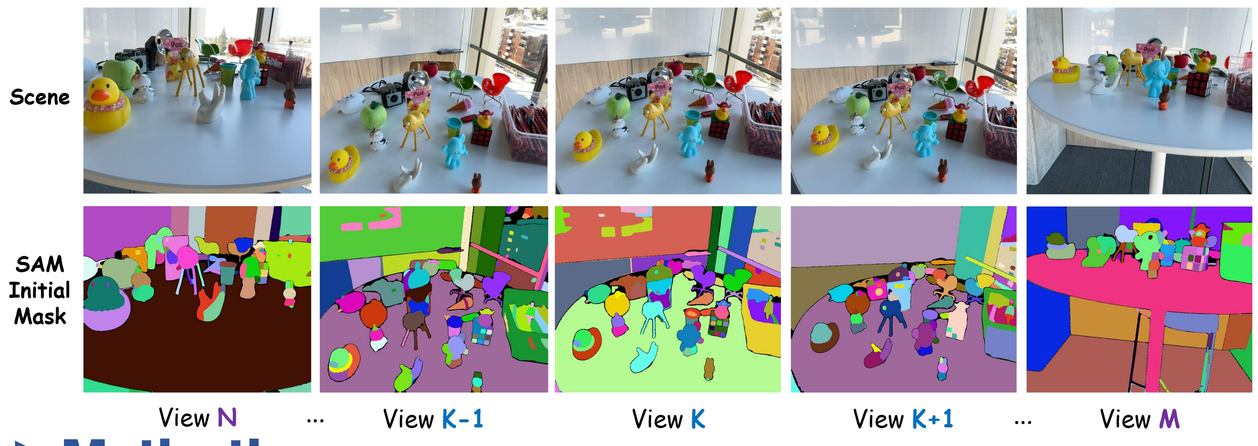
Consistent Scene Understanding in 3D Gaussian Splatting via Multi-Cue Mask Refinement

Hyunjoon Park, Donghyeon Cho
Department of Computer Science, Hanyang University

LinkedIn Project Page

ICPR 2026 International Conference on Pattern Recognition LYON

INTRODUCTION



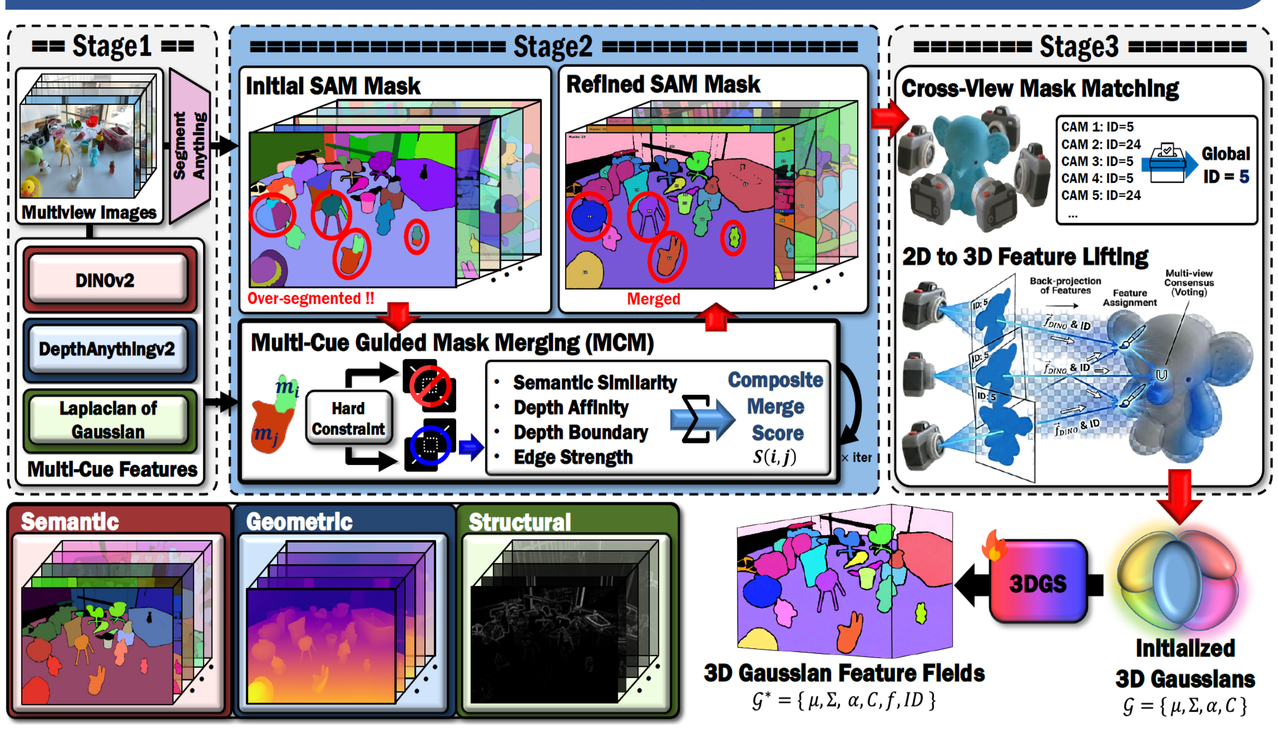
▶ Motivation

- **Cross-view Inconsistency**
- **Over-segmentation of SAM automatic mask generation**
  - ➔ Lack 3D spatial awareness and geometric context
  - ➔ Merge & Split segmentation mask of a single object erroneously

▶ Contribution

- **Multi-Cue Mask Refinement framework**
  : Refine SAM-induced over-segmented masks
- **Cross-view Mask Matching**
  : Establishes globally consistent object identities and suppresses view-dependent label inconsistencies
- **Multiview Consensus Lifting**
- **Performance Enhancement**

METHOD



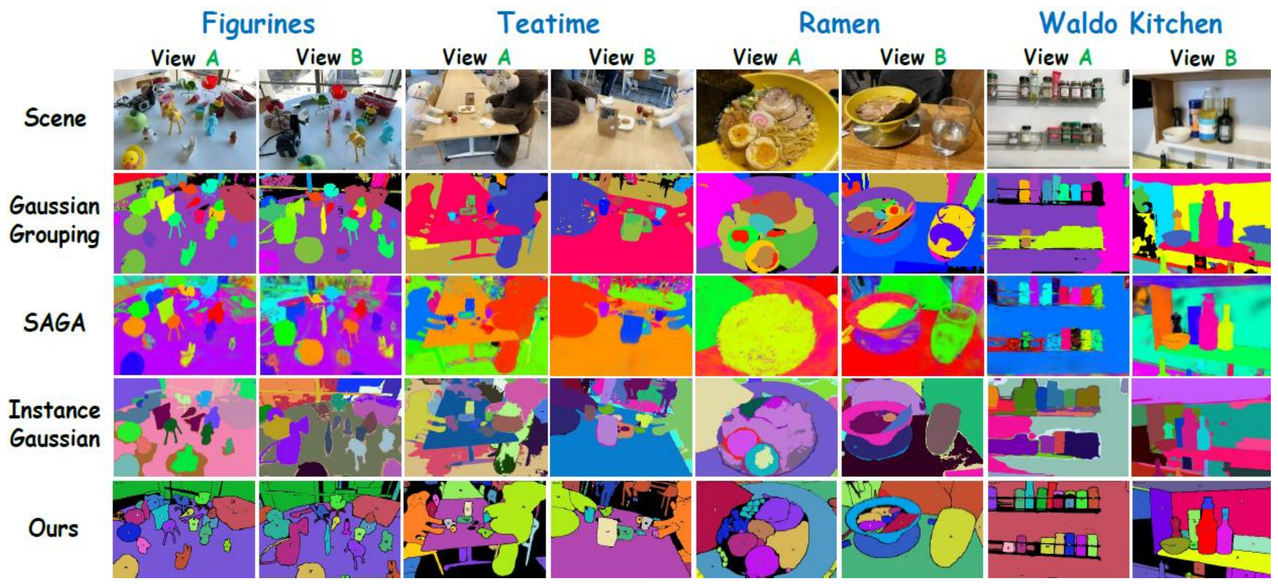
- **Stage1**
  Extract initial SAM mask and multi-cue features, which are semantics (DINOv2), geometric (DepthAnythingv2), and structural (Laplacian of Gaussian Edge Filter).
- **Stage2**
  Refine initial SAM mask by performing score-based Multi-Cue-Guided Mask Merging (MCM) algorithm

$$S(i,j) = w_{sem} \cdot \cos(f_i, f_j) + w_{depth} \cdot \exp\left(-\frac{|\overline{d_i} - \overline{d_j}|^2}{2\sigma_d^2}\right) - w_{dbound} \cdot \max(\nabla_D \cap \partial m_{ij}) - w_{edge} \cdot \max(\nabla_{edge} \cap \partial m_{ij})$$

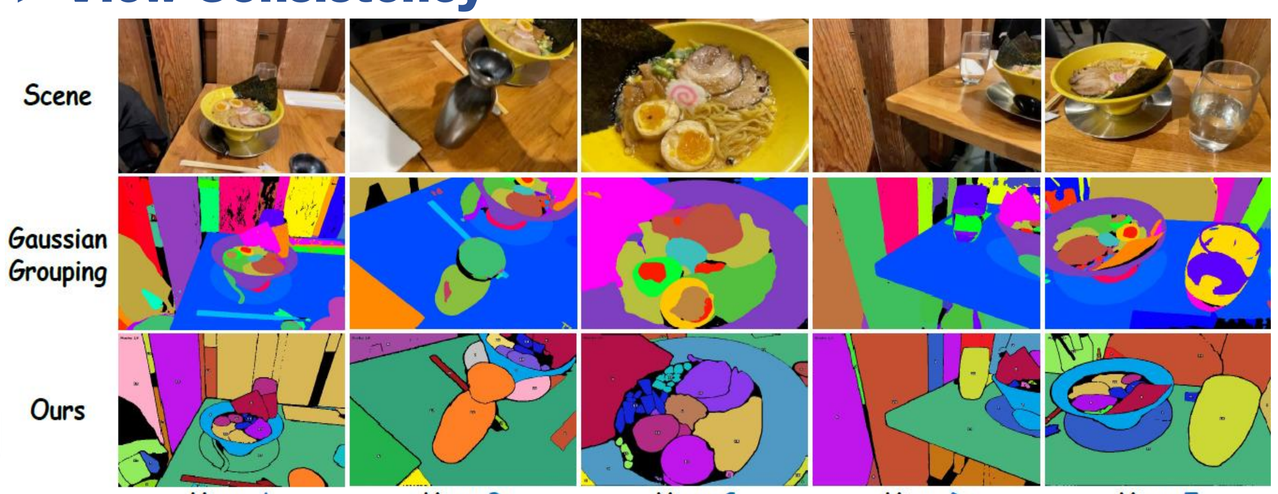
- **Stage3**
  After refinement, to unify view-independent mask IDs into globally consistent ones, Cross-View Mask Matching is performed, and then the 2D features are lifted to 3D space.

EXPERIMENTS

▶ Qualitative Results



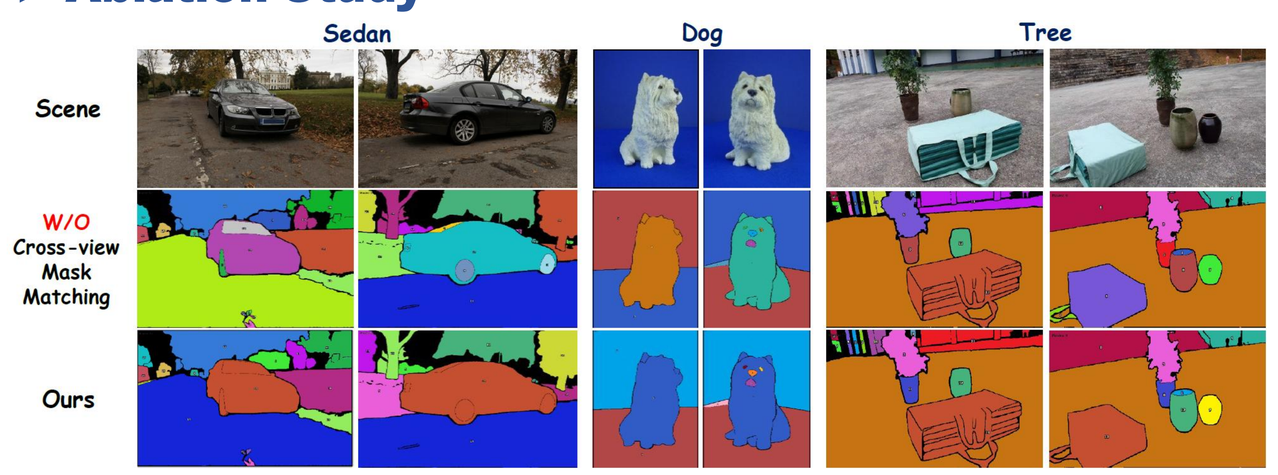
▶ View Consistency



▶ Quantitative Results

Category	Model	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	mBIoU ↑	#Mask ↓
Standard	3DGS [7]	28.6	0.88	**0.11**	-	-	-
3D Segmentation	Gau-Group [32]	28.6	0.88	0.12	0.685	0.642	9,627
	SAGA [2]	28.2	0.87	0.15	0.712	0.658	178
	InstanceGaussian [15]	**28.8**	0.87	0.15	0.702	0.662	101
Feature Field	Feature3DGS [37]	27.5	0.84	0.18	0.538	0.485	-
	GARField [9]	28.1	0.87	0.16	0.710	0.615	-
	CF3 [14]	28.3	0.87	0.14	0.524	0.582	-
Proposed	**Ours**	28.6	**0.89**	0.13	**0.728**	**0.677**	**67**

▶ Ablation Study



Configuration	Signals			Metrics				
	DINOv2	LoG	DAV2	mIoU ↑	# Masks ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline [32]	–	–	–	0.685	9,627	**28.6**	0.88	0.13
Ours (1)	✓	–	–	0.721	245	28.4	0.89	0.14
Ours (2)	✓	✓	–	0.729	112	28.5	**0.91**	**0.12**
Ours (Full)	✓	✓	✓	**0.741**	**67**	**28.6**	**0.91**	**0.12**

* DAV2 denotes DepthAnythingV2.

CONCLUSION

- Propose unified framework for 3D scene decomposition that transforms 2D priors into structured and view-consistent object primitives
- Address over-segmentation of SAM, while ensuring global identity coherence
- Provide robust foundation for structured 3D scene understanding and downstream applications