INFORMATIK

augmented VISION

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Master Thesis

Object Detection Using Transformer Fusion And Detection Transformer On Multi-Sensor Data For Automated Driving

Anshu Garg

a_garg19@cs.uni-kl.de

Supervisor: David Michael Fürst

Professor: Prof. Dr. Didier Stricker

Outline

- Motivation
- Related Work
- Research Gaps
- Methodology
- Experiments
- Qualitative Results
- Conclusion and Future Work
- References

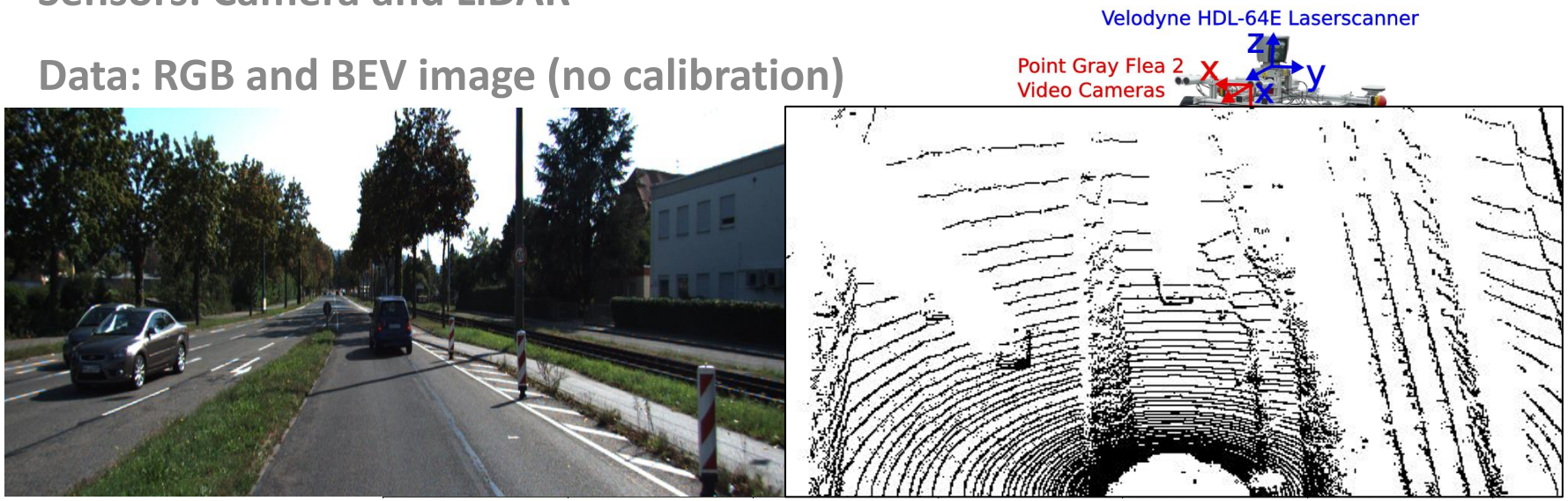
Motivation

Model: TransFuserDeTr (TransFuser[1] + DeTr[2])

Task: End-to-end 2D Object Detection

Sensors: Camera and LiDAR

Data: RGB and BEV image (no calibration)

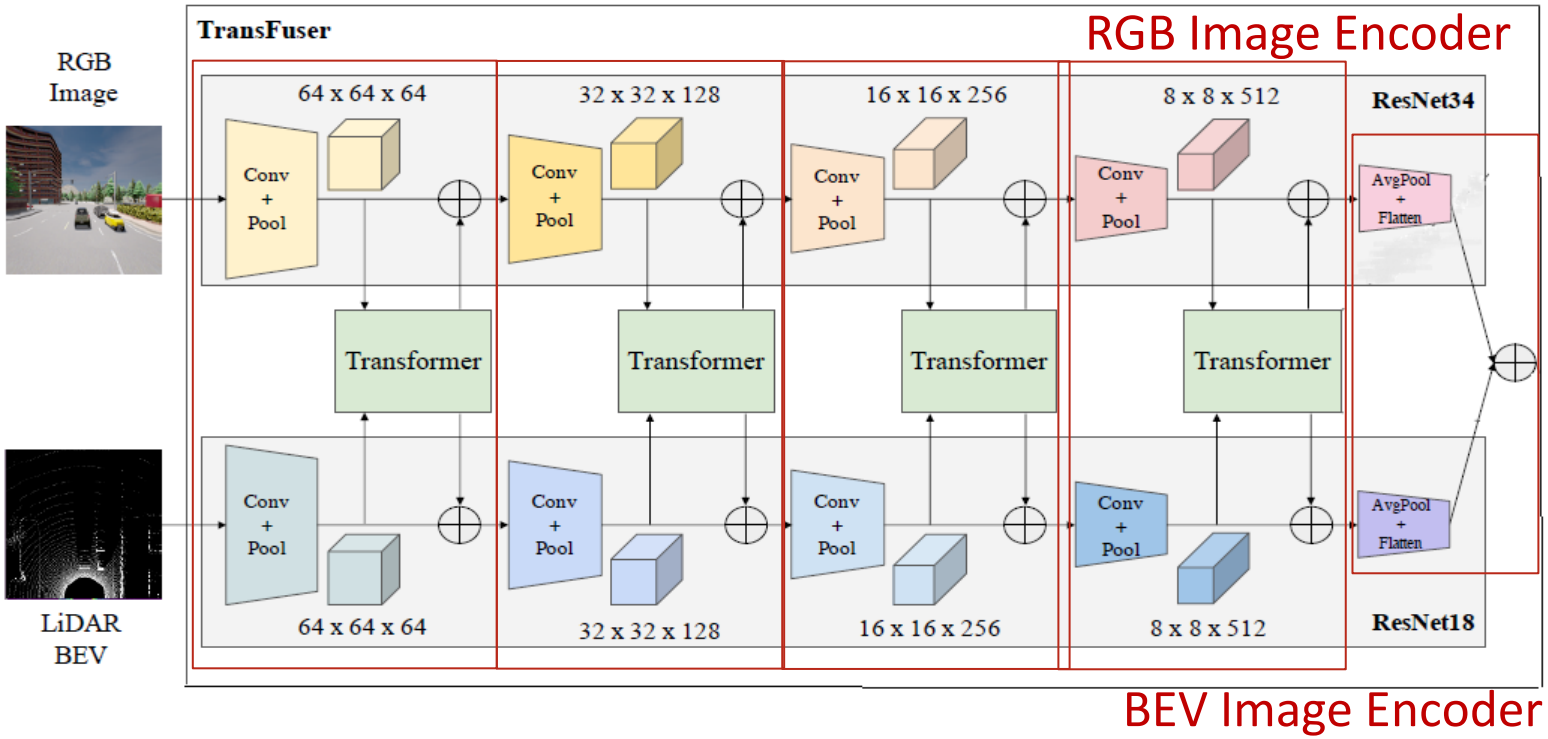


1. A.Prakash et al. “Multimodal fusion transformer for end-to-end autonomous driving” (CVPR), 2021.
2. N.Carion et al. “End-to-end object detection with transformers” (ECCV), 2020.
3. A.Geiger et al. “Vision meets robotics: The kitti dataset” (IJRR), 2013.

Outline

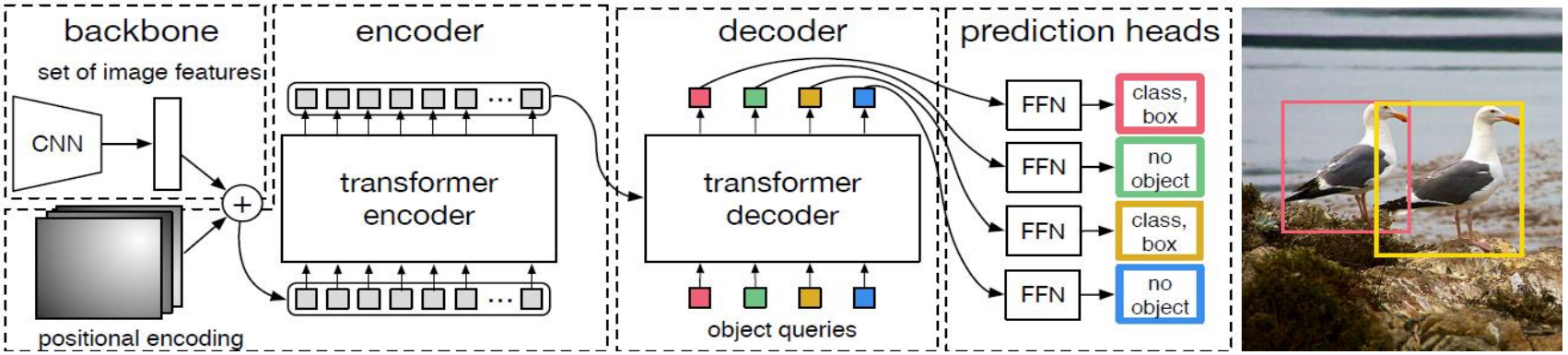
- Motivation
- Related Work
- Research Gaps
- Methodology
- Experiments
- Qualitative Results
- Conclusion and Future Work
- References

Transformer Fusion (TransFuser)



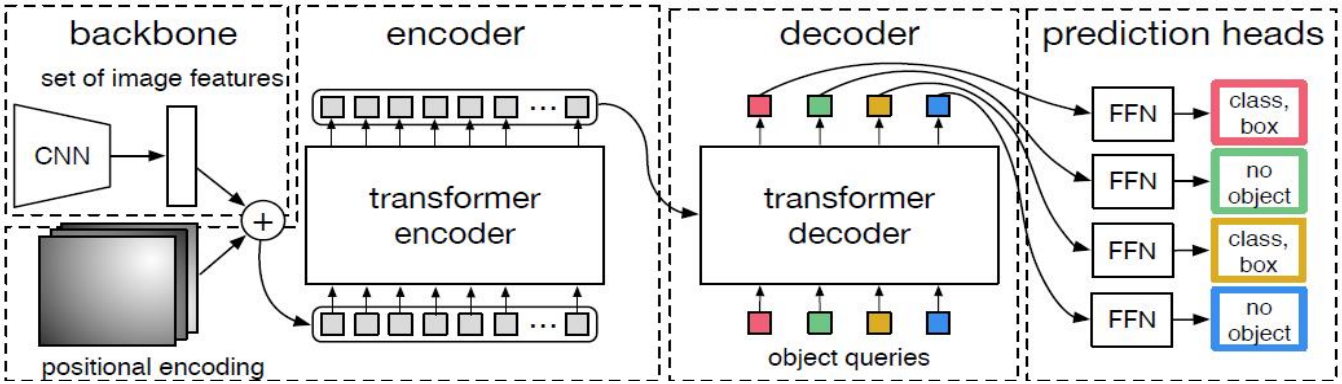
- Task: Waypoint Prediction
- Deep and slow fusion on multi-sensor data by attention menchanism using Transformer.
- Outputs global context feature vector.

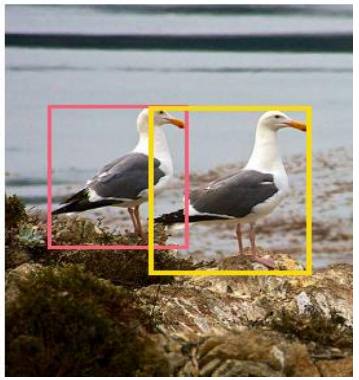
Detection Transformer (DeTr)



- Task: End-to-end 2D Object Detection.
- Idea: Considers object detection as set prediction problem.
- CNN: ResNet50.
- 1x1 conv. reduces feature map size (input_proj).
- Encoder‘s positional encoding: Sine/ Learned.
- Standard Encoder and Decoder Transformer (decodes N objects in parallel).

Detection Transformer (DeTr)





- Decoder‘s Learned positional encoding called Object Queries.
- Decoder‘s Embedding layer (query_embed).
- FFN: label classification head (class_embed) and bounding box coordinates prediction head (bbox_embed).
- Loss: Cross Entropy (label classification), L1 and GIoU (bounding box regression).

Outline

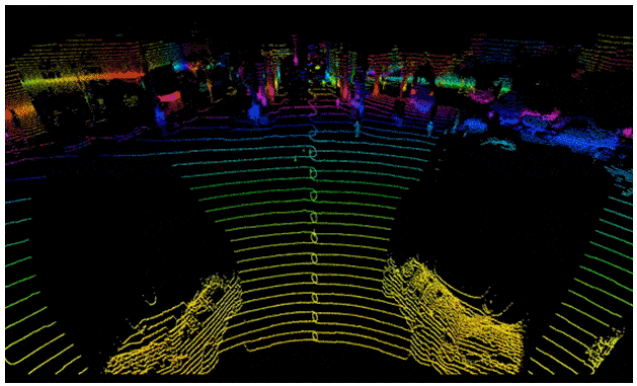
- Motivation
- Related Work
- Research Gaps
- Methodology
- Experiments
- Qualitative Results
- Conclusion and Future Work
- References

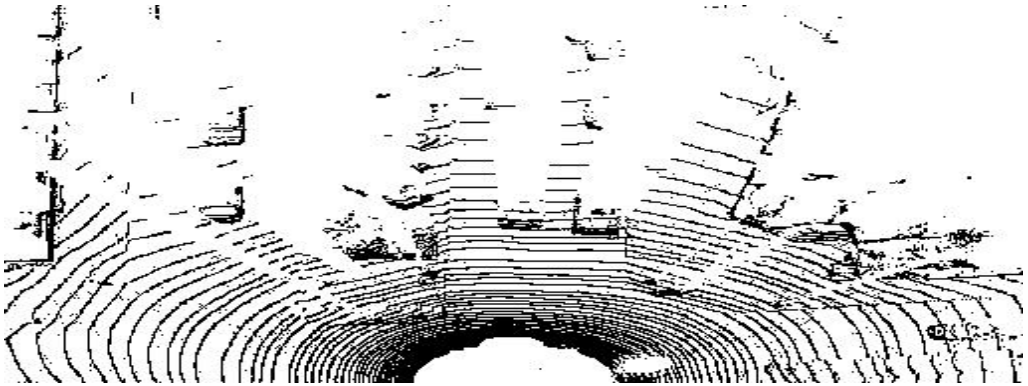
Research Gaps

Sensor Modalities:

- **Camera:** 2D image.

- **LiDAR:** 3D point cloud and 2D BEV image.





Use of multi-sensor data might be beneficial for object detection.

Research Gaps

Multi sensor modality: data fusion.

- **Late Fusion** for e.g., CLOCs[7]. Little exchange of information.
- **Slow Fusion:** for e.g., AVOD[8]. Dependency on good calibration.
- **Deep and Slow Fusion:** for e.g., TransFuser.

Research Gap: TransFuser doesn‘t solve object detection task.

7. Su Pang et al. “CLOCs:Camera-LiDAR object Candidates fusion for 3D object detection” (IROS), 2020.
8. Jason Ku et al. “Joint 3d proposal generation and object detection from view aggregation” (IROS), 2018.

Research Gaps

Object Detection Approaches:

- **Conventional Approaches:** One and Two stage object detection models (for e.g., YOLO[5], Faster-RCNN[6]).
  - **Drawbacks:** Prior knowledge, Post-processing steps.
- **Recent Approaches:** direct set prediction task (for e.g., DeTr).
  - **Advantage:** direct predictions.

  Research Gap: DeTr doesn‘t use multi-sensor data.

5. R.Joseph “You only look once: Unified, real-time object detection” (CVPR) 2016.
6. S.Ren et al. “Faster R-CNN: Towards real-time object detection with region proposal networks” (NeurIPS), 2015.

Research Gaps

- Extension of DeTr and TransFuser model.
- TransFuserDeTr: TransFuser + DeTr model.
- Evaluation on Kitti dataset (7.4K images).
- Pre-training on NuScenes[4] dataset (34K images).

4. C.Holger et al. “Nuscenes: A multimodal dataset for autonomous driving” (CVPR), 2020.

Outline

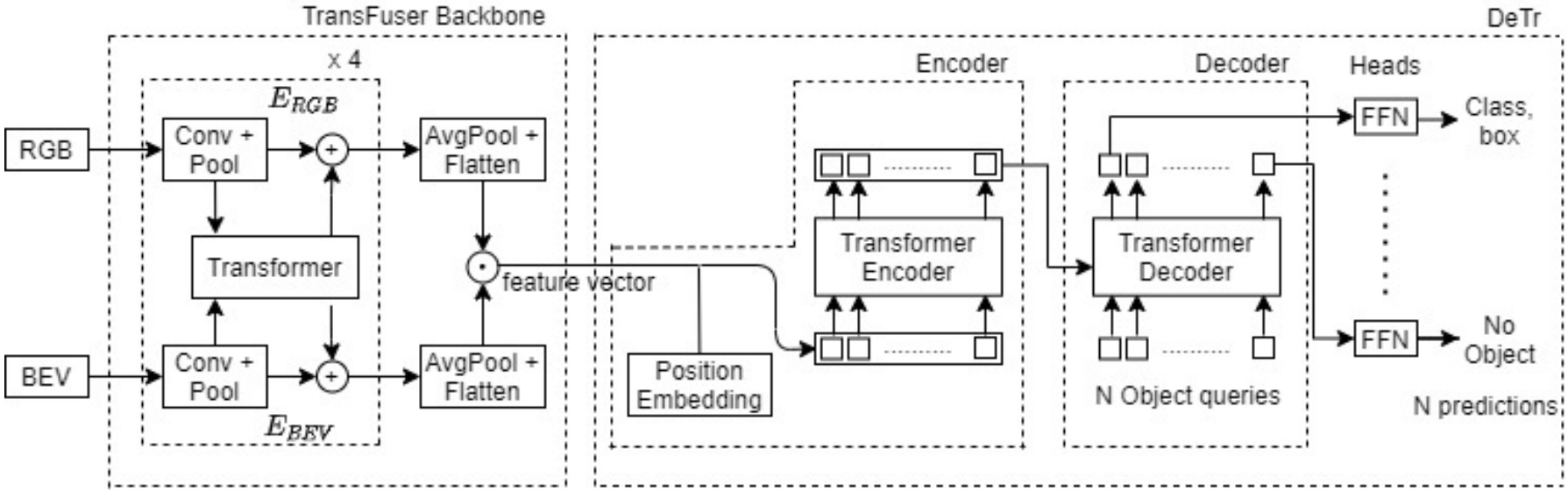
- Motivation
- Related Work
- Research Gaps
- Methodology
- Experiments
- Qualitative Results
- Conclusion and Future Work
- References

Baseline DeTr Model

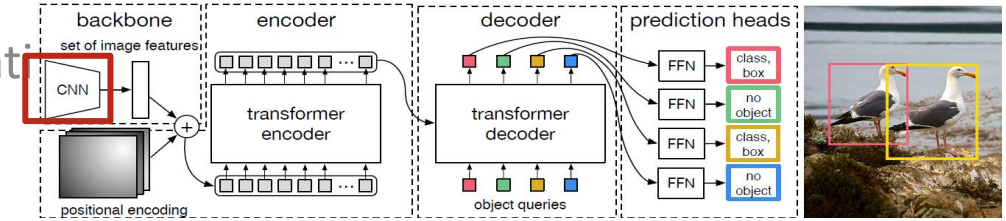
- Training DeTr on Kitti dataset for object detection task.
- Kitti to COCO[12] data format conversion.
- Baseline set up: code migration.
- Evaluation on Kitti dataset (7481 images).
- Prediction selection using threshold value (during evaluation).
- Transfer Learning from COCO to Kitti dataset for DeTr model.
- Different experimental set up to find the best setting.

12. Tsung-Yi.Lin et al. “Microsoft COCO: Common objects in context” (ECCV), 2014.

TransFuserDeTr Model

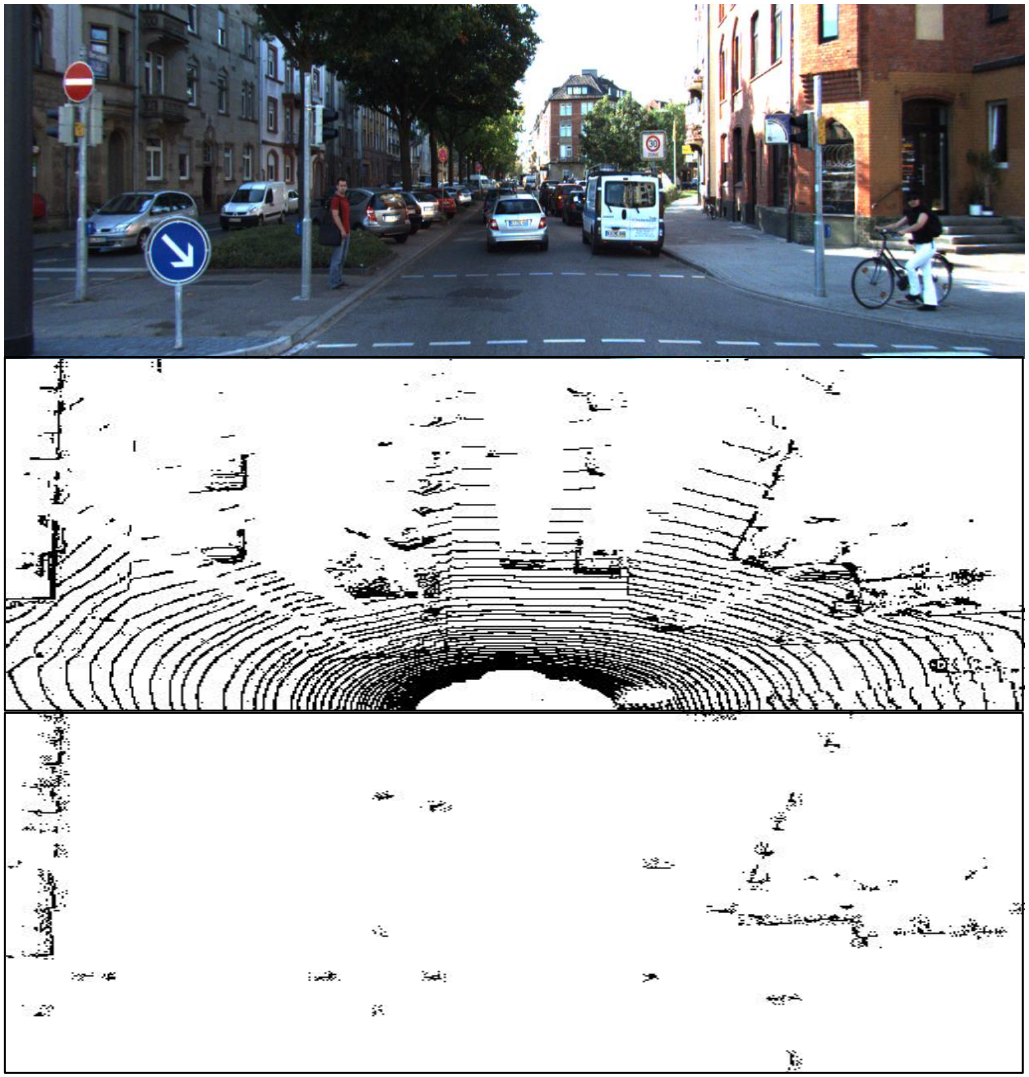


- Replace DeTr backbone by TransFuser.
- TransFuser and DeTr customisat
- BEV generation.

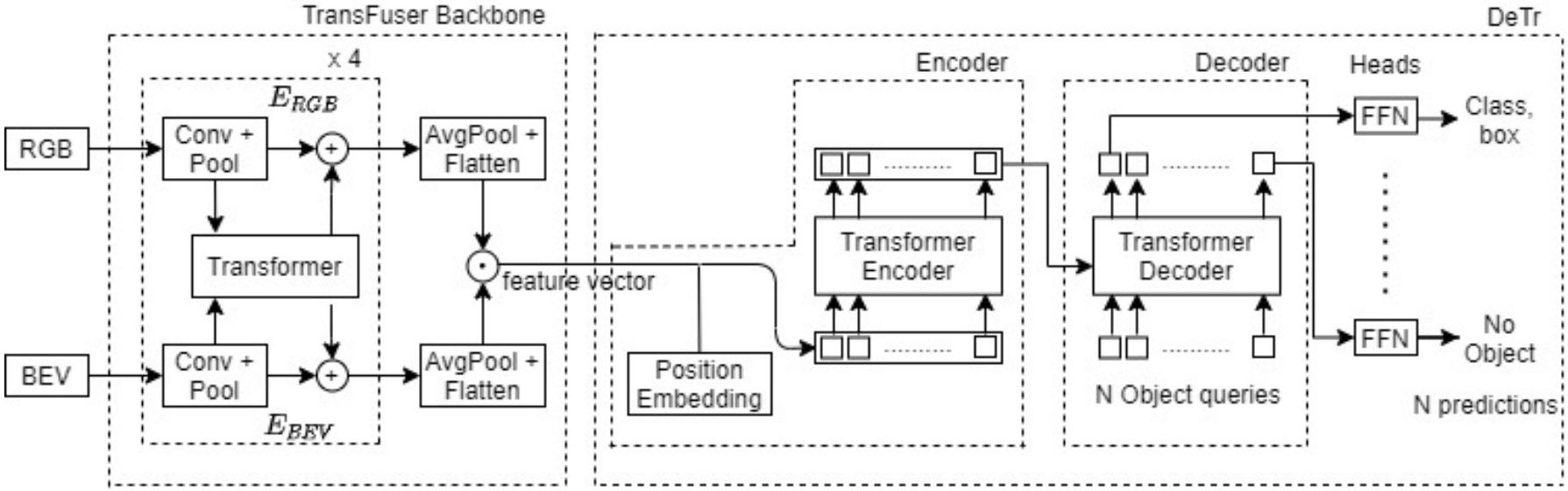


TransFuserDeTr Model: BEV Generation

2 channel BEV capturing the camera view.



TransFuserDeTr Model



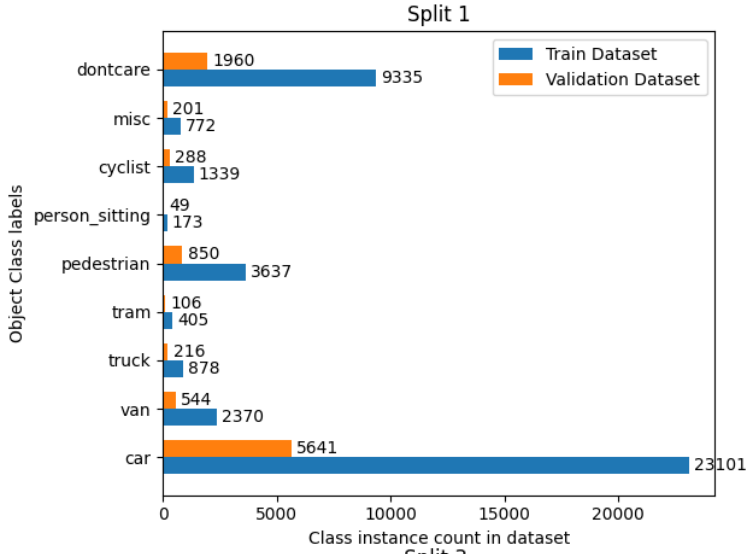
- Replace DeTr backbone by TransFuser.
- TransFuser and DeTr customisations.
- BEV generation.
- Find the best experimental set up.

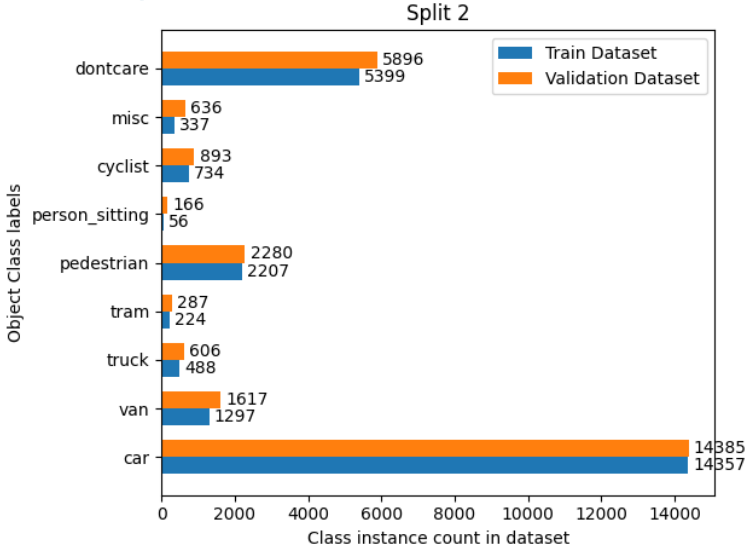
Kitti Dataset: Different Splits

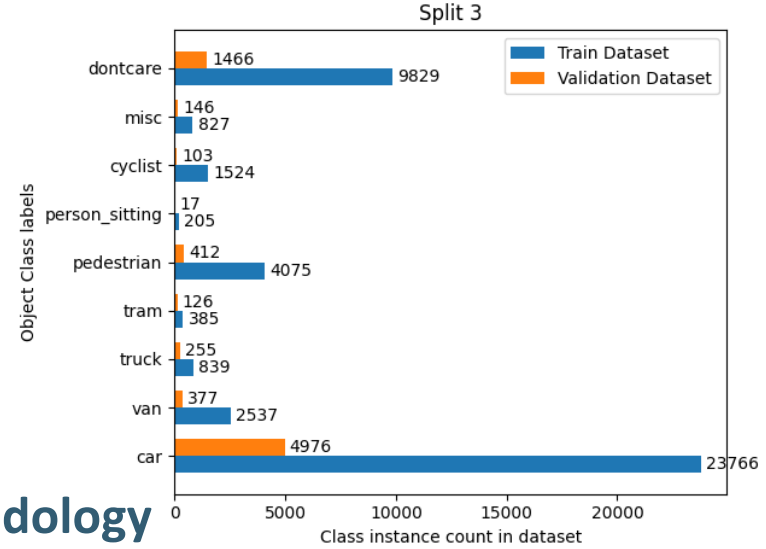
- TransFuserDeTr performance testing on different data splits:
  - Split 1
  - Split 2[13]
  - Split 3

13. C.Xiaozhi et al. “3d object proposals for accurate object class detection” (NeurIPS), 2015.

Kitti Dataset: Different Splits







Split 1
- 80:20 (Ours) (used so far)
- No image sequence.

Split 2
- 50:50 (less training data)
- Considers image sequence.

Split 3
- 80:20 (Ours) (used further)
- Considers image sequence.

Transfer Learning using NuScenes

Overfitting challenge on split 3.

Possible Solutions: L1, L2 Reg., Weight decay, Auxiliary Loss, Dropout, Data Augmentation (not effective)

Training on bigger dataset

Model pre-training

Pre-training on NuScenes dataset (34k images).

Fine tuning on Kitti dataset (split 3).

Evaluation using Kitti evaluator on both datasets.

Outline

- Motivation
- Related Work
- Research Gaps
- Methodology
- Experiments
- Qualitative Results
- Conclusion and Future Work
- References

Baseline DeTr Model

Using Kitti dataset (split 1).

- DeTr pretrained weights versus scratch training.

Pretrained weights	Backbone weights	Object Queries	Validation Loss
No	No	100	1.021
Yes	Yes	30	0.1948
Yes	Yes	100	0.1885
Yes	**No**	**100**	**0.1172**

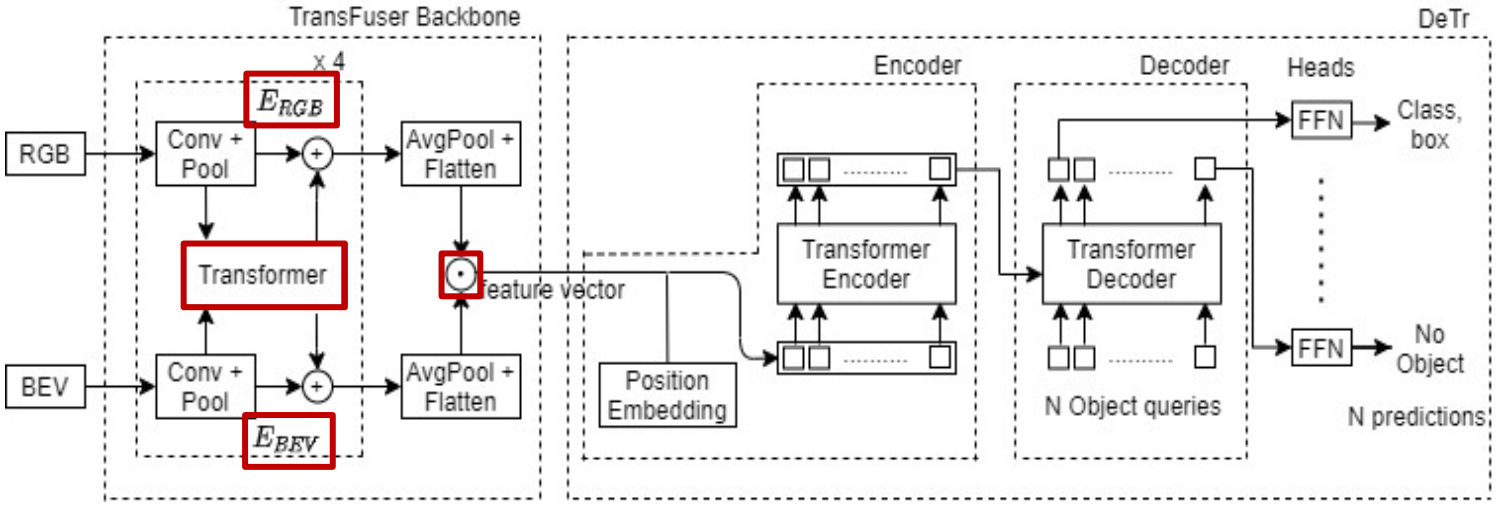
- Hyper-parameter tuning

Threshold	Pre-Norm	Car (AP)	Ped (AP)	Cyc (AP)
70%	T	89.14	72.41	72.68
70%	**F**	**89.39**	**77.56**	**86.08**
80%	F	89.38	76.00	79.30
90%	F	89.36	69.76	79.09

TransFuserDeTr Model

Using Kitti dataset (split 1).

row	ip pw	Hyper Params	Encoder	DA	Fusion, PG	Size	Car	Ped	Cyc



TransFuserDeTr Model

Using Kitti dataset (split 1).

- Pretrained weights for RGB and BEV Image Encoder (Row 1).
- TransFuser Backbone hyper parameter tuning (Row 2).
- TransFuser RGB and BEV image Encoder (Row 3).

row	ip pw	Hyper Params	Encoder	DA	Fusion, PG	Size	Car	Ped	Cyc
1	Y	(8,4,4,200)	(R50,R50)	N	(+, 1)	604M	7.5	9.0	-
2	Y	(1,2,4,200)	(R50,R50)	N	(+, 1)	202M	84.7	61.7	63.9
3	Y	(1,8,4,200)	(R34,R18)*	N	(+, 1)	87M	87.6	70.4	74.5

TransFuserDeTr Model

Using Kitti dataset (split 1).

- Weight Initialisation techniques (no significant improvement).
- TransFuserDeTr Model optimisation (Row 4 and 5).
- Data Augmentation (Row 6, also reduced generalisation gap).

row	ip pw	Hyper Params	Encoder	DA	Fusion, PG	Size	Car	Ped	Cyc
1	Y	(8,4,4,200)	(R50,R50)	N	(+, 1)	604M	7.5	9.0	-
2	Y	(1,2,4,200)	(R50,R50)	N	(+, 1)	202M	84.7	61.7	63.9
3	Y	(1,8,4,200)	(R34,R18)*	N	(+, 1)	87M	87.6	70.4	74.5
4	N	(1,8,4,200)	(R34,R18)	N	(+, 1)	84M	84.4	63.5	67.2
5	N	(1,2,1,200)	(R34,R18)	N	(+, 1)	55M	86.8	71.9	78.6
6	N	(1,2,1,200)	(R34,R18)	Y	(+, 1)	55M	88.9	70.7	82.2

TransFuserDeTr Model

Using Kitti dataset (split 1).

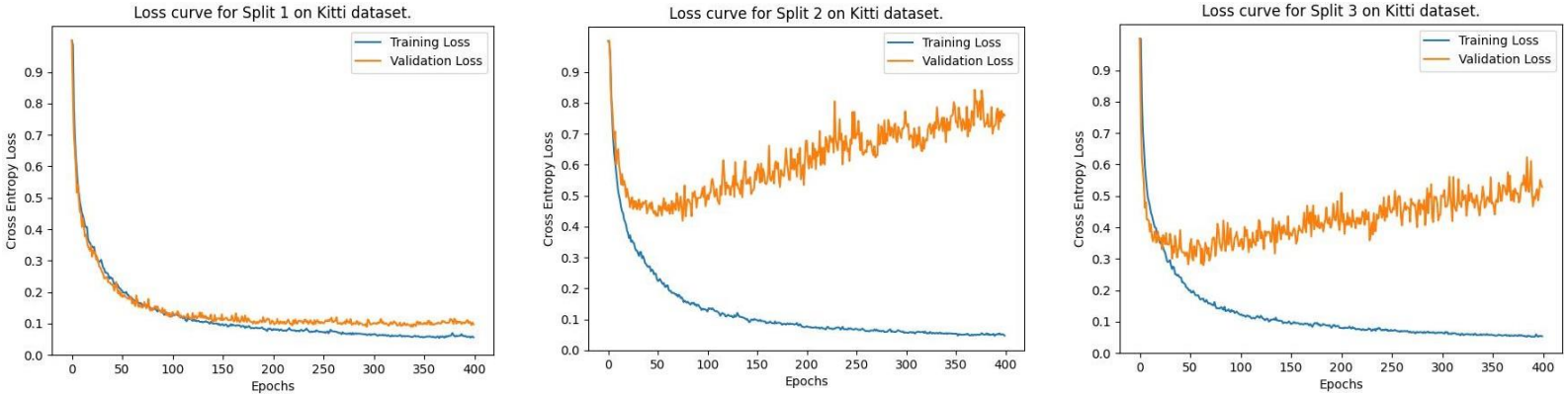
- Feature fusion technique & different parameter groupings (Row 7, 8, 9).
- Miscellaneous Experiments: different image resolution, LR Scheduler, lr, normalize BEV, position encoding etc (no improvement).

row	ip pw	Hyper Params	Encoder	DA	Fusion, PG	Size	Car	Ped	Cyc
1	Y	(8,4,4,200)	(R50,R50)	N	(+, 1)	604M	7.5	9.0	-
2	Y	(1,2,4,200)	(R50,R50)	N	(+, 1)	202M	84.7	61.7	63.9
3	Y	(1,8,4,200)	(R34,R18)*	N	(+, 1)	87M	87.6	70.4	74.5
4	N	(1,8,4,200)	(R34,R18)	N	(+, 1)	84M	84.4	63.5	67.2
5	N	(1,2,1,200)	(R34,R18)	N	(+, 1)	55M	86.8	71.9	78.6
6	N	(1,2,1,200)	(R34,R18)	Y	(+, 1)	55M	88.9	70.7	82.2
7	N	(1,2,1, -)	(R34,R18)	Y	(+, 2)	55M	88.8	70.2	84.0
8	N	(1,2,1, -)	(R34,R18)	Y	(· , 3)	55M	**89.4**	74.2	77.1
9	N	(1,2,1, -)	(R34',R18)	Y	(· , 4)	38M	89.3	**82.9**	**86.8**

TransFuserDeTr Model: Different Splits

Model evaluation for split 1, 2 and 3.

Split	Best Epoch	Car (AP)	Ped (AP)	Cyc (AP)
1	390	89.3	82.9	86.8
2	240	81.4	55.3	38.0
3	360	**86.7**	**70.1**	**58.1**



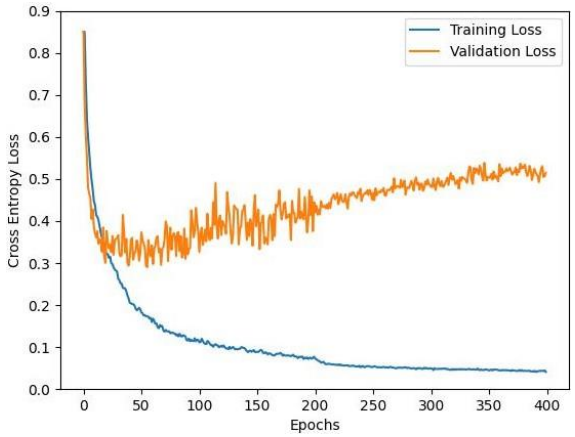
- More training data in split 3 than 2: Less overfitting.
- Car enough training data: AP_{70} is not affected much among three splits.
- Ped & Cyc less training data: least AP_{70} & varies between split 2 and 3

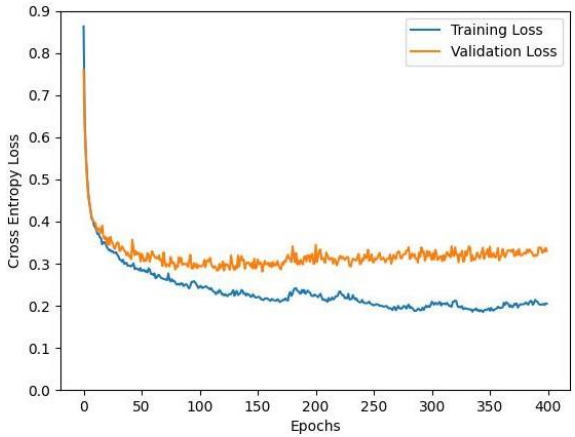
Transfer Learning using NuScenes

TransFuserDeTr model evaluation on NuScenes dataset using Kitti evaluator.

Row	Resolution	Camera	Set up	Car	Ped	Cyc
1	(1024, 512)	front only	1	48.1	49.2	31.2
2	(1600, 800)	front only	1	50.1	**53.1**	**40.4**
3	(1024, 512)	front and back	2	**52.4**	51.5	39.4

Kitti (split 3) versus NuScenes Classification Loss.





SOTA Comparison

Comparison between TransFuserDeTr model and SOTA on Kitti dataset.

Model	Car (AP)	Ped (AP)	Cyc (AP)
AVOD [8]	89.8	39.4	52.6
AVOD-FPN [8]	88.9	57.8	60.7
PFF3D [9]	92.1	52.5	66.2
PointPainting [10]	92.5	53.7	78.0
Fast-CLOCs [11]	95.7	62.5	75.0
DeTr(w)*	85.4	79.7	71.9
TransFuserDeTr [Ours]	**86.7**	70.1	58.1
TransFuserDeTr + Pre-train [Ours]	86.3	**71.0**	**61.4**

For Pedestrian: TransFuserDeTr outperforms all SOTA (except DeTr).

For Car: TransFuserDeTr outperforms DeTr.

For Cyclist: TransFuserDeTr outperforms AVOD and AVOD-FPN.

Pretraining TransFuserDeTr on NuScenes, improved AP_{70} for Pedestrian and Cyclist classes.

Outline

- Motivation
- Related Work
- Research Gaps
- Methodology
- Experiments
- Qualitative Results
- Conclusion and Future Work
- References

Qualitative Results

TransFuserDeTr model prediction visualisation on Kitti dataset (split 3).

Outline

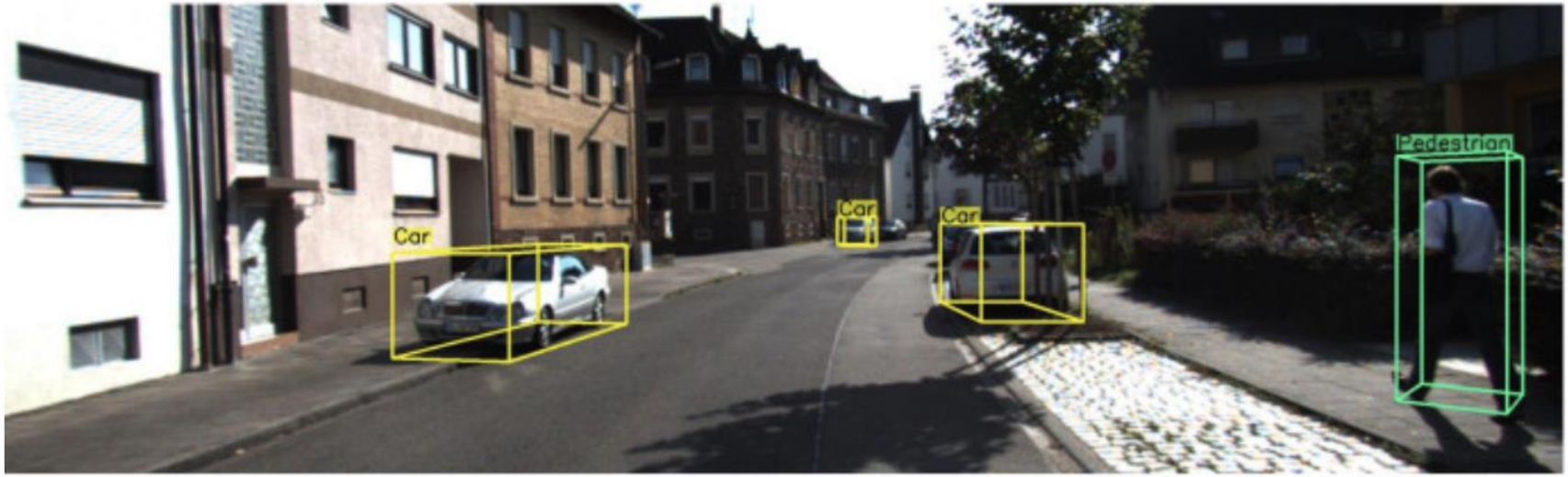
- Motivation
- Related Work
- Research Gaps
- Methodology
- Experiments
- Qualitative Results
- Conclusion and Future Work
- References

Conclusion

- Transfer Learning on DeTr from COCO to Kitti dataset.
- Novel architecture TransFuserDeTr successfully performs end-to-end object detection task using mulit-sensor data.
- Do not need any calibration between sensor data.
- More training data: better model performance (AP value) and lower the loss, overfitting, generalisation gap challenges.
- Pre-training TransFuserDeTr enhances performance for Pedestrian and Cyclist, though couldn‘t address overfitting.
- TransFuserDeTr performs better than DeTr for Car prediction.
- TransFuserDeTr performs better than other SOTA using multi-sensor data, for Pedestrian (except DeTr) and Cyclist (AVOD, AVOD-FPN).

Future Work

- Use TransFuserDeTr model for 3D Object Detection task.



References

1. Aditya Prakash, Kashyap Chitta, and Andreas Geiger. “Multimodal fusion transformer for end-to-end autonomous driving”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021

2. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, and N.Usunier et al. “End-to-end object detection with transformers”. In: European Conference on Computer Vision (ECCV). Springer. 2020.

3. Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. “Vision meets robotics: The kitti dataset”. In: The International Journal of Robotics Research (IJRR) 32.11 (2013), pp. 1231–1237.

4. Holger Caesar, Varun Bankiti, Alex H Lang, and S.Vora et al. “Nuscenes: A multimodal dataset for autonomous driving”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

5. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016.

6. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: Advances in Neural Information Processing Systems (NeurIPS) 28 (2015).

7. Su Pang, Daniel Morris, and Hayder Radha. “CLOCs: Camera-LiDAR object candidates fusion for 3D object detection”. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. 2020.

8. Jason Ku, Melissa Mozifian, Jungwook Lee, and A.Harakeh et al. “Joint 3d proposal generation and object detection from view aggregation”. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2018.

9. Li-Hua Wen and Kang-Hyun Jo. “Fast and accurate 3D object detection for lidar-camera-based autonomous vehicles using one shared voxel-based backbone”. In: IEEE Access 9 (2021), pp. 22080–22089.

10. Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. “Pointpainting: Sequential fusion for 3d object detection”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

References

11. Su Pang, Daniel Morris, and Hayder Radha. “Fast-CLOCs: Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection”. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

12. Tsung-Yi Lin, Michael Maire, Serge Belongie, and J.Hays et al. “Microsoft COCO: Common objects in context”. In: European Conference on Computer Vision (ECCV). Springer. 2014.

13. Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, and A.G.Berneshawi et al. “3d object proposals for accurate object class detection”. In: Advances in Neural Information Processing Systems (NeurIPS) 28 (2015).

14. Ashish Vaswani, Noam Shazeer, Niki Parmar, and J.Uszkoreit et al. “Attention is all you need”. In: Advances in Neural Information Processing Systems (NeurIPS) 30 (2017).

Thank You

Happy Researching ☺