LeTFuser: Light-weight End-to-end Transformer-Based Sensor Fusion for Autonomous Driving with Multi-Task Learning

Pedram Agand*, Mohammad Mahdavian*, Manolis Savva, and Mo Chen

SFU

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

MARS LAB

Problem Definition

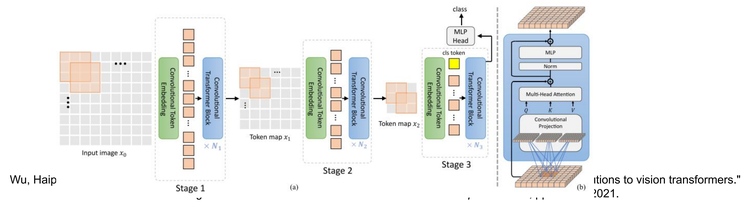
- End-to-end autonomous driving using light-weight transformer networks
- Training a light-weight model based on imitation learning method
- Using multi-task learning for better training different parts of the model
- Inputs: Three RGB-D Cameras / Outputs: Steering, Throttle, Brake values

Perception

- Using CvT network for better extracting image features
- A Semantic Depth Cloud (SDC) for better recognizing the environment

Convolutional vision Transformers:

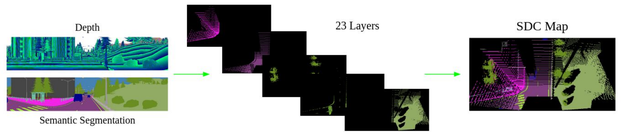
- Combining convolutions and attention to extract features locally and globally
- 2D convolutions better learn the image features with respect to 1D tensors



Wu, Haip itions to vision transformers." 2021.

Semantic Depth Cloud:

- Using depth image and semantic segmentation to create a 23 layer Bird Eye View map

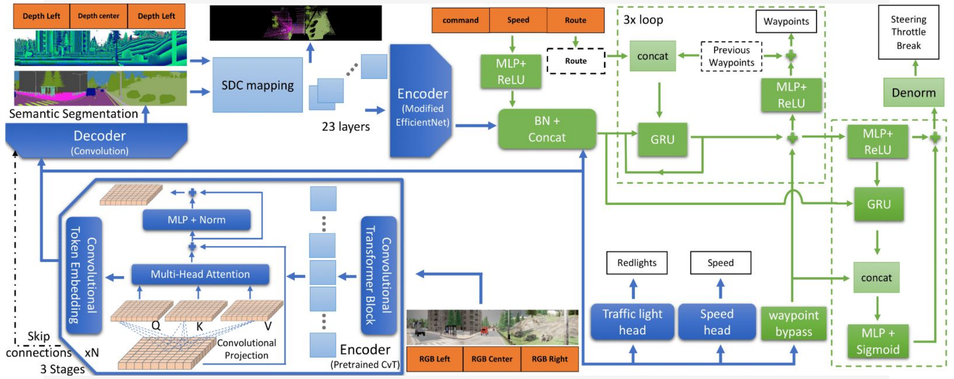


* equal contribution

Control

Predicting vehicular control (throttle, break, and steering) and waypoints

- **Fused features:** Measurements + RGB features + SDC map
- **Waypoint branch:** Fused features >> environment-agent static knowledge
- **Dynamic branch:** Static knowledge + Fused features >> Vehicular control
- **Multi agent control:** PID on-the-fly to track waypoints + Imitation learning to obtain vehicular control



Perception module is blue while controller is green. Inputs are orange, while color box are separate learned tasks. Trainable and non-trainable in light and dark-colored items.

Dataset

CARLA 0.9.10

Train:

8 publicly available towns, Short scenarios, 2500 routes, All weather

Test:

Town 05, short (32) and long (10) scenarios, Adversarial and Normal condition

Results

Metrics:

- IP: Infraction Penalty
- RC: Route Completion
- DS: Driving Score = IP x RC

Results:

Experiment	Model	Inputs	Normal Clear Noon (1WN)			Adversarial Clear Noon (1WA)		
Town5			DS	RC	IPS	DS	RC	IPS
short	X13-F	1 RGB-D	32.814	68.284	0.468	28.359	58.792	0.480
	X13-A	3 RGB-D	48.833	75.824	0.588	37.263	73.986	0.466
	TF-F	1 RGB-L	17.800	19.864	0.942	23.641	24.373	0.953
	TF-A	3 RGB-L	12.494	16.315	0.886	11.349	14.675	0.843
	Ours	3 RGB-D	**66.012**	**99.717**	0.663	**51.669**	**91.918**	0.574
	Expert	*	99.919	99.919	1.00	79.675	95.349	0.833
long	X13-F	1 RGB-D	8.381	63.633	0.194	7.601	48.114	0.246
	X13-A	3 RGB-D	7.670	48.039	0.291	11.866	52.424	0.456
	TF-F	1 RGB-L	22.456	24.509	0.950	12.964	15.393	0.910
	TF-A	3 RGB-L	7.111	7.351	0.963	8.829	9.001	0.971
	Ours	3 RGB-D	13.943	**63.685**	0.215	**20.584**	47.186	0.493
	Expert	*	60.808	100	0.608	23.344	58.872	0.619

Model	Total Parameters	GPU memory
X13-F	20985934	2920 MB
X13-A	20985934	4958 MB
TF-F	66218754	3898 MB
TF-A	66401154	5015 MB
Ours	31331865	3761 MB

Model	Acc_{TL}	MAE_{SP}	BCE_{SEG}	MAE_{WP}	MAE_{ST}	MAE_{TH}	MAE_{BR}	$epoch^*$
X13-F	0.9846	NA	0.1591	0.0792	0.0173	0.0482	0.0236	30
X13-A	**0.9882**	NA	0.0648	0.07878	0.0195	0.04254	0.01989	19
Ours no side SDC	0.9812	0.2786	0.0647	0.08275	0.02285	0.0531	0.03323	15
Ours no CvT	0.988	**0.13443**	0.0634	0.07307	**0.0173**	0.0510	0.0199	11
Ours no VC	0.9839	0.3051	0.0621	0.0817	0.01879	0.0531	0.02567	19
Ours	0.987	0.2524	**0.0620**	**0.0729**	0.0182	**0.0445**	**0.0185**	21

Link to our Github Repository

Contact info: pagand@sfu.ca