基于信息论的智能驾驶可解释多模态感知

张新钰[1,2,3*]，国纪龙[1,3]，李骏[1,2,3]，李德毅[4]，张世焱[1,3]，沈思甜[1,3]，吴凡[1,3]，刘华平[4]

1. 清华大学智能绿色车辆与交通全国重点实验室, 北京 100084
2. 北京航空航天大学交通学院, 北京 100191
3. 清华大学车辆与运载学院, 北京 100084
4. 清华大学计算机科学与技术系, 北京 100084
* 通信作者. E-mail: xyzhang@tsinghua.edu.cn

国家重点研发计划（2018YFE0204300）和国家自然科学基金（62273198、U1964203）

Background

(a) 2016年特斯拉未能识别道路清扫车

(b) 2018年Uber未能准确识别行人

(c) 2021年特斯拉再次撞上白色货车

(d) 2021年蔚来汽车未能识别静止车辆

目标检测错误导致的智能驾驶车辆碰撞事故

智能驾驶感知安全问题主要集中于感知任务
感知任务目标：更高的精度、更快的速度
现有方案：

1）单模态感知器， 但局限于传感器性能，漏检误检概率高。

2）多模态融合， 但多数使用结果级融合，容难以克服检测结果中目标数量或类别差异下的有效匹配，导致漏检误检。另外大量使用深度学习，解释性差。

Contributions

- 基于特征融合的多模态融合，利用多头注意力融合模块融合不同模态之间的特征信息，避免了结果级融合存在的目标漏检问题。由于多模态特征之间的相互补充与矫正，使模型即使面临车辆遮挡以及光线骤变等特殊场景，依然能够保证感知安全与准确

- 基于信源信道联合编码 (Joint Source Encoding and Channel Coding)的理论对感知模型中的特征提取和特征融合理论进行解释, 在增强复杂场景下感知能力的同时保证模型的可解释性

- 构建新的评价指标平均熵变稳定性（Average Entropy Variation, AEV）对模型与外界的感知交互过程进行实时评估

Method

- 基于联合编码的可解释性多模态感知模型设计
- 基于联合编码的多模态特征深度融合网络
- 基于信源信道联合编码的多模态特征融合

1) 基于联合编码的可解释性多模态感知模型设计

Assumptions:

1) 多模态融合网络的信息熵建模：熵值变化的稳定程度作为网络稳定性的量化度量。

2) 熵值越低，信息量越大。而在一个可信度较高的通信模型中，信息压缩网络各层信息变化是稳定的，保证了信息压缩的平滑性，在信息穿过网络各层时，保持熵变不变，可以防止信息的突然失真。反映出模型进行特征提取过程的稳定程度，为感知模型的可解释性提供了量化度量

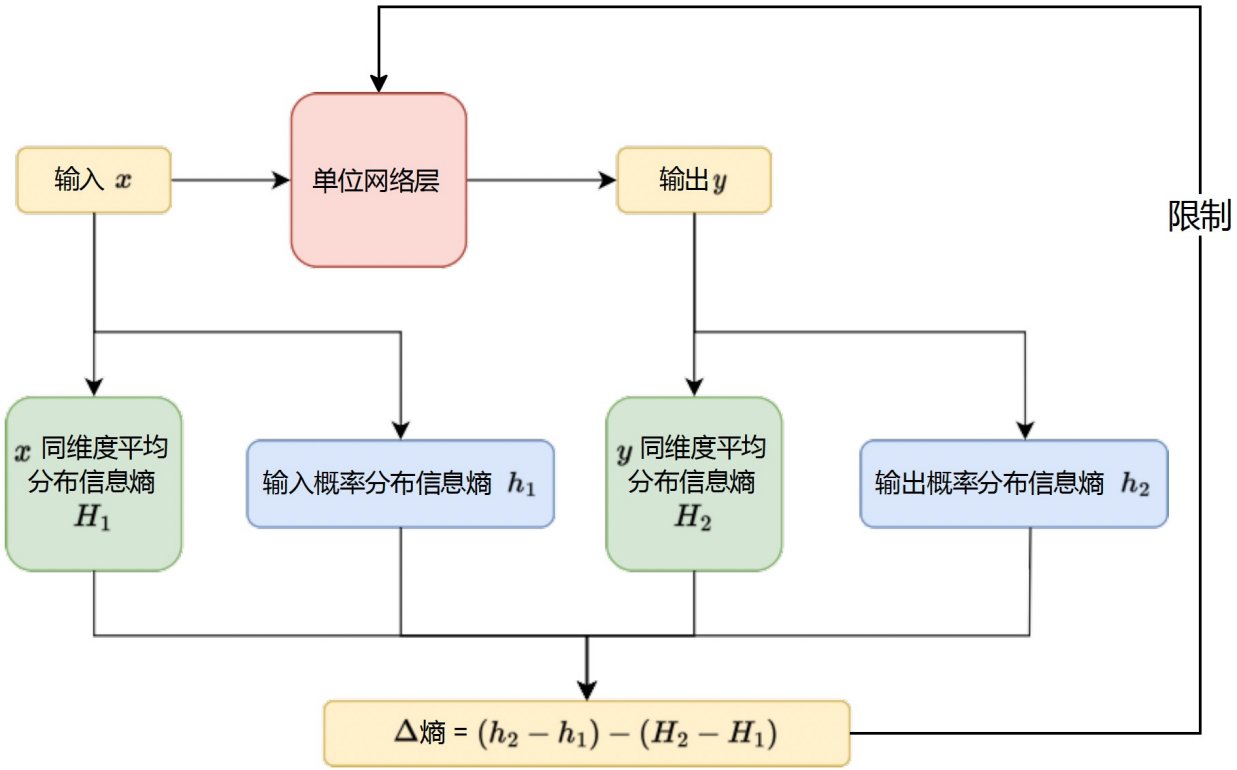


图 2 熵变指标

Entropy change indicator:

Difference between input and output features

1) 基于联合编码的可解释性多模态感知模型设计

Implementation:

- **微分熵**: f (x) 为 X 的概率密度函数, 任何网络的每一层输出都可以被认为是连续型随机变量 X，而任何层的输出都可以被选为 X 的一组样本 x_i

$$h(X) = -\int f(x) \log f(x) dx$$

Differential Entropy (Continuous Entropy): a measure of average of a random variable, to continuous probability distributions.

- **概率密度函数:** 概率分布中只有有限数量的样本值可用时，可以使用KNN来估计PDF.

$$p(x_i) = [(n-1) \cdot r_d(x_i)^d] \cdot V_d]^{-1}$$

n 为样本个数，$r_d(x_i)$ 为样本 x_i 与其最近样本点之间的 d 维欧氏距离，V_d 为 d 维空间中单位球面的体积

- 随机变量 X 的熵的估计值为(Samples are discrete):

$$H(X) = \frac{1}{n} \sum_{i=1}^{n} [-\log p(x_i)] + \gamma$$

$\gamma \approx 0.5772$, Euler Constant

- KNN熵估计方法：与其最近的第 k 个样本点之间的距离

$$H(X,k) = -\psi(k) + \psi(n) + \log V_d + \frac{d}{n} \sum_{i=1}^{n} \log r_{d,k}(x_i)$$

- 平均信息熵变AEV Average Entropy Variation 定义

$$AEV(t) = \frac{\sum_{n=1}^{N} (\Delta H_n(t) - \widehat{\Delta H}(t))^2}{N}$$

1) 基于联合编码的可解释性多模态感知模型设计

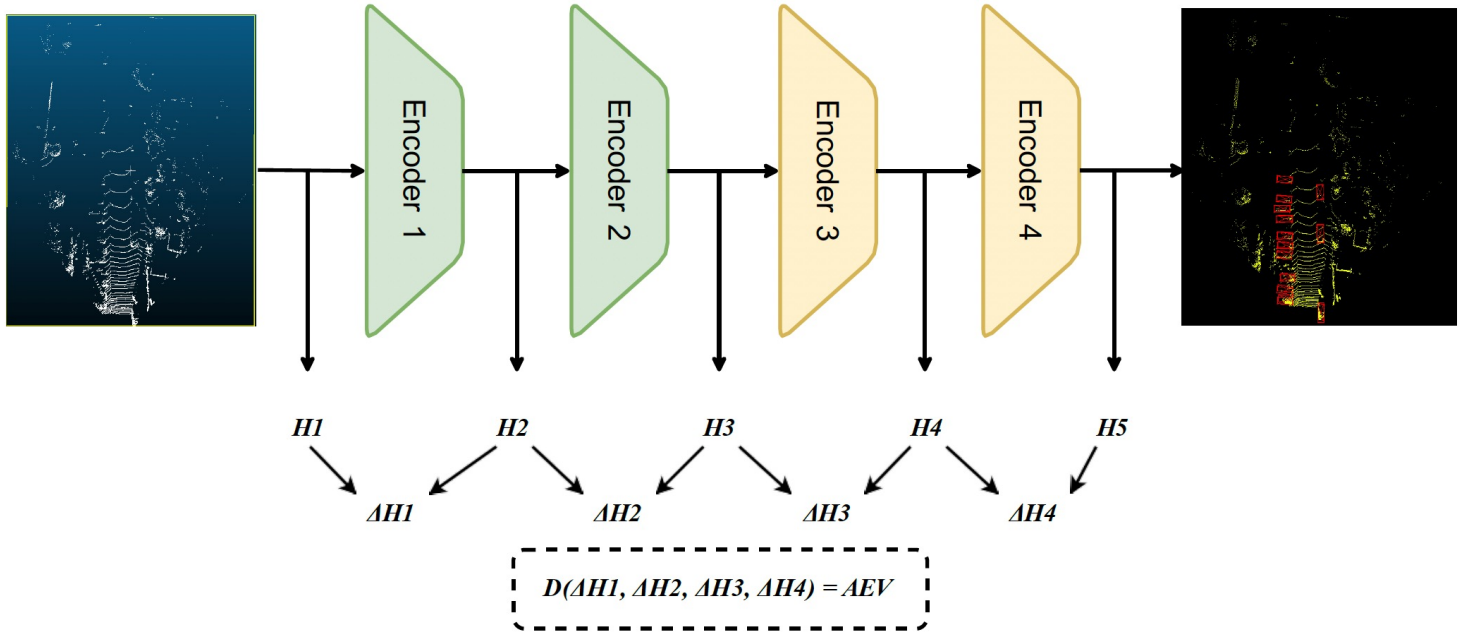


图 3　提出的评价指标 AEV 的详细说明

Figure 3　An detailed Illustration of our proposed AEV method

- H(t) 网络各层实时输出的熵值
- 网络各层的熵变量$\Delta H_n(t)$，$\Delta H_n(t) = H_{n+1}(t) - H_n(t)$，n denotes index.
- 当 AEV (t) 的数值越低时，模型的感知过程就越稳定

(2) 基于联合编码的多模态特征深度融合网络

Source Coding with **Autoencoders**:
信源符号由最短的码字表示，各码元所载荷的平均信息量最大，能保证无失真地恢复原来的符号序列

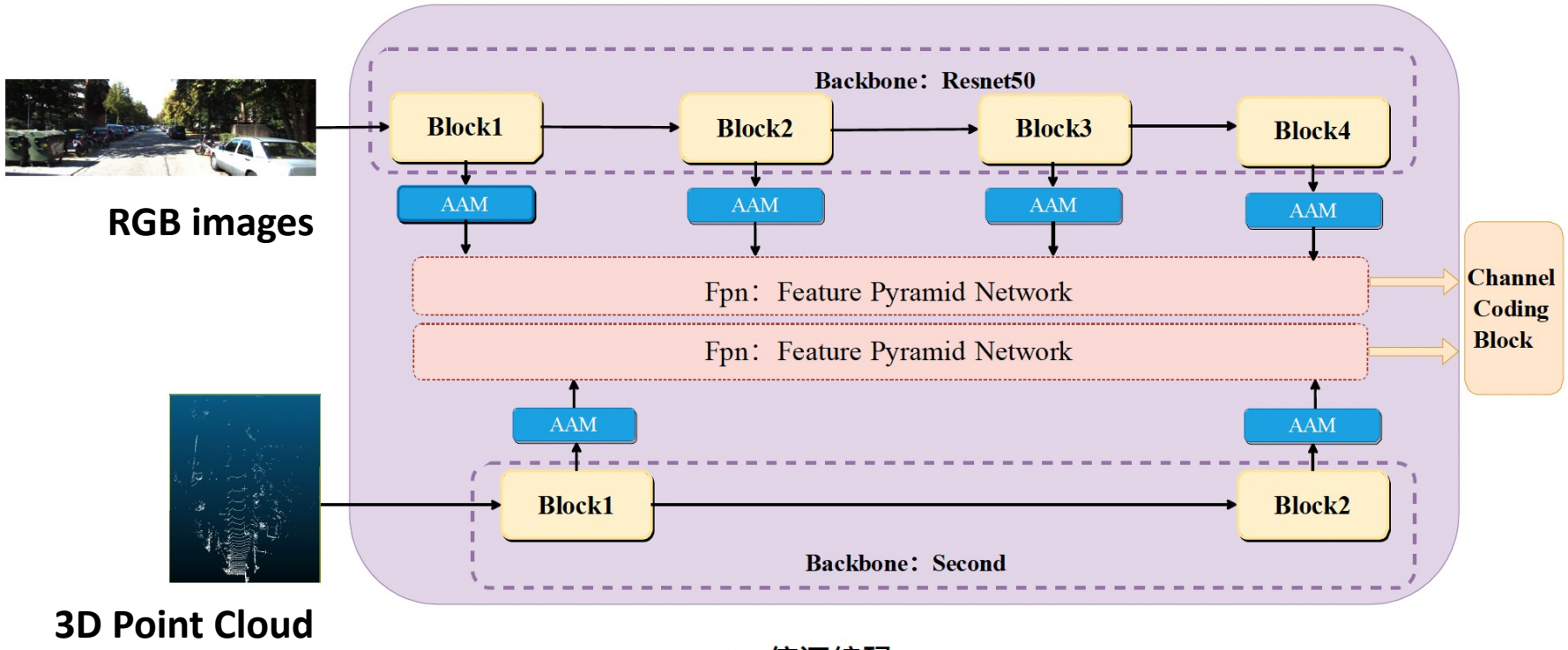


图 5　信源编码
Figure 5　Source Coding

Axial Attention Model: Ho, Jonathan, et al. "Axial attention in multidimensional transformers." *arXiv preprint* (2019).

Yan, Yan, Yuxing Mao, and Bo Li. "**Second: Sparsely embedded convolutional detection**." Sensors (2018).

(2) 基于联合编码的多模态特征深度融合网络

Axial Attention Model: Ho, Jonathan, et al. "Axial attention in multidimensional transformers." *arXiv preprint* (2019).

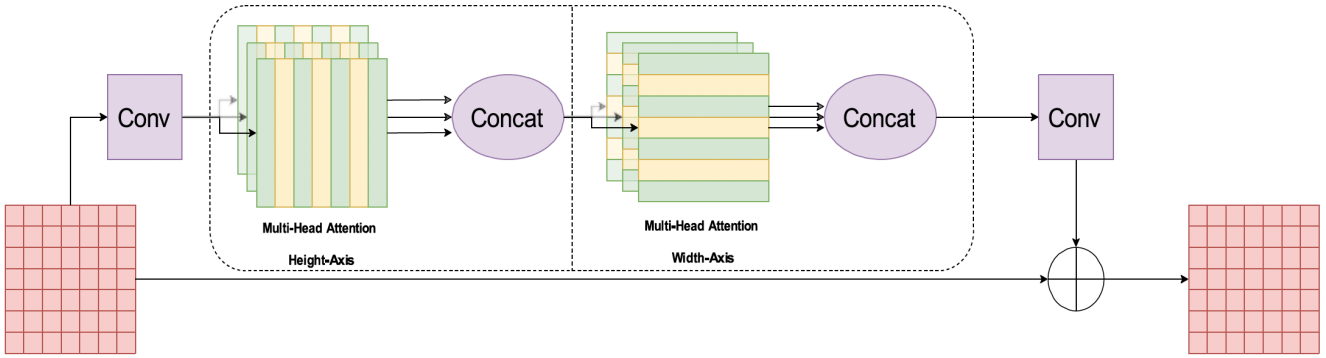


图 7 轴向注意力机制

Figure 7 Axial attention mechanism

Yan, Yan, Yuxing Mao, and Bo Li. "**Second: Sparsely embedded convolutional detection**." Sensors (2018).

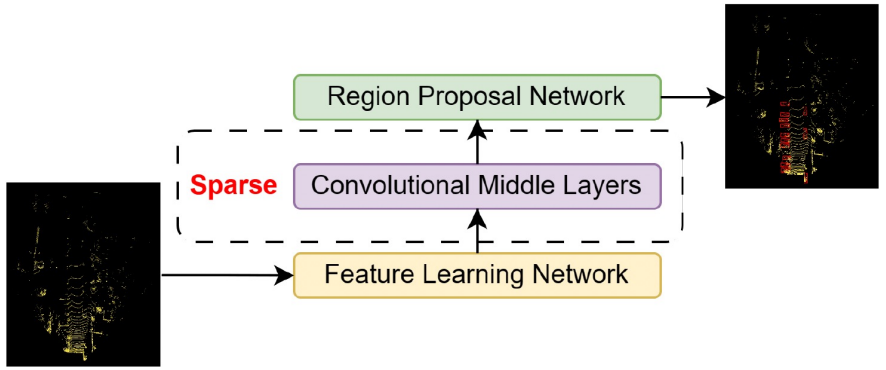


图 6 基于 VoxelNet 改进的 SECOND 网络框架

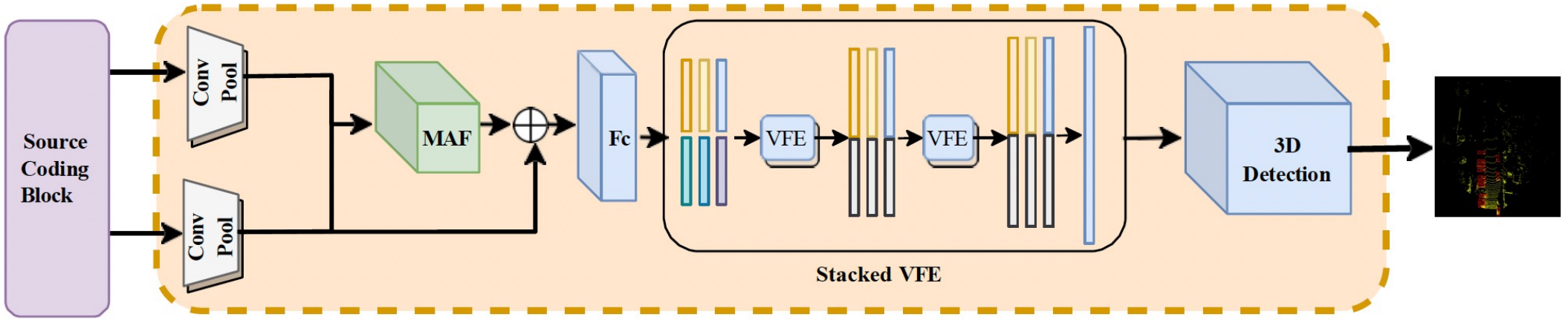
Figure 6 Based on VoxelNet's improved SECOND network framework

Improve computational efficiency on sparse voxels.

(3)基于信源信道联合编码的多模态特征融合

Joint Source-Channel Coding

Stacked **V**oxel
Feature **E**ncoding



Feature maps from pyramid networks

Multi-Head
Attention **F**usion

图 8 信道编码

Figure 8 Channel Coding

Evaluation

- MMDetection3D Framework, PyTorch
- Result analysis

  (1) AEV 评价指标的可靠性分析 (Validation of AEV)

  (2) 感知模型的结构分析 (Ablation study on each block)

  (3) 感知模型的性能比较 (Comparison between baseline models)

(1) AEV 评价指标的可靠性分析

- 添加噪声后的置信度低于不添加噪声时的置信度, 可以识别出异常感知过程
- 当噪声添加率不同时，结果的置信度变化很小, 不能判断噪声添加率

表 1 AEV 有效性

Table1 Effectiveness of AEV

Setting		Value	AEV			Confidence		
Model	Dataset		Clean	noise1	noise2	Clean	noise1	noise2
VoxelNet	KITTI	Mean	0.015	0.008	0.009	0.495	0.248	0.248
		Change	0.0%	-48.5%	-39.1%	0.0%	-49.9%	-49.9%
PointPillars	KITTI	Mean	0.012	2.086	0.008	0.487	0.344	0.344
		Change	0.0%	17475.6%	-36.1%	0.0%	-29.3%	-29.3%
PointPillars	nuSenes	Mean	0.034	1.918	0.016	0.168	0.128	0.128
		Change	0.0%	5494.7%	-54.5%	0.0%	-23.7%	-23.7%

(2) 感知模型的结构分析

表 2　消融实验

Table2 Ablation Study

model	AAM	FPN	MAF	mAP		AEV		
				bbox	bev	AEV-channel	AEV-source	AEV-joint
without-AAM	×	√	√	93.3073	90.0549	0.628824	0.131481	0.760305
without-FPN	√	×	√	92.8058	87.9710	0.710557	0.007787	0.718344
without-MAF	√	√	×	93.0922	89.1565	0.630544	0.003330	0.633874
ours	√	√	√	95.0187	90.7390	0.587259	0.003189	0.590448

- AAM: 在数据压缩过程能够有效的朝着冗余信息减少的方向进行，保证了数据压缩过程数据信息的准确全面。
- FPN: 通过对不同层次数据信息的拼接，避免了模型在数据压缩过程中的失真，保证了模型的感知精度以及感知过程的稳定性。
- MAF: 主模态的语义信息进行了检错与矫正，通过冗余信息与互补信息的加入，保证了主模态语义信息的在模型感知过程中的准确与稳定。

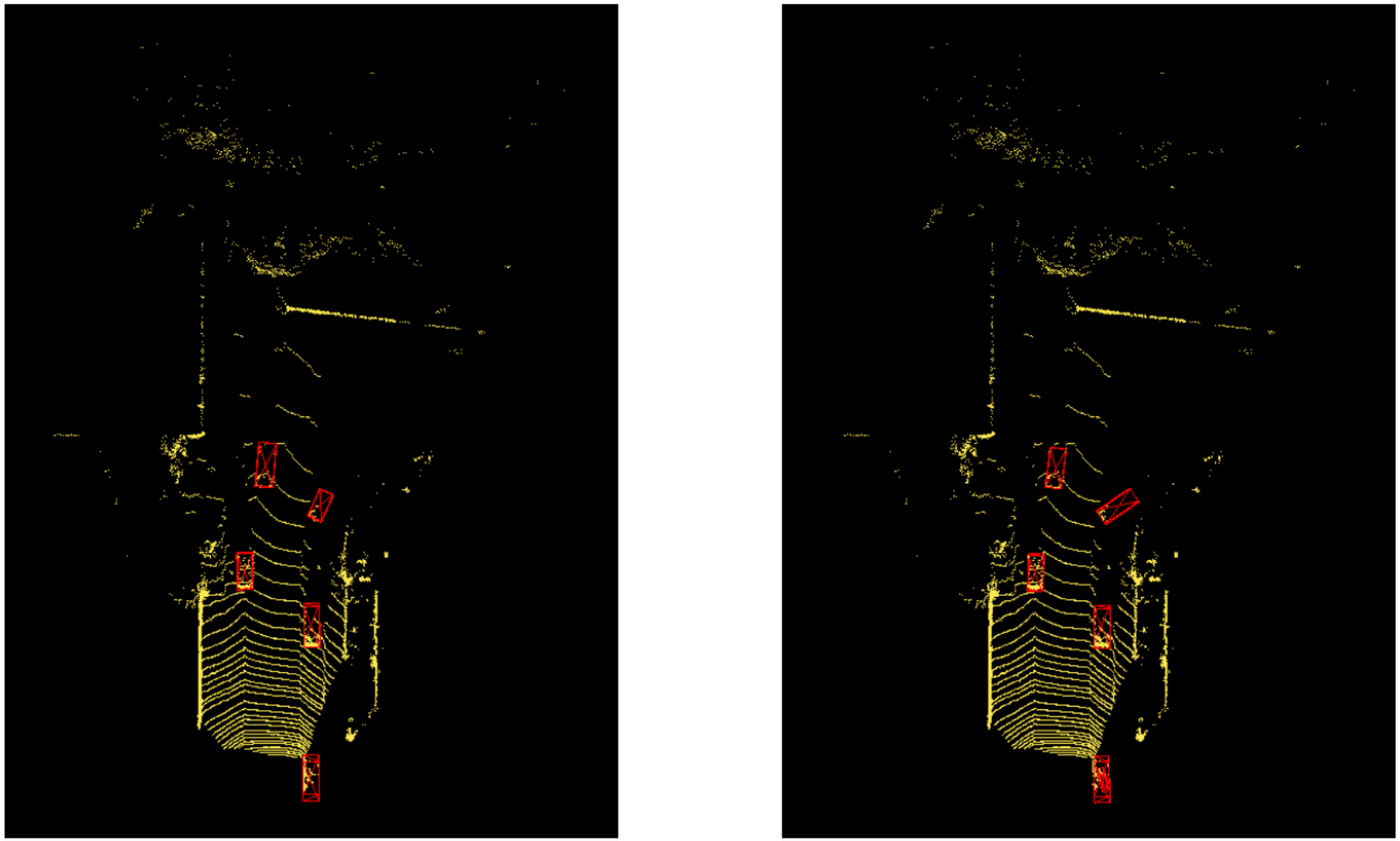
(3) 感知模型的性能比较

表 3　不同模型性能比较

Table3 Comparison of performance of different models

Result	mAP		AEV		
	bbox	bev	AEV-channel	AEV-lidar	AEV-joint
Smoke	90.8558	17.6157	\	0.34768	0.3476
Second	95.2876	90.0579	\	5.515044	5.5150
Mvxnet	96.2173	92.293	2.203327	1.184249	3.3875
EPNet	97.7698	94.364	1.061605	0.00275	1.0643
Sourse coding	93.0922	89.1565	0.63054	0.00333	0.6338
Channel coding	93.4052	91.6867	0.75995	0.09278	0.8527
ours	95.0187	90.739	0.587259	0.003189	0.5904

Example



(a) ground truth

(b) predict

图 10 感知结果

Figure 10 Perceptual outcome