DMS: Diffusion-Based Multi-Baseline Stereo Generation for Improving Self-Supervised Depth Estimation

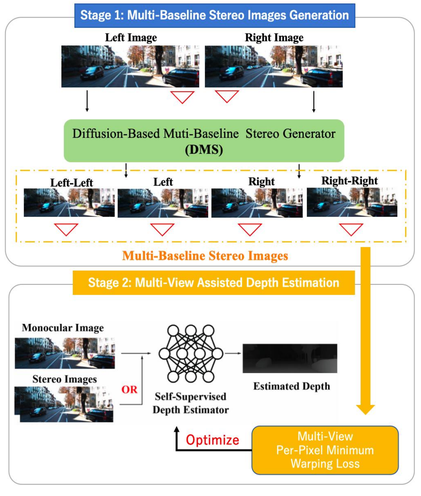
Zihua Liu[1], Yizhou Li[2], Songyan Zhang[3] and Masatoshi Okutomi[1]

1.Institute of Science Tokyo 2. Sony Semiconductor Solutions Group 3.Nanyang Technological University

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Method Overview

- **Two-Stage Training Pipeline.**

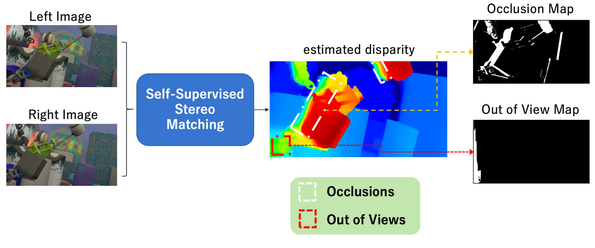


Stage 1: Start from given views, using diffusion model to generated multi-baseline images.

Stage 2: Use multi-baseline images to provide extra geometry clues to assist the self-supervised depth estimation.

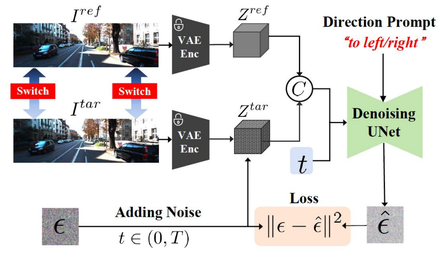
Limitations of Existing Methods

In self-supervised depth estimation, the photometric warping loss is bounded by the two-views information, **yielding only limited gains in occluded and out-of-view regions.**



DMS: Diffusion-Based Multi-Baseline Stereo Generation

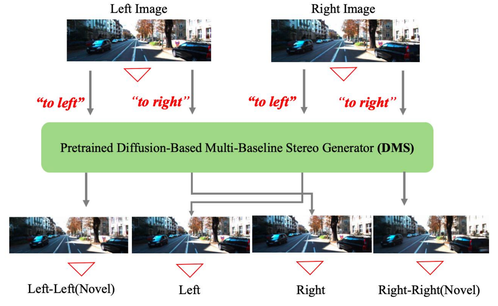
- **Training Phase of the DMS**



(a) Training Phase of the DMS

- Finetune Stable Diffusion to produce the missing stereo view from a reference image, either left→right or right→left.
- Guide the model with directional prompts—'to left' or 'to right'—to indicate view offsets

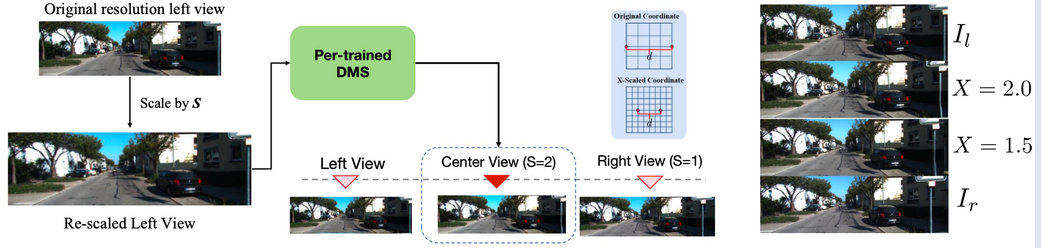
- **Inference Phase of the DMS**



Input View	Direction Prompt	Target View
Left Image	*to left*	Left-Left Image (**Novel View**)
Left Image	*to right*	Right Image (Existing View)
Right Image	*to left*	Left Image (Existing View)
Right Image	*to right*	Right-Right Image (**Novel View**)

- At inference stage, extend the baseline by prompting: left + 'to left' → **left-left,** right + 'to right' → **right-right.**

Intermediate View Approximation with Scaling Operations



We observe that scaling the reference view makes the view shift inversely proportional to the original baseline (**d′ ≈d/scale**), enabling an approximate estimate of intermediate views

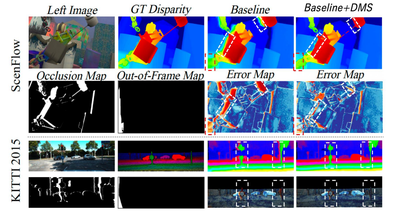
Experimental Results

- **Multi-Baseline Stereo Images Generation.**

* The white line marking the same horizontal position.

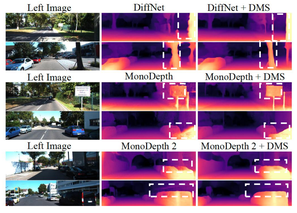
- **Improving Self-Supervised Stereo Matching**

Method	SceneFlow[47] EPE↓ All	Occ	Oof	SceneFlow[47] >3px%↓ All	Occ	Oof	KITTI 2015 [48] EPE↓ All	Occ	Oof	KITTI 2015 [48] >3px%↓ All	Occ	Oof
Baseline	4.09	22.66	10.83	13.5	83.6	42.8	1.48	4.38	9.26	7.7	39.6	64.2
+ *ll* + *rr*	2.45	11.45	6.16	9.0	51.0	25.0	1.34	3.83	7.82	6.5	34.4	42.8
+ *c*	3.75	21.70	7.73	12.5	81.3	38.4	1.36	4.14	7.64	6.7	37.0	49.6
+ *ll* + *rr* +*c*	**2.32**	**11.15**	**5.57**	**8.4**	**49.2**	**23.4**	**1.24**	**3.56**	**7.31**	**5.8**	**32.3**	**39.7**



Method	All(%)↓ D1-bg	D1-fg	D1-all
Non-Learning Based Method			
SGM† [28]	8.95	20.55	10.88
SGM† + DMS	**7.96(11.1%↓)**	**16.68(18.8%↓)**	**9.41(13.5%↓)**
Learning-Based Methods			
Zhou *et.al* [95]	-	-	9.91
SegStereo [82]	-	-	8.79
OASM[36]	6.89	19.42	8.98
Flow2Stereo[41]	5.01	14.62	6.61
PASMnet[71]	5.41	16.36	7.23
PASMnet + DMS	**5.24(3.1%↓)**	**13.96(14.7%↓)**	**6.69(7.5%↓)**
StereoNet*[34]	7.31	17.77	9.05
StereoNet* + DMS	**4.68(36.0%↓)**	**12.06(32.1%↓)**	**5.91(34.7% ↓)**
CFNet* [59]	7.22	18.54	9.11
CFNet* + DMS	**4.64(35.7%↓)**	**12.33(33.5%↓)**	**5.92(35.0%↓)**
RaftStereo* [39]	3.38	13.62	5.08
RaftStereo* + DMS	**2.95(12.7%↓)**	**6.88(50.9%↓)**	**3.60(29.1%↓)**
IGEVStereo* [80]	3.76	11.14	4.98
IGEVStereo* + DMS	**2.80(25.5%↓)**	**6.37(45.27%↓)**	**3.40(31.72%↓)**
MCStereo* [18]	3.01	13.38	4.73
MCStereo* + DMS	**2.83(6.0%↓)**	**7.03(47.46%↓)**	**3.67(22.41%↓)**

- **Improving Self-Supervised Monocular Depth Estimation**



Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE_log↓
MonoDepth [21]	0.120	1.041	5.272	0.217
MonoDepth + DMS	**0.109**	**0.860**	**5.004**	**0.202**
MonoDepth2 [22]	0.109	0.873	4.960	0.209
MonoDepth2 + DMS	**0.105**	**0.811**	**4.850**	**0.200**
SDFANet* [97]	0.104	0.997	4.583	0.186
SDFANet* + DMS	**0.097**	**0.643**	**4.218**	**0.181**
DiffNet [96]	0.104	0.809	4.766	0.201
DiffNet + DMS [96]	**0.098**	**0.726**	**4.606**	**0.191**