Depth Estimation from Single Image and Semantic Prior

Praful Hambarde, Akshay Dudhane, Prashant W. Patil, Subrahmanyam Murala and Abhinav Dhall

Computer Vision and Pattern Recognition Lab

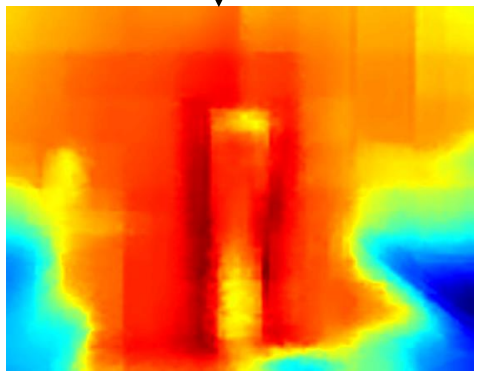
Indian Institute of Technology Ropar, India and Monash University, Australia

Motivation

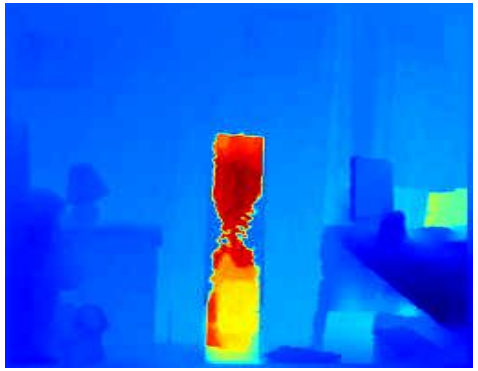
Single Image Depth Estimation (SIDE)

- The SIDE methods are extremely unreliable since a single RGB image does not provide depth clue on its own.
- The state-of-the-art SIDE methods produce a high prediction error.
  - Indoor : RMSE=50cm
  - Outdoor : RMSE=7m

RGB Image



Predicted Depth-map

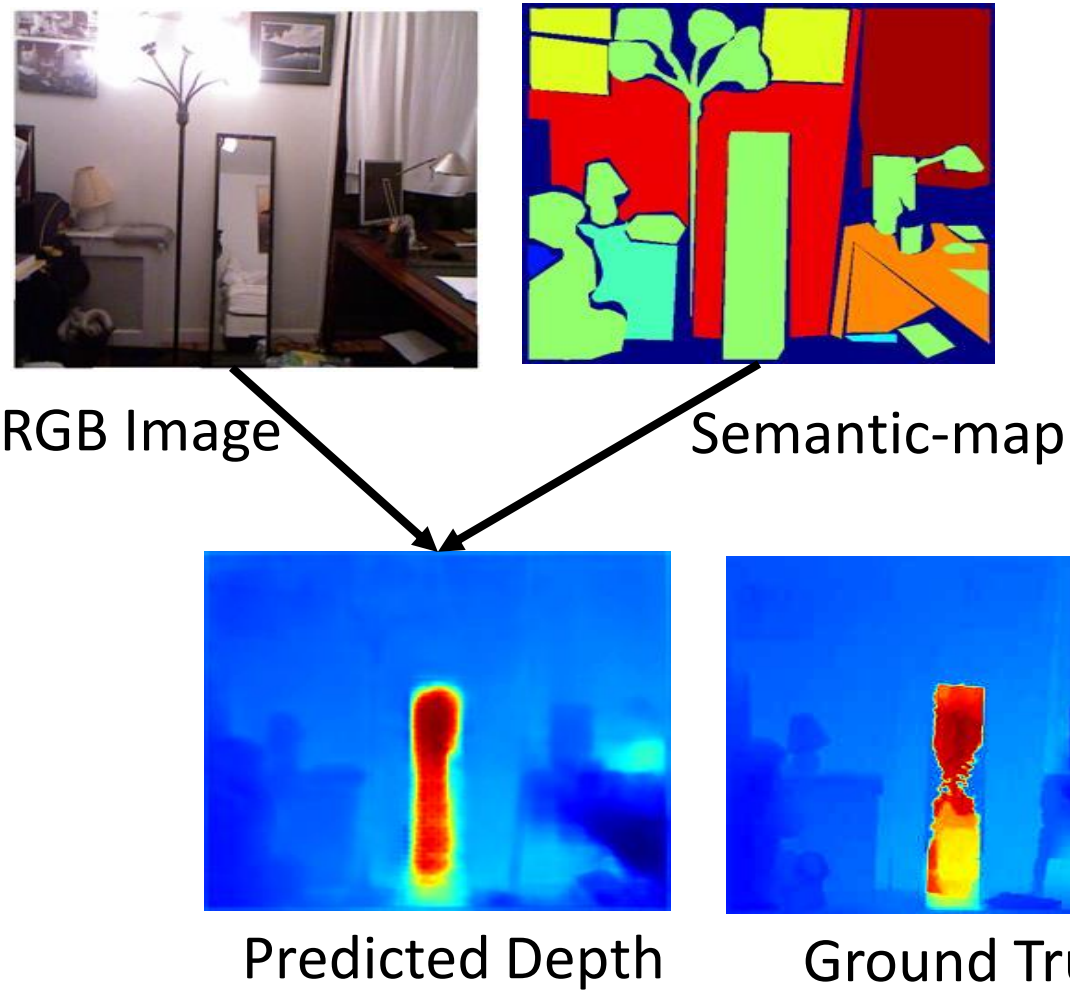


Ground Truth Depth-map

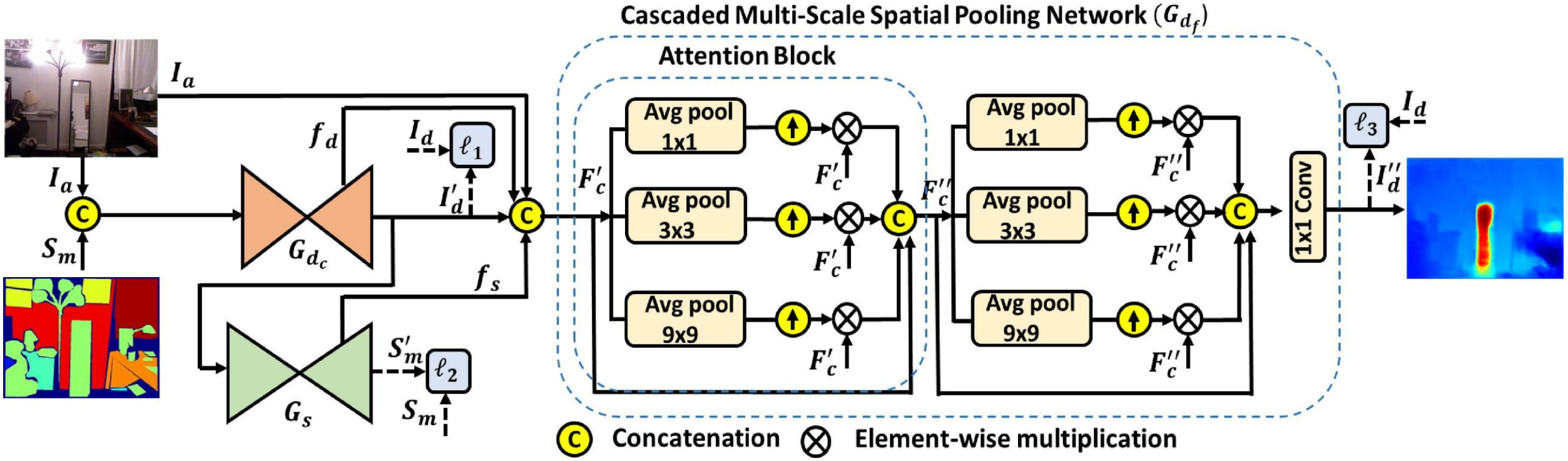
Single Image and Semantic-map Depth Estimation

- To address glossy, crystal-clear, and delicate surfaces limitations of SIDE, in this paper, we make use of single image with semantic prior for depth estimation.

- Semantic maps are readily available from semantic segmentation algorithms.



Proposed Network



Network Objective

- The proposed S2D-GAN for depth estimation consist three generators namely, G_{d_c}, G_s, G_{d_f} and a joint discriminator D.
- Task of the joint discriminator is to discriminate between the generated depth maps I'_d , I''_d from the real depth map I_d .
- To optimize the network parameters, we have considered traditional L1 loss along with the adversarial loss.

$$L_{GAN}(I_a, I'_d) = E_{I_a,I_d}[logD(I_a, I_d)] + E_{I_a,I'_d}[\log(1 - D(I_a, I'_d))]$$

$$L_{GAN}(I_a, I''_d) = E_{I_a,I_d}[logD(I_a, I_d)] + E_{I_a,I''_d}[\log(1 - D(I_a, I''_d))]$$

$$\min_{G_{d_c},G_s,G_{d_f}} \max_D L_{GAN} = \sum_{j=1}^{3} \beta_j L_1^j + L_{GAN}(I_a, I'_d) + \beta L_{GAN}(I_a, I''_d)$$

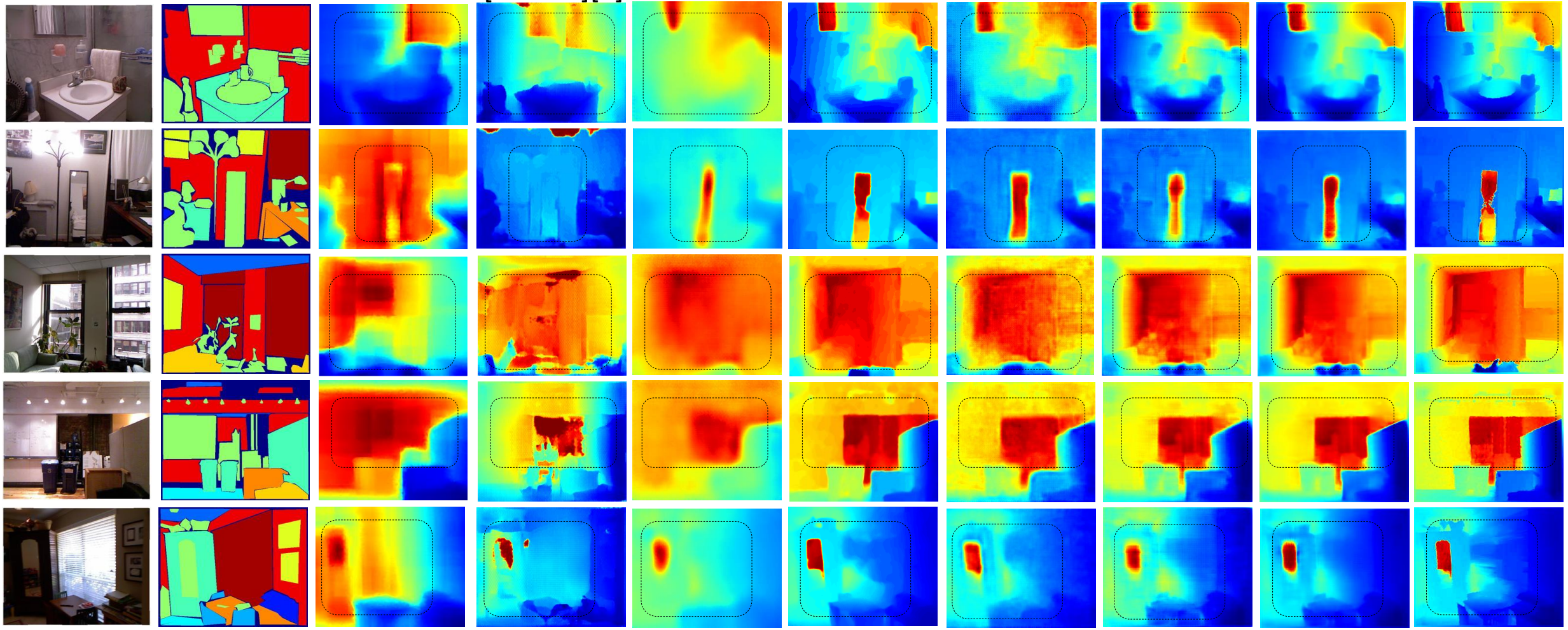
Training Details

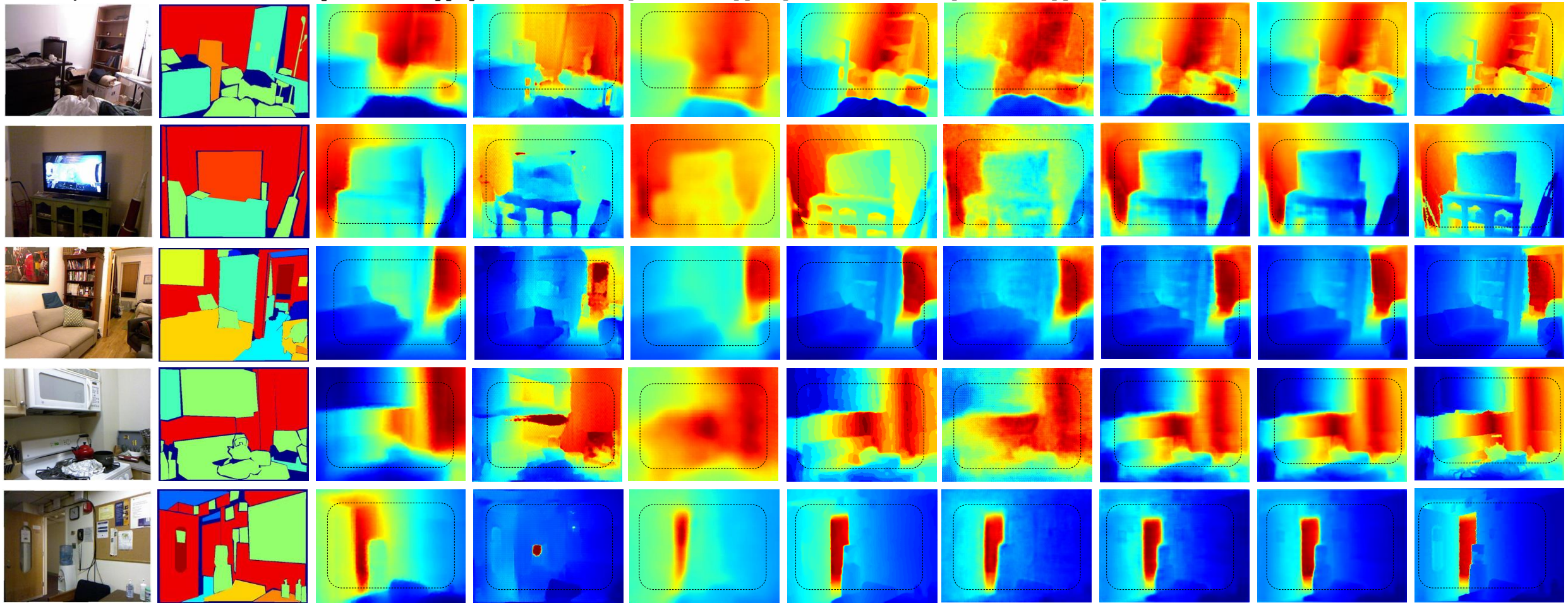
- Indoor scene images and their respective depth and semantic maps from existing benchmark NYU-Depth-V2 database are considered to train the proposed S2D-GAN for depth map estimation.
- NYU-Depth-V2 database consists of 1449 labeled indoor images with 640x480 resolution and their respective depth-map. We have considered pre-defined training (795 images) and testing (654 images) spilts for the experimentation.
- The proposed S2D-GAN is trained using ADAM optimization algorithm for 200 epochs with learning rate of 0.0001 on NVIDIA DGX station.

Results Analysis

Table 1. Comparative depth estimation evaluation results of the proposed (S2CD-GAN, S2D-GAN), and existing methods on NYU-Depth-V2 dataset.

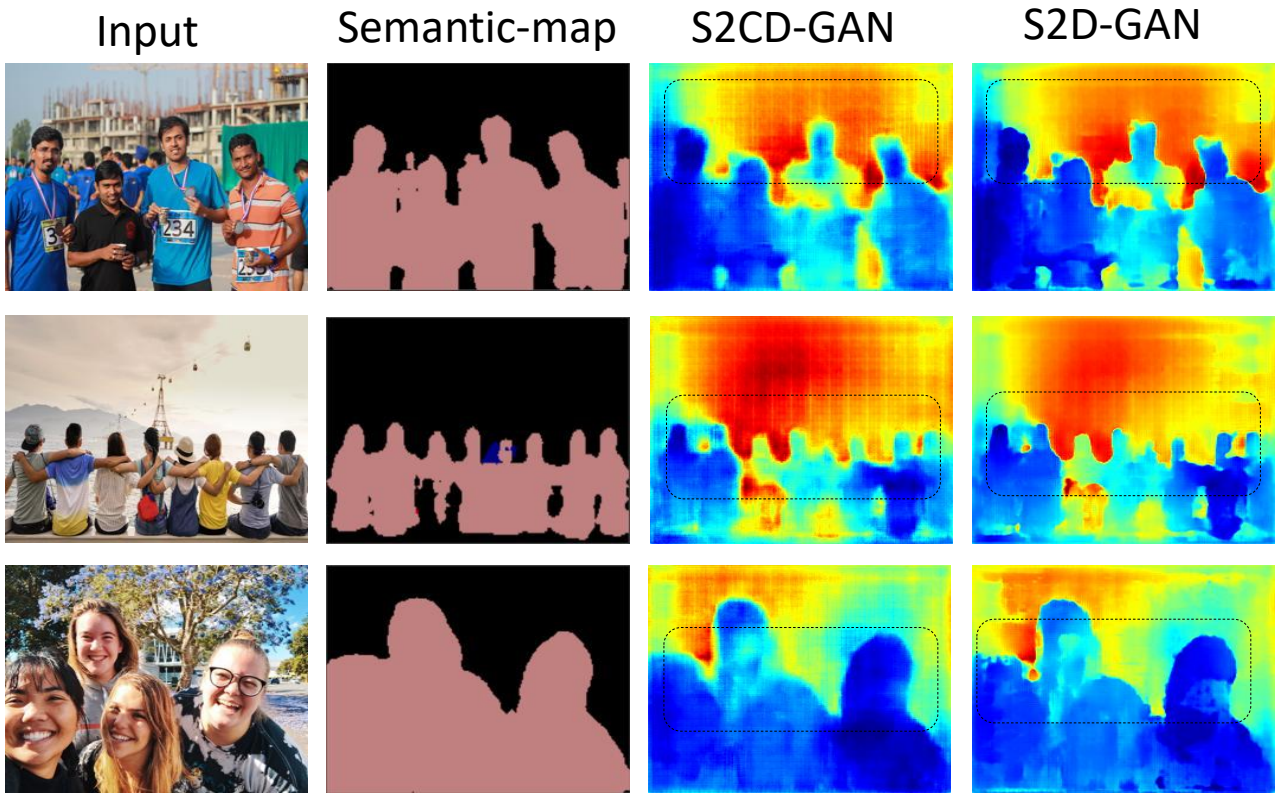
Method	RMSE↓	REL↓	δ_1↑	δ_2↑	δ_3 ↑
CVPR-18 [3]	0.572	0.139	81.5	96.3	99.1
CVPR-18 [13]	0.547	0.116	85.6	96.1	98.6
CVPR-19 [4]	0.538	0.131	83.7	97.1	99.4
ICIP-19 [5]	0.509	0.142	80.6	95.5	98.8
ICRA-17(225) [14]	0.442	0.104	87.8	96.4	98.9
ECCV-18(200) [6]	0.221	**0.040**	97.0	99.1	99.3
ITSC-19(200) [17]	0.203	**0.040**	**97.6**	99.2	99.7
S2CD-GAN	0.219	0.055	96.2	99.4	99.4
S2D-GAN	**0.196**	0.048	96.7	**99.3**	**99.7**


Input
Semantic-map
[CVPR-19][4]
[ICIP-19][5]
[ICRA-18][1]
[ECCV-18][6]
[ITSC-19][17]
S2CD-GAN
S2D-GAN
Ground Truth

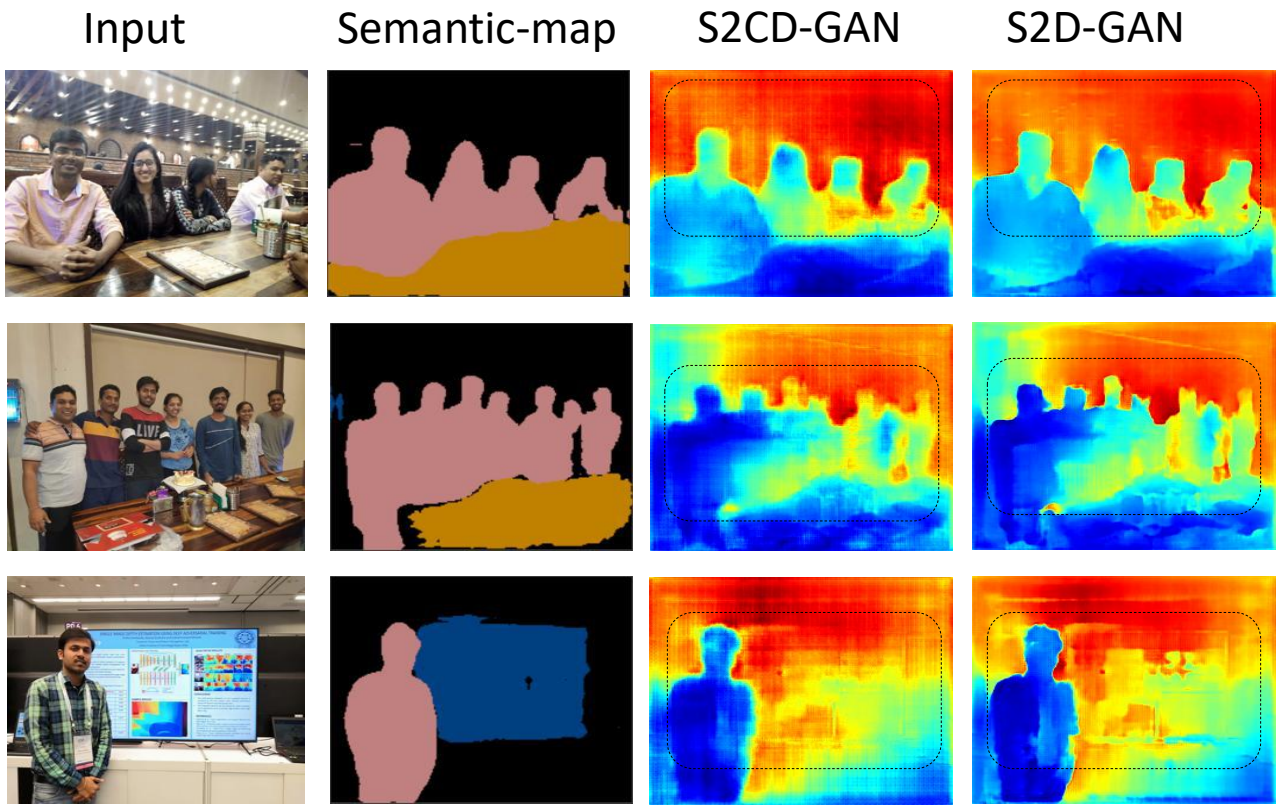

Input
Semantic-map
[CVPR-19][4]
[ICIP-19][5]
[ICRA-18][11]
[ECCV-18][6]
[ITSC-19][17]
S2CD-GAN
S2D-GAN
Ground Truth

Cross Dataset Evaluation

Outdoor Scene:



Indoor Scene:

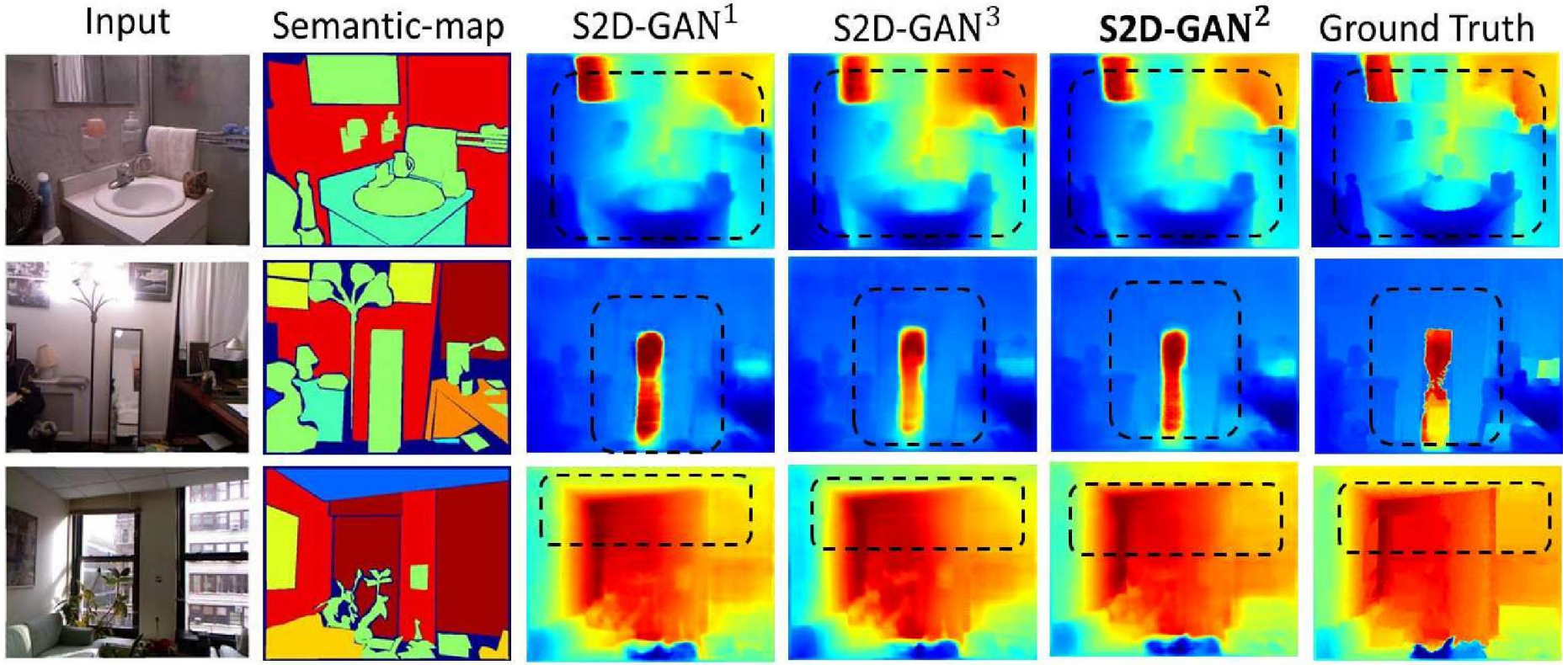


Ablation Study

Table 2. Ablation Study about the number of attention blocks in the proposed S2D-GAN.

Method	RMSE↓	REL↓	δ_1↑	δ_2↑	δ_3 ↑
S2D-GAN^1	0.246	0.063	94.8	98.7	99.6
S2D-GAN^3	0.263	0.069	94.2	98.7	99.6
S2D-GAN^2	**0.196**	**0.048**	**96.7**	**99.3**	**99.7**

Ablation Study



Conclusion

- We propose a novel method of S2D-GAN for depth estimation from an input RGB image and its semantic-map.

- In the first stage S2D-GAN, predicts the coarse level depth-map followed by the cascaded multi-scale spatial pooling network which reduces the pixel-level discontinuity present in the coarse-level depth map.

- Experiments show that the proposed S2D-GAN effectively handles the illumination problem as well as repetitive patterns and obtained a fine-level depth map as compared with the existing state-of-the- art methods.