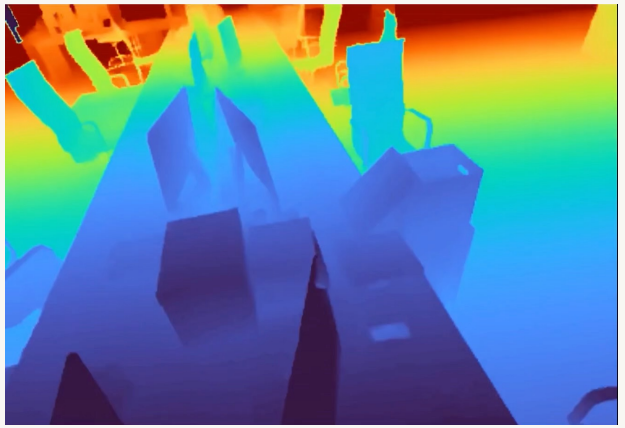
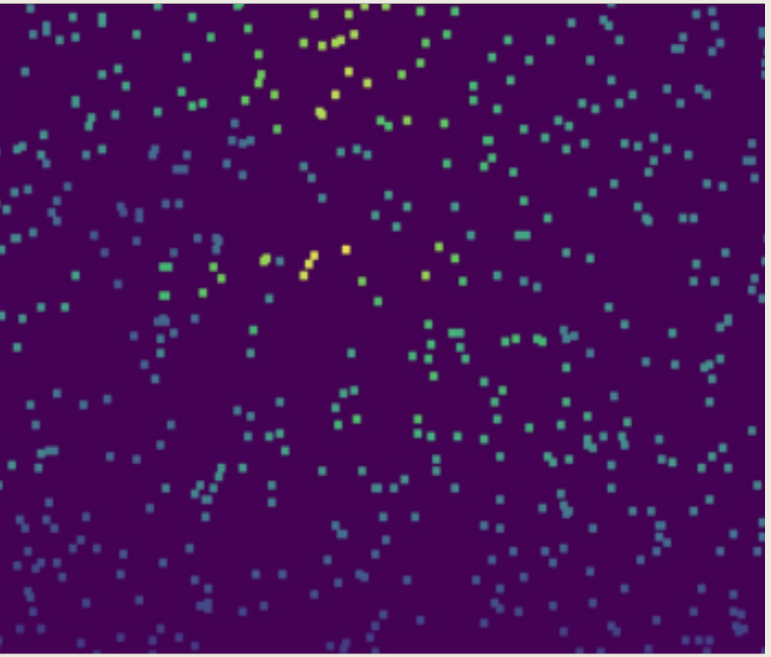


NTRANS-NET: A MULTI-SCALE NEUTROSOPHIC-UNCERTAINTY GUIDED TRANSFORMER NETWORK FOR INDOOR DEPTH COMPLETION

Akshat Ramachandran, Ankit Dhiman, Basavaraja Shanthappa Vandrotti, Jooyoung Kim

Samsung Research Institute Bangalore, India

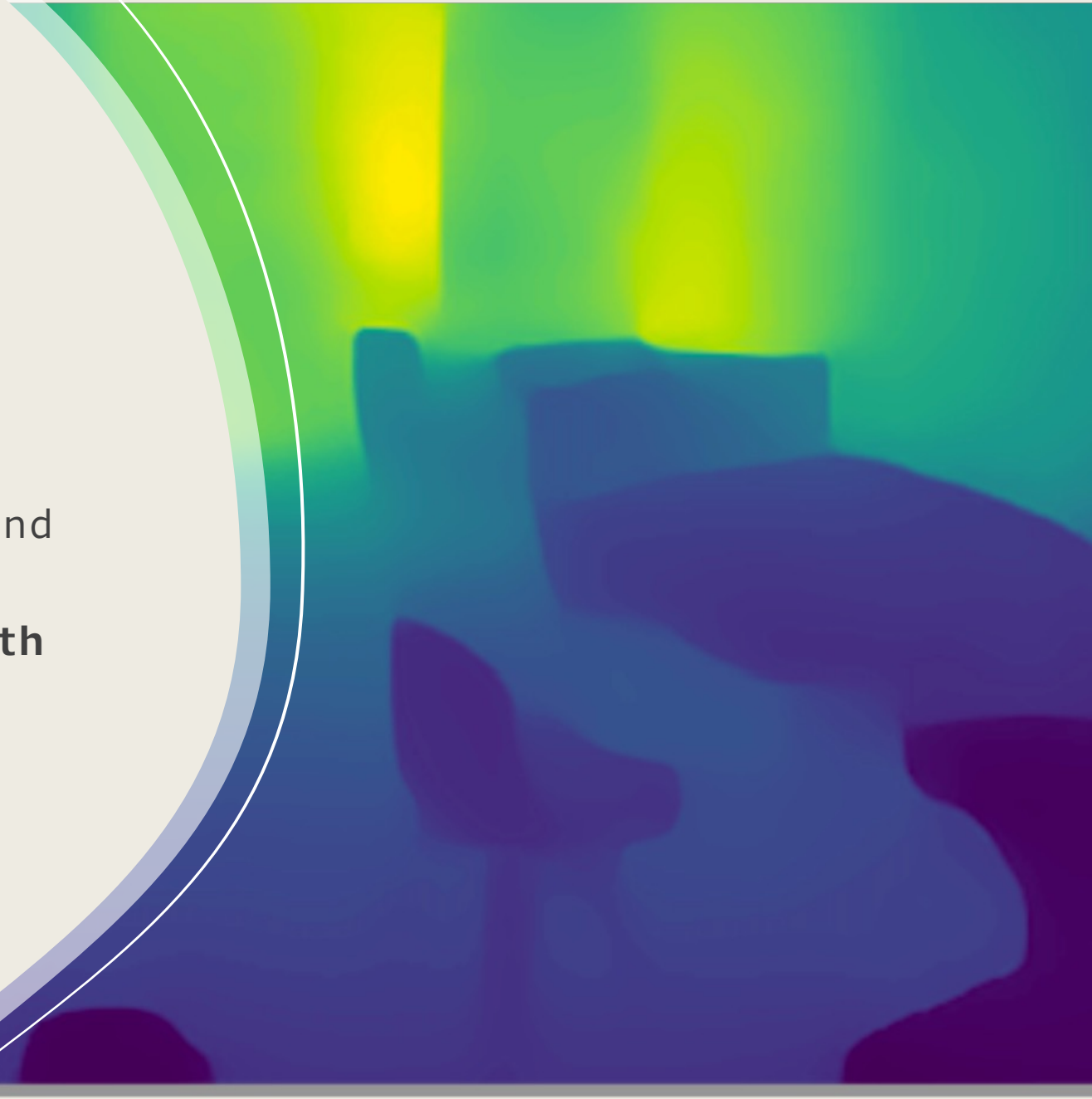
Samsung Electronics Suwon, Korea



Challenges of depth data captured by off-the-shelf sensors

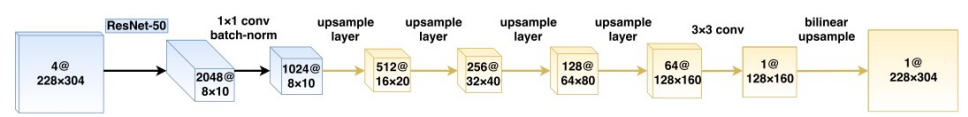
1. Sparse in nature.
2. Incomplete (semi-dense) depth data with holes and artifacts of varying distributions and sizes due to characteristics of the captured scene such as
    a. Transparent
    b. Reflective and,
    c. Dark surfaces

But **dense depth cues** are used in a wide range of applications such as augmented reality, 3D reconstruction and robotics, to provide **reliable** 3D spatial information of a scene. Therefore, **depth completion as a task is employed to generate a dense depth estimation from an input sparse and/or incomplete depth map for reliable utilization in downstream tasks.**

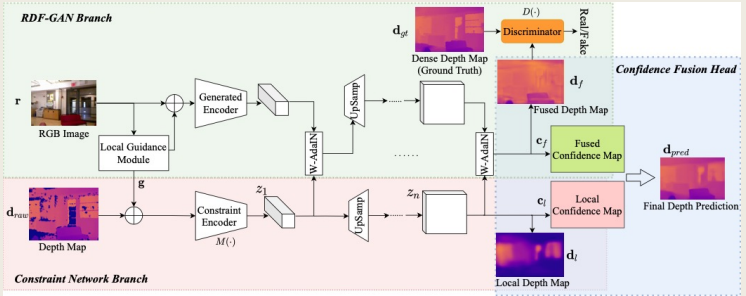


Typical Approaches to solve the Depth Completion Problem

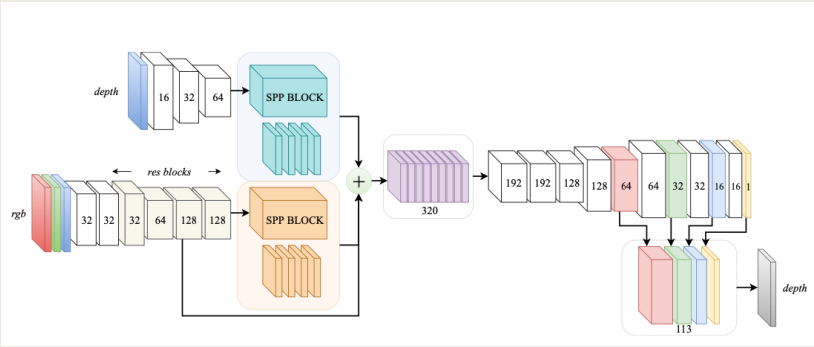
Convolutional Neural Networks based only on the sparse input. (Fangchang Ma et al.) etc.



Complex learning methodologies to adapt to the diverse and disparate spatial contexts in indoor depth. (Wang et al.) etc.



Learning based techniques leveraging multi-modal information like RGB information in addition to the corrupted depth map. (Shivakumar et al.) etc.



Broad Categories of Solving Depth Completion Problem

REGRESSION

Most of the Regression based techniques model depth completion by regressing the depth through weighting each depth hypothesis.

- Advantages
  - Helps achieve sub-pixel level depth compeletion
  - Accurate completion in holes and artifact regions.
- Disadvantages
  - Greater risk of overfitting and impact robustness

CLASSIFICATION

Classification methods predict the probability of the depth hypotheses for each pixel and take the depth hypothesis with the maximal probability as the final estimation.

- Advantages
  - Constrain the network to achieve superior completeness and hence improve the robustness of completion.
- Disadvantages
  - Greater risk of overfitting

Our Proposal

We seek to unify the advantages of regression and classification in a novel model representation, that is, we enable the model to accurately predict the depth while maintaining robustness.

Intuition and Reasoning: It is well established that the depth hypothesis close to the ground-truth has more potential knowledge, over the other remaining hypotheses (or depth levels) due to the wrong induction of multi-modal. Motivated by this, we propose that estimating the weights for all depth hypotheses is redundant, and the model only needs to do regression on the optimal depth hypothesis for the representative depth interval that contains the ground truth depth.

$$\{O_i\}_{i=0}^{K-1} = \begin{cases} 1 - e^{(-|D^{x,y} - r_i^{x,y}|)}, & \text{if } D^{x,y} \geq r_i^{x,y} \\ 0, & \text{otherwise} \end{cases}$$

Assuming a depth interval is discretized into K discrete ordinal labels $\Lambda = \{r_0, r2, ..., r_{K-1}\}$, each pixel in the depth map $D \in R^{H \times W}$ is represented as a K-length UOV, $\{Oi \in R^{H \times W}\}^{K-1}$

How?

We propose a unified novel representation for depth, termed UOVs, that can leverage the unified benefits of both these techniques.

1. The motivation behind using ordinality is that depth values are naturally ordered, and ordinal vectors are able to naturally encapsulate explicit order relations among depth hypotheses.
2. With the natural and unified representation capability of our proposal, we first adopt classification using ordinal labels to narrow the depth range of the final regression by enabling the model to classify which hypothesis is optimal and then regress to the actual depth within the optimal hypothesis.
3. Therefore, the model using our unified representation framework is able to estimate ac- curate depth like regression methods, while also directly constraining the network like classification methods.

Algorithm 2: Ordinal Decoding

Input : Predicted ordinal vectors $\{O_i \in \mathbb{R}^{H\times W}\}_{i=1}^{K}$, K ordinal categories produced by SID $\Lambda = \{r_0, r_2, ..., r_{K-1}\}$

Output : Regressed Depth $D \in \mathbb{R}^{H\times W}$

for $(x, y) = (0, 0)$ *to* (H, W) **do**

  `// Select Optimal Index`

  $o \leftarrow \arg\max_{i\in\Lambda} O_i^{x,y}$;

  `// Regress depth value`

  $D^{x,y} \leftarrow -\ln(1 - O_o) + r_o^{x,y}$;

end

return D

Algorithm 1: Ordinal Encoding

Input : Ground-truth depth $D_{gt} \in \mathbb{R}^{H\times W}$, K ordinal categories produced by SID $\Lambda = \{r_0, r_2, ..., r_{K-1}\}$

Output : Ground-truth ordinal vectors $\{O_i \in \mathbb{R}^{H\times W}\}_{i=1}^{K}$

for $(x, y) = (0, 0)$ *to* (H, W) **do**

  for $i = 0$ *to* $K - 1$ **do**

    if $D^{x,y} \geq r_i^{x,y}$ **then**

      $O_i \leftarrow 1 - \exp\{-|D^{x,y} - r_i^{x,y}|\}$;

    end

    else

      $O_i \leftarrow 0$;

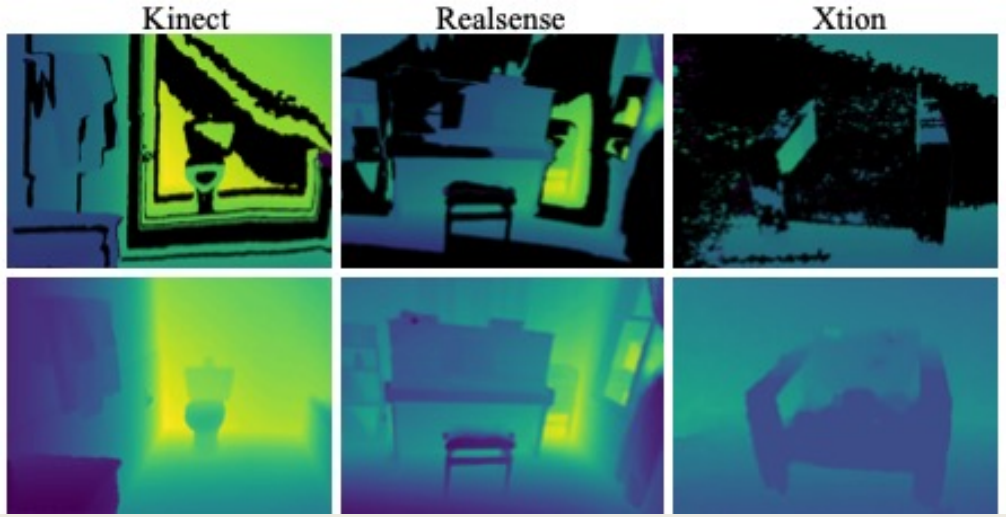
    end

  end

end

return $\{O_i\}_{i=1}^{K}$

Unified Ordinal Vector Encoding and Decoding



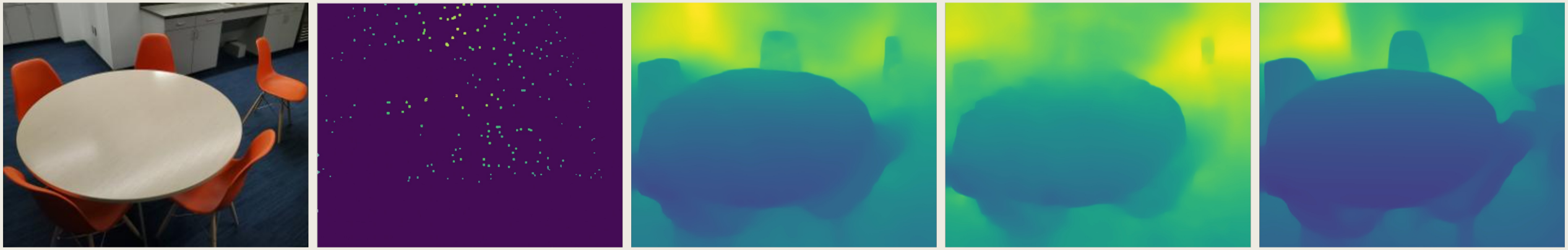
Addressing real-world depth completion quality

While many existing depth completion techniques have shown remarkable performance in completing uniformly sampled sparse depth maps from a singular sensor configuration, that performance is not representative of the performance on real-world depth maps with large missing regions and semantic missing patterns. Since indoor depth is dynamic in nature and has varying sensor-dependent data distributions a completion network has to properly model the ambiguity in these depth maps to truly generalize to the multitudinous possibilities of noise and artefacts present in indoor depth sensors.

How we overcome this drawback in several SoTA techniques

The complexity we leverage to enable better and robust depth completion is uncertainty. In this work, we handle uncertainty via a concept known as Neutrosophic Sets (NS) owing to their superior capability to highlight, extract and handle uncertainty information in learning-based techniques.

RGB | Sparse Depth | NLSPN | CSPN | **Ours**



Methodology: Neutrosophic Sets

A NS is made of three elements namely, Truth (T), False (F) and Indeterminacy (I). In general, the neutrosophic set N can be represented as a composition of the three sub-sets as follows,

$$N = \{(T, I, F) : T, I, F \in [0, 1]\}$$

To transform an image to the neutrosophic domain, each pixel P (i, j) in the original image domain is represented as P (t, l, f) in the neutrosophic domain where t, l, f indicate that the pixel is t% true, l% indeterminate and f% false using the following transformations.

$$t(i,j) = \frac{\bar{g}(i,j) - g_{min}}{\bar{g}_{max} - \bar{g}_{min}}$$

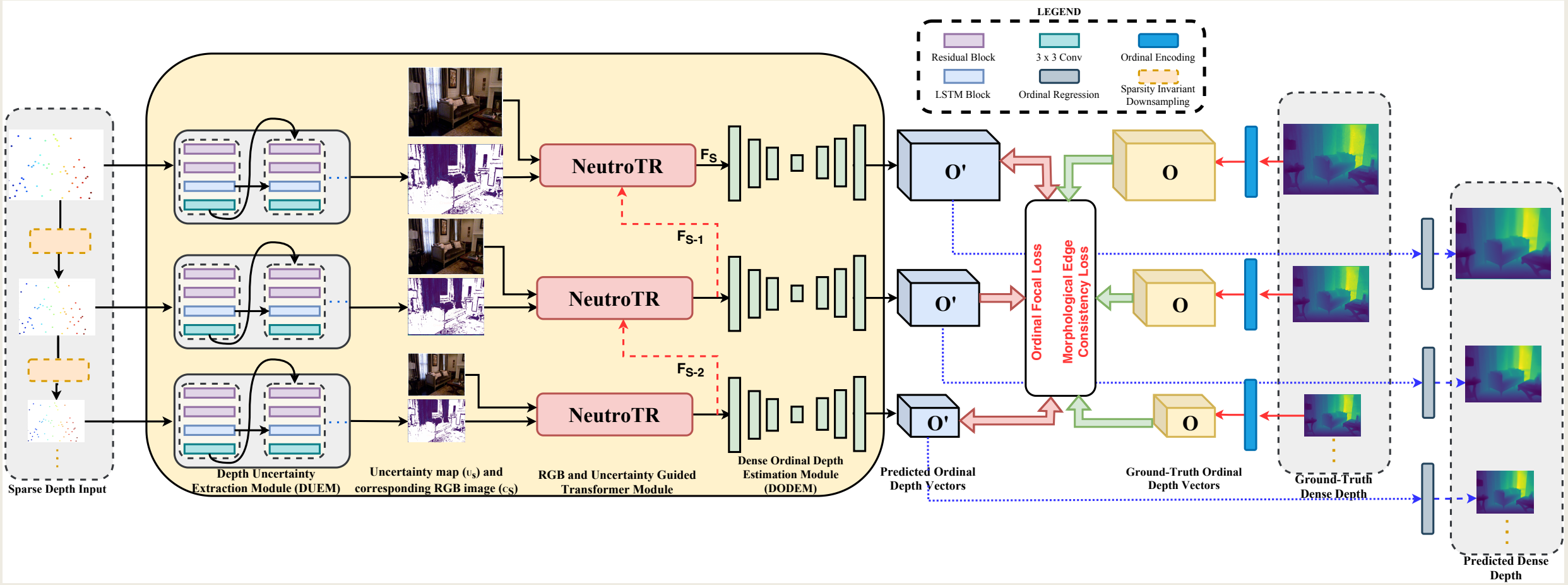
$$l(i,j) = \frac{\delta(i,j) - \delta_{min}}{\delta_{max} - \delta_{min}}$$

$$f(i,j) = 1 - t(i,j) - l(i,j)$$

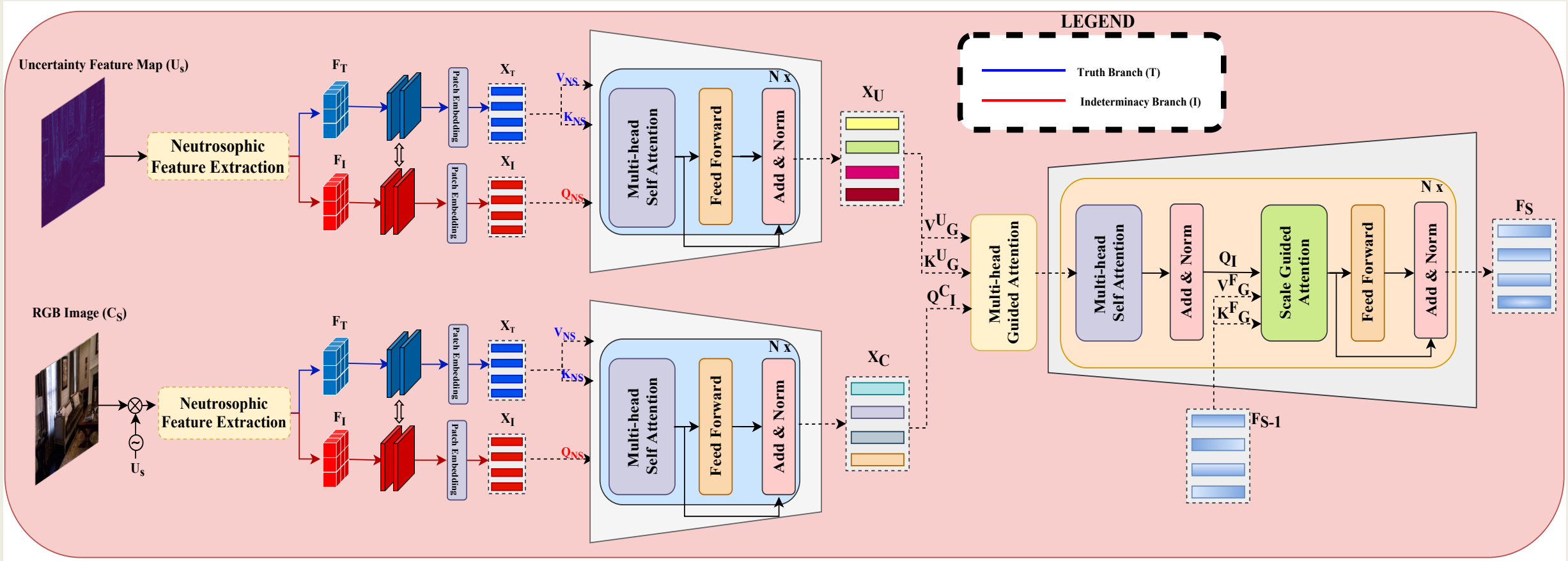
$$\bar{g}(i,j) = \frac{1}{w * w} \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} g(m,n)$$

$$\delta(i,j) = \boldsymbol{abs}(g(i,j) - \bar{g}(i,j))$$

Proposed Architecture



Proposed Dual-encoder transformer architecture: NeutroTR



Data Preparation

1. Specular Masking: We identify regions of pos- sible specular highlights and reflective regions in the corresponding RGB images using an adaptive color balance threshold compared to a smoothed non-specular surface color at each pixel position. We then mask these identified segments in the raw depth map in the raw depth.

2. Black Masking: We randomly mask surfaces and objects with pixel intensities in the range [0, 10] to emulate absence of depth in dark surfaces in real-world depth maps.

3. Semantic Masking: Usually there are random re- gions with large contiguous holes in depth maps, we try to emulate the same by masking off certain objects with a probability of 0.35 identified from the corresponding segmentation map.

4. Random Noise Injection: We randomly add Gaus- sian noise to 5 – 10% of the valid pixels, since in real-world use cases many pixels in the depth map may have noise due to superfluous reflections from objects.

5. Depth Range Clipping: We also randomly clip 25 – 30% of depth values farther than 8m, since commodity depth sensors present in most mobile devices give unreliable depth for long ranges.

Loss Function

We make use of a Multiscale Ordinal Focal Loss (Lo) as the primary training loss.

$$L_o = \frac{1}{N}\sum_{(x,y)}\sum_i \begin{cases} \alpha|q_i - u_i|^\gamma BCE(u_i, q_i), & \text{if } q_i > 0 \\ (1-\alpha)u_i^\gamma BCE(u_i, q_i), & \text{otherwise} \end{cases}$$

$$L = L_o + \lambda L_e$$

where, $BCE(u_i, q_i) = -q_i\log(u_i) - (1-q_i)\log(1-u_i)$

To emphasize depth discontinuities at object boundaries and to enhance smoothness in homogeneous regions, we use a Morphological Edge Consistency Loss (Le) on the regressed depth as

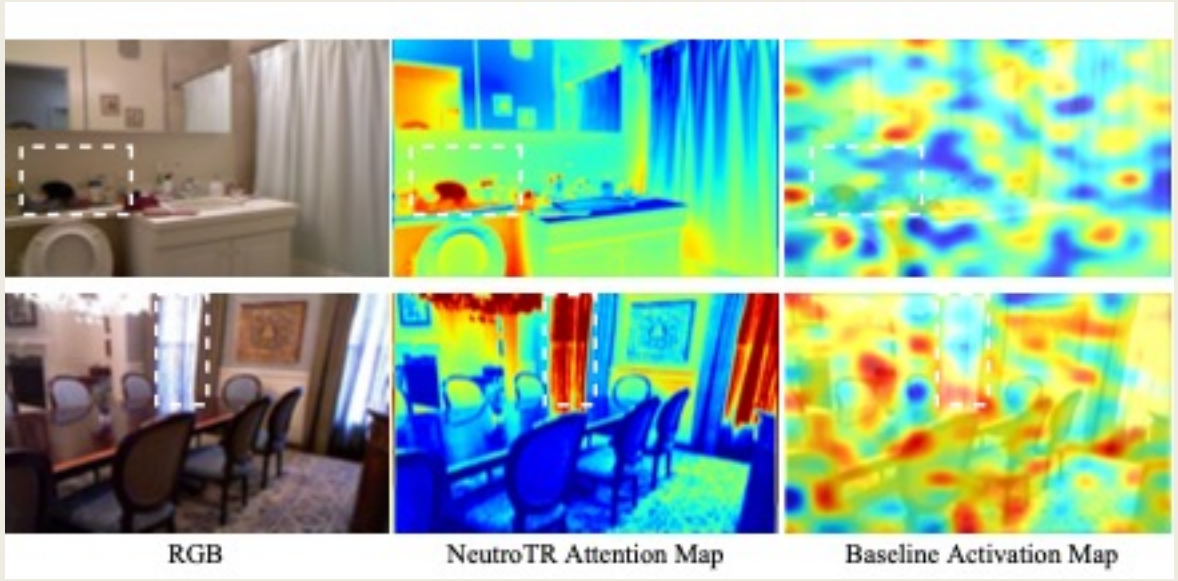
$Le = ||Morph(D_{x,y}) - Morph(d_{x,y})||$, where an improved canny edge detection with morphological gradients (Morph()) is applied to extract refined depth edges and suppress noisy textural edges that are typically extracted with basic edge detectors like Sobel.

Ablation Study on NYUv2

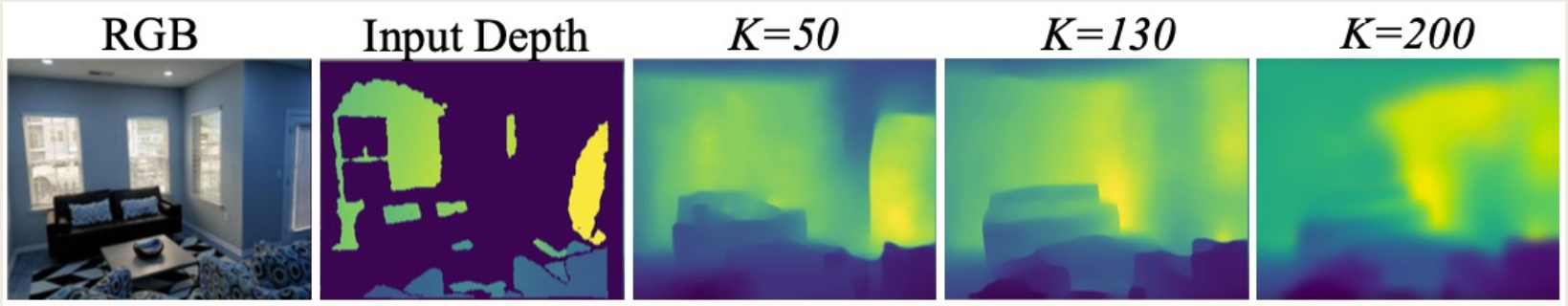
Model	MS	DUEM	NTR	L_1	L_o	L_e	RMSE (m)↓	REL (m)↓
S_base				✓			0.221	0.040
M_base	✓			✓			0.196	0.028
S_U		✓			✓	✓	0.165	0.019
M_U	✓	✓			✓	✓	0.142	0.019
w/o L_o	✓	✓	✓	✓		✓	0.125	0.017
w/o L_e	✓	✓	✓		✓		0.098	0.013
Ours(reg)	✓	✓	✓	✓		✓	0.103	0.013
Ours(cla)	✓	✓	✓		✓	✓	0.106	0.015
Ours	✓	✓	✓		✓	✓	**0.091**	**0.012**

MS: Multi-scale, NTR: NeutroTR

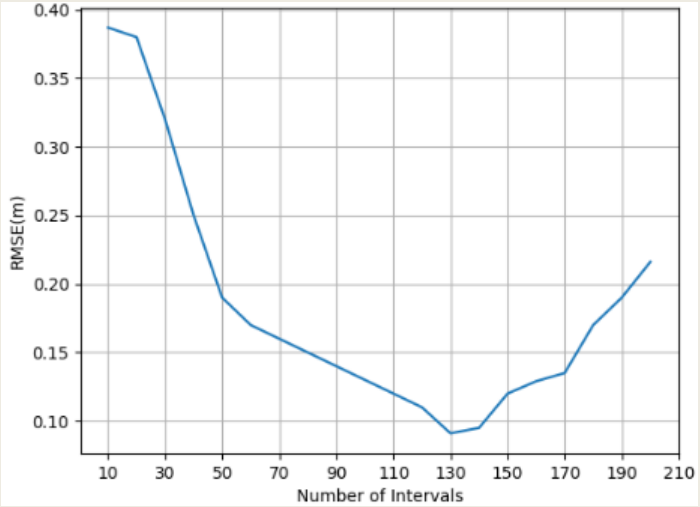
Ablation results on NYUv2 test set. In the table, S, M, Reg, Cla, Base denote single-scale, multi-scale, regression, classification and baseline respectively.



Layer response of NeutroTR compared with U-Net like baseline, higher response is denoted with red and lower tending to blue. NTrans-Net implicitly learns to gives higher response to regions where depth is prone to errors like at dark and reflective surfaces as compared to the global response of CNNs.

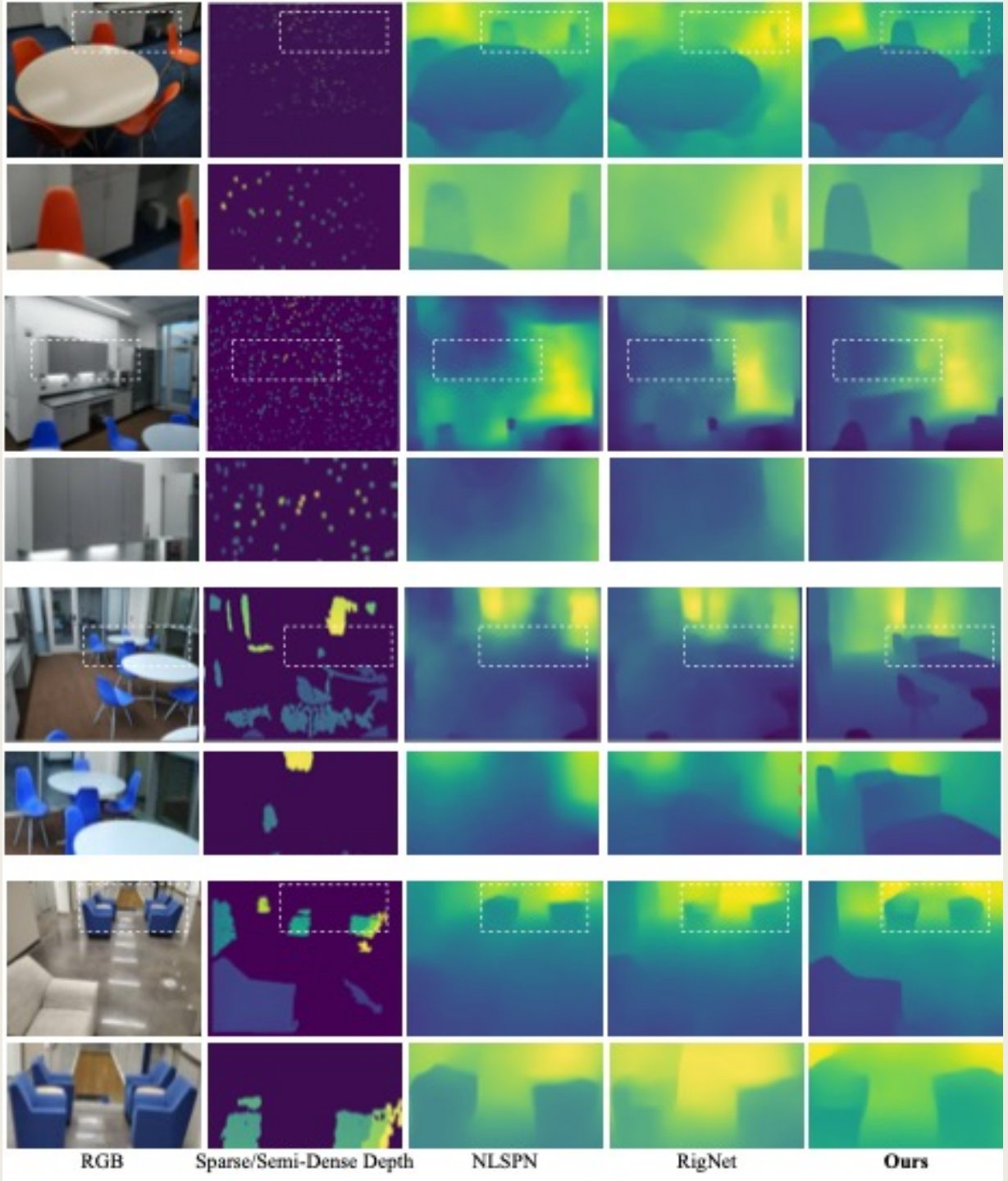


Visualization of the effects of choosing different discretization intervals(K) on the completion quality. Lower values of K, introduce quantization errors, and large values of K produce overly smooth output, whereas K=130 results in the optimal completion quality.

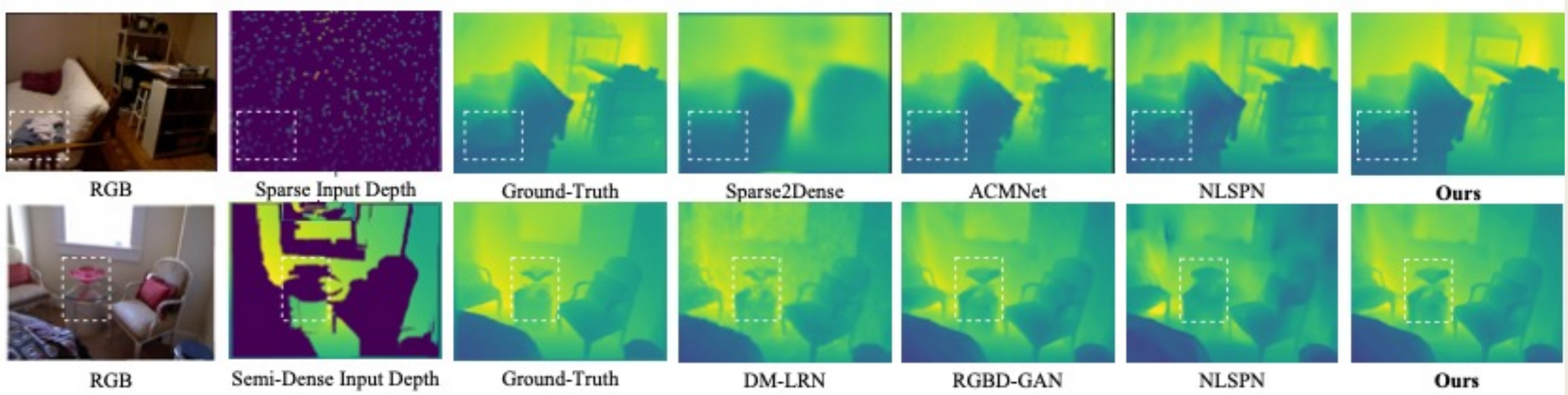


UOV performance on different intervals. We observed that UOVs acquire the best results at K=130 with the lowest RMSE of 0.087m.

Results on real world ToF18K


RGB
Sparse/Semi-Dense Depth
NLSPN
RigNet
Ours

Results on NYuv2 dataset



Quantitative Evaluation

$\mathcal{S}\rightarrow\mathbf{D}$						$\mathcal{S}^{*}\rightarrow\mathbf{D}$					
Method	RMSE (m)↓	REL (m)↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Method	RMSE (m)↓	REL (m)↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
DC-BCS [7]	0.268	0.017	98.2	99.1	99.3	Sparse2Dense [17]	0.227	0.043	97.1	99.4	99.8
RGB-GU [33]	0.253	0.019	98.0	99.1	99.3	RGBD-GAN [34]	0.103	0.016	99.4	99.9	100.0
DM-LRN [23]	0.203	0.017	98.2	99.3	99.9	ACMNet [39]	0.102	0.014	99.5	99.9	100.0
RigNet [35]	0.164	0.016	98.6	99.6	99.9	GuideNet [31]	0.101	0.015	99.5	99.9	100.0
NLSPN [18]	0.161	0.015	98.6	99.6	99.9	NLSPN [18]	0.092	0.012	99.6	99.9	100.0
RGBD-GAN [34]	0.139	**0.013**	98.7	99.6	99.9	RigNet [35]	0.090	0.013	99.6	99.9	100.0
Ours	**0.134**	**0.013**	98.7	99.6	99.9	**Ours**	**0.087**	**0.010**	99.6	99.9	100.0

NYUv2

Method	RMSE (m)↓	REL (m)↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Sparse2Dense [17]	0.247	0.048	96.1	98.5	99.3
GuideNet [31]	0.234	0.044	96.3	98.7	99.3
ACMNet [39]	0.230	0.041	96.3	98.7	99.3
RigNet [35]	0.218	0.038	96.8	98.9	99.4
RGBD-GAN [34]	0.212	0.031	97.1	99.2	99.7
NLSPN [18]	0.210	0.028	97.1	99.4	99.8
Ours	**0.197**	**0.021**	97.3	99.4	99.8

ToF18k

Method	RMSE (m)↓	REL (m)↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Sparse2Dense [17]	0.329	0.074	93.9	97.0	98.1
GuideNet [31]	0.295	0.135	93.8	97.2	98.2
ACMNet [39]	0.274	0.071	94.5	97.7	98.3
NLSPN [18]	0.267	0.063	97.3	98.1	98.5
RigNet [35]	0.265	0.063	97.1	98.1	98.5
RGBD-GAN [34]	0.255	0.059	96.9	98.4	99.0
Ours	**0.244**	**0.051**	97.2	98.6	99.4

SUN RGB-D

Conclusion

We present NTrans-Net, a novel multi-scale network which combines the advantages of regression and classification techniques by proposing a UOV representation for depth values. Further, to make the framework robust to different sensor-dependent input distributions, we propose NeutroTR, a dual encoder-decoder transformer with data indeterminacy handling in the neutrosophic domain. Through our experiments, we demonstrate the flexibility of our framework to adapt to the diverse and disparate spatial contexts and artefacts present in depth maps. Extensive experiments demonstrate that our proposed framework achieves state-of-the-art performance on the NYUv2 benchmark dataset, while achieving superior generalization on the real-world ToF18K dataset–captured using Samsung Galaxy Note10 device in indoor scenarios. Moreover, our exhaustive ablation studies show the effectiveness of each proposed component in the framework.