Fusion-based 3D Object Detection Methods: Literature Review

Wei Hai Cui

Institute for Aerospace Studies
UNIVERSITY OF TORONTO

TORONTO ROBOTICS + AI LAB

Overview

- Lidar offers depth information, but sparse
- Images offer dense color and texture, but lack depth information
- Fusion theoretically offers best of both modalities

- Majority of fusion-based methods underperform compared to Lidar-only
  - Attributed to difficulty in training with different modalities
    - More parameters
    - Higher chance of overfitting
    - Different backbones with different learning rates
    - **Lack of Data Augmentation**

Papers Reviewed:

- MVX-Net (ICRA 2019)
- 3D-CVF (ECCV, 2020)
- Dense Voxel Fusion (CVPR 2022)
- Sparse Fuse Dense (CVPR 2022)

MVX-Net (ICRA 2019)

Vishwanath A. Sindagi, Yin Zhou and Oncel Tuzel
John Hopkins University

- Builds upon VoxelNet by incorporating high-level features from 2D Detection Networks such as Faster R-CNN
- Proposes Two Single-Stage Methods :
  - **Point Fusion,** where 3D points are aggregated by an image feature to create a dense context before the Voxel Fusion Encoding Layers (VFE)
  - **Voxel Fusion,** a later fusion method where image features are appended on a voxel level.

MVX-Net (ICRA 2019)

Point Fusion

- **Extracts high-level features** from pre-trained **2D Detection Network**
- **Projects 3D points** onto **image features** using calibration matrix
- Appends **original points** with corresponding **combined features** for the **VFEs**

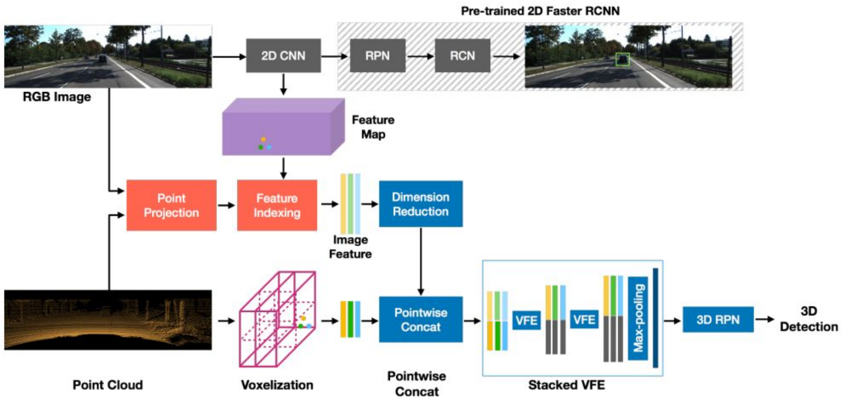


Fig 1. Point Fusion Architecture

MVX-Net (ICRA 2019)

Voxel Fusion

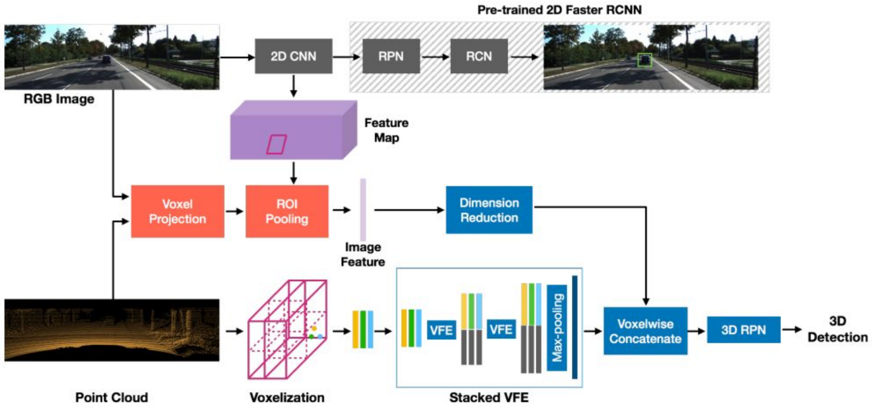


Fig 2. Voxel Fusion Architecture

- **Non-empty voxels** are **projected** onto i**mage plane** to be combined with the 2D RoI, which is then fused after the Stacked VFEs

MVX-Net (ICRA 2019)

Summary

- Voxel Fusion:
  - **Early Fusion** allows VFEs to obtain image information
  - *Pointwise Concatenation of Features to Voxels results in **loss** of **dense image information***
- Point Fusion:
  - **Lower resource usage** compared to Voxel Fusion
  - Late Fusion potentially results in **better performance** with **low resolution lidar** clouds
    - Image information can be overlaid onto empty voxels
  - ***Lower overall performance** compared to Voxel Fusion on KITTI*
- ***Lack of synchronized data augmentation***
- ***Outdated Performance compared to current state of the art***

Institute for Aerospace Studies
UNIVERSITY OF TORONTO

TORONTO ROBOTICS + AI LAB

3D-CVF (ECCV, 2020)

Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi
Hanyang University

- Two-Stage Detector:
    1. **Calibrated** multi-camera and Lidar features are joined using the **Gated Camera-Lidar Feature Fusion**
    2. **RoI-based feature pooling** is used for further improvements through fusion.

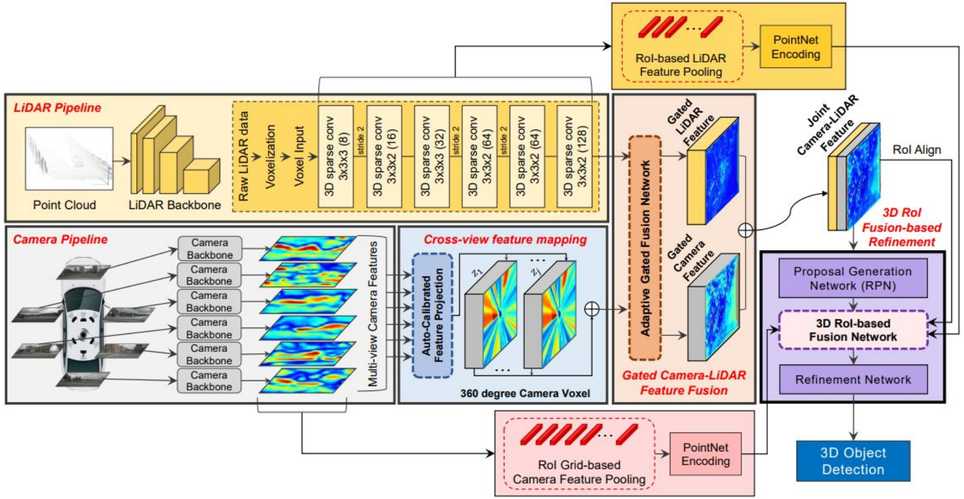


Fig 3. 3D-CVF Architecture

3D-CVF (ECCV, 2020)

Cross View Feature Mapping

- Camera voxels are generated to be **2x larger in the x & y** dimensions compared to Lidar voxels to allow for **spatially dense features**
- Camera features are **transformed into BEV** and assigned to Camera Voxel
- Lidar coordinates are then mapped to the Camera coordinates with a **calibration offset**
- **Neighbouring four feature pixels** are then assigned to **corresponding voxel**

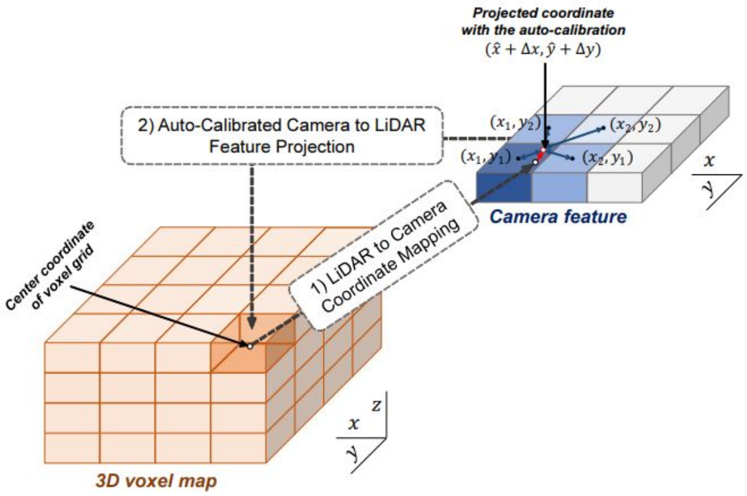


Fig 4. Auto-Calibrated Projection illustration

Institute for Aerospace Studies
UNIVERSITY OF TORONTO

TORONTO ROBOTICS + AI LAB

3D-CVF (ECCV, 2020)

Gated Camera-Lidar Feature Fusion

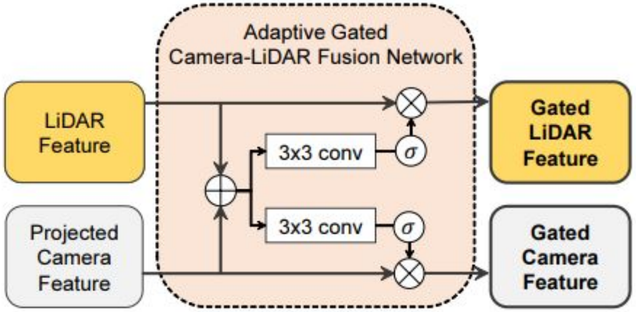


Fig 5. Gated Fusion Structure

- Adaptive Gated Fusion Network **selectively combines feature maps** depending on **relevance to detection task**

3D-CVF (ECCV, 2020)

3D RoI Fusion-Based Refinement

- Individual **Low-Level Lidar & Camera** features **are pooled** using 3D RoI-based pooling, and **combined** with **Camera-Lidar Features** from **RPN**
    - Low-level features retain **detailed spatial information** to **refine** region proposals
- RoI Grid-Based Pooling **projects points** from 3D RoI box to **Camera-View** Domain
- **Camera Features** associated with those points are **encoded** by **PointNet**

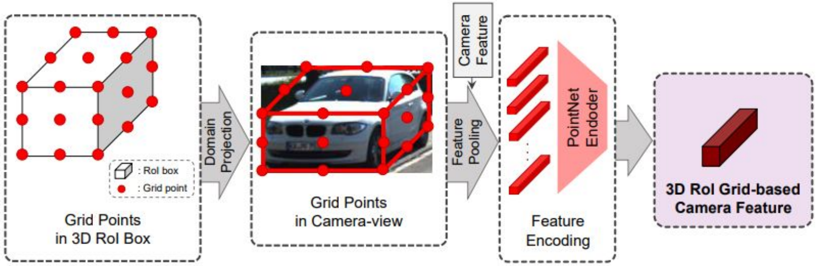


Fig 6. RoI grid- based pooling of camera features

Institute for Aerospace Studies
UNIVERSITY OF TORONTO

TRAIL
TORONTO ROBOTICS + AI LAB

3D-CVF (ECCV, 2020)

Summary

- Uses **Multi-View Images** & **Cross View Spatial Features** to resolve alignment between Lidar BEV and Camera
- **Two Stage Detector** allows for **additional refinement** using low-level lidar and camera features.

- ***Loss of information*** *as domains are transformed into* ***BEV representation***
- ***Lack of synchronized data augmentation***
- ***Outdated Performance compared to current state of the art***

Dense Voxel Fusion (CVPR 2022)

Anas Mahmoud, Jordan S. K. Hu and Steven L. Waslander
University of Toronto

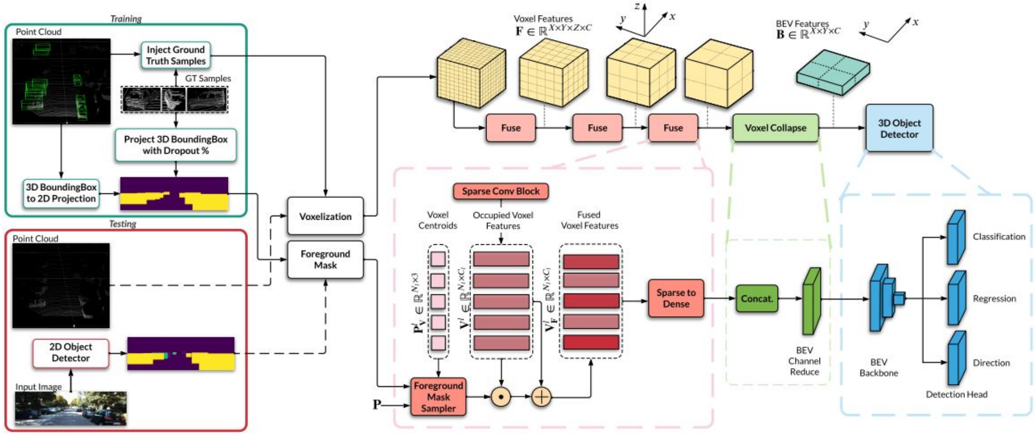


Fig 7. Dense Voxel Fusion Architecture

- Dense Voxel Fusion focuses on **improving expressiveness** in **low point density regions** through dense correspondence between pixels and points.
- Trains with **ground truth** rather than 2D Predictions

Dense Voxel Fusion (CVPR 2022)

Dense Voxel Fusion

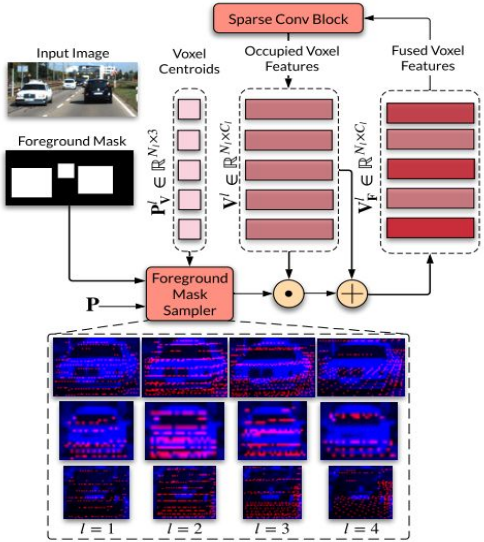


Fig 8. Dense Voxel Fusion Architecture (top), Example centroids at each block (bottom)

- **Foreground mask** created from any 2D detection is **fused** with voxel-based **lidar stream** between each Sparse Convolution Block
- Fused features from each block are then processed by the next block
- At each block, a **new set of centroids** samples the mask at **new pixel locations**, resulting in **dense correspondence**.

Dense Voxel Fusion (CVPR 2022)

Multi-Modal Training Strategy

- *Foreground Heatmap generation*
  - During Training, foreground mask is generated using **3D Ground Truth** which is **projected onto 2D space** to obtain **2D Detections**
    - Removes the requirement of labeled camera data & accurate 2D detections
- *Simulating False Detections*
  - A random subset of 3D bounding boxes are **not added** to the foreground heatmap
    - 3D Detector is **robust** to **missed 2D Object detections**.
  - **False positives** are added as **3D Ground Truth** may be **occluded** in the **2D Image**.

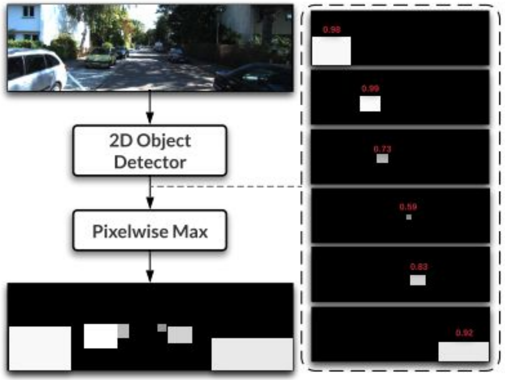


Fig 9. 2D Heatmap Generation

Institute for Aerospace Studies
UNIVERSITY OF TORONTO

TORONTO ROBOTICS + AI LAB

Dense Voxel Fusion (CVPR 2022)

Summary

- DVF can be **added to existing** Voxel-based **Lidar Training without** any **additional training parameters**
- Dense correspondence uses **dense image information**, **improving accuracy** at range
- **Independent** of **2D Object Detector Performance** with **Multi-Modal Training Strategy**
- Comparable Performance on KITTI dataset compared to state of the art

- *Lack of synchronized data augmentation between modalities*

Sparse Fuse Dense (CVPR 2022)

Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, Deng Cai
Zhejiang University

- Removes issue of different modalities by creating "pseudo 3D clouds" from images using Depth Completion

- RoI from Lidar Clouds with traditional methods
- RoI from Pseudo Clouds with **Color Point Convolution**
- RoI from Lidar & Pseudo are fused with **3D-GAF**
- Augments lidar & pseudo clouds using **Synchronized Augmentation**

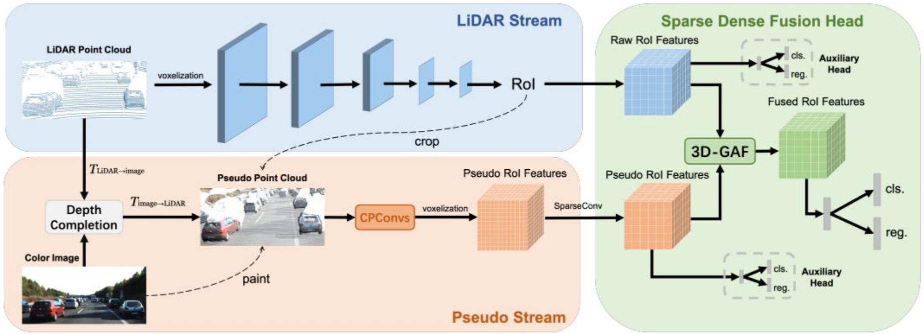


Fig 10. Sparse Fuse Dense Architecture

https://arxiv.org/pdf/2203.09780.pdf

Sparse Fuse Dense (CVPR 2022)

Color Point Convolution:

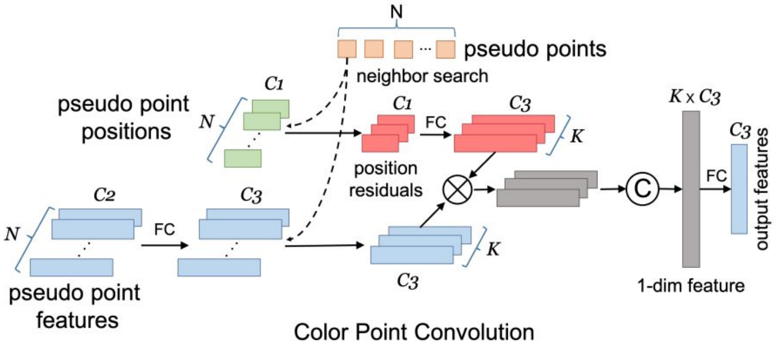


Fig 11. Color Point Convolution Architecture

- Searches **Neighbours** on the **image domain**
- Uses RoI-Aware Neighbour search to **prevent occluded** pseudo points **becoming neighbours**
- For kth neighbours of each point, **position residuals** are calculated:

$$h_i^k = (x_i - x_i^k, y_i - y_i^k, z_i - z_i^k, u_i - u_i^k, v_i - v_i^k, ||p_i - p_i^k||),$$

$$||p_i - p_i^k|| = \sqrt{(x_i - x_i^k)^2 + (y_i - y_i^k)^2 + (z_i - z_i^k)^2}$$

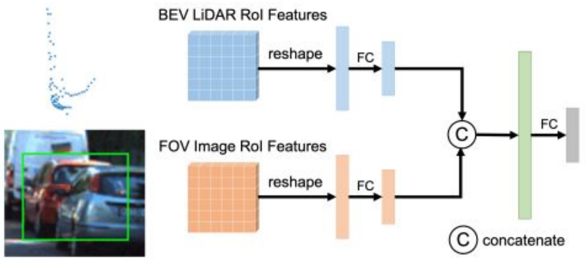
- Feature is **weighed** using the residuals, then concatenated

Sparse Fuse Dense (CVPR 2022)

3D Grid-Wise Attentive Fusion:

- Previous methods use a **coarse** RoI Fusion strategy:
    - Concatenating **2D Lidar BEV RoI** with **2D FOV image RoI**
        - Interference from Occluded & Background objects in 2D FOV Image RoI

- 3D GAF fuses 3D RoI from Lidar & Pseudo Cloud
    - Uses Attentive Fusion to fuse each grid pair separately:
        - Predicts & weighs **each grid pair** for the fused grid features



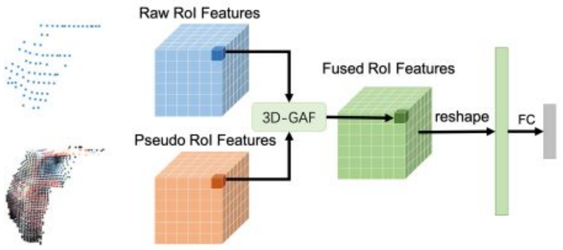


Fig 12. Previous Methods (top) vs 3D GAF (bottom)

Sparse Fuse Dense (CVPR 2022)

Synchronized Augmentation:

- Augments lidar & pseudo clouds together, only requiring lidar methods
- Removes need for additional synchronization between different modalities

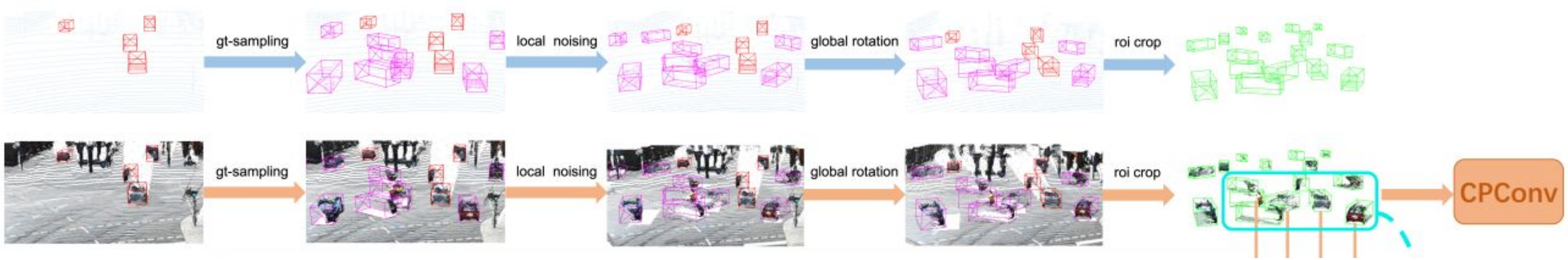


Fig 13. Synchronized Augmentation between Lidar & Pseudo Clouds

Institute for Aerospace Studies
UNIVERSITY OF TORONTO

TORONTO ROBOTICS + AI LAB

Sparse Fuse Dense (CVPR 2022)

Summary:

- Uses **dense image information** to creating **Pseudo Clouds** from image data for refinement
- **Simplifies training** through conversion to the same modalities
- **Removes need** for additional synchronization for data augmentation between different modalities with **Synchronized Augmentation**
- **Top current performance on KITTI Dataset**

- *Still relies on sparse lidar points to pick 3D RoI*

KITTI 3D Dataset Comparison:

Method	Easy	Medium	Hard
MVX-Net Voxel Fusion	82.3	72.2	66.8
MVX-Net Point Fusion	83.2	73.2	63.7
3D CVF	89.20	80.05	73.11
Dense Voxel Fusion	***90.99***	***82.40***	***77.31***
Sparse Fuse Dense	**91.73**	**84.76**	**77.92**

Institute for Aerospace Studies
UNIVERSITY OF TORONTO

TORONTO ROBOTICS + AI LAB