

ENTROPY-BASED CONTEXTUAL FUSION FOR OBJECT DETECTION IN ADVERSE DRIVING CONDITIONS

Zhengkao Li and Markus Flierl

Information Science and Engineering
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology, Stockholm, Sweden

1. SUMMARY

This work extends context-based multi-sensor fusion (contextual fusion) for 3D object detection in adverse operating conditions [1]. Contextual fusion combines LiDAR point cloud and RGB image information while considering domain knowledge about the sensors behaving differently across lighting and weather variations. However, contextual fusion is limited to channel-level gating and cannot consider spatially local variations that are of particular relevance in night and rainy conditions. Therefore, we propose a fine-grained entropy-enhanced contextual fusion network based on [2].

In our work, contextual fusion is achieved by residual gating. We define an environment descriptor and determine the channel-wise bias term $\Delta g \in \mathbb{R}^{B \times C}$ for C feature channels and batch size B by a fully connected layer. For efficient convergence during training, we work with the fusion weight $g = 1 + \alpha \tanh(\Delta g)$, where α is a tuneable hyperparameter.

In order to compute local entropy efficiently, the quantized grayscale values $I_q(b, h, w)$ of batch b at pixel position (h, w) are transformed into a sparse representation $I_{oh} \in \mathbb{R}^{B \times V \times H \times W}$ through one-hot encoding with V values and image size (H, W) .

$$I_{oh}(b, v, h, w) = \begin{cases} 1, & \text{if } I_q(b, h, w) = v \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Here, v denotes a quantized grayscale value. We then calculate the local grayscale histograms for each pixel neighborhood and normalize these to obtain the probability

$$p(v|b, h, w) = \frac{1}{l^2} \sum_{(i,j) \in \mathcal{N}(h,w)} I_{oh}(b, v, i, j), \quad (2)$$

where $\mathcal{N}(h, w)$ denotes a local neighborhood with side length l , centered at pixel (h, w) . The resulting entropy map $E \in \mathbb{R}^{B \times H \times W}$ is utilized for subsequent adaptive feature weighting. For the implementation, the entropy map is further downsampled and projected onto the corresponding N LiDAR points to match the feature dimension C , yielding the point-level entropy features $E_p \in \mathbb{R}^{B \times N \times C}$. Then, a feature weight mask $W_m \in \mathbb{R}^{B \times N \times C}$ is generated by applying the sigmoid activation to each element. This mask is now used for element-wise weighting of the fused features.

For the experiments, we use a balanced sample of the nuScenes dataset [3], where day, rain and night scenes are equally likely. We

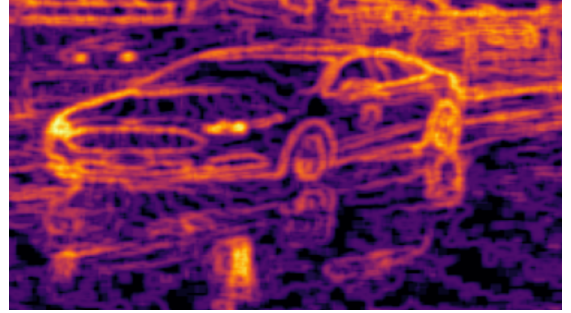


Fig. 1. Entropy-enhanced contour visualization in rainy conditions.

also consider levels of difficulty (easy, moderate, hard) according to size of bounding box and occlusion level. All presented average precision results are determined by using the AP_{R40} protocol [4].

According to **Table 1**, entropy-based contextual fusion can improve the average precision (AP) for hard rainy conditions from 34.9 to 38.2 (relative increase of almost 10%). The entropy enhancement gives a relative increase of about 3%. This is similar for hard night conditions. For day conditions, the entropy advantage is smaller, as the quality of the image data is more spatially uniform.

2. REFERENCES

- [1] S. Sural, N. Sahu, and R.R. Rajkumar, “ContextualFusion: Context-based multi-sensor fusion for 3D object detection in adverse operating conditions,” in *IEEE Intelligent Vehicles Symposium (IV)*, Jeju Island, South Korea, June 2024, pp. 1534–1541.
- [2] T. Huang, Z. Liu, X. Chen, and X. Bai, “EPNet: Enhancing point features with image semantics for 3D object detection,” in *European Conference on Computer Vision (ECCV)*, Glasgow, UK, Aug. 2020, pp. 35–52.
- [3] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, June 2020, pp. 11618–11628.
- [4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *The International Journal of Robotics Research*, vol. 32, pp. 1231–1237, Sept. 2013.

Model	Total			Day			Rain			Night		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Baseline [2]	59.05	41.08	39.02	63.45	45.22	41.17	56.06	35.90	34.94	60.00	51.32	50.00
Residual gating	60.50	43.60	41.53	63.43	47.13	43.08	59.57	38.69	37.32	61.87	55.14	52.18
Residual gating + entropy	61.25	44.47	42.33	64.65	47.16	43.00	59.63	39.68	38.20	62.52	56.82	53.77

Table 1. Average precision for 3D detection on nuScenes-balanced (AP@IoU=0.7, $\alpha = 0.2$).