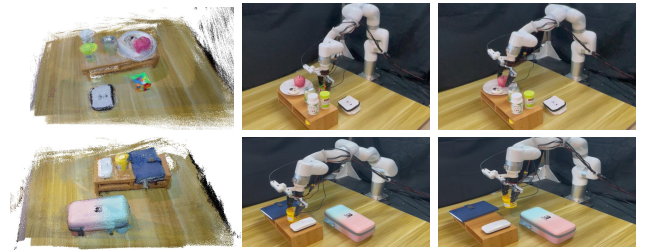


Figure 1. Robotic grasping demonstration. Left: reconstructed point clouds from LiteVGGT. Right: corresponding grasp execution snapshots. The full videos are provided in the supplementary material (video.mp4).

058 **4. Additional Visualizations**

059 In Fig. 2, we present additional reconstruction results
060 of LiteVGGT, and compare them against the ground-truth
061 (GT) point clouds as well as those produced by VGGT and
062 FastVGGT. In Fig. 3, we show further examples of cam-
063 era pose estimation, again comparing our results with GT,
064 VGGT, and FastVGGT. Finally, Fig. 4 provides additional
065 visualizations of pixel gradients (Grad map), token variance
066 (Variance map), and the fused Geometry-aware map (GA
067 map).



Figure 2. Additional 3D reconstruction visualizations from small indoor objects to large outdoor scenes.

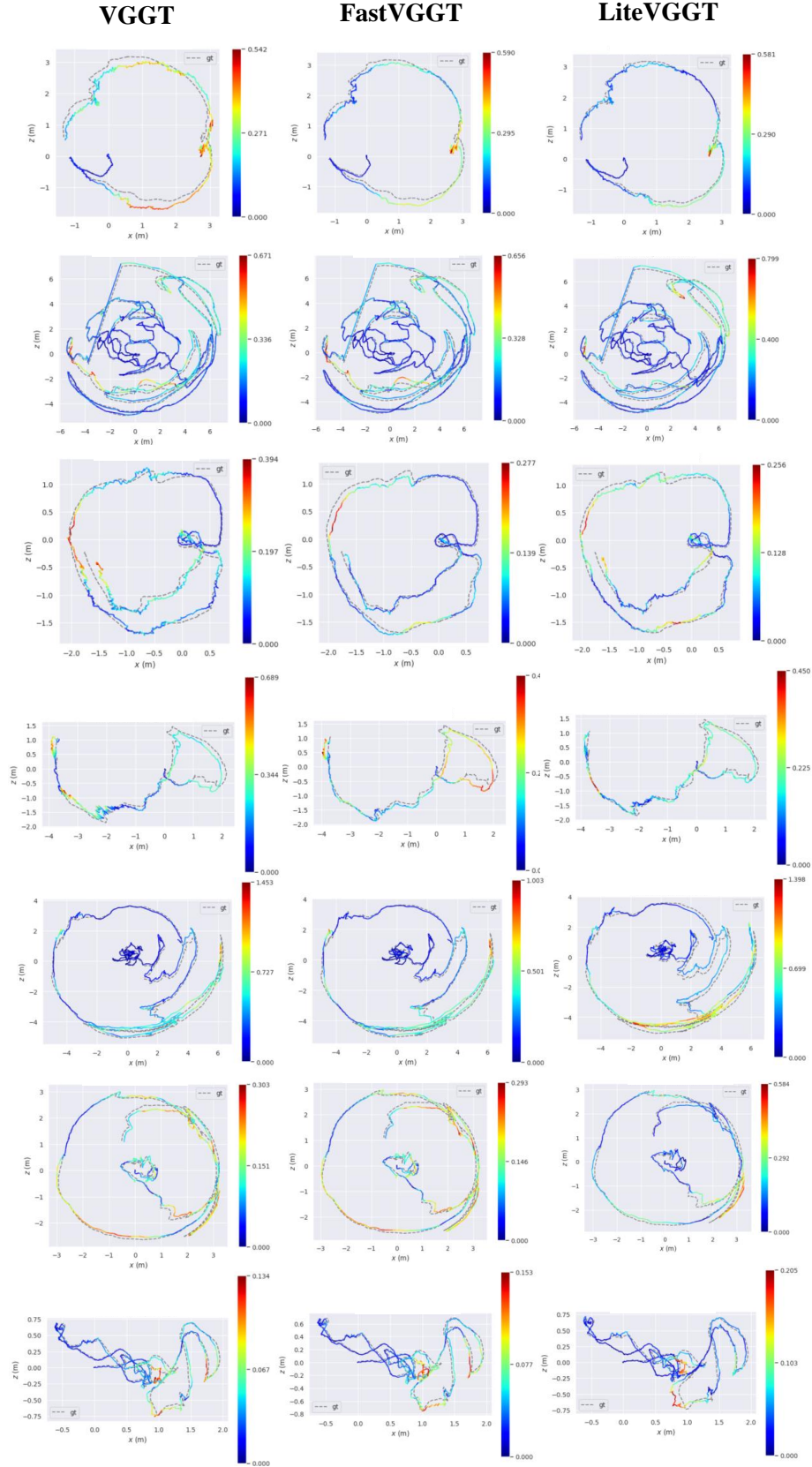


Figure 3. Additional visualizations of pose estimation results on the ScanNet dataset.

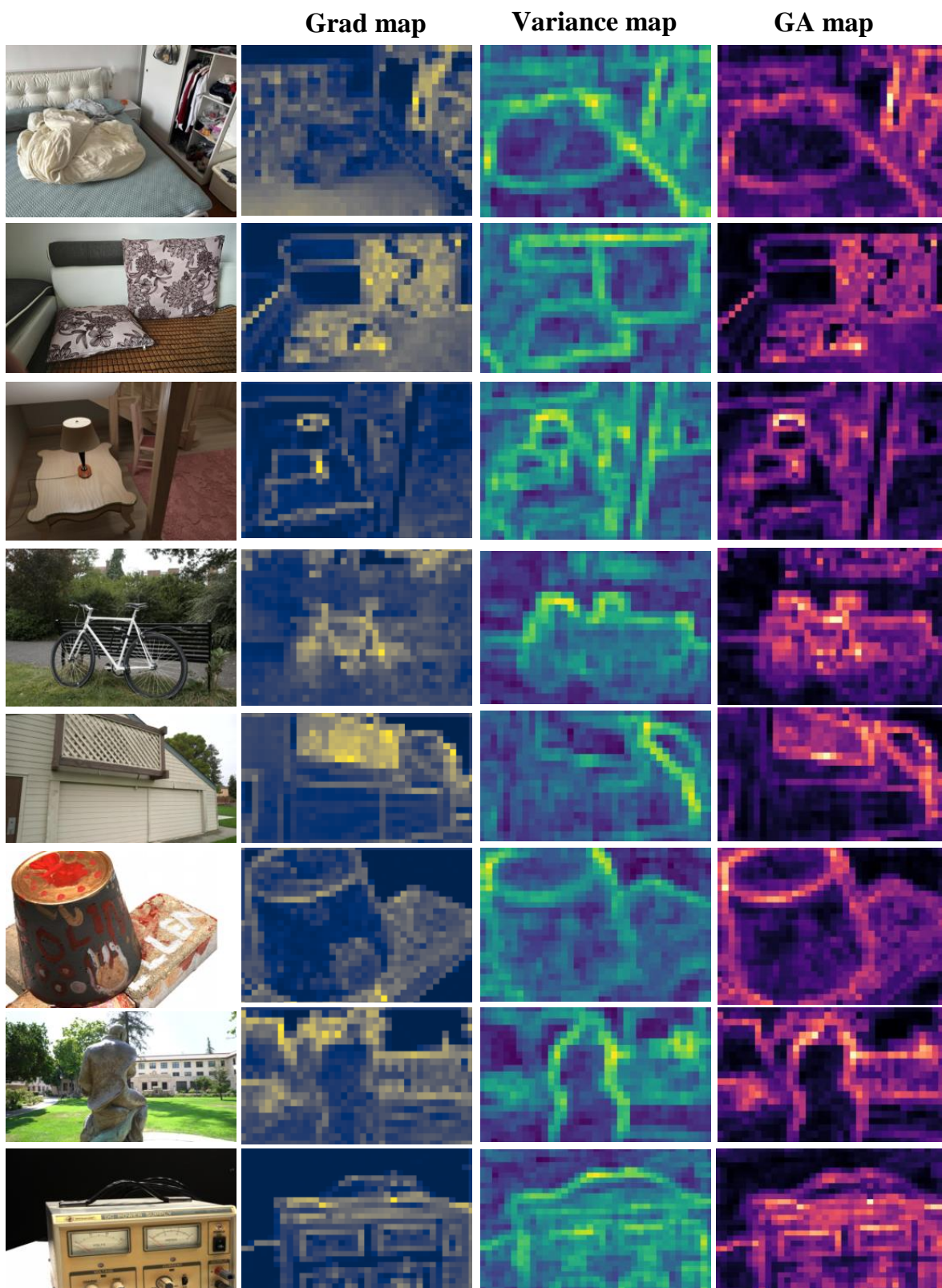


Figure 4. Additional visualizations of pixel gradients (Grad map), token variance (Variance map), and the fused Geometry-aware map (GA map).

References

- [1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 1
- [2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1
- [3] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2821–2830, 2018. 1
- [4] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 1
- [5] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1
- [6] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 1
- [7] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 1
- [8] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 1
- [9] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 1
- [10] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 1
- [11] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6229–6238, 2021. 1
- [12] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024. 1
- [13] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 1
- [14] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 1