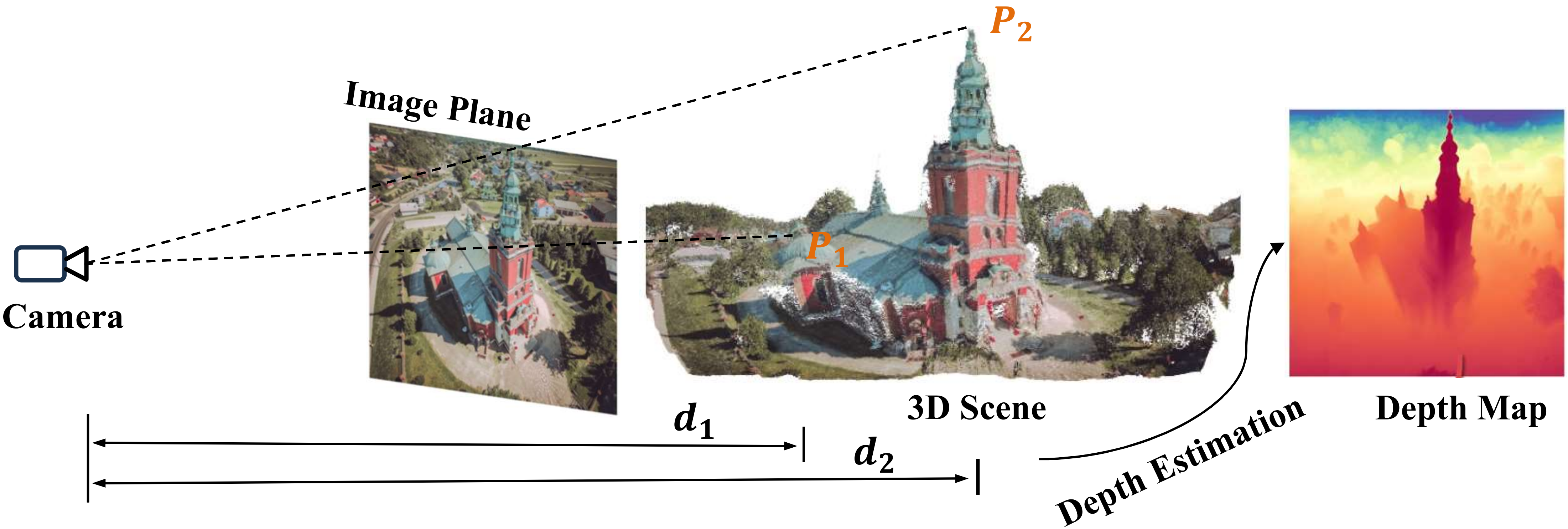
GeoDepth

From Point-to-Depth to Plane-to-Depth Modeling for Self-Supervised Monocular Depth Estimation

Haifeng Wu[1], Shuhang Gu[1], Lixin Duan[1], Wen Li[1,*]
[1]University of Electronic Science and Technology of China

Monocular Depth Estimation

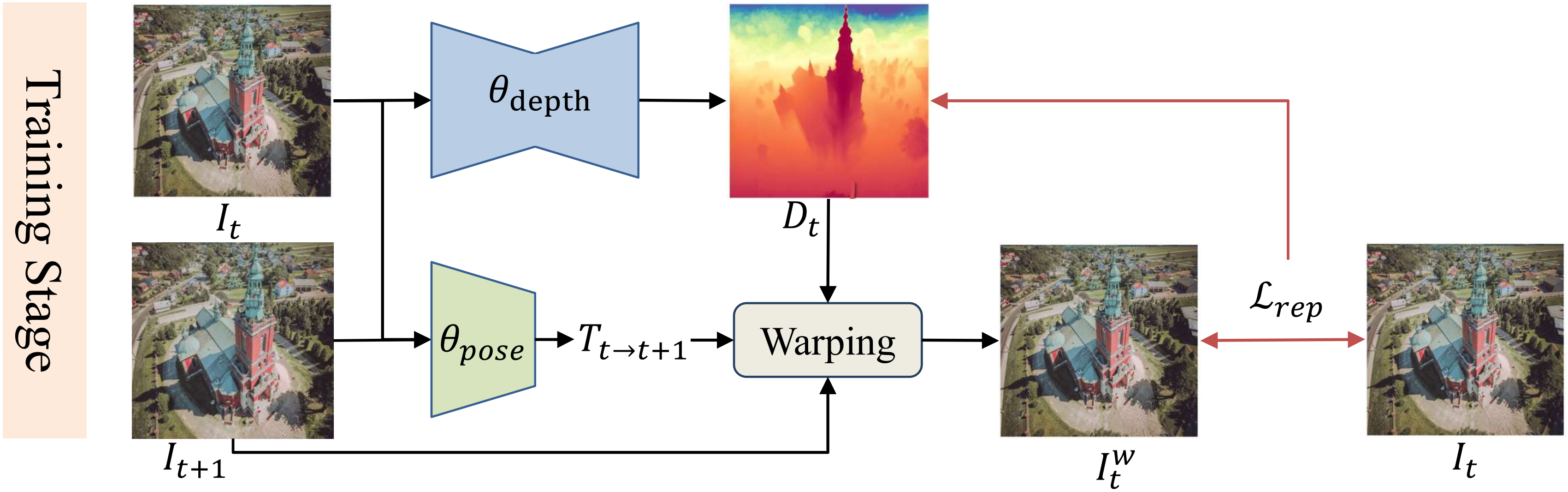
- **Problem Statement**:
  - Given an image, predict the depth information of each 3D point in the corresponding scene relative to the camera plane.



Self-Supervised Monocular Depth Estimation

Problem Statement :

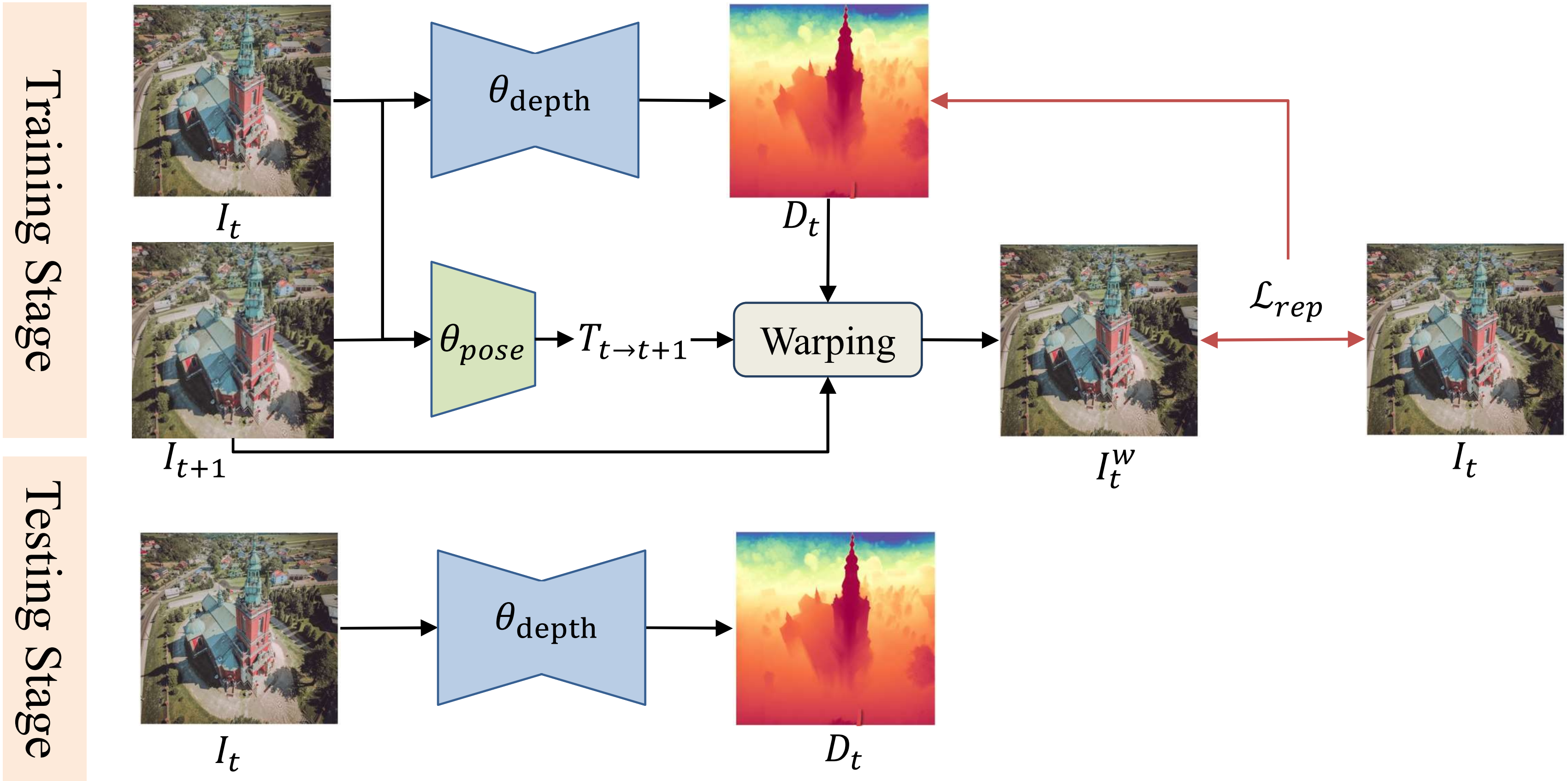
- Learn depth from disparity between neighbor frames without depth GTs (expensive and sparse)
- Reprojection loss $\mathcal{L}_{rep}$: photometric error I_t and I_t^w



Self-Supervised Monocular Depth Estimation

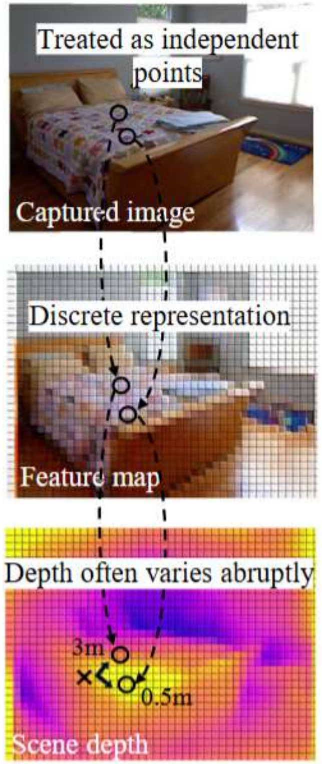
Problem Statement :

- Learn depth from disparity between neighbor frames without depth GTs (expensive and sparse)
- Reprojection loss $\mathcal{L}_{rep}$: photometric error I_t and I_t^w



Depth Jump Problem

- **Existing Modeling:**
  - ***Point-to-Depth*** Modeling (a point-wise prediction problem)
  - => Depth values for points located in the same region may jump dramatically

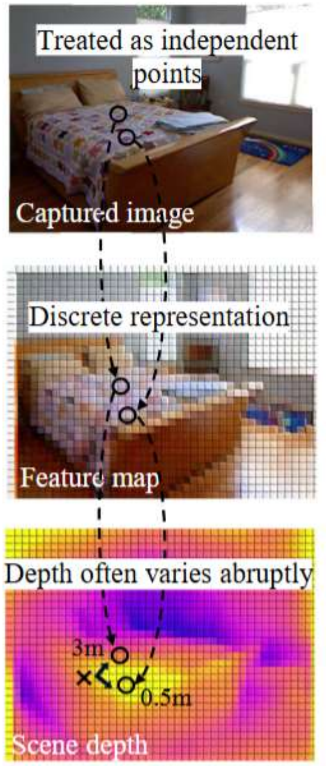


(a) Point-to-depth modeling

Depth Jump Problem

Existing Modeling:

- ***Point-to-Depth*** Modeling (a plane parameter prediction problem)
- => Depth values for points located in the same region may jump dramatically

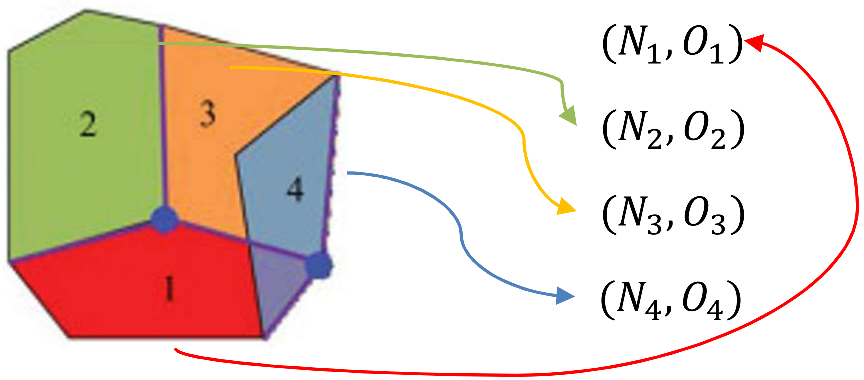


(a) Point-to-depth modeling

Ignoring the geometric structure of the scene?

Motivation

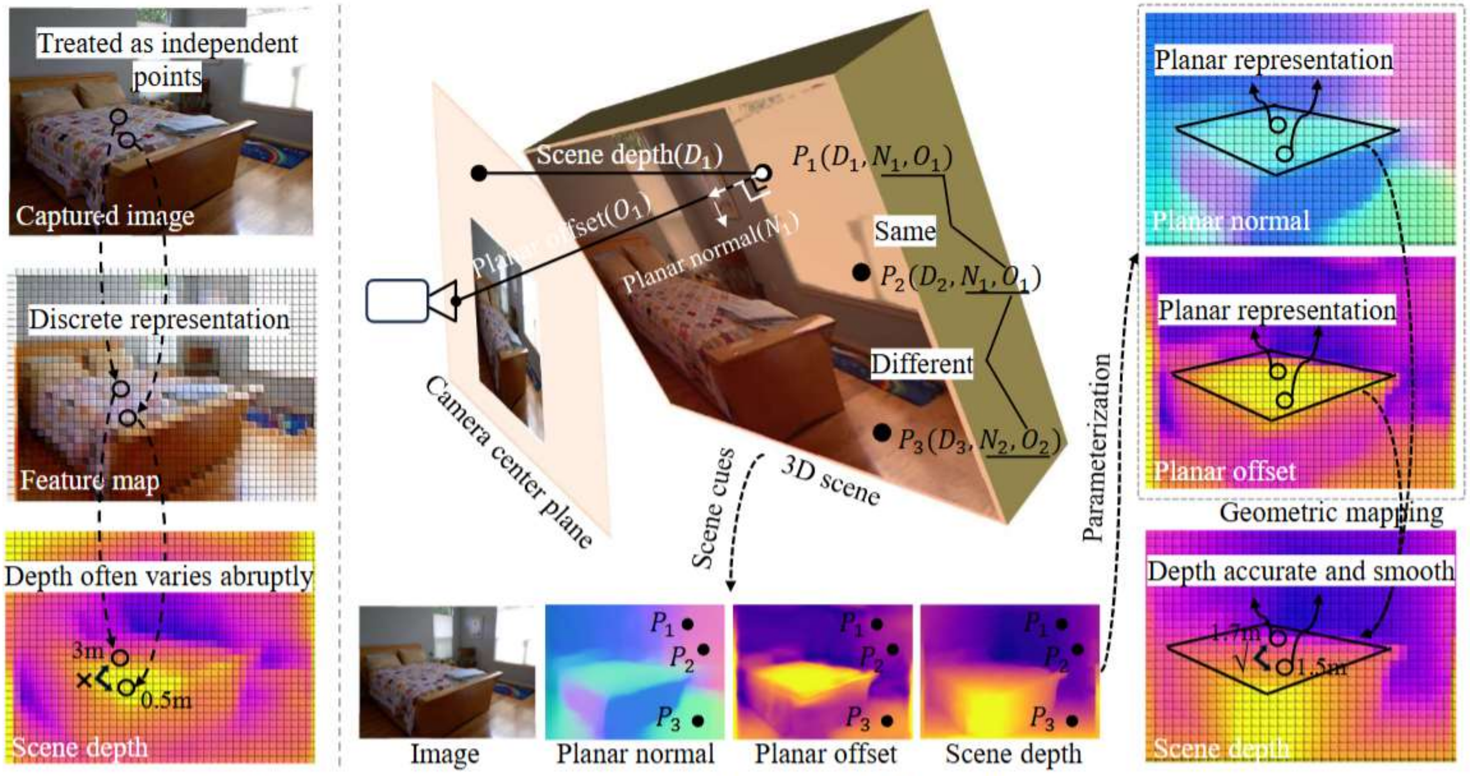
(Planar Normal, Planar Offset)



Geometric parameter modeling

Depth Jump Problem

- **Our Modeling:**
  - *Plane-to-Depth* Modeling (a point-wise prediction problem)
  - => Depth values for points located in the same region are accurate and smooth



(a) Point-to-depth modeling

(b) Plane-to-depth modeling (Ours)

The Proposed Method: GeoDepth

DIG

Plane-to-Depth Modeling:

- 3D scene plane parameterization:

$$\pi = \{\boldsymbol{n}_k, \boldsymbol{o}_k\}_{k=1}^{M}$$

$\boldsymbol{n}$: normal $\boldsymbol{o}$:offset

- Correlation between depth and the plane:

$$\boldsymbol{n}_i^T \boldsymbol{P} = \boldsymbol{o}_i$$ — **Point-normal form**

$$\widetilde{\boldsymbol{p}} = \boldsymbol{P}/(\boldsymbol{d}\boldsymbol{K}^{-1})$$ — **Projective geometry**

$$\boldsymbol{d} = \frac{\boldsymbol{o}_i}{\boldsymbol{n}_i^T \boldsymbol{K}^{-1}\widetilde{\boldsymbol{p}}}$$

$\boldsymbol{P}$: 3D point
$\boldsymbol{K}$: Camera intrinsic
$\widetilde{\boldsymbol{p}}$: 2D point
$\boldsymbol{d}$: Depth

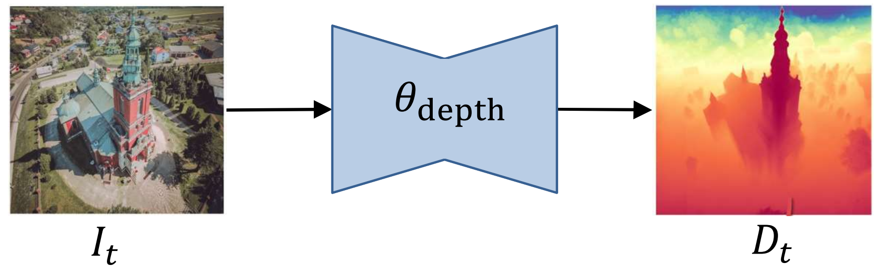
- Plane-to-Depth modeling:

$$\boldsymbol{D}(p) = \frac{\boldsymbol{O}(p)}{\boldsymbol{N}^T(p)\boldsymbol{K}^{-1}\widetilde{\boldsymbol{p}}}$$

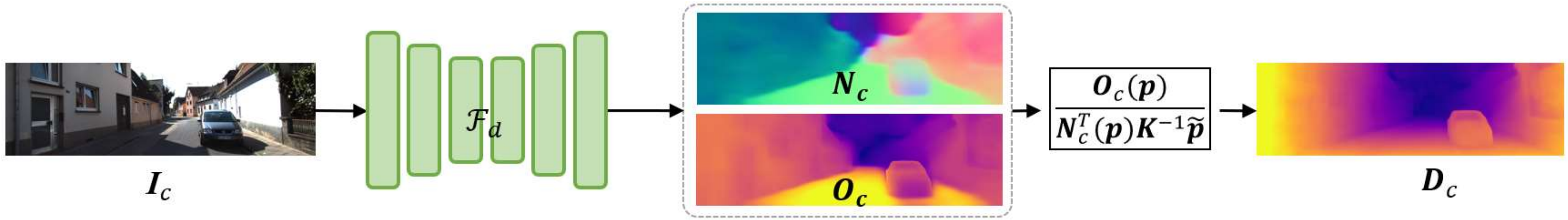
The Proposed Method: GeoDepth

Pipeline:

Previous



Ours

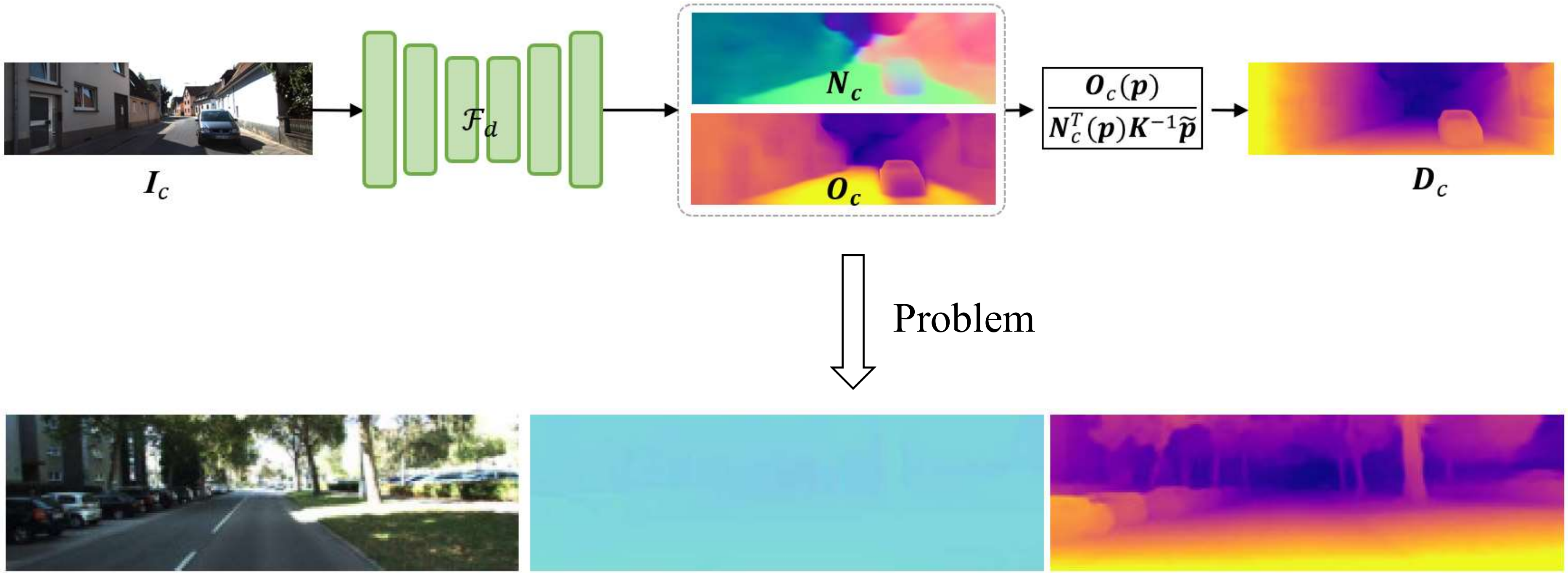


The Proposed Method: GeoDepth

- **Pipeline:**

Ours

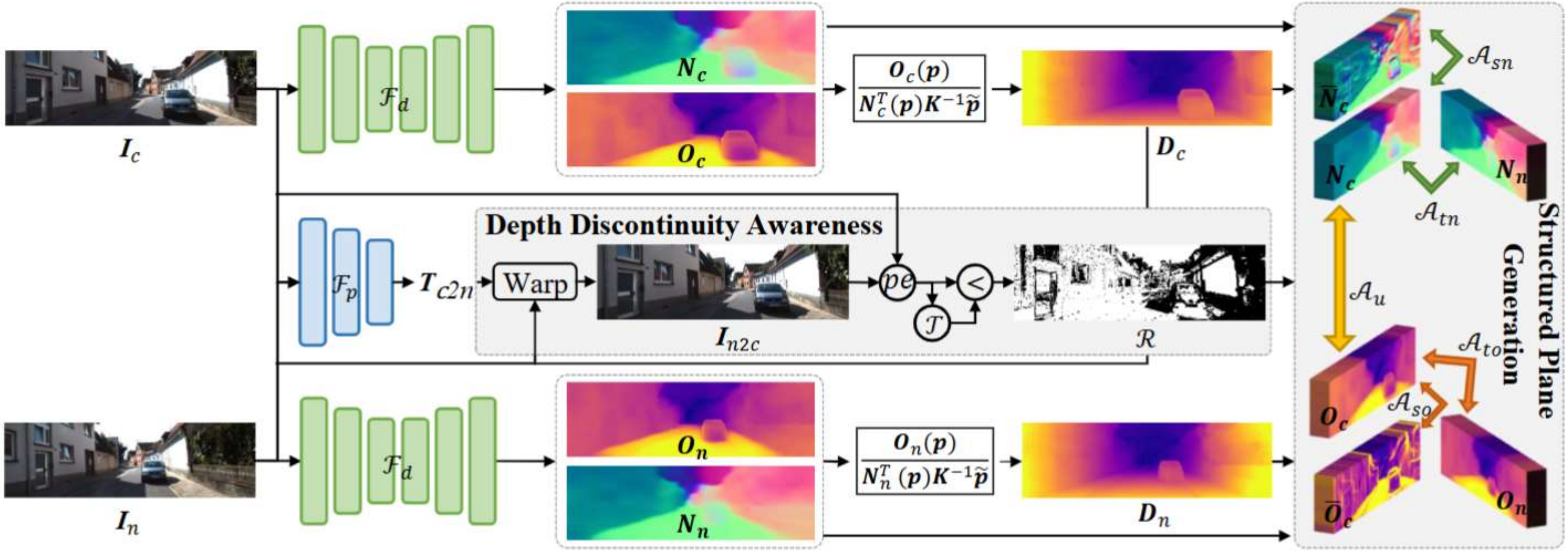
***w/o* geometric constraint**



Input sample | Planar normal | Planar offset

The Proposed Method: GeoDepth

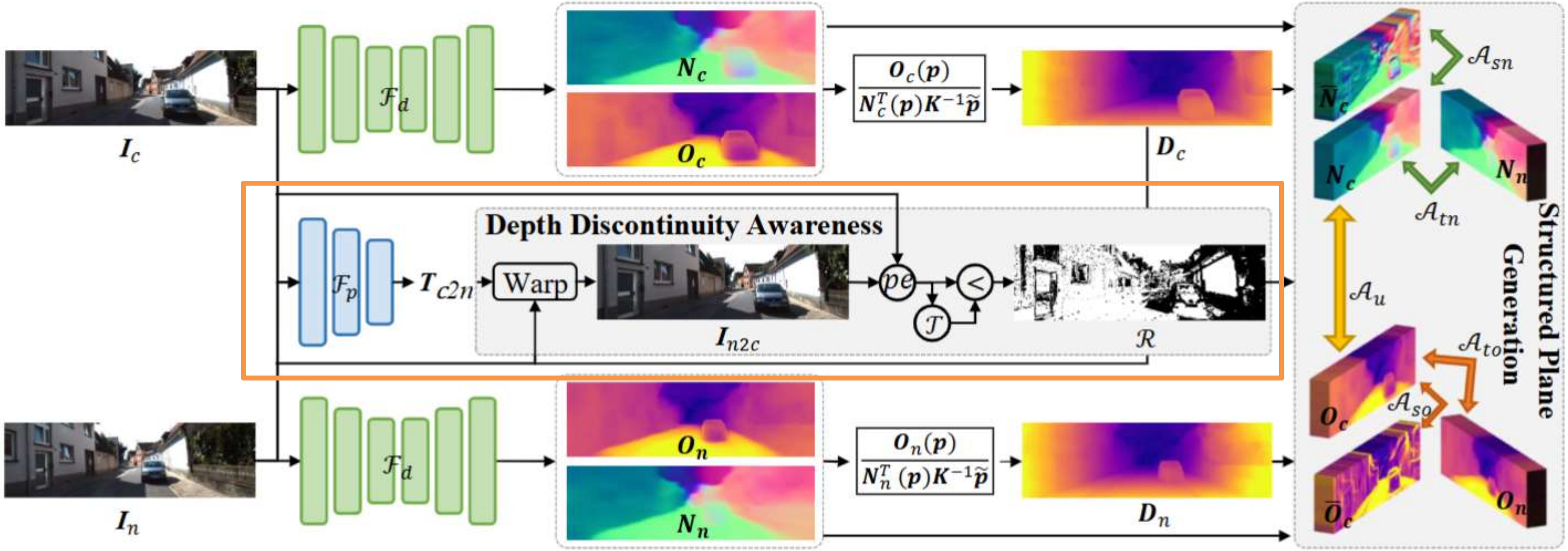
- **Pipeline:**
  - Depth Discontinuity Awareness Module: Identifying the primary planar regions.
  - Structured Plane Generation Module: 1. Utilizes spatio-temporal geometric cues to constraint the planar normal and planar offset of target image. 2. Jointly optimizes the planar normal and planar offset.



The Proposed Method: GeoDepth

Pipeline:

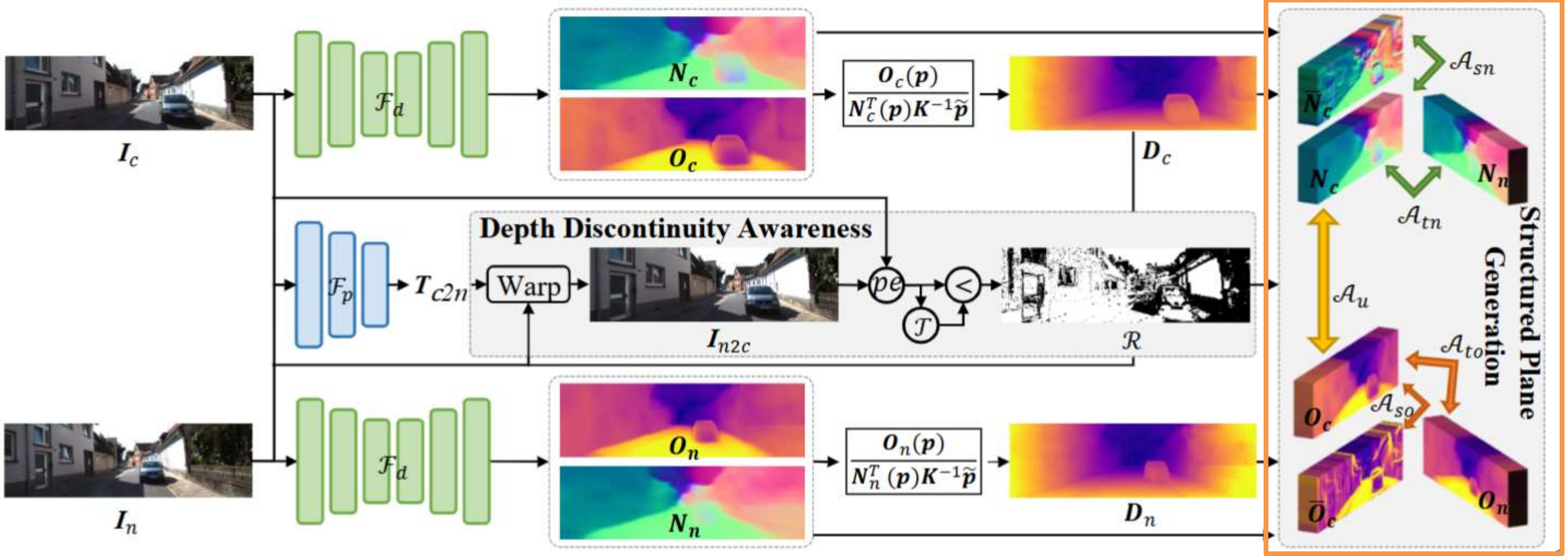
- Depth Discontinuity Awareness Module: Identifying the primary planar regions.
- Structured Plane Generation Module: 1. Utilizes spatio-temporal geometric cues to constraint the planar normal and planar offset of target image. 2. Jointly optimizes the planar normal and planar offset.



The Proposed Method: GeoDepth

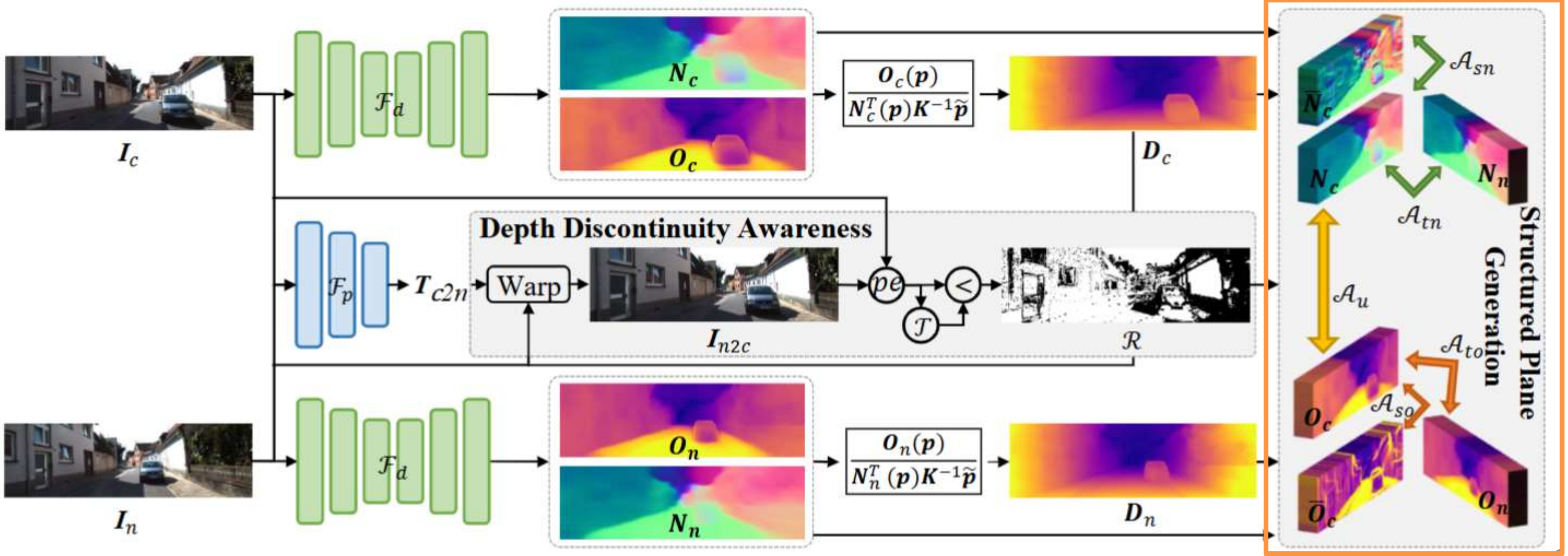
Pipeline:

- Depth Discontinuity Awareness Module: Identifying the primary planar regions.
- Structured Plane Generation Module: 1. Utilizes spatio-temporal geometric cues to constraint the planar normal and planar offset of target image. 2. Jointly optimizes the planar normal and planar offset.



The Proposed Method: GeoDepth

- **Pipeline:**
  - Depth Discontinuity Awareness Module: Identifying the primary planar regions.
  - Structured Plane Generation Module: 1. Utilizes spatio-temporal geometric cues to constraint the planar normal and planar offset of target image. 2. Jointly optimizes the planar normal and planar offset.



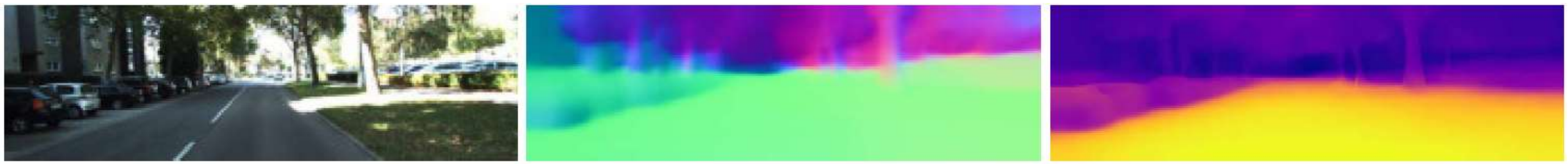
The Proposed Method: GeoDepth

□ Pipeline:

w/o
geometric constraint

Input sample | Planar normal | Planar offset

w
geometric constraint



Input sample | Planar normal | Planar offset

Experimental Results

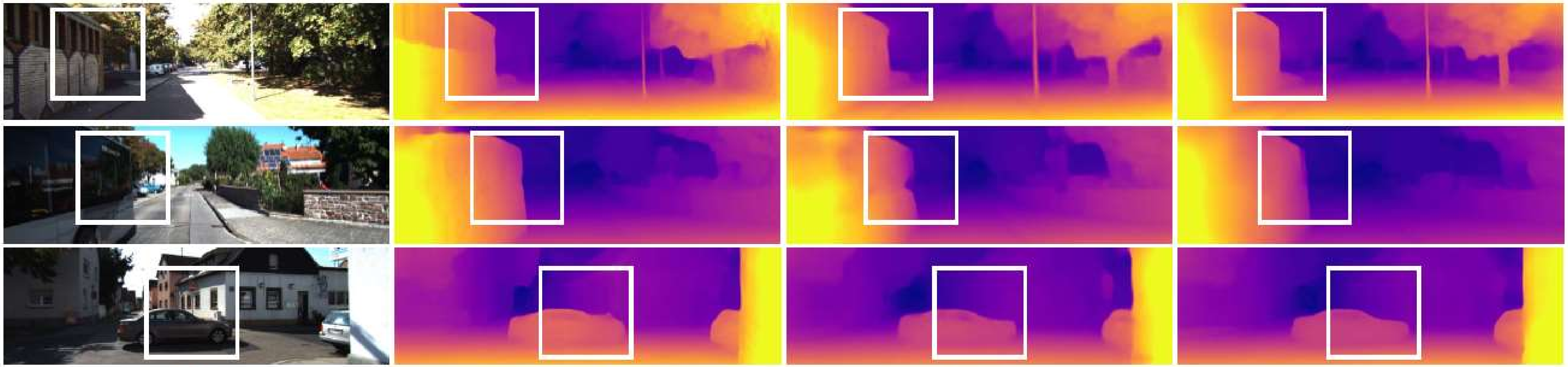
Quantitative Results on Outdoor Datasets (KITTI & Make3D):

Dataset	Method	Size	Mode	Train	Test	RMSE↓	RMSE log↓	Sq Rel↓	Abs Rel↓	$\delta<1.25$↑	$\delta<1.25^2$↑	$\delta<1.25^3$↑
KITTI	Monodepth [12]	512×256	S	✓	✓	5.927	0.247	1.344	0.148	0.803	0.922	0.964
	3Net [35]	512×256	S	✓	✓	5.888	0.208	1.201	0.119	0.844	0.941	0.978
	Monodepth2 [13]	640×192	S	✓	✓	4.960	0.208	0.873	0.109	0.864	0.948	0.975
	BRNet [19]	640×192	S	✓	✓	4.716	0.197	0.876	0.103	0.954	0.978	
	Monodepth2 [13]	640×192	MS	✓	✓	4.750	0.196	0.818	0.106	0.874	0.957	0.979
	DepthHints [48]	640×192	MS	✓	✓	4.627	0.189	0.769	0.105	0.875	0.959	0.982
	HR-Depth [32]	640×192	MS	✓	✓	4.612	0.185	0.785	0.107	0.887	0.962	0.982
	R-MSFM6 [67]	640×192	MS	✓	✓	4.625	0.189	0.787	0.111	0.882	0.961	0.981
	Monodepth2 [13]	640×192	M	✓	✓	4.863	0.193	0.903	0.115	0.877	0.959	0.971
	HR-Depth [32]	640×192	M	✓	✓	4.632	0.185	0.792	0.109	0.884	0.962	0.983
	CADepth-Net [53]	640×192	M	✓	✓	4.535	0.181	0.769	0.105	0.892	0.964	0.983
	DIFFNet [64]	640×192	M	✓	✓	4.483	0.180	0.764	0.102	0.896	0.965	0.983
	MonoFormer [1]	640×192	M	✓	✓	4.580	0.183	0.846	0.104	0.891	0.962	0.982
	SC-DepthV3 [45]	640×192	M	✓	✓	4.709	0.188	0.756	0.118	0.864	0.960	0.984
	SRD-Depth [30]	640×192	M	✓	✓	4.619	0.186	0.762	0.111	0.877	0.961	0.983
	Swin-Depth [40]	640×192	M	✓	✓	4.510	0.182	0.739	0.106	0.890	0.964	0.984
	Lite-Mono [60]	640×192	M	✓	✓	4.561	0.183	0.765	0.107	0.886	0.963	0.983
	ShuffleMono [29]	640×192	M	✓	✓	4.821	0.193	0.850	0.114	0.872	0.957	0.980
	Liu *et al.* [29]	640×192	M	✓	✓	4.724	0.187	0.747	0.114	0.863	0.960	0.984
	Dynamo-Depth [46]	640×192	M	✓	✓	4.505	0.183	0.758	0.112	0.873	0.959	0.984
	GeoDepth	640×192	M	✓	✓	**4.381**	**0.176**	**0.694**	**0.100**	**0.897**	**0.966**	**0.984**
Make3D	Monodepth2 [13]	640×192	M	×	✓	7.418	0.163	3.589	0.322	-	-	-
	HR-Depth [32]	640×192	M	×	✓	7.024	0.159	3.208	0.315	-	-	-
	CADepth-Net [53]	640×192	M	×	✓	7.066	0.159	3.086	0.312	-	-	-
	DIFFNet [64]	640×192	M	×	✓	7.008	0.155	3.313	0.309	-	-	-
	Lite-Mono [60]	640×192	M	×	✓	6.981	0.158	3.060	0.305	-	-	-
	Zhao *et al.* [62]	640×192	M	×	✓	7.095	0.158	3.200	0.316	-	-	-
	Xiong *et al.* [52]	640×192	M	×	✓	7.005	0.161	3.102	0.319	-	-	-
	GeoDepth	640×192	M	×	✓	**6.735**	**0.153**	**2.750**	**0.296**	-	-	-

- **KITTI**:
  - ✓ In-domain testing
  - ✓ Verifying robustness
- **Make3D**:
  - ✓ Cross-domain testing
  - ✓ Verifying generalization

Experimental Results

- **Qualitative Results on Outdoor Datasets (KITTI):**



(a) Input (b) Monodepth2 (c) Lite-Mono (d) Ours

- Existing Methods: Inconsistencies in planar regions and noticeable errors along object edges
- Ours: Preserving both planar structures and sharp boundaries.

Experimental Results

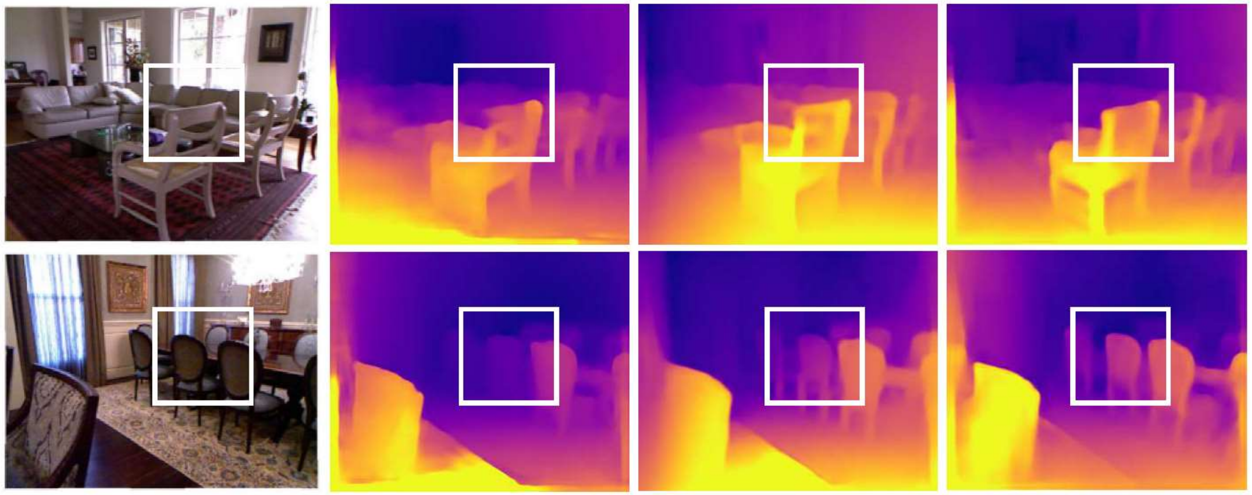
☐ Quantitative Results on Indoor Datasets (NYUv2 & ScanNet):

Dataset	Method	Size	Mode	Train	Test	RMSE↓	Abs Rel↓	$\delta<1.25$↑	$\delta<1.25^2$↑	$\delta<1.25^3$↑
NYUv2	MovingIndoor [65]	320×256	M	✓	✓	0.712	0.208	0.674	0.900	0.968
	Monodepth2 [13]	320×256	M	✓	✓	0.601	0.160	0.767	0.949	0.988
	P^2Net [57]	320×256	M	✓	✓	0.599	0.159	0.772	0.942	0.984
	SC-DepthV1 [4]	320×256	M	✓	✓	0.639	0.159	0.734	0.937	0.983
	PLNet [22]	320×256	M	✓	✓	0.562	0.151	0.790	0.953	0.989
	StructDepth [27]	320×256	M	✓	✓	0.540	0.142	0.813	0.954	0.988
	ADPDepth [43]	320×256	M	✓	✓	0.592	0.165	0.753	0.934	0.981
	F^2Depth [17]	320×256	M	✓	✓	0.569	0.153	0.787	0.950	0.987
	Guo *et al.* [18]	320×256	M	✓	✓	0.567	0.152	0.792	0.950	0.988
	Ours	320×256	M	✓	✓	**0.520**	**0.134**	**0.833**	**0.963**	**0.991**
ScanNet	MovingIndoor [65]	320×256	M	×	✓	0.483	0.212	0.650	0.905	0.976
	Monodepth2 [13]	320×256	M	×	✓	0.458	0.200	0.672	0.922	0.981
	TrainFlow [63]	320×256	M	×	✓	0.415	0.179	0.726	0.927	0.980
	P^2Net [57]	320×256	M	×	✓	0.420	0.175	0.740	0.932	0.982
	PLNet [22]	320×256	M	×	✓	0.414	0.176	0.735	0.939	0.985
	IFMNet [49]	320×256	M	×	✓	0.402	0.170	0.758	0.940	0.989
	SC-Depthv1 [4]	320×256	M	×	✓	0.392	0.169	0.749	0.938	0.983
	StructDepth [27]	320×256	M	×	✓	0.400	0.165	0.754	0.939	0.985
	GeoDepth	320×256	M	×	✓	**0.387**	**0.161**	**0.769**	**0.946**	**0.987**

- **NYUv2**:
  - ✓ In-domain testing
  - ✓ Verifying robustness
- **ScanNet**:
  - ✓ Cross-domain testing
  - ✓ Verifying generalization

Experimental Results

Qualitative Results on Indoor Datasets (NYUv2):



(a) Input (b) StructDepth (c) PLNet (d) Ours

- Existing Methods: Inconsistencies in planar regions and noticeable errors along object edges
- Ours: Preserving both planar structures and sharp boundaries.

Experimental Results

Ablation Study on Outdoor Datasets (KITTI):

- The effectiveness of each design choice

Method	P2D	SPG	DDA	Sq Rel↓	RMSE ↓	$\delta<1.25\uparrow$	#Params
Baseline				0.751	4.471	0.895	9.98M
+P2D	✓			0.740	4.436	0.896	10.0M
+P2D+SPG	✓	✓		0.722	4.412	0.896	10.0M
GeoDepth	✓	✓	✓	**0.694**	**4.381**	**0.897**	10.0M

P2D: Plane-to-Depth Modeling
SGP: Structured Plane Generation Module
DDA: Depth Discontinuity Awareness Module

- Like-for-like comparisons

Method	Backbone	Sq Rel↓	RMSE ↓	$\delta<1.25\uparrow$
CADepth-Net	ResNet50	0.769	4.535	0.892
GeoDepth	ResNet50	**0.745**	**4.478**	**0.896**
RA-Depth	HRNet18	0.632	4.216	0.903
GeoDepth	HRNet18	**0.624**	**4.169**	**0.904**
MonoViT	MPViT	0.708	4.372	0.900
GeoDepth	MPViT	**0.662**	**4.237**	**0.902**

- Integrating our idea with recent SOTA frameworks
- Our method consistently outperforms these frameworks across various backbones

Summary

GeoDepth: From Point-to-Depth to Plane-to-Depth Modeling for Self-Supervised Monocular Depth Estimation

Problem

- Self-supervised monocular depth estimation has long been treated as a point-wise prediction problem (***Point-to-Depth***).
- Artifacts are often observed in the estimated depth map, *e.g*. depth values for points located in the same region may jump dramatically

Solution

- We propose GeoDepth, a novel self-supervised monocular depth estimation framework, which develops a ***plane-to-depth*** modeling strategy to address the depth discontinuity issues inherent in ***point-to-depth*** methods.

Results

- State-of-the art results outdoor dataset KITTI and Make3D;
- State-of-the art results indoor dataset NYUv2 and ScanNet;

Thank You!