Estimating and Exploiting the Aleatoric Uncertainty in Surface Normal Estimation
Supplementary Material

Gwangbin Bae Ignas Budvytis Roberto Cipolla
University of Cambridge
{gb585,ib255,rc10001}@cam.ac.uk

In this supplementary material, we provide (1) the derivations for the proposed AngMF distribution, (2) quantitative evaluation with additional metrics, (3) cross-dataset evaluation on KITTI [4] and DAVIS [5], (4) discussion on failure modes and (5) additional qualitative comparison against the state-of-the-art.

1. Derivations for the proposed AngMF distribution

In the paper, we introduced a variant of the von Mises-Fisher distribution [3], such that its negative log-likelihood (NLL) is the angular loss with learned attenuation. We call this the Angular vonMF (AngMF) distribution. In this section, we provide the derivations for Eq. 4, Eq. 5 and Eq. 6 in the paper.

1.1. Probability density function (Eq. 4)

The NLL of the distribution should have the form of

$$\mathcal{L}_i = C(\kappa_i) + \kappa_i \cos^{-1} \boldsymbol{\mu}_i^T \mathbf{n}_i^{gt}, \tag{1}$$

where i is the pixel index and $\cos^{-1} \boldsymbol{\mu}_i^T \mathbf{n}_i^{gt}$ is the angle between the predicted mean direction $\boldsymbol{\mu}_i$ and the ground truth surface normal $\mathbf{n}_i^{gt}$. The angular error is weighted by the concentration parameter κ_i, which encodes the network's confidence in the predicted mean direction. The first term $C(\kappa_i)$ should be a monotonically decreasing function of κ_i in order to prevent the trivial solution where $\kappa_i = 0 \; \forall \; i$. Then, the probability density function (PDF) should look like

$$p(\mathbf{n}_i|\boldsymbol{\mu}_i, \kappa_i) = D(\kappa_i) \exp(-\kappa_i \cos^{-1} \boldsymbol{\mu}_i^T \mathbf{n}_i), \tag{2}$$

where $C(\kappa_i) = -\log(D(\kappa_i))$. We can then compute the cumulative probability of the angular error $\cos^{-1} \boldsymbol{\mu}_i^T \mathbf{n}_i$ being less than some threshold α^*. See Fig. 1-(a) for the axes orientation used for the integration.

$$\begin{aligned} P[\cos^{-1}(\boldsymbol{\mu}_i^T \mathbf{n}_i) \leq \alpha^*] &= \int_0^{2\pi} \int_0^{\alpha^*} D(\kappa_i) \exp(-\kappa_i \phi) \sin\phi d\phi d\theta \\ &= 2\pi D(\kappa_i) \int_0^{\alpha^*} \exp(-\kappa_i \phi) \sin\phi d\phi \\ &= 2\pi D(\kappa_i) \left[\frac{-\exp(-\kappa_i \phi)(\cos\phi + \kappa_i \sin\phi)}{\kappa_i^2 + 1} \right]_0^{\alpha^*} \\ &= 2\pi D(\kappa_i) \frac{1 - \exp(-\kappa_i \alpha^*)(\cos\alpha^* + \kappa_i \sin\alpha^*)}{\kappa_i^2 + 1} \end{aligned} \tag{3}$$

Solving $P[\cos^{-1}(\boldsymbol{\mu}_i^T \mathbf{n}_i) \leq \pi] = 1$ gives

$$D(\kappa_i) = \frac{1}{2\pi} \frac{\kappa_i^2 + 1}{1 + \exp(-\kappa_i \pi)}. \tag{4}$$

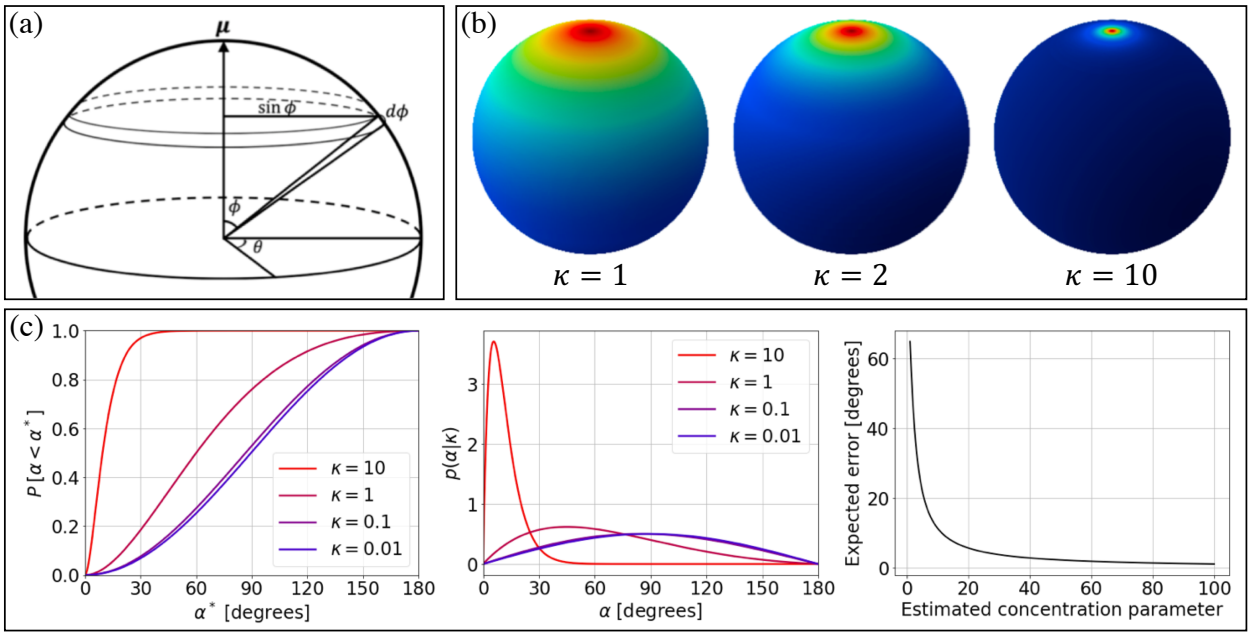

Figure 1. (a) The axes orientation used for the integrations in Eq. 3 and Eq. 9. The mean direction $\boldsymbol{\mu}$ is aligned with the z-axis, and is thus excluded in the integration. (b) Visualization of Eq. 5 for different values of κ. κ determines how concentrated the distribution is towards the mean direction. (c) Eq. 7, Eq. 8 and Eq. 9 plotted for different values of κ. The expected error decreases as the confidence κ increases.

Inserting Eq. 4 into Eq. 2 gives

$$p_i(\mathbf{n}_i|\boldsymbol{\mu}_i, \kappa_i) = \frac{(\kappa_i^2 + 1)\exp(-\kappa_i \cos^{-1} \boldsymbol{\mu}_i^T \mathbf{n}_i)}{2\pi(1 + \exp(-\kappa_i \pi))}, \tag{5}$$

which is Eq. 4 in the paper. Fig. 1-(b) visualizes the distribution for different values of κ. As κ increases, the distribution becomes more concentrated around the mean direction.

1.2. Negative log-likelihood (Eq. 5)

The network is trained by minimizing the NLL of the ground truth normal. The training loss can thus be written as

$$\mathcal{L}_i = -\log(\kappa_i^2 + 1) + \log(1 + \exp(-\kappa_i \pi)) + \kappa_i \cos^{-1} \boldsymbol{\mu}_i^T \mathbf{n}_i^{\text{gt}}, \tag{6}$$

where we drop the constant term, $\log 2\pi$. This is Eq. 5 in the paper.

1.3. Measure of uncertainty (Eq. 6)

Inserting Eq. 4 to Eq. 3 gives

$$P[\cos^{-1}(\boldsymbol{\mu}_i^T \mathbf{n}_i) \leq \alpha^*] = \frac{1 - \exp(-\kappa_i \alpha^*)(\cos \alpha^* + \kappa_i \sin \alpha^*)}{1 + \exp(-\kappa_i \pi)}. \tag{7}$$

From this, we can calculate the probability density function for the angular error α via differentiation.

$$\begin{aligned} p(\alpha|\boldsymbol{\mu}_i, \kappa_i) &= \frac{d}{d\alpha}\left(\frac{1 - \exp(-\kappa_i \alpha)(\cos \alpha + \kappa_i \sin \alpha)}{1 + \exp(-\kappa_i \pi)}\right) \\ &= \frac{-\exp(-\kappa_i \alpha)(-\sin \alpha + \kappa_i \cos \alpha) + \kappa_i \exp(-\kappa_i \alpha)(\cos \alpha + \kappa_i \sin \alpha)}{1 + \exp(-\kappa_i \pi)} \\ &= \frac{\exp(-\kappa_i \alpha) \sin(\alpha)(\kappa_i^2 + 1)}{1 + \exp(-\kappa_i \pi)} \end{aligned} \tag{8}$$

Then, the expected value of α can be obtained as

$$
\begin{aligned}
E[\alpha] &= \int_0^{\pi} \alpha \frac{\exp(-\kappa_i \alpha)\sin(\alpha)(\kappa_i^2+1)}{1+\exp(-\kappa_i \pi)} d\alpha \\
&= \frac{\kappa_i^2+1}{1+\exp(-\kappa_i \pi)} \int_0^{\pi} \alpha \exp(-\kappa_i \alpha)\sin\alpha d\alpha \\
&= \frac{\kappa_i^2+1}{1+\exp(-\kappa_i \pi)} \left[-\frac{\exp(-\kappa_i \alpha)((\kappa_i((\kappa_i^2+1)\alpha+\kappa_i)-1)\sin\alpha + ((\kappa_i^2+1)\alpha+2\kappa_i)\cos\alpha)}{(\kappa_i^2+1)^2} \right]_0^{\pi} \\
&= \frac{\kappa_i^2+1}{1+\exp(-\kappa_i \pi)} \left[\frac{2\kappa_i(1+\exp(-\kappa_i\pi)) + \exp(-\kappa_i\pi)(\kappa_i^2+1)\pi}{(\kappa_i^2+1)^2} \right] \\
&= \frac{2\kappa_i}{\kappa_i^2+1} + \frac{\exp(-\kappa_i\pi)\pi}{1+\exp(-\kappa_i\pi)},
\end{aligned}
\tag{9}
$$

which is Eq. 6 in the paper. This quantity is used as a measure of the aleatoric uncertainty. Fig. 1-(c) visualizes Eq. 7, Eq. 8 and Eq. 9 for different values of κ. The expected error decreases as κ increases. For $\kappa = 0$, the distribution is uniform and the expected error is $\pi/2$.

2. Quantitative evaluation with additional metrics

In this section, we provide the quantitative evaluation of our method with additional metrics. Tab. 1, Tab. 2 and Tab. 3 are extensions of Tab. 4, Tab. 5 and Tab. 6 in the paper, respectively. Fig. 2 and Fig. 3 are extensions of Fig. 8 in the paper.

2.1. Comparison against TiltedSN

Tab. 1 provides comparison against TiltedSN [2] on ScanNet [1] with additional metrics. Note that the difference in the accuracy (% of pixels with error less than t°) increases for lower thresholds.

Method	mean	median	rmse	5.0°	7.5°	11.25°	22.5°	30°
TiltedSN[2]	12.6	6.0	21.1	42.8	57.5	69.3	83.9	88.6
Ours	**11.8**	**5.7**	**20.0**	**45.1**	**59.6**	**71.1**	**85.4**	**89.8**
Difference	-0.8	-0.3	-1.1	+2.3	+2.1	+1.8	+1.5	+1.2

Table 1. Quantitative comparison against TiltedSN [2] on ScanNet [1].

2.2. Quality of the estimated uncertainty

Tab. 2 and Tab. 3 compare different methods of estimating the surface normal uncertainty. "*Drop*" (making 8 inferences with dropout enabled), "*Aug*" (making 2 inferences by flipping the image) and "*Drop+Aug*" (making 8×2 inferences by applying both) are task-independent approaches which does not require the output to be distributional. The proposed pipeline, trained with the NLL losses, significantly outperforms other approaches across all metrics, suggesting that the estimated uncertainty better correlates with the prediction error.

Method	AUSC ↓						AUSE ↓					
	mean	median	rmse	11.25°	22.5°	30.0°	mean	median	rmse	11.25°	22.5°	30.0°
Drop	9.01	4.91	15.84	19.32	8.66	6.07	4.02	0.91	9.61	10.23	6.10	4.76
Aug	8.64	4.68	15.08	18.75	8.26	5.64	3.93	0.97	9.14	10.25	5.84	4.42
Drop + Aug	8.16	4.68	14.32	16.73	7.18	4.97	3.22	0.73	8.15	7.75	4.65	3.68
Ours *(NLL-vonMF)*	7.03	4.47	10.96	14.24	5.51	3.53	**2.11**	**0.56**	**4.80**	5.10	2.92	2.24
Ours *(NLL-AngMF)*	**6.83**	**4.25**	**10.92**	**13.47**	**5.27**	**3.45**	2.13	**0.56**	4.98	**5.01**	**2.86**	**2.22**

Table 2. Quantitative evaluation of uncertainty on NYUv2 [7].

Method	AUSC ↓						AUSE ↓					
	mean	median	rmse	11.25°	22.5°	30.0°	mean	median	rmse	11.25°	22.5°	30.0°
Drop	7.25	4.35	12.51	13.95	5.49	3.60	3.24	1.02	7.55	8.58	4.14	2.94
Aug	7.06	4.08	12.58	13.72	5.36	3.48	3.32	1.03	7.92	8.81	4.13	2.87
Drop + Aug	6.87	4.17	12.07	12.73	4.82	3.13	2.93	0.92	7.20	7.49	3.51	2.49
Ours *(NLL-vonMF)*	5.84	3.92	9.30	10.31	3.21	1.94	**1.85**	**0.64**	**4.38**	4.69	**1.86**	1.30
Ours *(NLL-AngMF)*	**5.64**	**3.73**	**9.07**	**9.48**	**3.11**	**1.90**	1.88	0.66	**4.38**	**4.47**	1.88	**1.29**

Table 3. Quantitative evaluation of uncertainty on ScanNet [1].

2.3. Sparsification curves

Fig. 2 and Fig. 3 provide the sparsification curves for NYUv2 [7] and ScanNet [1], respectively. When evaluated on all pixels, all methods perform similarly. However, as the pixels with high uncertainty are removed, our method gets significantly more accurate than the others, suggesting that our uncertainty correlates better with the prediction error. For "Ours (*NLL-AngMF*)", we also show the ideal sparsification (oracle) by sorting the pixels by the error.

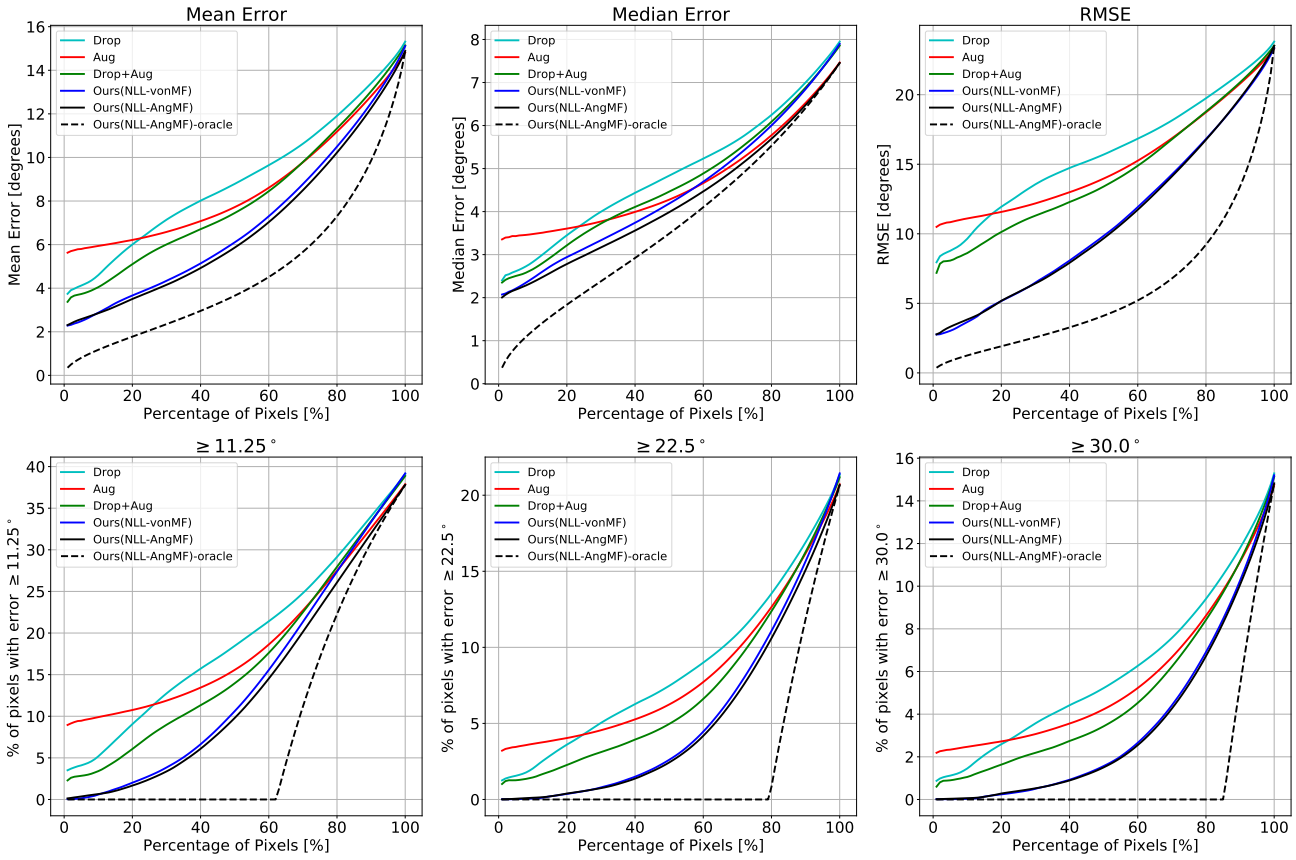

Figure 2. Sparsification curves for NYUv2 [7].

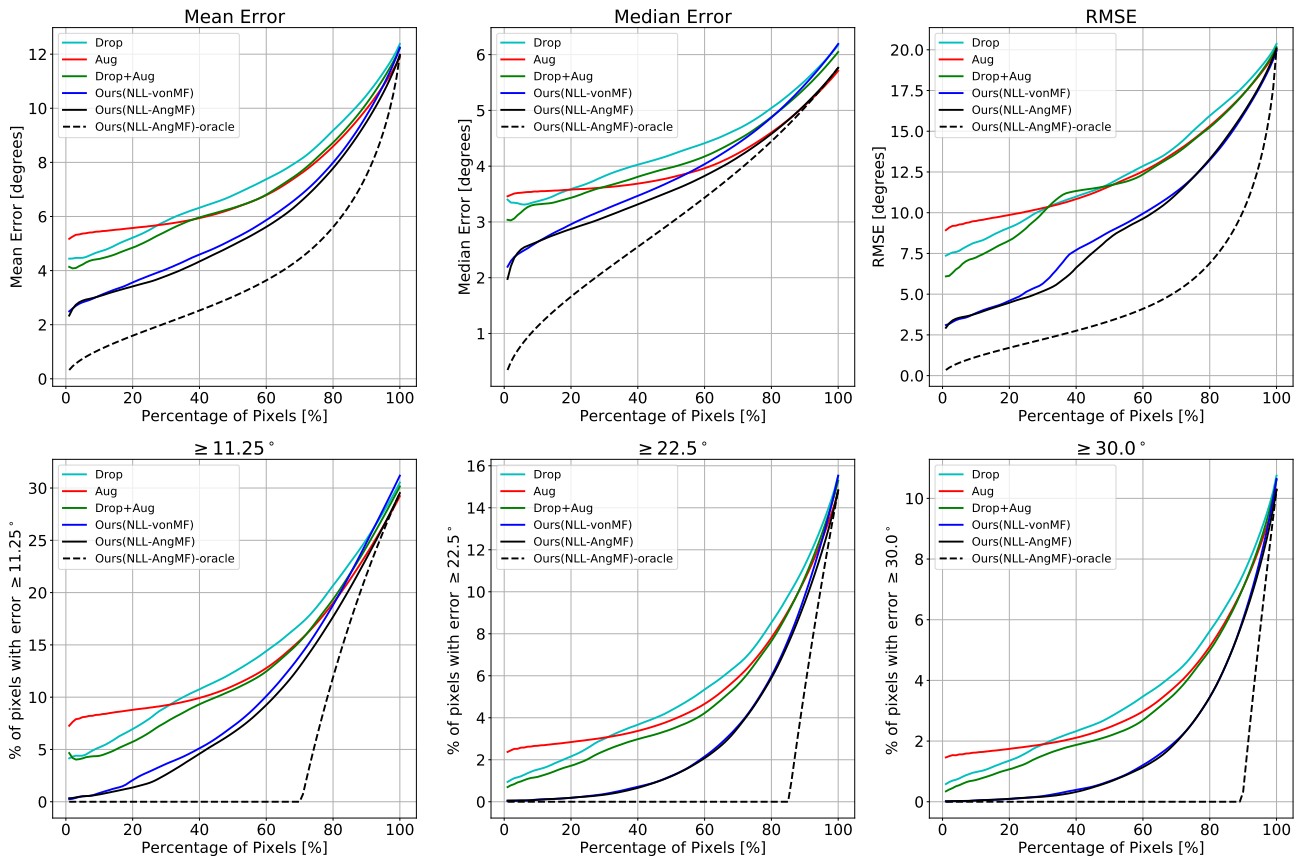

Figure 3. Sparsification curves for ScanNet [1].

3. Cross-dataset evaluation on KITTI and DAVIS

In the paper, we performed a cross-dataset evaluation by training the network on ScanNet [1] and testing it on NYUv2 [7] without fine-tuning. However, this is not a challenging task as both datasets contain images of indoor scenes with similar visual features. In this section, we further demonstrate the generalization ability of our method by testing the network (trained only on ScanNet) on two challenging datasets - KITTI [4] and DAVIS [5]. The results are provided in Fig. 4 and Fig. 5. For comparison, we also provide the predictions made by TiltedSN [2].

The ground truth surface normal for ScanNet is calculated from a 3D mesh that is obtained by fusing thousands of depth-maps. For this reason, the ground truth generally does not exist for dynamic objects such as humans. As the dataset is collected in indoor scenes, it also does not contain instances of cars and buildings. Nonetheless, the network can generalize well for such unseen objects. We believe that this is because the network utilizes low-level features, such as edges and shades, which are universal in most datasets. Fig. 7, which will be discussed in Sec. 4, supports such argument. Even when the input image only contains edges or shades, the network can infer the 3D structure.

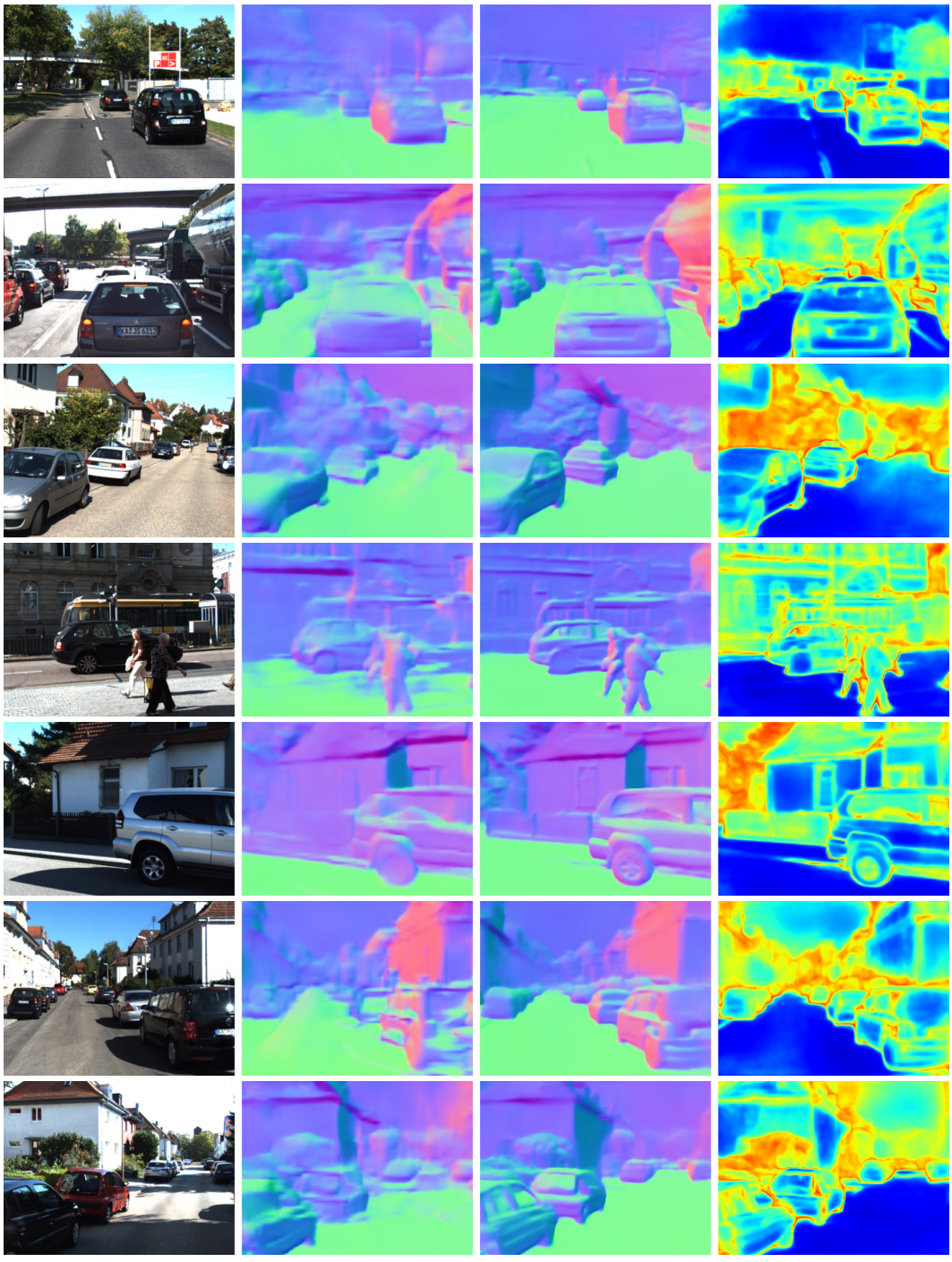

Figure 4. Cross-dataset evaluation on KITTI [4].

Figure 5. Cross-dataset evaluation on DAVIS [5].

4. Failure modes

In this section, we discuss the failure modes of the proposed method.

4.1. Tilted images

Fig. 6 shows the predictions made for tilted images. The network is robust against mild rotations ($\sim 20^\circ$), but suffers when the images are tilted severely ($30^\circ \sim$). Nevertheless, the expected error (clamped between 0° and 60° in all images) also increases for such images, demonstrating the usefulness of the estimated uncertainty. Tilted images can be handled by using a spatial rectifier to warp the images such that its surface normal distribution matches to that of the training images, as done in [2]. This will be investigated in our future work.

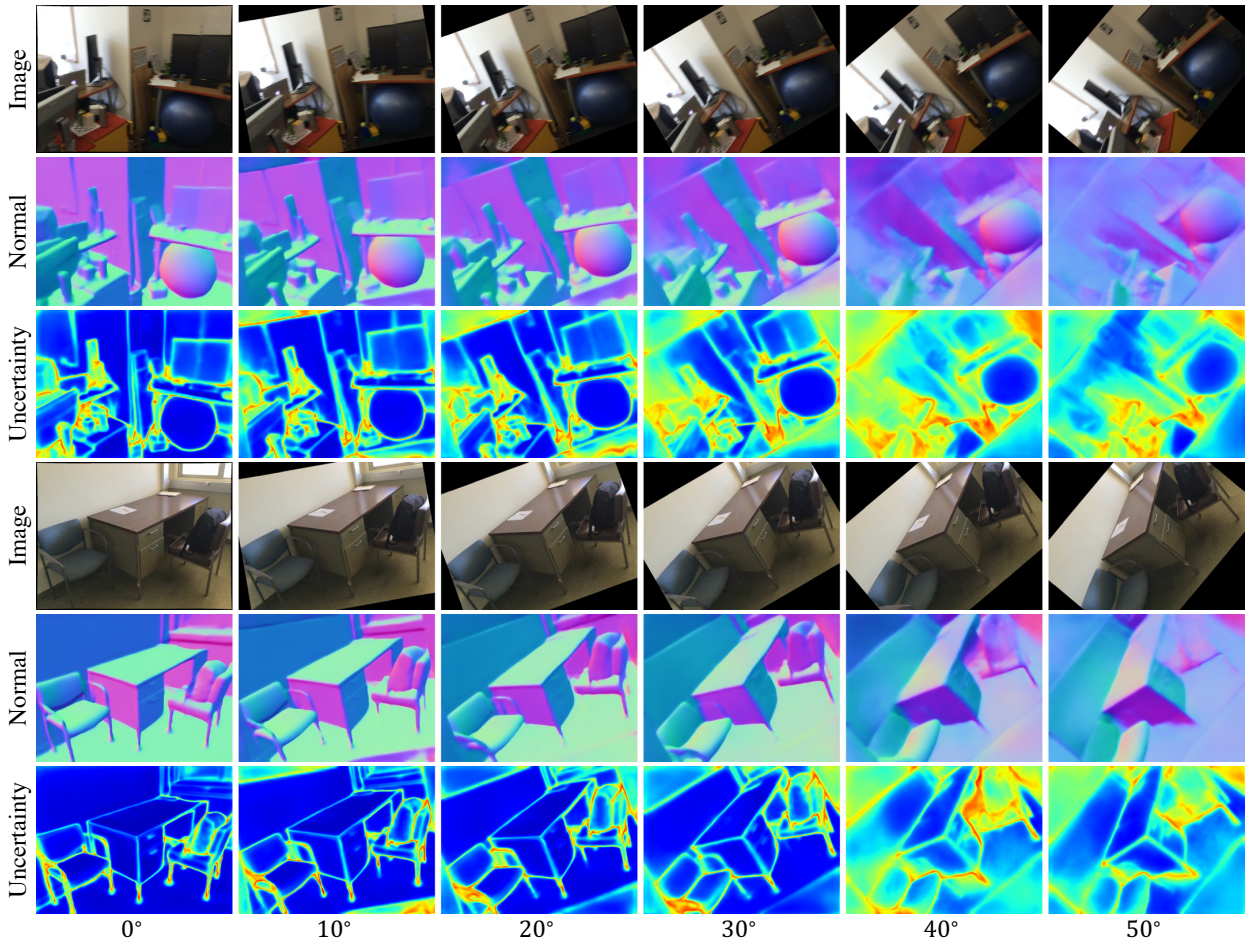

Figure 6. Predictions made for tilted images.

4.2. Inherent ambiguity of the problem

To understand the visual cues used by the network, we created artificial images consisting only of edges and shades. Fig. 7 shows the predictions made by the network. The first three images show "Y"-shaped structures and the other three are their vertically flipped versions. Note that the depth of each pixel can have any arbitrary value, meaning that the surface can have any form. It can be a concave (or convex) corner or even a drawing on a flat wall.

For the last three images, the network thinks that it is a concave corner. This is because such structure was mostly seen in the lower corners of cuboid-shaped rooms. However, the prediction is not clear for the "Y"-shaped structures. We believe that this is because such structure was seen in both concave corners (upper corners of rooms) and convex corners (external corners of furnitures). To handle such ambiguity, the network should estimate a *multi-modal* surface normal distribution, which consists of multiple uni-modal distributions with mixing coefficients. This will be investigated in our future work.

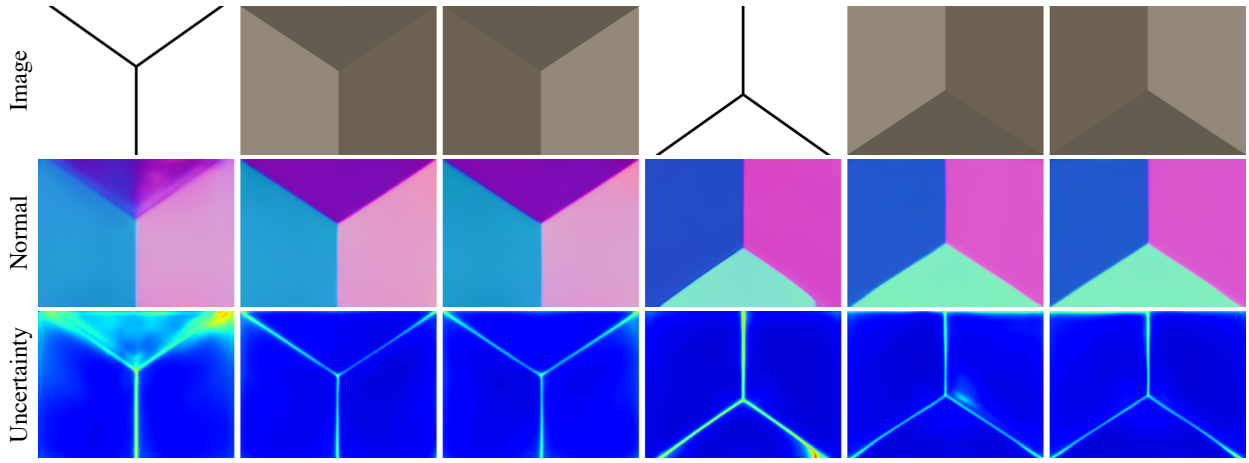

Figure 7. Predictions made for artificial images consisting only of edges and shades.

5. Additional comparison against the state-of-the-art

Lastly, we provide additional qualitative comparison against GeoNet++ [6] (in Fig. 8) and TiltedSN [2] (in Fig. 9).

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 3, 4, 5, 11

[2] Tien Do, Khiem Vuong, Stergios I Roumeliotis, and Hyun Soo Park. Surface normal estimation of tilted images via spatial rectifier. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 265–280. Springer, 2020. 3, 5, 8, 9, 11

[3] Nicholas I Fisher, Toby Lewis, and Brian JJ Embleton. *Statistical analysis of spherical data*. Cambridge university press, 1993. 1

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013. 1, 5, 6

[5] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016. 1, 5, 7

[6] Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip HS Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 9, 10

[7] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 3, 4, 5, 10

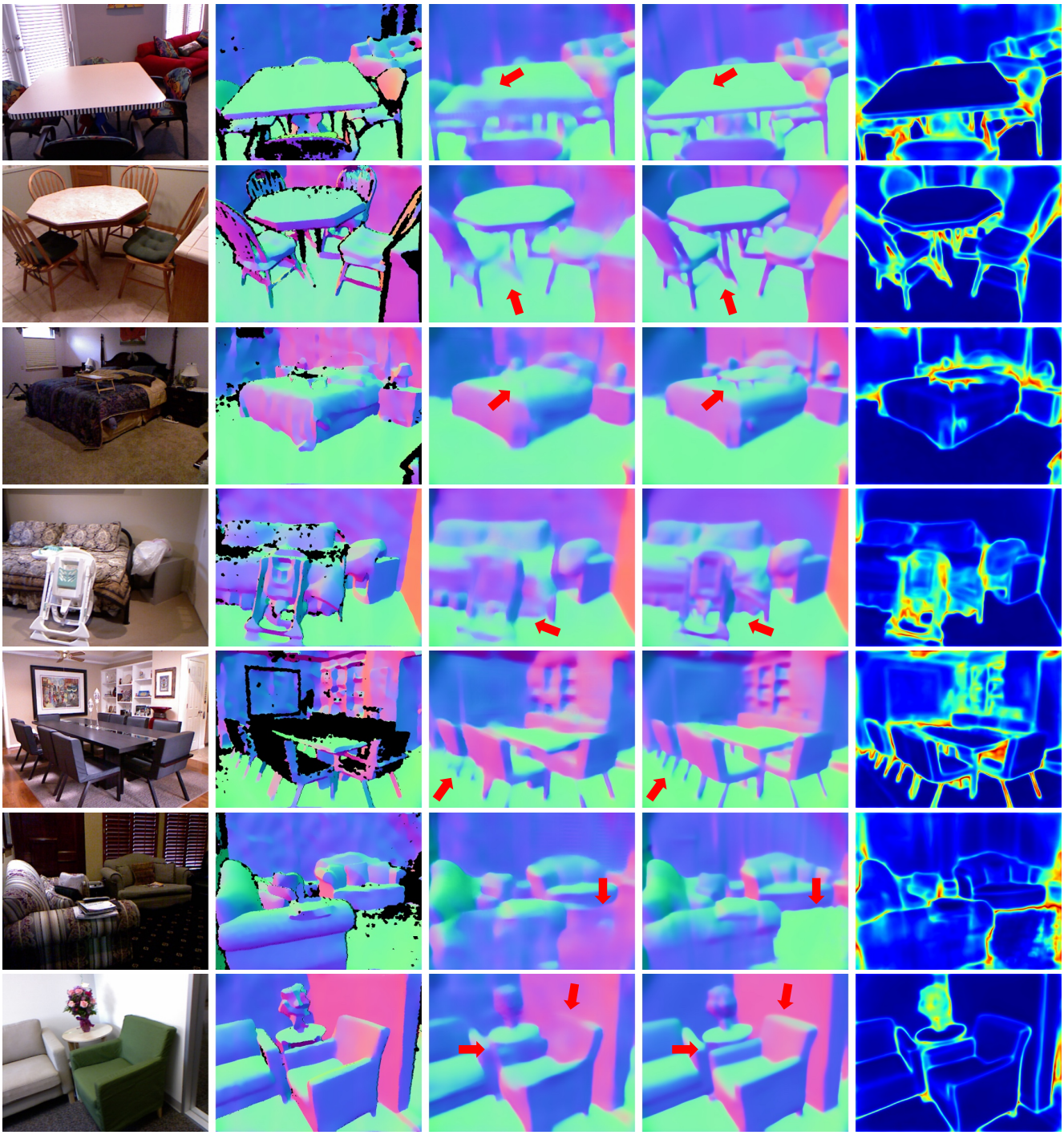

Figure 8. Additional qualitative comparison against GeoNet++ [6] on NYUv2 [7]. Despite the poor quality of the ground truth, our method can recover the fine details of the scene geometry (see the areas pointed by the red arrows).

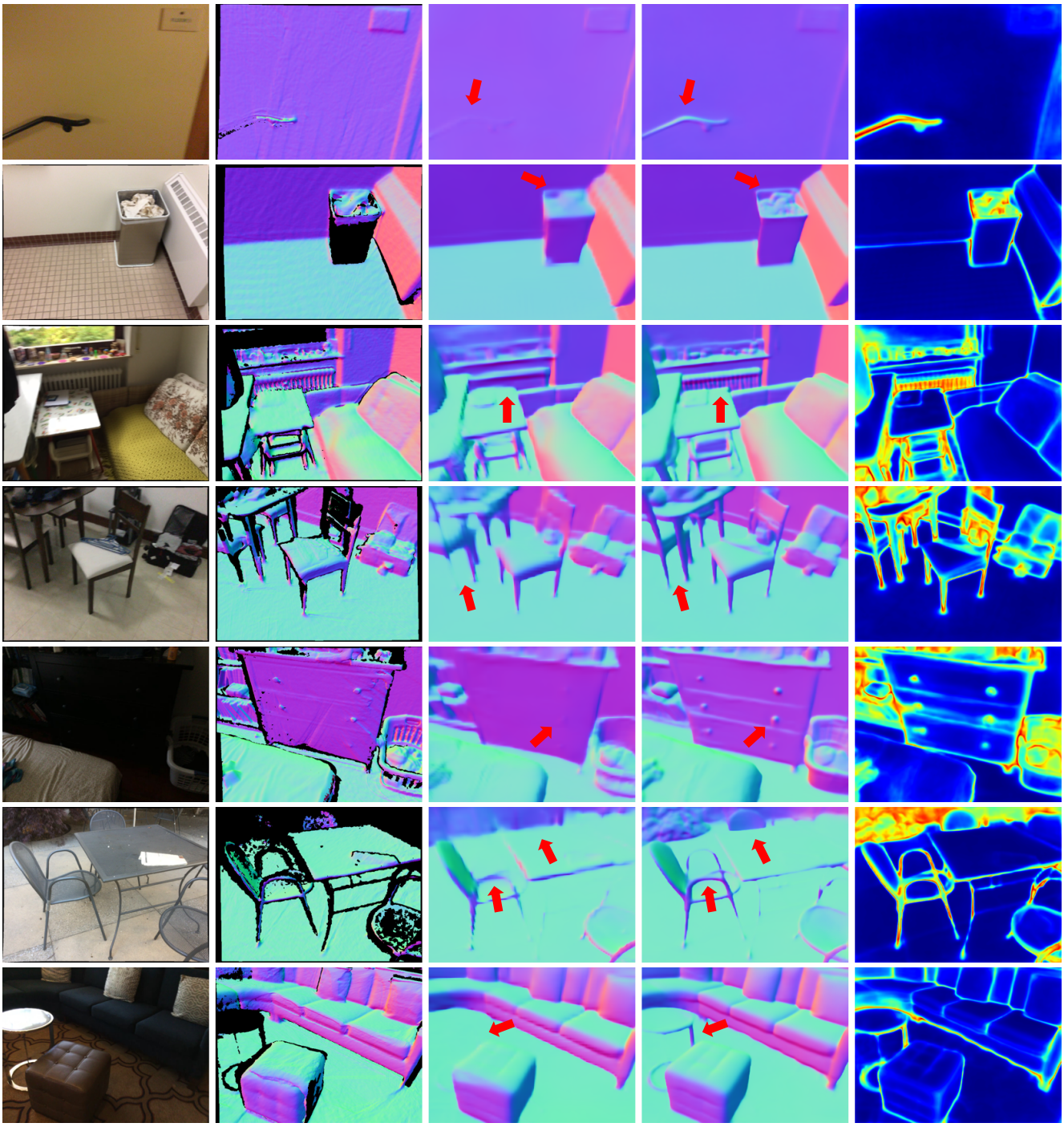

Figure 9. Additional qualitative comparison against TiltedSN [2] on ScanNet [1]. The predictions made by our method contain higher level of detail (see the areas pointed by the red arrows).