UniFusion: Unified Multi-view Fusion Transformer for Spatial-Temporal Representation in Bird's-Eye-View Paper Review 05/16

Bo Han, Chen

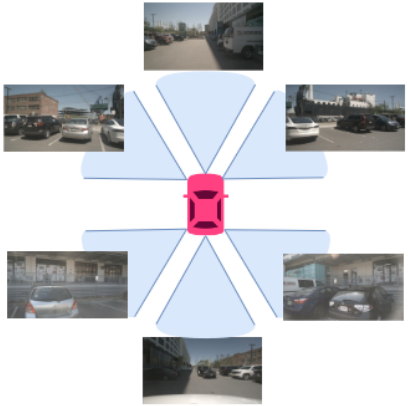
National Yang Ming Chiao Tung University, Taiwan
bhchen312551074.cs12@nycu.edu.tw

May 16, 2024

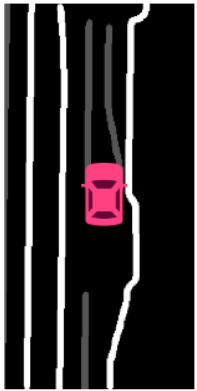
Info

- Title: UniFusion: Unified Multi-view Fusion Transformer for Spatial-Temporal Representation in Bird's-Eye-View [1]
- Authors: Qin, Zequn and Chen, Jingyu and Chen, Chao and Chen, Xiaozhi and Li, Xi
- Conference: IEEE/CVF International Conference on Computer Vision
- Year: 2023

Bird's-Eye-View



(a) Inputs with surrounding images.



(b) Map.

Figure 1: Bird's-Eye-View

Motivation

- Limitations of temporal fusion
  - Warp-based methods
  - Long-Range fusion

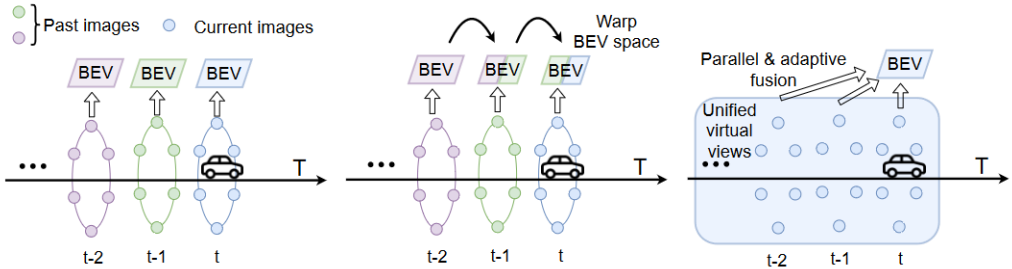


Figure 2: Temporal Fusion Methods

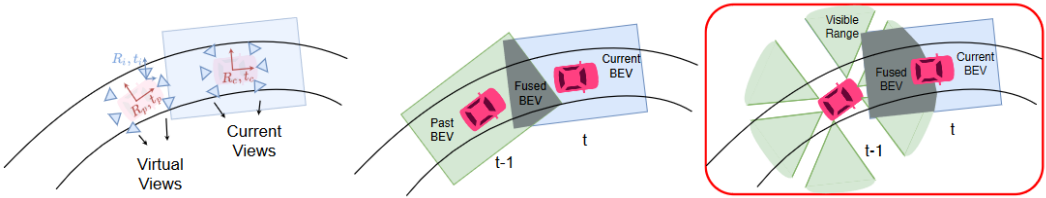
Proposed Method

- Multi-view perspective for BEV representation
  - unify spatial and temporal fusion
- Virtual view for temporal fusion
- Propose new experiment setting
  - more realistic
  - avoid overfitting

Proposed Method

Virtual View

- Why warp-based methods are not good?
  - all features are organized in pre-defined BEV space
  - cause information loss



(a) Illustration of virtual views. (b) Warp-based BEV fusion. (c) Actual BEV space that can be fused.

Figure 3: Warped-based Methods Comparison

Proposed Method

Virtual View

- Map the views that are not presented in current time step
- Using rotation, translation and intrinsic matrix
- Each time step can be done in parallel

Proposed Method

Network Design

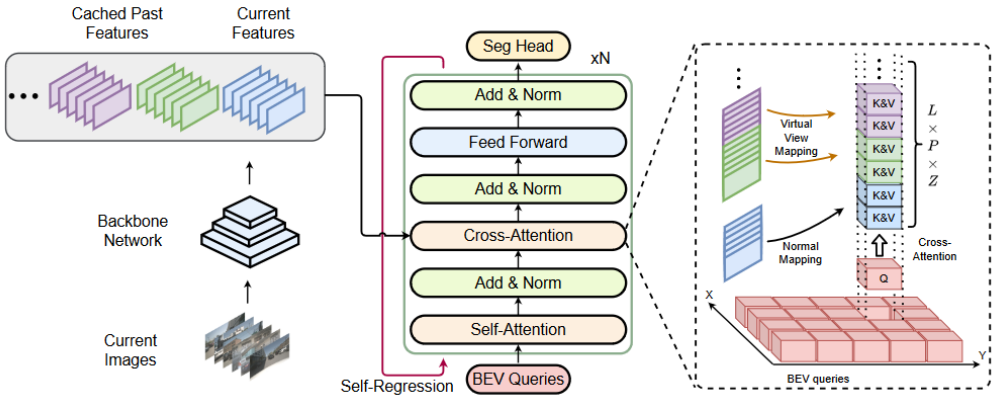


Figure 4: Network Design

Proposed Method

Network Design - Backbone

- Extract features from camera images
- Features of past images can be maintained and reused
- Model used: ResNet50, Swin-Tiny, VoVNet

Proposed Method

Network Design - Fusion Transformer

- Fusion features from all views
- Including 4 major parts
  - BEV query
  - self-attention
  - cross-attention
  - self-regression

Proposed Method

Network Design - BEV Query & Self-Attention

- BEV query
  - 2-D grid for BEV representation
  - Unified the spatial and temporal fusion
  - Pre-defined size for each experiment
- Self-attention
  - exchange information between different BEV views
  - capture long-range spatial dependencies

- Map spatial-temporal features to BEV space
  - virtual view for temporal fusion
  - provide adaptive weights for each view
  - all features can directly access

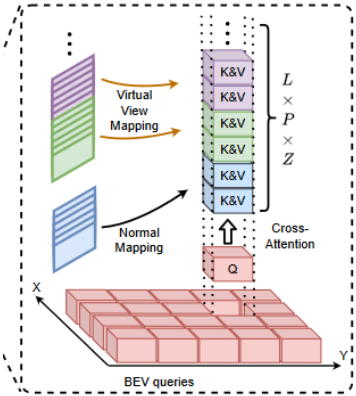


Figure 5: Cross-Attention

- Concatenate previous BEV feature & queries
- Can be viewed as multiple grafted Transformer layer
- Implicitly deepen the number of layers

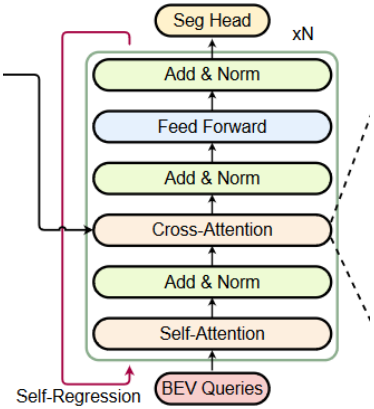


Figure 6: Self-Regression

Experiments

- NuScenes dataset
- New setting with
  - wider visible range: 160m x 160m
  - split avoid overfitting
- Evaluation metrics
  - mIoU

- Parameters & FPS

Method	Years	Backbone	Parameters	FPS	mIoU (Vanilla / City-based)		
					Road mIoU	Lane mIoU	All
LSS	ECCV20	EffNetb0	-	-	72.9 / -	20.0 / -	46.5 / -
VPN*	IROS20	Res101DCN	-	-	76.9 / -	19.4 / -	48.2 / -
LSS*	ECCV20	Res101DCN	-	-	77.7 / -	20.0 / -	48.9 / -
M2BEV	-	ResNeXt101	112.5	1.4	77.2 / -	40.5 / -†	58.9 / -†
BEVFormer	ECCV22	Res101DCN	68.7	1.7	80.1 / -	25.7 / -	52.9 / -
UniFusion	-	ResNet50	42.4	2.6	**82.0 / 42.6**	**25.8 / 11.2**	**53.9 / 26.9**
UniFusion		VoVNet99	84.0	2.7	**85.4 / 47.9**	**31.0 / 11.6**	**58.2 / 29.8**

Figure 7: Experiment Results

Experiments

Results

Method	Years	Backbone	mIoU (Easy)				mIoU (Hard)			
			Divider	Crossing	Boundary	All	Divider	Crossing	Boundary	All
VPN	IROS20	ResNet50	25.4 / 8.3	6.7 / 0.5	25.3 / 14.6	19.1 / 7.8	13.4 / 2.9	4.3 / 0.0	13.1 / 6.5	10.3 / 3.1
LSS	ECCV20	ResNet50	11.3 / 6.4	0.3 / 0.2	10.8 / 4.4	7.5 / 3.7	6.0 / 1.2	0.4 / 0.2	6.2 / 1.1	4.2 / 0.8
BEVFormer	ECCV22	ResNet50	42.2 / 16.1	26.9 / 7.6	42.1 / 18.6	37.1 / 14.1	27.3 / 7.8	17.5 / 2.3	26.3 / 10.0	23.7 / 6.7
UniFusion	-	ResNet50	**46.3 / 18.5**	**30.5 / 10.5**	**45.8 / 21.0**	**40.9 / 16.7**	**28.1 / 8.8**	**17.6 / 2.7**	**26.9 / 10.2**	**24.2 / 7.2**

Figure 8: Experiment Results

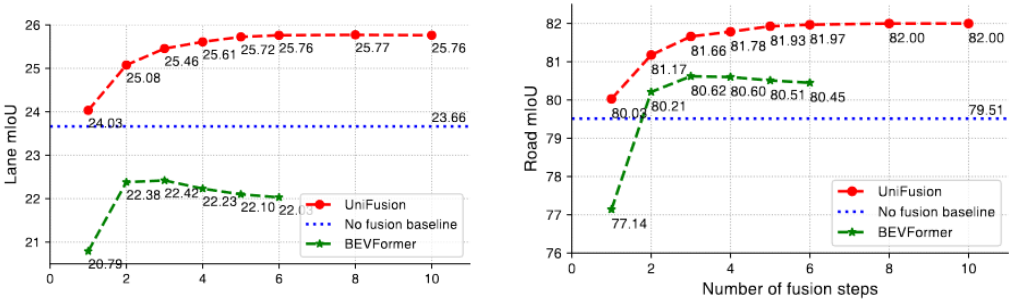
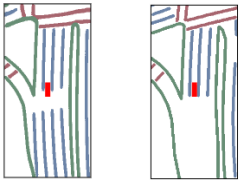


Figure 9: Long-range fusion ability

Conclusion

- Dataset quality
- Virtual view with dynamic scene?
- Memory consumption

Surrounding Images



Pred

GT

References

[1] Zequn Qin et al. "Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 8690–8699.

Thanks for Listening

Q & A