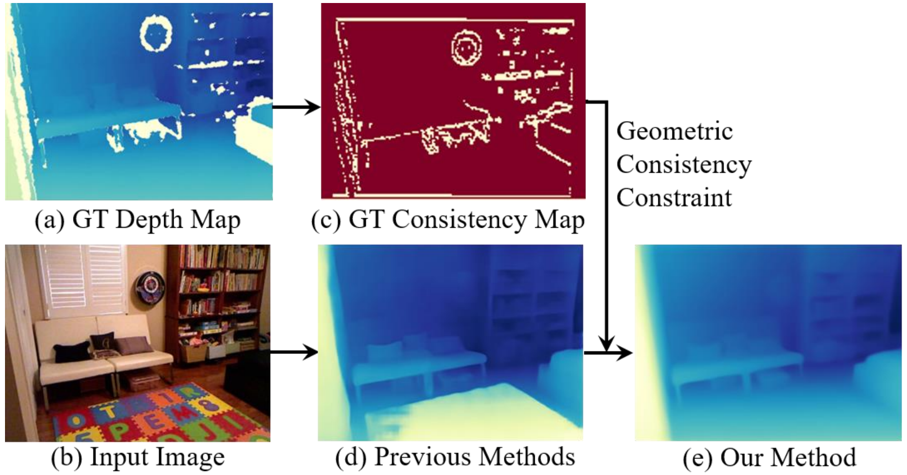
Hyper-Depth: Hypergraph-based Multi-Scale Representation Fusion for Monocular Depth Estimation

*Lin Bie†, Siqi Li†, Yifan Feng, Yue Gao**

biel20@mails.tsinghua.edu.cn

Introduction

- Monocular depth estimation (MDE) is a classical task in computer vision. However, the problem is inherently ill-posed and ambiguous.
- Existing transformer-based methods mitigate the substantial computational overhead by employing local attention mechanisms, which lead to depth estimation errors caused by over-fitting the image's local textures, as figure shows.
- Furthermore, transformer-based methods still struggle to effectively utilize multi-scale visual features despite the critical importance of multi-scale information in MDE.
- To address these challenges, we introduce a semantic consistency enhancement (SCE) module, which effectively improves the representation of multi-scale features by leveraging hypergraph convolution (HyperConv) in the semantic space.
- Furthermore, we propose a geometric consistency constraint (GCC) module, which provides geometric guidance to reduce over-fitting to local features, as figure shows.



Contribution

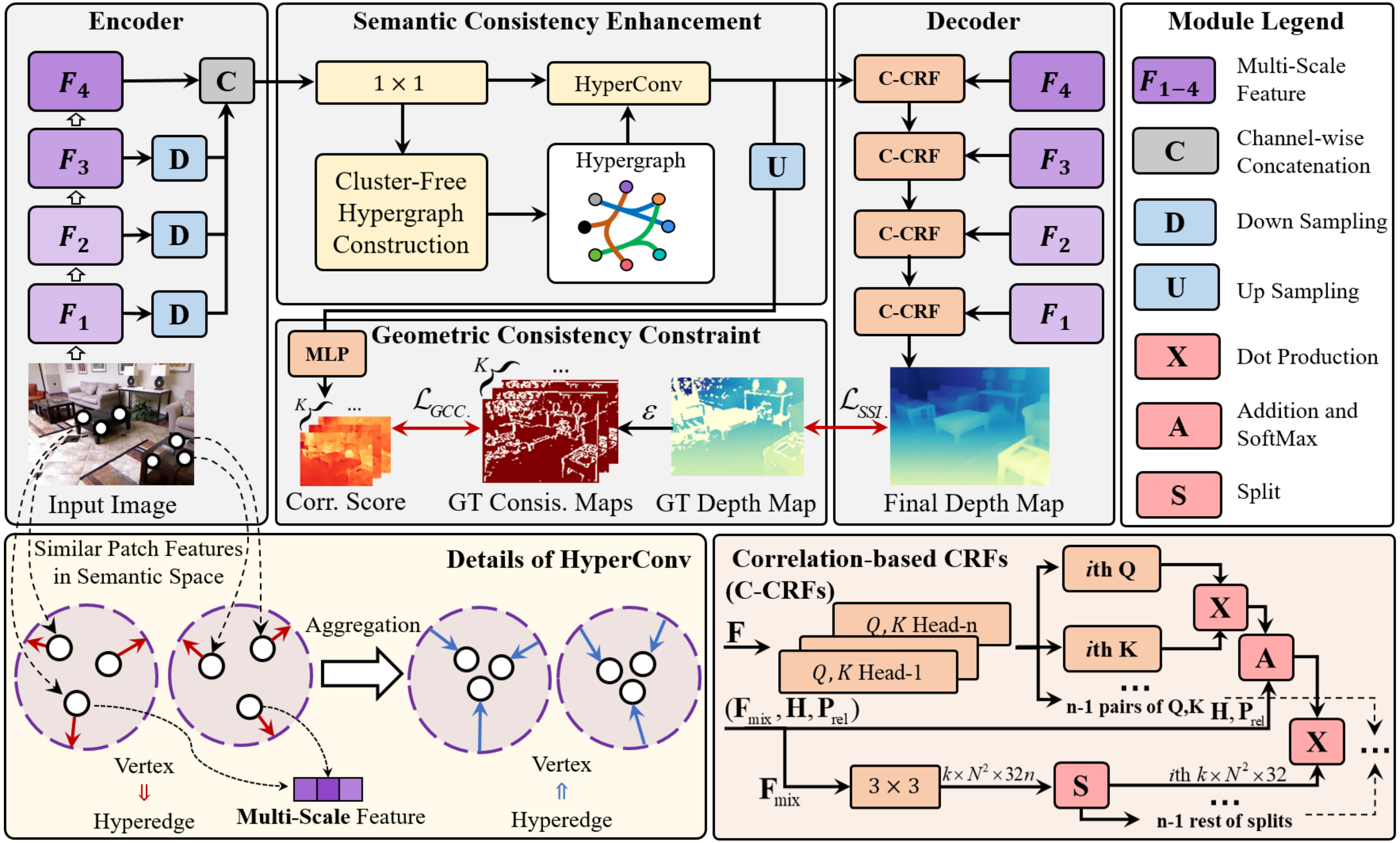
- We propose a hypergraph-based multi-scale representation fusion framework that effectively aggregates cross-position patch features and attention weights generated by a transformer-based backbone while maintaining an acceptable computational cost through a semantic consistency enhancement (SCE) module.

- We introduce a geometric consistency constraint (GCC) module that learns the correlations between patches and their surrounding context to reduce overfitting errors caused by excessive reliance on local features. Additionally, we introduce a GCC loss, based on cross-entropy loss, to supervise the training process.

- We design a correlation-based CRFs (C-CRFs) module to filter relevant patches for attention computation without being constrained by fixed window sizes utilized in previous works.

- Extensive experiments on four widely used datasets demonstrate that our method significantly outperforms state-of-the-art approaches while exhibiting superior generalizability in zero-shot testing.

Method



Encoder

F_4 C F_3 D F_2 D F_1 D

Input Image

Semantic Consistency Enhancement

1 × 1

HyperConv

Cluster-Free Hypergraph Construction

Hypergraph

U

Decoder

C-CRF F_4 C-CRF F_3 C-CRF F_2 C-CRF F_1

Final Depth Map

Module Legend

F_{1-4}	Multi-Scale Feature
C	Channel-wise Concatenation
D	Down Sampling
U	Up Sampling
X	Dot Production
A	Addition and SoftMax
S	Split

Geometric Consistency Constraint

MLP

K ...

$\mathcal{L}_{GCC}.$

$\mathcal{E}$

$\mathcal{L}_{SSI}.$

Corr. Score

GT Consis. Maps

GT Depth Map

Similar Patch Features in Semantic Space

Details of HyperConv

Aggregation

Vertex ⇓ Hyperedge

Multi-Scale Feature

Vertex ⇑ Hyperedge

Correlation-based CRFs (C-CRFs)

$\mathbf{F}$

Q, K Head-n

Q, K Head-1

$(\mathbf{F}_{\text{mix}}, \mathbf{H}, \mathbf{P}_{\text{rel}})$

ith Q

ith K

n-1 pairs of Q,K

$\mathbf{H}, \mathbf{P}_{\text{rel}}$

X A X

$\mathbf{F}_{\text{mix}}$

3 × 3

$k \times N^2 \times 32n$

S

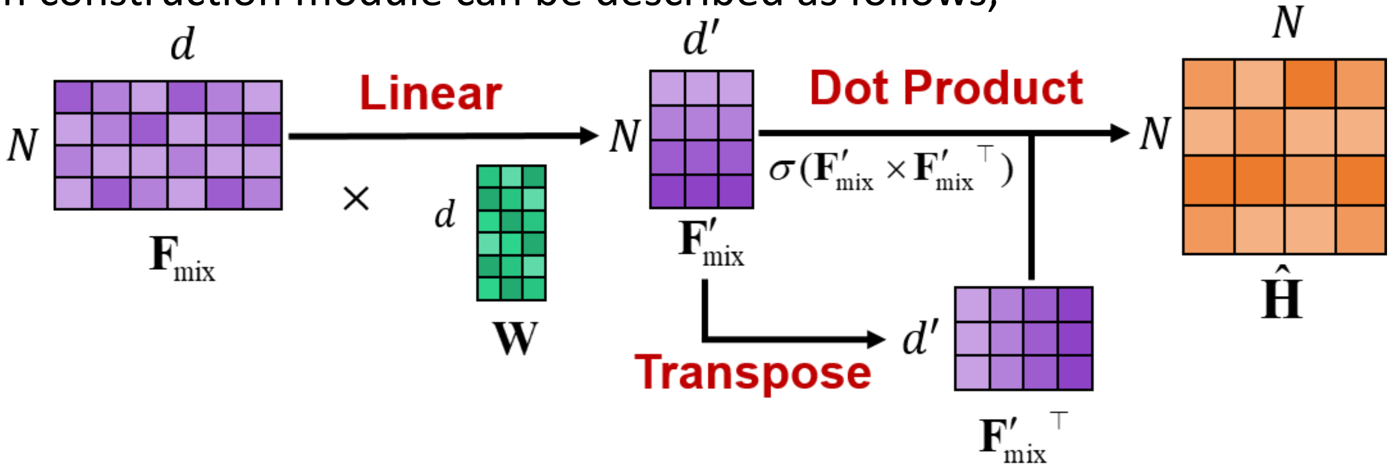
ith $k \times N^2 \times 32$

n-1 rest of splits

Method

Semantic Consistency Enhancement

Cluster-free hypergraph construction: We construct the hypergraph to model the correlation between the multi-scale cross-position patch features. Specifically, our proposed module can be summarized as two stages: cluster-free hypergraph construction and hypergraph convolution for semantic consistency enhancement. The details of the cluster-free hypergraph construction module can be described as follows,


d
N
$\mathbf{F}_{\text{mix}}$
Linear
×
d
$\mathbf{W}$
d'
N
$\mathbf{F}'_{\text{mix}}$
Dot Product
$\sigma(\mathbf{F}'_{\text{mix}} \times \mathbf{F}'^{\top}_{\text{mix}})$
N
N
$\hat{\mathbf{H}}$
Transpose
d'
$\mathbf{F}'^{\top}_{\text{mix}}$

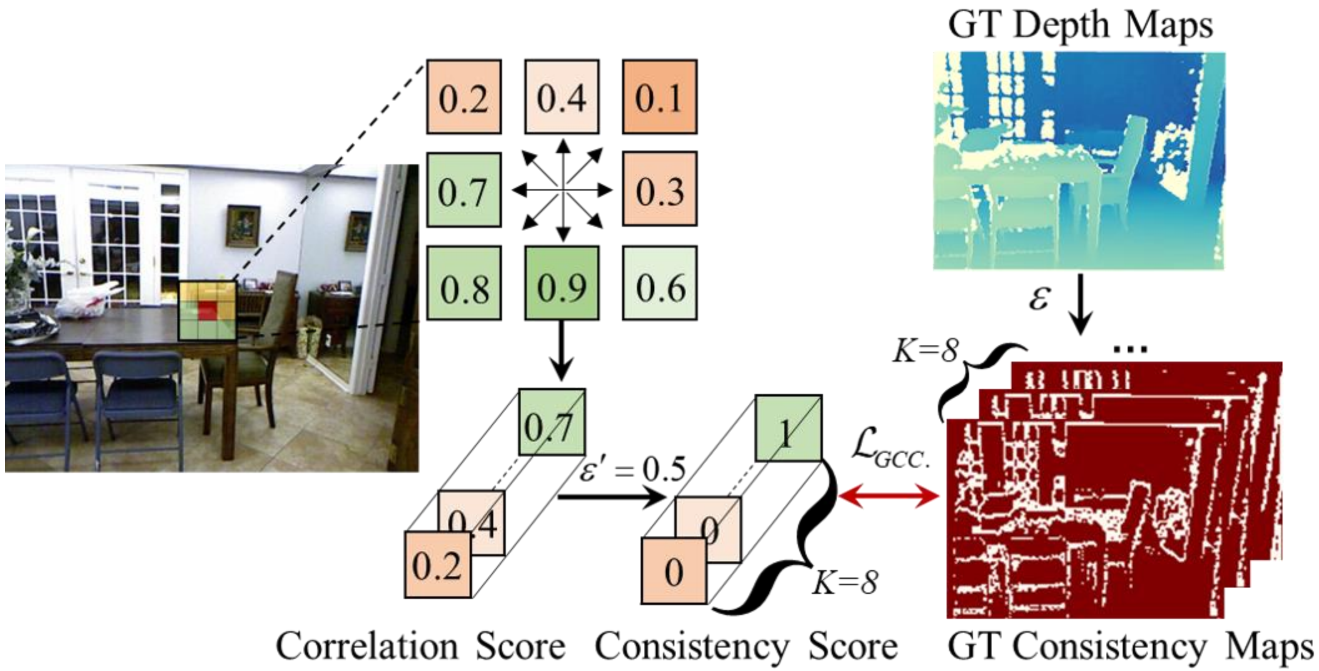
Hypergraph convolution for semantic consistency enhancement: We design a HyperConv head to aggregate the cross-position information of the whole image with extra residual connection to perform high-order learning on vertex features $f_v \in \mathbf{F}_{mix}$ and hyperedge feature f_e as follows,

$$\begin{cases} f_e = \frac{1}{|N(e)|} \sum_{v \in N(e)} f_v \Theta \\ f'_v = f_v + \frac{1}{|N(e)|} \sum_{e \in N(e)} f_e \end{cases},$$

Method

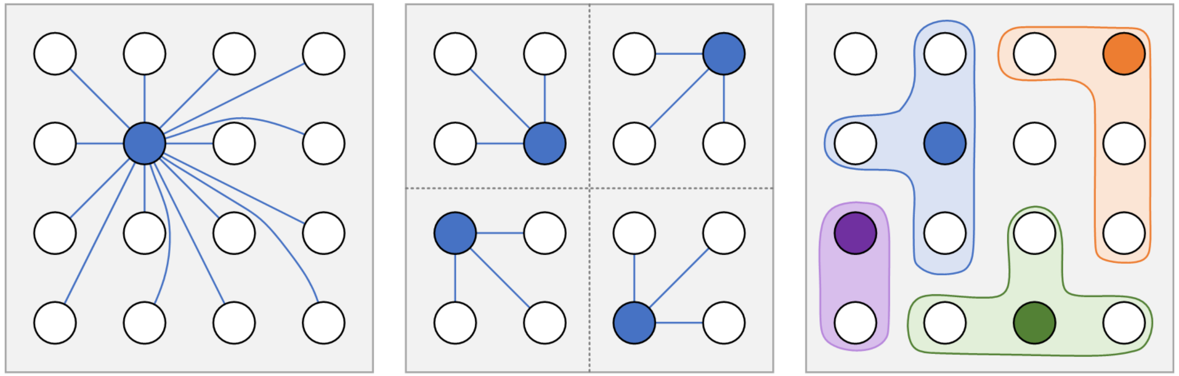
Geometric Consistency Constraint

We further introduce a GCC module to supervise the cross-position patch correlation hypergraph learning process. Intuitively, the similar patches in the whole image belonging to the same object or plane tend to have similar depth. The geometric information provided by GT depth is crucial for the MDE task. Therefore, we generate a ground truth consistency score for each patch relative to its K=8 surrounding patches based on GT depth.



Method

Correlation-based Conditional Random Fields (CRFs)



(a) FC-CRFs (b) Window FC-CRFs (c) C-CRFs

We propose a correlation-based CRFs method, which calculates the query vector $\mathbf{q}$ and key vector $\mathbf{k}$ between related patches among the whole image based on the hypergraph incidence matrix $\mathbf{H}$. Therefore, the energy function of our correlation-based CRF is defined as follows,

$$E = \sum_i \phi_c(f_i, \{f_j \mid j \in \mathcal{V}_i\}),$$

Furthermore, we add relative position embedding $\mathbf{P}_{rel}$ and generate a potential function for CRF as follows,

$$\sum_i \phi_c = \mathrm{SoftMax}(\mathbf{H}(\mathbf{Q} \cdot \mathbf{K}^\top + \mathbf{P}_{\mathrm{rel}}))$$

□ Method

■ Loss Function

For end-to-end training process, we propose a combined loss function consisting of a Scaled Scale-Invariant (SSI) loss and a **Geometric Consistency Constraint (GCC) Loss** as follows:

$$\mathcal{L}_{Comb} = \alpha\mathcal{L}_{SSI} + \gamma\mathcal{L}_{GCC}$$

Specifically, SSI loss L_{SSI} is defined as follows,

$$\mathcal{L}_{SSI} = \sqrt{\frac{1}{T}\sum_i \Delta d_i^2 - \frac{\lambda}{T^2}(\sum_i \Delta d_i)^2}$$

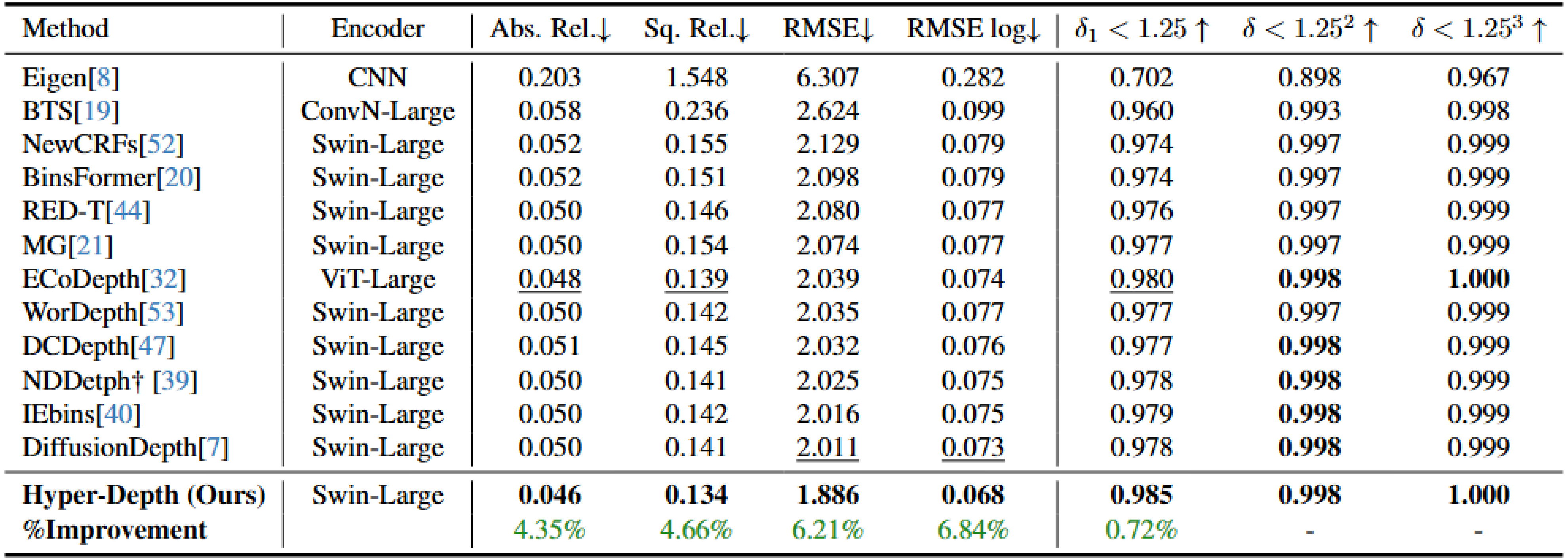
Where $\Delta d_i = \log\hat{d}_i - \log d_i^*$ with the predicted depth $\hat{d}_i$ and the d_i^* ground truth depth. *T* denotes the number of the pixels having valid ground truth values. Meanwhile, we mitigate errors caused by an over-reliance on image features by constructing a **Geometric Consistency Constraint (GCC) Loss** based on a cross-entropy loss function as follows,

$$\mathcal{L}_{GCC} = -\frac{1}{N}\sum_{i=1}^{N}\left[S_{\mathrm{gt}}^{(i)}\cdot\log(S_{\mathrm{pred}}^{(i)}) + \left(1-S_{\mathrm{gt}}^{(i)}\right)\cdot\log\left(1-S_{\mathrm{pred}}^{(i)}\right)\right]$$

Experiment

Quantitative Comparison

To evaluate the performance of our proposed Hyper-Depth framework, we have conducted experiments on KITTI and NYU-Depth-v2 datasets compared with the state-of-the-art, such as DiffusionDepth (ECCV 2024), DCDepth (NIPS 2024) and ECoDepth (CVPR 2024).



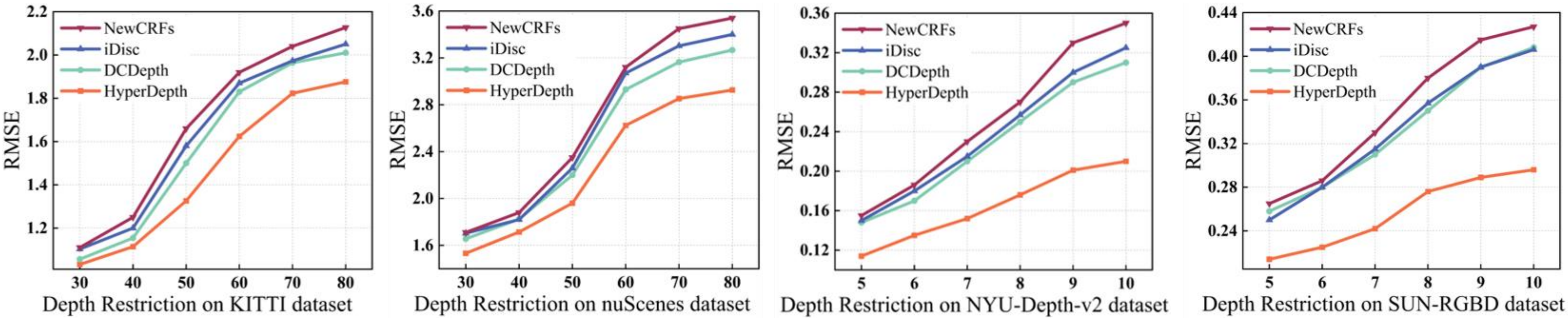
Method	Encoder	Abs. Rel.↓	Sq. Rel.↓	RMSE↓	RMSE log↓	$\delta_1 < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Eigen[8]	CNN	0.203	1.548	6.307	0.282	0.702	0.898	0.967
BTS[19]	ConvN-Large	0.058	0.236	2.624	0.099	0.960	0.993	0.998
NewCRFs[52]	Swin-Large	0.052	0.155	2.129	0.079	0.974	0.997	0.999
BinsFormer[20]	Swin-Large	0.052	0.151	2.098	0.079	0.974	0.997	0.999
RED-T[44]	Swin-Large	0.050	0.146	2.080	0.077	0.976	0.997	0.999
MG[21]	Swin-Large	0.050	0.154	2.074	0.077	0.977	0.997	0.999
ECoDepth[32]	ViT-Large	0.048	0.139	2.039	0.074	0.980	**0.998**	**1.000**
WorDepth[53]	Swin-Large	0.050	0.142	2.035	0.077	0.977	0.997	0.999
DCDepth[47]	Swin-Large	0.051	0.145	2.032	0.076	0.977	**0.998**	0.999
NDDetph† [39]	Swin-Large	0.050	0.141	2.025	0.075	0.978	**0.998**	0.999
IEbins[40]	Swin-Large	0.050	0.142	2.016	0.075	0.979	**0.998**	0.999
DiffusionDepth[7]	Swin-Large	0.050	0.141	2.011	0.073	0.978	**0.998**	0.999
Hyper-Depth (Ours)	Swin-Large	**0.046**	**0.134**	**1.886**	**0.068**	**0.985**	**0.998**	**1.000**
%Improvement		4.35%	4.66%	6.21%	6.84%	0.72%	-	-

Table 2. Quantitative depth estimation performance comparison on KITTI Eigen split. The maximum depth is capped at 80m. The best result is indicated in bold, and the second is underlined. The RMSE is the main ranking metric. (†): Training with extra data.

□ Experiment

■ Quantitative Comparison

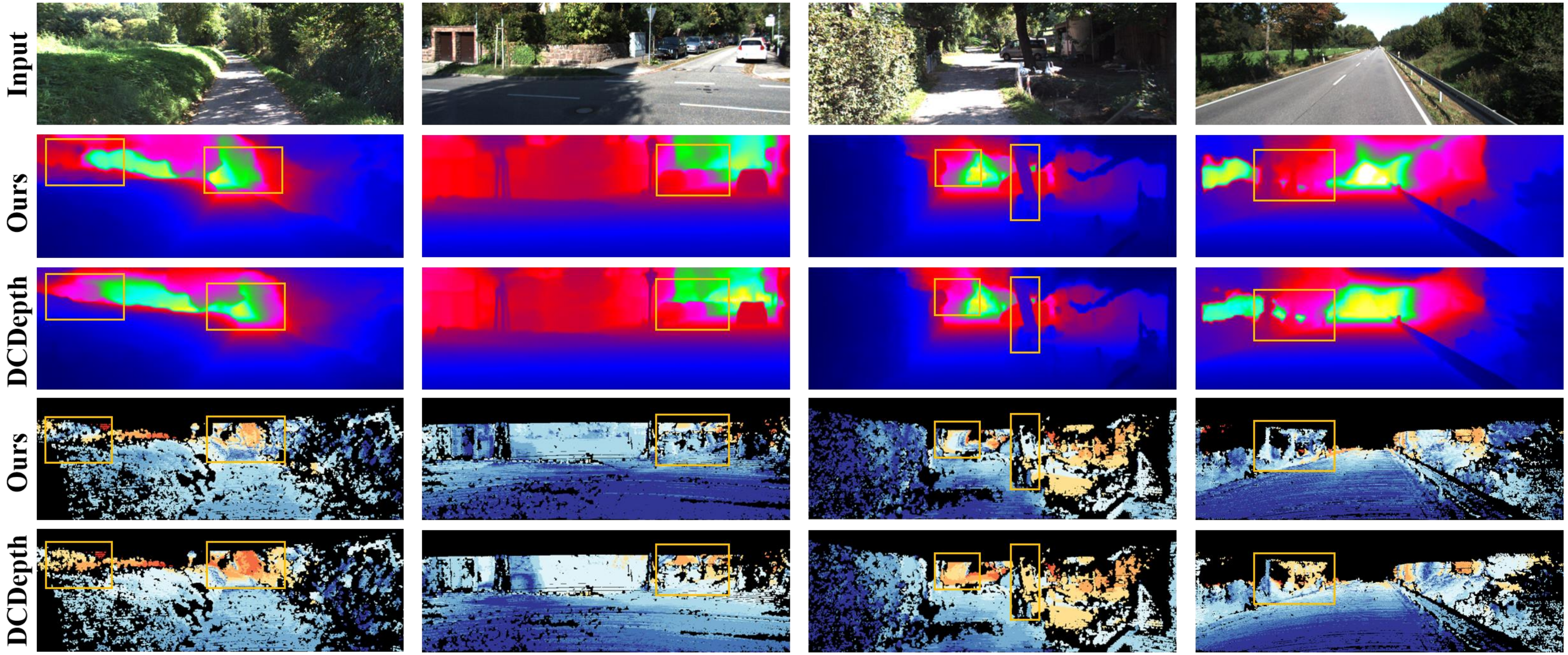
We can observe that our method surpasses other previous leading approaches at some stage by a large margin when the maximum depth is capped from 30m to 80m for outdoor scenes and from 5m to 10m for indoor scenes. This result demonstrates that our method is effective at all distances, especially when the object is far away from the camera location.



Experiment

Qualitative Comparison

For visual comparison, we provide a visual comparison of the KITTI official online benchmark for outdoor scenes. The results obtained from the official online server indicate that our method offers significant advantages over DCDepth (NIPS 2024), particularly at greater distances.



Thank you!

biel20@mails.tsinghua.edu.cn